# Mathematical Statistics for Economics and Business

Springer
*New York*
*Berlin*
*Heidelberg*
*Barcelona*
*Hong Kong*
*London*
*Milan*
*Paris*
*Singapore*
*Tokyo*

Ron C. Mittelhammer

# Mathematical Statistics for Economics and Business

With 83 Illustrations

Springer

Ron C. Mittelhammer
Program in Statistics
and
Department of Agricultural Economics
Washington State University
Pullman, WA 99164-6210
USA

Library of Congress Cataloging-in-Publication Data
Mittelhammer, Ron.
Mathematical statistics for economics and business / Ron Mittelhammer.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-1-4612-8461-1 e-ISBN-13: 978-1-4612-3988-8
DOI: 10.1007/978-1-4612-3988-8
1. Commercial statistics. 2. Mathematical statistics. I. Title.
HF1017.M53 1995
519.5—dc20 95-37686

Printed on acid-free paper.

Softcover reprint of the hardcover 1st edition 1996

Production managed by Laura Carlson; manufacturing supervised by Jacqui Ashri.
Typeset by The Bartlett Press, Inc.

9 8 7 6 5 4 3 (Corrected third printing, 1999)

To my wife Linda,
and to Carl, Edith, Dolly and Ralph

# Preface

This book is designed to provide beginning graduate students and advanced undergraduates with a rigorous and accessible foundation in the principles of probability and mathematical statistics underlying statistical inference in the fields of business and economics. The book assumes no prior knowledge of probability or statistics and effectively builds the subject "from the ground up." Students who complete their studies of the topics in this text will have acquired the necessary background to achieve a mature and enduring understanding of statistical and econometric methods of inference and will be well equipped to read and comprehend graduate-level econometrics texts. Additionally, this text serves as an effective bridge to more advanced study of both mathematical statistics and econometric theory and methods. The book will also be of interest to researchers who desire a decidedly business and economics-oriented treatment of the subject in terms of its topics, depth, breadth, examples, and problems.

Without the unifying foundations that come with training in probability and mathematical statistics, students in statistics and econometrics classes too often perceive the subject matter as a potpourri of formulae and techniques applied to a collection of special cases. The details of the cases and their solutions quickly fade for those who do not understand the reasons for using the procedures they attempt to apply. Many institutions now recognize the need for a more rigorous study of probability and the mathematical statistics principles in order to prepare students for a higher level and a longer lasting understanding of the statistical techniques employed in the fields of business and economics. Furthermore, quantitative analysis in these fields has progressed to the point where a deeper understanding of the principles of probability and statistics is now virtually necessary for one to read and contribute successfully to quantitative research in economics and business. Contemporary students themselves know this and need little convincing from advisors that substantial statistical training must be acquired in order to compete successfully with their peers and to become effective researchers. Despite these observations, there are very few rigorous books on probability and mathematical statistics foundations that are also written with the needs of business and economics student in mind.

This book is the culmination of 15 years of teaching graduate level statistics and econometrics classes for students who are beginning graduate programs in business (primarily finance, marketing, accounting, and decision sciences), economics, and agricultural economics. When I originally took on the teaching assignment in this area, I cycled through a number of very good texts in mathematical statistics searching for an appropriate exposition for beginning graduate students. With the help of my students, I ultimately realized that the available textbook presentations were optimizing the wrong objective functions for our purposes! Some books were too elementary, other presentations did not cover multivariate topics in sufficient detail, and proofs of important results were omitted occasionally because they were "obvious" or "clear" or "beyond the scope of the text." In most cases they were neither obvious nor clear to students, and in many cases, useful and accessible proofs of the most important results can and should be provided at this level of instruction. Sufficient asymptotic theory was often lacking and/or tersely developed. At the extreme, material was presented in a sterile mathematical context at a level that was inaccessible to most beginning graduate students while nonetheless leaving notable gaps in topic coverage of particular interest to business and economics students. I then began to teach the course from lecture notes that I had created and iteratively refined them as I interacted with scores of students who provided me with feedback regarding what was working—and what wasn't—with regard to topics, proofs, problems, and exposition. I am deeply indebted to the hundreds of students who persevered through, and contributed to, the many revisions and continual sophistication of my notes. Their influence has had a substantial impact on the text: It is a time-tested and class-tested product. Other students at a similar stage of development should find it honest, accessible, and informative.

Instructors attempting to teach a rigorous course in mathematical statistics soon learn that the typical new graduate student in economics and business is thoroughly intelligent, but often lacks the sophisticated mathematical training that facilitates understanding and assimilation of the mathematical concepts involved in mathematical statistics. My experience has been that these students can understand and become functional with sophisticated concepts in mathematical statistics if their backgrounds are respected and the material is presented carefully and thoroughly, using a realistic level of mathematics.

Furthermore, it has been my experience that most students are actually eager to see proofs of propositions, as opposed to merely accepting statements on faith, so long as the proofs do not insult the integrity of the nonmathematician. Additionally, students almost always remark that the understanding and the long-term memory of a stated result is enhanced by first having worked through a formal proof of a proposition, and then working through examples and problems that require the result to be applied.

With the preceding observations in mind, the prerequisites for the book include only the usual introductory college-level courses in basic calculus (including univariate integration and differentiation, partial differentiation, and multivariate integration of the iterated integral type) and basic matrix algebra. The text is largely self-contained for students with this preparation. A significant effort has been made to present proofs in ways that are accessible. Care has been taken to choose methods and types of proofs that exercise and extend the learning process regarding statistical results and concepts learned prior to the introduction of the proof. A generous number of examples are presented with a substantial amount of detail to illustrate the application of major theories, concepts, and methods. The problems at the end of the chapters are chosen to provide an additional perspective to the learning process. The majority of the problems are word problems designed to challenge the reader to become adept at what is generally the most difficult hurdle—translating descriptions of statistical problems arising in business and economic settings into a form that lends itself to solutions based on mathematical statistics principles. I have also warned students through the use of asterisks (*) when a proof, concept, example, or problem may be stretching the bounds of the prerequisites so as not to frustrate the otherwise diligent reader, and to indicate when the help of the instructor or additional readings may be useful.

The book is designed to be versatile. The course that inspired this book is a 4-credit, semester-long, intensive mathematical statistics foundation course. I do not lecture on all of the topics contained in the book in the 50 contact hours available in the semester. The topics that I do not cover are taught in the first half of a subsequent semester-long 3-credit course in statistics and econometric methods. I have tended to treat chapters 1 through 4 in detail, and I recommend that this material be thoroughly understood before venturing into the statistical inference portion of the book. Thereafter, the choice of topics is flexible. For example, the instructor can control the depth at which asymptotic theory is taught by her choice of whether the starred topics in chapter 5 are discussed. While random sampling, empirical distribution functions, and sample moments should be covered in chapter 6, the instructor has leeway in the degree of emphasis that she places on other topics in the chapter. Point estimation and hypothesis testing topics can then be mixed and matched with a minimal amount of back-referencing between the respective chapters.

Distinguishing features of this book include the care with which topics are introduced, motivated, and built upon one another; use of the appropriate level of mathematics; the generous level of detail provided in the proofs; and a familiar business and economics context for examples and problems. This text is a bit longer than some of the others in the field. The additional length comes from additional explanation, and the detail in examples, problems, and proofs and not from a proliferation of topics which are merely surveyed rather than fully developed. As I see it, a survey of statistical techniques is useful only after one has the fundamental statistical background to appreciate what is being surveyed. And this book provides the necessary background.

## Acknowledgments

I am indebted to large number of people for their encouragement and comments. Millard Hastay, now retired from the Washington State University economics faculty, is largely responsible for my unwavering curiosity and enthusiasm for the field of theoretical and applied statistics and econometrics. George Judge has been a constant source of encouragement for the book project and over the years has provided me with very valuable and selfless advice and support in all endeavors in which our paths have crossed. I thank Jim Chalfant for giving earlier drafts of chapters a trial run at Berkeley, and for providing me with valuable student and instructor feedback. Thomas Severini at Northwestern provided important and helpful critiques of content and exposition. Martin Gilchrist at Springer-Verlag provided productive and pleasurable guidance to the writing and revision of the text. I also acknowledge the steadfast support of Washington State University in the pursuit of the writing of this book. Of the multitude of past students who contributed so much to the final product and that are too numerous to name explicitly, I owe a special measure of thanks to Don Blayney, now of the Economic Research Service, and Brett Crow, currently a promising Ph.D candidate in Economics at WSU, for reviewing drafts of the text literally character by character and demanding clarification in a number of proofs and examples. I also wish to thank many past secretaries who toiled faithfully on the book project. In particular, I wish to thank Brenda Campbell, who at times literally typed morning, noon, and night to bring the manuscript to completion, and without whom completing the project would have been infinitely more difficult. Finally, I thank my wife Linda, who proofread many parts of the text, provided unwavering support, sustenance, and encouragement to me throughout the project, and, despite all of the trials and tribulations, remains my best friend.

# Contents

## 10. Hypothesis-Testing Methods 595

## Appendix A. Math Review: Sets, Functions, Permutations, Combinations, and Notation 677

# 1 Elements of Probability Theory

## 1.1 Introduction

The objective of this chapter is to define a quantitative measure of either the level of certainty (or uncertainty) associated with observing various outcomes of a chance situation, or the degree of belief in a stated proposition. The quantitative measure, called **probability**, is relevant for quantifying such things as how likely it is that a shipment of transistors contains less than 5 percent defectives, that a gambler will win a crap game, that next year's corn yields will exceed 80 bushels per acre, or that electricity demand in Los Angeles will exceed generating capacity on a given day. The probability concept will also be relevant for quantifying an individual's degree of belief in such propositions as it will rain tomorrow, Congress will raise taxes next year, and the United States will suffer another oil embargo in the coming year.

The value of such a quantitative measure of uncertainty or degree of belief for decision making in business, economics, government, and everyday life is substantial. In the absence of such a measure, all one can effectively say when faced with a chance situation whose outcome is unknown or with a proposition whose validity is unknown is, "I don't know what will happen" or "I don't know whether the proposition is true or false." A rational decision maker will most definitely prefer to narrow the uncertainty of a chance situation's outcome or a proposition's validity in any decision making context in which profits, utility, and/or welfare are affected. Indeed, the problem of increasing profit, utility, or welfare through appropriate choices of production and inventory levels and scheduling, product pricing, advertising effort, trade policy, tax strategy, input or commodity purchases, technology adoption, and/or capital investment is substantially more difficult when the results of one's choice are affected by factors that are simply unknown.

Probability is a tool for distinguishing likely from unlikely outcomes or states of affairs and provides business managers, economists, legislators, and consumers with information that can be used to rank the potential results of their decisions in terms of likelihood of occurrence. It then becomes possible to make choices that maximize the likelihood of a desired outcome, provide a high likelihood of avoiding disastrous outcomes, or achieve a desirable expected result (where "expected" will be rigorously defined in Chapter 2).

Four basic definitions have been involved in the development of probability theory: **classical probability**, **relative frequency probability**, **subjective probability**, and the **axiomatic approach** to the definition of probability. We briefly discuss the first three approaches and then concentrate on the modern axiomatic approach. We will see that in the modern mathematical theory, probabilities are values of set functions which have special properties that satisfy the requirements of certain probability axioms.

Prior to our excursion into the realm of probability theory, it is helpful to examine how the terms "experiment," "sample space," "outcome," and "event" will be used in our discussion. The next section provides the necessary information.

## 1.2 Experiment, Sample Space, Outcome, and Event

The term **experiment** is used very generally in the field of statistics to refer to any activity for which the outcome or final state of affairs cannot be specified in advance, but for which a set containing all potential outcomes or final states of affairs can be identified. Thus, determining the yield per acre of a new type of wheat, observing the quantity of a commodity sold during a promotional campaign, identifying the fat percentage of a hundredweight of raw farm milk, observing tomorrow's closing price of gold on the London exchange, or analyzing the underlying income elasticity affecting the demand for gasoline are all examples of experiments according to our usage of the term.

The final result, observation, or measurement from the experiment is referred to as the **outcome** of the experiment. Thus, referring to the above examples of experiments, 80 bushels per acre, 2,500 units sold during a week of promotions, 3.7 percent fat per hundredweight, $440 per ounce, and .25 are, respectively, possible outcomes.

Prior to analyzing probabilities of outcomes of an experiment, it is necessary to identify what outcomes are *possible*. This leads to the definition of the **sample space** of an experiment.

**Definition 1.1**
*Sample space*

A set that contains all possible outcomes of a given experiment.

Note that our definition of sample space does not necessarily identify a unique set since we require only that the sample space *contain* all possible outcomes of an experiment. In many cases, the set of all possible outcomes will be readily identifiable and not subject to controversy, and in these cases it will be natural to refer to this set as the sample space. For example, the experiment of rolling a die and observing the number of dots facing up has a sample space that can be rather uncontroversially specified as $\{1, 2, 3, 4, 5, 6\}$. However, defining the collection of possible outcomes of an experiment may also require some careful deliberation. For instance, in our example of measuring the fat percentage of a given hundredweight of raw farm milk, it is clear that the outcomes must reside in the set $A = \{x : 0 \leq x \leq 100\}$. However, the accuracy of our measuring device might only allow us to observe differences in fat percentages up to hundredths of a percent, and thus a smaller set containing all possible *measurable* fat percentages might be specified as $B = \{x : x = (.01)n, n = 0, 1, 2, \ldots, 10, 000\}$, where $B \subset A$. It might be argued further that fat percentages of greater than 20 percent and less than 1 percent will not occur in raw farm milk, and thus the smaller set $C = \{x : x = (.01)n, n = 100, 101, 102, \ldots, 2, 000\}$, where $C \subset B \subset A$ could represent the sample space of the fat-measuring experiment. Fortunately, as the reader will come to recognize, the only concern of practical importance is that the sample space be specified large enough to contain the set of all possible outcomes of the experiment as a subset. The sample space need not be identically equal to the set of all possible outcomes. The reader may wish to suggest appropriate sample spaces for the remaining four example experiments described above.

We will use the capital letter $S$ to refer to sample space in the remainder of our study. Consistent with set theory terminology, each outcome in a sample space is called an **element** (or member) **of the sample space**. Elements of sample spaces are also sometimes called **sample points**. The sample space, as all sets, can be classified according to whether the number of elements in the set is finite, countably infinite, or uncountably infinite. It is customary to refer to a sample space that is finite or countably infinite as a **discrete sample space**. An uncountably infinite sample space that consists of a continuum of points, such as all of the points on an interval of the real line or all of the points in a rectangle, is often referred to as a **continuous sample space**.

The fundamental entities to which probabilities will be assigned are subsets of the sample space called events.

**Definition 1.2**
***Event***

An **event** is a subset of the sample space.

Thus, events are collections of outcomes of an experiment.[1] In the special case where the event consists of a single element or outcome, we will use the special

[1] For now, we will act as if every subset of a sample space is an event. A technical problem arises when we deal with uncountably infinite sample spaces, such that certain complicated sets will not be considered events because it will be impossible to assign probability to them in a consistent manner. We will investigate this technical problem in Section 1.6. As a practical matter, all sets to which the reader would be interested in assigning probability will be events.

term **elementary event** to refer to the event. We say that the event $A$ has occurred if the experiment results in an outcome that is an element of the set $A$. The real-world meaning of the statement "**the occurrence of event $A$**" will be provided by the real-world definition of the set $A$. That is, verbal or mathematical statements that are utilized in the verbal or mathematical method of defining set $A$, or the collection of elements or description of elements placed in brackets in the exhaustive listing method of defining set $A$, provide the meaning of "the occurrence of event $A$." The following examples will illustrate the meaning of both "event" and the "occurrence of an event."

**Example 1.1** An experiment consists of rolling a die and observing the number of dots facing up. The sample space is defined to be $S = \{1, 2, 3, 4, 5, 6\}$. Examine two subsets of $S : A_1 = \{1, 2, 3\}$, $A_2 = \{2, 4, 6\}$.

Event $A_1$ has occurred if the outcome, $x$, of the experiment (the number of dots facing up) is such that $x \in A_1$. Then $A_1$ is an event whose occurrence means that after a roll the number of dots facing up on the die is three or less.

Event $A_2$ has occurred if the outcome, $x$, is such that $x \in A_2$. Then $A_2$ is an event whose occurrence means that the number of dots facing up on the die is an even number. □

**Example 1.2** An experiment consists of observing the percentage of a large group of consumers, representing a consumer taste panel, that prefer Schpitz beer to its closest competitor, Nickelob beer. The sample space for the experiment is specified as $S = \{x : 0 \leq x \leq 100\}$. Examine two subsets of $S : A_1 = \{x : x < 50\}$, $A_2 = \{x : x > 75\}$.

Event $A_1$ has occurred if the outcome, $x$, of the experiment (the actual percentage of the consumer panel preferring Schpitz beer) is such that $x \in A_1$. Then $A_1$ is an event whose occurrence means that less than 50 percent of the consumers preferred Schpitz to Nickelob or, in other words, the group of consumers preferring Schpitz were in the minority. Event $A_2$ has occurred if the outcome $x \in A_2$. Then $A_2$ is an event whose occurrence means that greater than 75 percent of the consumers preferred Schpitz to Nickelob. □

Examples 1.1 and 1.2 can be used to illustrate the concept of **mutually exclusive or disjoint events**. The concept is identical to the concept of mutually exclusive or disjoint sets. In Ex. 1.1, it is recognized that events $A_1$ and $A_2$ are *not* mutually exclusive events, since $A_1 \cap A_2 = \{2\} \neq \emptyset$. Events that are not mutually exclusive can occur simultaneously. Events $A_1$ and $A_2$ will occur simultaneously (which cannot be the case for mutually exclusive events) only if $x \in A_1 \cap A_2 = \{2\}$. In Ex. 1.2, events $A_1$ and $A_2$ are mutually exclusive events since $A_1 \cap A_2 = \emptyset$. Events $A_1$ and $A_2$ cannot occur simultaneously since if the outcome is such that $x \in A_1$, then it follows that $x \notin A_2$, or if $x \in A_2$, then it follows that $x \notin A_1$.

We should emphasize that in applications it is the researcher who specifies which events are of interest; i.e., the researcher defines the subsets of the sample space whose occurrence or lack thereof provides useful information from the researcher's viewpoint. Thus, referring to Ex. 1.2, if the researcher were employed by Schpitz Brewery, the identification of which beer was preferred by a majority of the participants in a taste comparison would appear to be of significant interest in a report to the management of the brewery, and thus event $A_1$ would be of great importance. Event $A_2$ in that example might be considered important if the advertising department of Schpitz Brewery wished to utilize an advertising slogan such as "Schpitz beer is preferred to Nickelob by more than 3 to 1."

## 1.3 Nonaxiomatic Probability Definitions

There are three major nonaxiomatic definitions of probability that have been suggested in the course of the development of probability theory. The definitions seek to identify a quantitative measure of the likelihood of occurrence, or degrees of belief in, the various events associated with an experiment. We briefly discuss each of these alternative probability definitions. In the definition below, $N(\cdot)$ is the size-of-set function (see Def. A.21).

**Definition 1.3**
***Classical probability***

Let $S$ be the finite sample space for an experiment having $N(S)$ equally likely outcomes, and let $A \subset S$ be an event containing $N(A)$ elements. Then the probability of the event $A$, denoted by $P(A)$, is given by $P(A) = N(A)/N(S)$.

In the classical definition, probabilities are images of sets generated by a set function, $P$, having as its domain the collection of all subsets of a finite (and thus discrete) sample space and having a range that is contained in the real interval $[0,1]$. The following example illustrates the application of the classical probability concept.

**Example 1.3** Reexamine the die-rolling experiment of Ex. 1.1. We assume that the die is fair so that the outcomes in the sample space $S = \{1,2,3,4,5,6\}$ are equally likely. The number of elements in the sample space is given by $N(S) = 6$. Let $E_i$, $i = 1,\ldots,6$, represent the elementary events in the set $S$. Then according to the classical probability definition, $P(E_i) = N(E_i)/N(S) = 1/6$ for all $i = 1,\ldots,6$, so that the probability of each elementary event is 1/6. Referring to events $A_1$ and $A_2$ of Ex. 1.1, note that

$$P(A_1) = \frac{N(A_1)}{N(S)} = \frac{3}{6} = \frac{1}{2} \quad \text{and} \quad P(A_2) = \frac{N(A_2)}{N(S)} = \frac{3}{6} = \frac{1}{2}.$$

Therefore, the probability of rolling a three or less and the probability of rolling an even number are both 1/2. Note finally that $P(S) = N(S)/N(S) = 6/6 = 1$, which states that the probability of the event that the outcome of the experiment is an element of the sample space is 1, as it intuitively should be if the number 1 is to be associated with an event that will occur with certainty. □

The classical definition has two major limitations that preclude its use as the foundation on which to build a general theory of probability. First, the sample space must be finite or else $N(S) = \infty$ and possibly $N(A) = \infty$. Thus probability in the classical sense is not useful for defining the probabilities of events contained in a countably infinite or uncountably infinite sample space. Another limitation of the classical definition is that outcomes of an experiment must be equally likely. Thus, for example, if we are engaged in a coin-tossing experiment and it cannot be assumed that the coin is fair, then the classical probability definition provides us with no information as to how probabilities should be defined. In order to relax these restrictions, we examine the relative frequency approach.

**Definition 1.4** ***Relative frequency probability***

Let $n$ be the number of times that an experiment is repeatedly performed under identical conditions. Let $A$ be an event in the sample space $S$, and define $n_A$ to be the number of times in $n$ repetitions of the experiment that the event $A$ occurs. Then the probability of the event $A$ is given by the limit of the relative frequency $n_A/n$, as

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n}.$$

It is recognized that in the relative frequency definition, the probability of an event $A$ is the image of $A$ generated by a set function $P$, where the image is defined as the limiting fraction of the total number of outcomes of the sequence of experiments that are observed to be members of the set $A$ (for a rigorous definition of limits, see Section 5.2). It is clear that the range of the set function must be contained in $[0, 1]$ since $0 \leq n_A \leq n$. The following example illustrates the application of the relative frequency concept of probability.

**Example 1.4** Consider the following collection of coin-tossing experiments, where a coin was tossed various numbers of times and, following the prescription in the relative frequency definition of probability, the fraction of the tosses resulting in heads was recorded for each collection of experiments.[2]

| No. of Tosses | No. of Heads | Relative Frequency |
|---|---|---|
| 100 | 48 | .4800 |
| 500 | 259 | .5180 |
| 1,000 | 489 | .4890 |
| 5,000 | 2,509 | .5018 |
| 75,000 | 37,447 | .4993 |

[2] These experiments were actually performed by the author, except the author did not actually flip coins to obtain the results listed here. Rather, the coin flips were simulated by the computer. In the coming chapters, the reader will come to understand exactly how the computer might be used to simulate the coin-flipping experiment, and how to simulate other experiments as well.

It would *appear* that as $n \rightarrow \infty$, the observed relative frequency of heads is approaching 0.5. □

The relative frequency definition enjoys some advantages over the classical definition. For one, the sample space can be an infinite set, since the ability to form the relative frequency $n_A/n$ does not depend on the underlying sample space's being finite. Also, there is no need to assume that outcomes are equally likely, since the concept of $\lim_{n\rightarrow\infty} n_A/n$ representing the limiting fraction of outcomes of the experiments that are members of set $A$ does not depend on the outcomes' being equally likely.

Unfortunately, there are problems with the relative frequency definition that reduce its appeal as a foundation for the development of a general theory of probability. First of all, while it is an *empirical fact* that for many types of experiments, such as the coin-tossing experiment in Ex. 1.4, the relative frequencies tend to stabilize as $n$ increases, how do we know that $n_A/n$ will actually converge to a limit in all cases? Indeed, how could we ever observe the limiting value if an infinite number of repetitions of the experiment are required? Furthermore, even if there is convergence to a limiting value in one sequence of experiments, how do we know that convergence to the same value will occur in another sequence of the experiments? Lacking a definitive answer to these conceptual queries, we refrain from using the relative frequency definition as the foundation for the probability concept.

A third approach to defining probability involves personal opinion, judgments, or educated guesses and is called subjective probability.

**Definition 1.5**
***Subjective probability***

> The subjective probability of an event $A$ is a real number, $P(A)$, in the interval $[0, 1]$, chosen to express the degree of personal belief in the likelihood of occurrence or validity of event $A$, the number 1 being associated with certainty.

Like the preceding definitions of probability, subjective probabilities can be viewed as images of set functions. Note that the subjective probability assigned to an event can obviously vary depending on who is assigning the probabilities and depending on the personal beliefs of the individual assigning the probabilities. Even supposing that two individuals possess exactly the same information regarding the characteristics of an experiment, the way in which each individual interprets the information may result in differing probability assignments to an event $A$.

Unlike the relative frequency approach, subjective probabilities can be defined for experiments that cannot be repeated. For example, one might be assigning probability to the proposition that a war in the Middle East will break out during the year 2000. Defining the probability of the event "war in the year 2000" does not conveniently fit into the relative frequency definition of probability since one can only observe the outcome of war or peace in the year 2000 *once*. In addition, the classical definition would not apply unless war and peace were equally likely a priori. Similarly, assigning probability to the event that one or the other team will win in a Superbowl game is commonly done in various ways by many individuals, and a considerable amount of betting is based on those probability assignments. However, the particular Superbowl "experiment" cannot be repeated, nor is there usually any a priori reason to suspect that the outcomes are equally likely so that neither relative frequency nor classical definitions apply.

In certain problem contexts the assignment of probabilities solely on the basis of personal beliefs may be considered undesirable. For example, if an individual is betting on some game of chance, that individual would prefer to know the "true" likelihood of the game's various outcomes and not rely merely on his or her personal perceptions. For example, after inspecting a penny, suppose you consider the coin to be fair and (subjectively) assign a probability of 1/2 to each of the outcomes "heads" and "tails." However, should the penny have been supplied by a ruthless gambler who has altered the penny in such a way that an outcome of heads is twice as likely to occur as tails, the gambler could induce you to bet in such a way that you would lose money in the long run if you adhered to your initial subjective probability assignments—the game would not be "fair." (Explain how you could be induced into betting in a way that makes you a loser in the long run.)

Given that "objective" (classical, relative frequency approaches) and subjective probability concepts might both be useful, depending on the problem situation, we seek a probability theory that is general enough to accommodate all of the concepts of probability discussed heretofore. Such an accommodation can be achieved by defining probability in axiomatic terms.

## 1.4 Axiomatic Definition of Probability

Our objective is to devise a quantitative measure of the likelihood of occurrence of, or the degree of belief in, various events contained in a sample space. How should one go about defining such a measure? A useful approach is to define the measure by stating mathematical properties that we feel our measure should possess. If the properties imposed on the measure are straightforwardly acceptable or appear clearly useful in devising our means of measurement, and if the assumptions are not contradictory, we could collect these properties into a set of axioms (see the definition in Section A.2) on which to build the theory of probability.

Note, as an aside, that the approach of using a set of axioms as the foundation for a body of theory should be familiar to students of business and economics. For example, the theory of the consumer is founded on a set of behavioral assumptions, i.e., a set of axioms. The reader might recall that the axioms of comparability, transitivity, and continuity of preferences are sufficient for the existence of a utility function, the maximization of which, subject to an income constraint, depicts consumption behavior in the neoclassical theory.[3]

What mathematical properties should our measure possess? First of all, it seems useful for our measure to be in the form of a real-valued set function, since this would allow probabilities of events to be stated in terms of real numbers. Thus, we begin with a set function, say $P$, which has as its domain all of the events in a sample space, $S$, and has as its range a set of real numbers, i.e., we have $P : \Upsilon \rightarrow R$, where $\Upsilon$ is the set of all events in $S$. The set $\Upsilon$ is called the **event space**.

**Definition 1.6**
***Event space***

The set of all events in the sample space $S$ is called the event space.

We have in mind that the images of events under $P$ will be probabilities of the events, i.e, $P(A)$ will be the probability of the event $A \in \Upsilon$. Now, what type of properties seem appropriate to impose on the real-valued set function $P$?

Reviewing the three definitions of probability presented in Section 1.3, it is recognized that in each case, probability was defined to be a *nonnegative* number. Since at this point we are completely free to choose the properties we desire of our measure, and since each of the previous nonaxiomatic definitions of probability possesses some intuitive appeal as measures of the likelihood of occurrence or degree of belief in an event (despite our recognition of some conceptual difficulties), let us agree that our measure should be nonnegative valued. By doing so, we will have defined the first axiom to which our measure must adhere while remaining consistent with all of our previous probability definitions. Since we decided that our measure would be generated by a set function, $P$, our assumption requires that our set function be such that the image of any event $A$, $P(A)$, be a nonnegative number. Our first axiom is thus

**Axiom 1.1**

For any event $A \subset S$, $P(A) \geq 0$.

Now that we have committed ourselves to a measure that is nonnegative, what nonnegative number should our measure associate with the **certain event**, $S$?[4] Again we are free to choose the property we desire of our measure. However, there are some advantages to choosing the number 1 to denote the likelihood of, or the degree of belief in, the certain event. First, it is consistent with all of our nonaxiomatic definitions of probability discussed earlier. Second, it allows

[3] See G. Debreu (1959), "Theory of Value: An Axiomatic Analysis of Economic Equilibrium." Cowles Monograph 17. New York: John Wiley, pp. 60–63. Note that additional axioms are generally included that are not needed for the existence of a utility function per se but that lead to a simplification of the consumer maximization problem. See L. Phlips (1983), *Applied Consumption Analysis*. New York: North-Holland, pp. 8–11.

[4] By definition, since $S$ contains all possible outcomes of the experiment, the event $S$ is then certain to occur.

the probability of any event $A \subset S$ to be directly interpreted as indicating a proportion of certainty. That is, if we assume that our set function is such that $P(S) = 1$, and if $P(A) = k$, say, then the measure of event $A$ relative to the measure of the certain event $S$ is $P(A)/P(S) = k/1 = k$, so that $P(A) = kP(S)$, and thus the event $A$ is assigned a proportion, $k$, of certainty. Our second axiom is then

**Axiom 1.2**

$$\boxed{P(S) = 1.}$$

Dwelling on the value of $k$ in the preceding discussion for a moment, it is clear that what we intuitively had in mind for the value of $k$ was a number $k \in [0, 1]$. Our intuitive reasoning would be that if $S$ is the certain event, then the occurrence of $A \subset S$ surely cannot be "more than certain." That is, unless $A = S$, there are sample points in $S$ that are not in $A$ (i.e., $S - A \neq \emptyset$). Thus, while the event $S$ will always occur, $A$ may or may not, and thus surely $A$ is no more certain to occur than $S$. This line of reasoning then suggests that our measure must be such that $P(A) \leq 1$. However, we can proceed further and extend this argument. If $A$ and $B$ are any two events such that $A \subset B$, then we would require that $P(A) \leq P(B)$, since every element of $A$ is also in $B$, but $B$ may contain sample points that are not in $A$, and thus $A$ can surely be no more likely to occur than $B$.

We would all agree that if our measure is to quantify meaningfully the degree of belief in, or the likelihood of, an event, it must possess the property suggested in the argument above. We are thus tempted to add to our collection another axiom that states "if $A \subset B$, then $P(A) \leq P(B)$." Before doing this, however, let us investigate the argument still further. If indeed $A \subset B$, and thus $P(B) \geq P(A)$, then to what should we ascribe the remaining portion, $P(B) - P(A)$, of the probability of event $B$? An intuitively obvious answer comes to mind. The set $B - A$ represents the collection of elements remaining in the set $B$ after we remove the elements in $B$ that are contained in the set $A$. Since $A$ and $B - A$ are mutually exclusive, and since $B = A \cup (B - A)$, the event $B$ is thus partitioned[5] into two subsets. Represented this way, $B$ can occur iff either $A$ or $(B - A)$ occurs. If $P(B) > P(A)$, the added probability, $P(B) - P(A)$, of event $B$ occurring compared to event $A$ must be due to the probability of the occurrence of event $B - A$. Thus, it is logical to attribute the increment in the probability measure, $P(B) - P(A)$, to the event $B - A$.

Then our measure should have the property that for events $A$ and $B$ for which $A \subset B$, $P(B) = P(A) + P(B - A)$. Note that since $P(B - A) \geq 0$ by Axiom 1.1, this *implies* our previous requirement that if $A \subset B$, $P(B) \geq P(A)$. However, we have discovered much more than just another way of stating a potential third axiom. We have actually found that for *any* two mutually exclusive events $A_1$ and $A_2$, our measure should have the property that $P(A_1 \cup A_2) = P(A_1) + P(A_2)$.

[5] A **partition** of a set $B$ is a collection of disjoint subsets of $B$, say $\{B_i, i \in I\}$, such that $B = \cup_{i \in I} B_i$.

To see this, define the set $B = A_1 \cup A_2$, where $A_1 \cap A_2 = \emptyset$. Then $B - A_1 = (A_1 \cup A_2) - A_1 = A_2$ because $A_1$ and $A_2$ are mutually exclusive. Since we have argued that $P(B) = P(A_1) + P(B - A_1)$, substituting $A_2$ for $(B - A_1)$ then yields $P(A_1 \cup A_2) = P(A_1) + P(A_2)$. Thus, we have demonstrated that probability should be additive across any two disjoint events.

The preceding additivity argument can be easily extended to three *or more* disjoint events. To motivate the extension, first examine the case of three disjoint events, $A_1$, $A_2$, and $A_3$. Note that the two events $A_1 \cup A_2$ and $A_3$ are disjoint events, since

$$(A_1 \cup A_2) \cap A_3 = (A_1 \cap A_3) \cup (A_2 \cap A_3) = \emptyset \cup \emptyset = \emptyset$$

because $A_i \cap A_j = \emptyset$ for $i \neq j$ by the disjointness of $A_1$, $A_2$, and $A_3$. Then, applying our probability additivity result for the case of two disjoint events results in

$$P(A_1 \cup A_2 \cup A_3) = P(A_1 \cup A_2) + P(A_3).$$

But since $A_1$ and $A_2$ are disjoint, $P(A_1 \cup A_2) = P(A_1) + P(A_2)$, so that by substitution for $P(A_1 \cup A_2)$,

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3),$$

and we have the implication that probability is additive across any three disjoint events. Recognizing the sequential logic of the extension of probability additivity from two to three disjoint events, the reader can no doubt visualize the repetition of the argument *ad infinitum* to establish that probability should be additive across an arbitrary number of disjoint events. A concise, formal way of establishing the extension is through the use of **mathematical induction**, which we will now review.

The process of mathematical induction is a method for proving the validity of each proposition in a sequence of propositions. Implementation of the method begins with establishing the validity of the first proposition, $P_1$, in the sequence. Then an argument is constructed to demonstrate that *if* the $k$th proposition in the sequence were true, where the symbol $k$ represents an unspecified, arbitrary positive integer, then the $(k+1)$th proposition is necessarily true, i.e., $P_k$ is a sufficient condition for $P_{k+1}$. It is recognized that the latter argument establishes a chain of sequential truths, where the truth of every proposition $P_{k+i}$, $i = 1, 2, \ldots$ follows from the truth of $P_k$. Since in the method the truth of $P_1$ is established, the truth of $P_j$, $j = 2, 3, \ldots$ is thus demonstrated.

**Lemma 1.1**
***Mathematical Induction Principle***

Let $P_1, P_2, P_3, \ldots$ be a sequence of propositions. Each of the propositions in the sequence is true provided

**a.** $P_1$ is true, and

**b.** for an arbitrary positive integer $k$, if $P_k$ were true, it would necessarily follow that $P_{k+1}$ is true.

Returning to our probability additivity argument, the first proposition in the sequence of propositions we are interested in is "$P(A_1 \cup A_2) = \sum_{i=1}^{2} P(A_i)$ for disjoint events $A_i$, $i = 1, 2$." We have already defended the validity of this proposition. Now examine the proposition that for some $k$, "$P(\cup_{i=1}^{k} A_i) = \sum_{i=1}^{k} P(A_i)$ for disjoint events $A_i$, $i = 1, 2, \ldots, k$." Using the method of mathematical induction, we tentatively act as if this proposition were true, and we attempt to demonstrate that the truth of the next proposition in the sequence follows from the truth of the previous proposition, i.e., is "$P(\cup_{i=1}^{k+1} A_i) = \sum_{i=1}^{k+1} P(A_i)$ for disjoint events $A_i$, $i = 1, 2, \ldots, k+1$" then true? Note that the two events $\cup_{i=1}^{k} A_i$ and $A_{k+1}$ are disjoint, since

$$\left(\bigcup_{i=1}^{k} A_i\right) \cap A_{k+1} = \bigcup_{i=1}^{k} (A_i \cap A_{k+1}) = \bigcup_{i=1}^{k} \emptyset = \emptyset,$$

where $A_i \cap A_{k+1} = \emptyset \; \forall \; i \neq k+1$ by the disjointness of the $k+1$ events $A_1, \ldots, A_{k+1}$. But then by additivity for the two-event case,

$$P\left(\left(\bigcup_{i=1}^{k} A_i\right) \bigcup A_{k+1}\right) = P\left(\bigcup_{i=1}^{k} A_i\right) + P(A_{k+1}) = \sum_{i=1}^{k+1} P(A_i),$$

where the last equality follows from the assumed validity of probability additivity in the $k$-disjoint event case. Then by mathematical induction, we have demonstrated that

$$P\left(\bigcup_{i=1}^{m} A_i\right) = \sum_{i=1}^{m} P(A_i)$$

for disjoint events $A_1, \ldots, A_m$, $\forall$ positive integer $m$, i.e., probability is additive across an arbitrary number of disjoint events.

We finally state our probability additivity requirement as a third probability axiom, where we generalize the representation of the collection of disjoint events by utilizing an index set of subscripts.[6]

[6] Technically, our mathematical induction approach demonstrates probability additivity for any *number* of disjoint sets, which is referred to as **finite additivity.** However, $\infty$ is not a number per se. We appeal to intuition in extending probability additivity from an *arbitrary* number of disjoint sets to a countably infinite collection of disjoint sets. Such an extension is needed to answer certain probability questions concerning limit operations when dealing with countably infinite and uncountably infinite sample spaces. To illustrate "countably infinite additivity," let the set of positive integers code calendar years starting from 1997, i.e., 1 denotes 1997, 2 denotes 1998, 3 denotes 1999, and so on. Suppose you were interested in the probability that human life on Earth ceases to exist in an odd year. This draconian event would be represented by the countably infinite set $A = \{x : x \text{ is an odd positive integer}\}$, and the probability we seek is $P(A)$. Note that the event can be conceptualized as the union of countably infinite elementary events, $\cup_{i \in A} \{i\}$, the elementary events being disjoint by definition. It is natural (and intuitively reasonable given our additivity discussion) to represent this probability as $P(A) = \sum_{i \in A} P(\{i\})$ i.e., the probability of event $A$ is the sum of the probabilities of each of the countably infinite elementary events comprising event $A$. This property is referred to as **countable additivity** in the probability literature. It should be noted that a minority of statisticians argue that the theory of probability should be founded on the notion of *finite* additivity, and not countable additivity. For example, see B. deFinetti (1974), *Theory of Probability, Vol. I.* Chicester: Wiley, pp. 116-119.

**Axiom 1.3** Let $I$ be a finite or countably infinite index set of positive integers, and let $\{A_i : i \in I\}$ be a collection of disjoint events contained in $S$. Then, $P(\cup_{i \in I} A_i) = \sum_{i \in I} P(A_i)$.

The Russian mathematician A. N. Kolmogorov suggested that Axioms 1.1 to 1.3 provide an axiomatic foundation for probability theory.[7] As it turns out, sufficient information concerning the behavior of probability is contained in the three axioms to be able to derive from them the modern theory of probability. We begin deriving some important probability results in the next section.

In summary, we have defined the concept of probability by defining a number of properties that probability should possess. Specifically, probability will be generated by a set function that has the collection of events of a sample space, i.e., the event space, as its domain; its range will be contained in the interval $[0,1]$; the image of the certain event $S$ will be 1; and the probability of a countable union of disjoint events of $S$ will be equal to the sum of the probabilities of the individual events comprising the union. Any set function, $P(\cdot)$, that satisfies the three Axioms 1.1, 1.2, and 1.3 will be called a **probability measure** or **probability set function**. The image of an event $A$ generated by the probability set function $P$ is called the **probability of event** $A$.

The following examples provide illustrations of probability set functions for finite, countably infinite, and uncountably infinite sample spaces.

**Example 1.5** Let $S = \{1, 2, 3, 4, 5, 6\}$ be the sample space for rolling a fair die and observing the number of dots facing up. Then $P(A) = N(A)/6$, for $A \subset S$, defines a probability set function on the events in $S$. We can verify that $P(\cdot)$ is a probability measure by noting that $P(A) \geq 0$ for all $A \subset S$, $P(S) = N(S)/6 = 6/6 = 1$ and $P(A_1 \cup \cdots \cup A_n) = P(A_1) + \cdots + P(A_n)$ for any collection $\{A_1, \ldots, A_n\}$ of disjoint subsets of $S$. For example, if $A_1 = \{1, 2\}$ and $A_2 = \{4, 5, 6\}$, then

$$P(A_1) = \frac{N(A_1)}{6} = \frac{2}{6},$$

$$P(A_2) = \frac{N(A_2)}{6} = \frac{3}{6},$$

$$P(A_1 \cup A_2) = \frac{N(A_1 \cup A_2)}{6} = \frac{5}{6},$$

and thus $P(A_1 \cup A_2) = P(A_1) + P(A_2)$. More generally, if $A_1, \ldots, A_n$ are disjoint events, then

$$N\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} N(A_i),$$

[7] See A. N. Kolmogorov (1956), *Foundations of the Theory of Probability*, 2nd ed. New York: Chelsea.

and thus

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} N(A_i)/6 = \sum_{i=1}^{n} P(A_i). \qquad \square$$

**Example 1.6** Let $S = \{x : x \text{ is a positive integer}\}$, and examine the set function defined by $P(A) = \sum_{x \in A} (1/2)^x$ for $A \subset S$. The set function, so defined, is a probability set function since, first of all, $P(A) \geq 0$ because $P(A)$ is defined as the sum of a collection of nonnegative numbers. To verify that $P(S) = 1$, recall the following results from real analysis:

**Lemma 1.2**

$$\sum_{j=1}^{n} ar^j = \frac{a\left(r - r^{n+1}\right)}{1 - r}.$$

For $|r| < 1$,

$$\sum_{j=1}^{\infty} ar^j = \lim_{n \to \infty} \frac{a(r - r^{n+1})}{1 - r} = \frac{ar}{1 - r}.$$

In the case at hand, $a = 1$ and $r = 1/2$, so that

$$P(S) = \sum_{x=1}^{\infty} \left(\frac{1}{2}\right)^x = \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1.$$

Finally, by definition of the summation operation, if $A_i$, $i \in I$, are disjoint subsets of $S$, then

$$P\left(\bigcup_{i \in I} A_i\right) = \sum_{\substack{x \in (\cup A_i) \\ i \in I}} (1/2)^x = \sum_{i \in I} \sum_{x \in A_i} (1/2)^x = \sum_{i \in I} P(A_i). \qquad \square$$

**Example 1.7** Let $S = \{x : 0 \leq x < \infty\}$ be the sample space corresponding to the experiment of observing the operating life, in hours, of computer memory chips produced by a chip manufacturer. Let the probability set function be given by $P(A) = \int_{x \in A} \frac{1}{2} e^{-x/2}\, dx$, with the event space, $\Upsilon$ (the domain of $P$), being the collection of all interval subsets of $S$ together with any sets that can be formed by a countable number of union, intersection, and/or complement operations applied to the interval subsets.[8]

We can verify that $P(A) \geq 0 \; \forall \; A \in \Upsilon$, since $P(A) = \int_{x \in A} \frac{1}{2} e^{-x/2}\, dx$ has a nonnegative integrand and the integral of a nonnegative integrand is nonnegative

[8] Unlike the previous example, which used a countable sample space, when the sample space is uncountable, not all subsets of $S$ can technically be considered events, i.e., there may be subsets of $S$ to which probability cannot be assigned (recall Footnote 1). The collection of subsets defined here are the Borel sets contained in $S$, all of which can be considered events in $S$. We discuss this technical question further in Section 1.6.

valued.[9] It is also true that $P(S) = 1$, since $P(S) = \int_0^\infty \frac{1}{2} e^{-x/2}\,dx = -e^{-x/2}\,|_0^\infty = 1$. Finally, if $A = \cup_{i=1}^n A_i$, with the sets $A_1, A_2, \ldots, A_n$ being disjoint, it follows from the additivity property of the Riemann integral that

$$P\left(\bigcup_{i=1}^n A_i\right) = \int_{x \in \cup_{i=1}^n A_i} \frac{1}{2} e^{-x/2}\,dx = \sum_{i=1}^n \int_{x \in A_i} \frac{1}{2} e^{-x/2}\,dx = \sum_{i=1}^n P(A_i),$$

and so countable additivity holds. □

All problems involving the assignment of probabilities to the various events in a sample space will formally share a common mathematical structure given by a 3-tuple of objects, collectively referred to as the **probability space** of an experiment.

**Definition 1.7**
***Probability space***

*A* probability space is the 3-tuple $\{S, \Upsilon, P\}$, where $S$ is the sample space of an experiment, $\Upsilon$ is the event space, and $P$ is a probability set function having domain $\Upsilon$.

In any probabilistic analysis of an experiment, we will seek to establish

1. a universal set, $S$, that contains all of the potential outcomes or elementary events of an experiment;
2. a set of sets, $\Upsilon$, representing the collection of events or subsets of $S$ on which probability will be defined; and
3. a probability set function, $P$, that can be used to assign the appropriate probabilities to the events in $S$.

Once the probability space is defined, all of the information is established that is needed to assign probabilities to the various events of interest related to an experiment. As one might suspect, it is the discovery of the appropriate probability set function that represents a major challenge in the application of probability and statistics, and we examine the discovery problem in the latter half of the text when we discuss topics in inferential statistics. Our immediate goal in the remaining sections of this chapter is to establish a number of useful results in probability theory that are implied by the probability axioms.

## 1.5 Some Probability Theorems

The three axioms governing the behavior of a probability set function, together with results (lemmas) from set theory, can be utilized to prove a number of theorems that will provide insights into the behavior of probability. In proving the

[9] We will tacitly assume, unless explicitly stated otherwise, that the orientation of integral ranges is from lowest to highest values in defining the integral over the set $A$.

theorems, we begin our development of some fundamental tools of probability theory.

**Theorem 1.1** *Let $A$ be an event in the sample space $S$. Then $P(A) = 1 - P(\bar{A})$.*

**Proof** By the definition of the complement of $A$, $A \cup \bar{A} = S$. Thus, by substitution, and by Axiom 1.2, $P(S) = 1 = P(A \cup \bar{A})$. However, since $A \cap \bar{A} = \emptyset$, Axiom 1.3 allows us to state that $1 = P(A) + P(\bar{A})$. Subtracting $P(\bar{A})$ from both sides obtains the result. ■

**Theorem 1.2** $P(\emptyset) = 0$.

**Proof** Let $A = \emptyset$ in Theorem 1.1. Then, since $\bar{A} = S$, we have immediately that $P(\emptyset) = 1 - P(S) = 1 - 1 = 0$, since $P(S) = 1$ by Axiom 1.2. ■

**Theorem 1.3** *Let $A$ and $B$ be two events in a sample space such that $A \subset B$. Then $P(A) \le P(B)$ and $P(B - A) = P(B) - P(A)$.*

**Proof** Since $A \subset B$, $B = A \cup (B - A)$. The sets $A$ and $B - A$ are disjoint, and thus by Axiom 1.3, $P(B) = P(A) + P(B - A)$. Since $P(B - A) \ge 0$ by Axiom 1.1, dropping $P(B - A)$ from the probability equation implies $P(B) \ge P(A)$. Subtracting $P(A)$ from both sides of the probability equation yields the second result of the theorem. ■

**Theorem 1.4** *Let $A$ and $B$ be two events in a sample space $S$. Then $P(A) = P(A \cap B) + P(A \cap \bar{B})$.*

**Proof** $A = A \cap S = A \cap (B \cup \bar{B}) = (A \cap B) \cup (A \cap \bar{B})$ since the intersection operation is distributive and $S = B \cup \bar{B}$. The events $A \cap B$ and $A \cap \bar{B}$ are mutually exclusive since $(A \cap B) \cap (A \cap \bar{B}) = \emptyset$, and thus by Axiom 1.3, $P(A) = P(A \cap B) + P(A \cap \bar{B})$. ■

**Theorem 1.5** *Let $A$ and $B$ be two events in a sample space $S$. Then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.*

**Proof** $A \cup B = (A \cup B) \cap S = (A \cup B) \cap (B \cup \bar{B}) = B \cup (A \cap \bar{B})$ since the union operation is distributive and $S = B \cup \bar{B}$. Events $B$ and $(A \cap \bar{B})$ are disjoint since $(A \cap \bar{B}) \cap B = A \cap (\bar{B} \cap B) = A \cap \emptyset = \emptyset$, where we have used associativity and the fact that $B \cap \bar{B} = \emptyset$. Then by Axiom 1.3, $P(A \cup B) = P(B) + P(A \cap \bar{B})$. However, Theorem 1.4 implies that $P(A \cap \bar{B}) = P(A) - P(A \cap B)$, and thus by substitution, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. ■

**Corollary 1.1** (Boole's Inequality)[10] $P(A \cup B) \le P(A) + P(B)$. (*This follows directly from Theorem 1.5 since $P(A \cap B) \ge 0$.*)

**Theorem 1.6** *Let $A$ be an event in a sample space $S$. Then $P(A) \in [0,1]$.*

[10] Named after the English mathematician and logician George Boole.

**Proof** $\emptyset \subset A$ implies $P(\emptyset) \leq P(A)$ by Theorem 1.3, and $A \subset S$, implies $P(A) \leq P(S)$ by Theorem 1.3. Since $P(S) = 1$ by Axiom 1.2, and $P(\emptyset) = 0$ by Theorem 1.2, we have $0 \leq P(A) \leq 1$. ■

**Theorem 1.7** (Bonferroni's Inequality (2-event case))[11] *Let $A$ and $B$ be two events in a sample space $S$. Then $P(A \cap B) \geq 1 - P(\bar{A}) - P(\bar{B})$.*

**Proof** By Theorem 1.1, $P(A \cap B) = 1 - P(\overline{A \cap B})$. DeMorgan's law indicates that $\overline{A \cap B} = \bar{A} \cup \bar{B}$, and thus $P(A \cap B) = 1 - P(\bar{A} \cup \bar{B})$ by substitution. Theorem 1.5 indicates that $P(\bar{A} \cup \bar{B}) = P(\bar{A}) + P(\bar{B}) - P(\bar{A} \cap \bar{B})$, and thus $P(A \cap B) = 1 - [P(\bar{A}) + P(\bar{B}) - P(\bar{A} \cap \bar{B})] = 1 - P(\bar{A}) - P(\bar{B}) + P(\bar{A} \cap \bar{B})$, again by substitution. Finally, since $P(\bar{A} \cap \bar{B}) \geq 0$ by Axiom 1.1, we have $P(A \cap B) \geq 1 - P(\bar{A}) - P(\bar{B})$. ■

**Theorem 1.8** (Bonferroni's Inequality (General)) *Let $A_1, \ldots, A_n$ be events in a sample space $S$. Then*

$$P\left(\bigcap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} P(\bar{A}_i).$$

**Proof** We have already proven the validity of the proposition when $n = 2$ by Theorem 1.7. Suppose for purposes of invoking the induction principle (recall Lemma 1.1) that

$$P\left(\bigcap_{i=1}^{k} A_i\right) \geq 1 - \sum_{i=1}^{k} P(\bar{A}_i)$$

is true. Using Theorem 1.7, we know that

$$P\left(\bigcap_{i=1}^{k+1} A_i\right) = P\left(\left(\bigcap_{i=1}^{k} A_i\right) \cap A_{k+1}\right) \geq 1 - P\left(\left(\overline{\bigcap_{i=1}^{k} A_i}\right)\right) - P\left(\bar{A}_{k+1}\right).$$

Theorem 1.1 allows us to rewrite the inequality as

$$P\left(\bigcap_{i=1}^{k+1} A_i\right) \geq P\left(\bigcap_{i=1}^{k} A_i\right) - P\left(\bar{A}_{k+1}\right).$$

By the assumption that the Bonferroni inequality is valid for $k$ events, we conclude that

$$1 - \sum_{i=1}^{k} P(\bar{A}_i) \leq P\left(\bigcap_{i=1}^{k} A_i\right),$$

and thus we may write

$$P\left(\bigcap_{i=1}^{k+1} A_i\right) \geq 1 - \sum_{i=1}^{k} P(\bar{A}_i) - P(\bar{A}_{k+1}),$$

[11] Named for the Italian mathematician C. E. Bonferroni.

which implies that Bonferroni's inequality is valid for $k+1$ events. The proposition in Theorem 1.8 is thus proved by induction. ■

**Theorem 1.9** (Classical Probability) *Let $S$ be the finite sample space for an experiment having $N(S)$ equally likely outcomes, and let $A \subset S$ be an event containing $N(A)$ elements. Then the probability of the event $A$ is given by $N(A)/N(S)$.*

**Proof** Let $E_1, \ldots, E_n$ represent the $n = N(S)$ outcomes (or elementary events) in the sample space $S$. Since all outcomes are equally likely, $P(E_1) = \ldots = P(E_n) = k$, and since the outcomes are mutually exclusive and $S = \left(\cup_{i=1}^{n} E_i\right)$, we have by Axioms 1.2 and 1.3 that $P(S) = \sum_{i=1}^{n} P(E_i) = \sum_{i=1}^{n} k = nk = 1$. It follows that $P(E_i) = k = 1/n$, for $i = 1, \ldots, n$. Let $I \subset \{1, 2, \ldots, n\}$ be the index set identifying the $N(A)$ number of outcomes (or elementary events) that define the event $A$, i.e., $A = \cup_{i \in I} E_i$. Then by Axiom 1.3,

$$P(A) = \sum_{i \in I} P(E_i) = \sum_{i \in I} \frac{1}{n} = \frac{N(A)}{n} = \frac{N(A)}{N(S)}. \qquad ■$$

By proving Theorem 1.9, we have shown that the classical probability definition is implied by the axiomatic definition of probability. Thus, whenever the conditions of the classical probability definition apply, we are free to follow the classical prescription for assigning probabilities to events. It can also be shown that the relative frequency definition of probability is implied by the axiomatic definition. Among other things, this implies that the axiomatic foundation for probability theory provides the rationale for the existence of the limit of relative frequencies referred to in the relative frequency definition of probability. We will need to develop results relating to asymptotic theory (Chapter 5) before a proof of this proposition is attempted. Finally, the subjective probability definition is implied by the axiomatic definition in the sense that the individual assigning subjective probabilities to events will be *required* to adhere to the axioms in making those assignments. The requirement is interpreted by subjective probabilists as a *consistency* condition for subjective probability assignments.

The next example illustrates the use of the preceding probability theorems (for an example of Theorem 1.9, recall Ex. 1.3).

**Example 1.8** An envelope manufacturing firm has three envelope machines. The probability that one or more of the machines must be shut down for repairs on any given day is characterized by the following probability set function defined for $A \subset S = \times_{i=1}^{3}\{0, 1\}$,

$$\begin{aligned} P(A) &= \sum_{(x_1,x_2,x_3) \in A} f(x_1, x_2, x_3) \\ &= \sum_{(x_1,x_2,x_3) \in A} (.1)^{x_1}(.9)^{1-x_1}(.05)^{x_2}(.95)^{1-x_2}(.2)^{x_3}(.8)^{1-x_3}, \end{aligned}$$

where $x_i = 0$ denotes that the $i$th machine does not have a breakdown during the day and $x_i = 1$ denotes a breakdown of the $i$th machine. The sample space for this problem is given by $S = \{(x_1, x_2, x_3) : x_i = 0 \text{ or } 1, i = 1, 2, 3\}$.

The elementary events collectively represent all of the possible shutdown possibilities for the three machines on a given day. Examine the following four events:

$$A_1 = \{(0,0,0), (1,0,0), (0,1,0), (0,0,1)\},$$

$$A_2 = \{(0,0,1), (1,0,1), (0,1,1), (1,1,1)\},$$

$$A_3 = \{(1,1,1)\},$$

$$A_4 = S - \{(1,1,1)\} = \bar{A}_3.$$

$A_1$ is the event that at most one machine breaks down on a given day; $A_2$ is the event that machine 3 breaks down on a given day; $A_3$ is the event that all three machines suffer breakdowns on a given day; and $A_4$ is the event that *at least* one machine does not suffer a breakdown on a given day. Note (please verify using the definition of $P$) that $P(A_1) = .967$, $P(A_2) = .2$, $P(A_3) = .001$, and $P(A_4) = .999$.

*Theorem 1.1* The probability that more than one machine breaks down on a given day is $P(\bar{A}_1) = 1 - P(A_1) = .033$. Note also that $P(A_3) = 1 - P(A_4) = .001$, which is as it should be, since $\bar{A}_3 = A_4$.

*Theorem 1.2* Examine the event $A_1 \cap A_3 = \emptyset$ (i.e., the event that all three machines break down and less than two machines break down on a given day). Then $P(A_1 \cap A_3) = P(\emptyset) = 0$, since $\sum_{(x_1,x_2,x_3)\in\emptyset} f(x_1, x_2, x_3) = 0$ by definition.

*Theorem 1.3* Note that $A_3 \subset A_2$. Theorem 1.3 then applies to these two events, where $P(A_3) = .001 \leq .200 = P(A_2)$. Now note that $(A_2 - A_3) = \{(0,0,1), (1,0,1), (0,1,1)\}$, so that $P(A_2 - A_3) = .199 = .2 - .001 = P(A_2) - P(A_3)$.

*Theorem 1.4* Note that $(A_1 \cap A_2) = \{(0,0,1)\}$ and $(A_1 \cap \bar{A}_2) = \{(0,0,0), (1,0,0), (0,1,0)\}$, and thus $P(A_1) = .967 = .171 + .796 = P(A_1 \cap A_2) + P(A_1 \cap \bar{A}_2)$.

*Theorem 1.5 and Corollary* Note that $(A_1 \cup A_2) = S - \{(1,1,0)\}$ and thus $P(A_1 \cup A_2) = .996 = .967 + .200 - .171 = P(A_1) + P(A_2) - P(A_1 \cap A_2)$. Also, $(A_2 \cup A_3) = A_2$, and Boole's inequality holds for these two events as $P(A_2 \cup A_3) = .2 \leq .2 + .001 \leq P(A_2) + P(A_3)$.

*Theorem 1.6* Note that $P(A_i) \in [0, 1]$ for any subset contained in $S$, given the definition of the set function $P$.

*Theorem 1.7* Note that $P(A_1 \cap A_2) = .171 \geq .167 = 1 - .033 - .800 = 1 - P(\bar{A}_1) - P(\bar{A}_2)$.

*Theorem 1.8* Note that $A_1 \cap \bar{A}_2 \cap \bar{A}_3 = \{(0,0,0),(1,0,0),(0,1,0)\}$, so that $P(A_1 \cap \bar{A}_2 \cap \bar{A}_3) = .796 \geq .766 = 1 - .033 - .2 - .001 = 1 - P(\bar{A}_1) - P(A_2) - P(A_3)$. □

## 1.6 A Digression on Events

When we first began our discussion of the axiomatic approach to the definition of probability, we stated that probability is generated by a set function whose domain consists of all of the events in a sample space. This collection of "all of the events in a sample space" was termed the *event space*. However, up to now we have not discussed which subsets of a sample space are events and which, if any, are not events. That is, we have said that events are subsets of a sample space, but we did *not* say that *all* subsets of a sample space are events. The issue here is whether a set function can have a domain (i.e, the event space) consisting of *all* of the subsets of a sample space and still adhere to the three axioms of probability.

In the case of a countable sample space, the event space *can* consist of all of the subsets of the sample space and still have the set function exhibit the properties required by the probability axioms. Henceforth, whenever we are dealing with a countable sample space, the event space will always be defined as the collection of all of the subsets of the sample space, unless explicitly stated otherwise.

The situation is more complicated in the case of an uncountably infinite sample space. In this case, the collection of all subsets of $S$ is, in a sense, so large that a set function cannot have this collection of sets for its domain and still have the probability axioms hold true for all possible applications. The problem is addressed in a field of mathematics called "measure theory" and is beyond the scope of our study. As a practical matter, essentially any subset of $S$ that will be of interest in real-world applications will be an event and thus will be in the domain of the probability set function. Put another way, the subsets of $S$ that are not in the event space are by definition so complicated that they will not be of interest in any real-world application.

While it takes a great deal of ingenuity to define a subset of an uncountably infinite sample space that is not an event, the reader may still desire a more precise definition of *event space* in the case of an uncountably infinite sample space so that one is certain to be referring to a collection of subsets of $S$ for which each subset can be assigned a probability. This can be done relatively straightforwardly so long as we restrict our attention to real-valued sample spaces (i.e., sample spaces whose sample points are all real numbers). Since we can always "code" the elements of a sample space with real numbers (this relates to the notion of random variables, which we address in the next chapter), restricting attention to real-valued sample spaces does not involve any loss of generality, and so we proceed on the assumption that $S \subset R^n$. Our characterization depends on the notion of Borel sets,[12] which we define next. The definition uses the concept of rectangles in $R^n$, which are generalizations of intervals.

**Definition 1.8**
***Rectangles in $R^n$***

Rectangles in $R^n$ are sets of points in $R^n$ defined as[13]

**a.** *closed rectangle*: $\{(x_1, \ldots, x_n) : a_i \leq x_i \leq b_i, i = 1, \ldots, n\}$

**b.** *open rectangle*: $\{(x_1, \ldots, x_n) : a_i < x_i < b_i, i = 1, \ldots, n\}$

**c.** *half-open/half-closed rectangle*:

$\{(x_1, \ldots, x_n) : a_i < x_i \leq b_i, i = 1, \ldots, n\}$

$\{(x_1, \ldots, x_n) : a_i \leq x_i < b_i, i = 1, \ldots, n\}$

where the $a_i$'s and $b_i$'s are real numbers, with $-\infty$ or $\infty$ being admissible for strong inequalities. Clearly, rectangles are intervals when $n = 1$.

The collection of Borel sets contained in a sample space $S$ will include all of the rectangle subsets of $S$ as well as an infinite number of other sets that can be formed from them via set operations as defined below.

**Definition 1.9**
***Borel sets in S***

Let $S \subset R^n$. The collection of Borel sets in $S$ consists of all closed, open, and half-open/half-closed rectangles contained in $S$, as well as any other set that can be defined by applying a countable number of union, intersection, and/or complement operations to these rectangles.[14]

The collection of Borel sets is extremely large and will contain any subset of the real-valued sample space that will be of interest in real-world applications. In particular, *all* open and *all* closed (rectangular or nonrectangular) sets are contained in the collection of Borel sets. Most importantly, probabilities can always be assigned to Borel sets. Consequently, we will use the collection of Borel sets identified in Definition 1.9 as our event space when dealing with real-valued sample spaces. The reader should attempt to convince herself that in the case of a discrete, real-valued sample space, the collection of Borel sets is equivalently the collection of all subsets of the sample space.

[12] Named after the French mathematician Emile Borel.

[13] One can also define rectangles that are represented as Cartesian products of any collection of closed, open, *and/or* half-open/half-closed intervals, rather than as Cartesian products of only closed intervals, *or* open intervals, *or* half-open/half-closed intervals as in the definition. These might also be referred to as nonopen/nonclosed rectangles.

[14] The collection of Borel sets in $S$ is an example of what is known in the literature as a **sigma-field** ($\sigma$-field), or a sigma-algebra ($\sigma$-algebra). A $\sigma$-field is a nonempty set of sets that is closed under countable union, intersection, and complement operations. The use of the word "closed" here means that if $A_i$, $i \in I$, all belong to the $\sigma$-field, any set formed by applying a countable number of unions, intersections, and/or complement operations to the $A_i$'s is also a set that belongs to the $\sigma$-field, where $I$ is any countable index set.

The following example illustrates the concept of Borel sets in $R^1$.

**Example 1.9** The following sets are examples of Borel sets in $R : A_1 = [0, .5]$, $A_2 = [.75, \infty)$, $A_3 = \{.6\}$, $A_1 \cup A_2$, $A_2 \cup A_3$, $A_1 \cup A_2 \cup A_3$, $\emptyset = A_1 \cap A_2 = A_2 \cap A_3 = A_1 \cap A_2 \cap A_3$, $\bar{A}_1, \bar{A}_2, \bar{A}_3$, $\overline{A_1 \cup A_2}$, $\overline{A_2 \cup A_3}$, $\overline{A_1 \cup A_2 \cup A_3}$, $S$. The reader can define other Borel sets in $R$ at her leisure, *ad infinitum*. In words, the collection of Borel sets in $R$ is the collection of all intervals in $R$ as well as any set that can be formed from the intervals by a countable number of union, intersection, and/or complement operations. □

To summarize our disposition concerning the meaning of *event space*, our general definition is still "the collection of events in a sample space." More specifically, however, it is the collection of *all* subsets of a sample space when the sample space is discrete, but it is "only" the collection of Borel sets in the case where the sample space is uncountably infinite and real valued.

## 1.7 Conditional Probability

When an experiment is conducted, the one event that is absolutely certain to occur is the event $S$, since the outcome of an experiment must be an element of the sample space. We now study the effect that additional information concerning the outcome of an experiment has on the probability of events. In particular, if it is known that the outcome of the experiment is an element of some subset, $B$, of the sample space, what is the effect of this additional information on the probabilities of events in $S$? As an example of such a situation, it appears intuitively plausible that the probability of a company earning \$10 million in annual profits would be higher if the company were randomly chosen from the list of Fortune 500 companies than if the company were chosen from among all companies in the United States. For another example, examine the experiment of tossing two fair coins in succession, and let the sample space for the experiment be defined by $S = \{(H, H), (H, T), (T, H), (T, T)\}$, where $H$ =heads and $T$ =tails. The probability (unconditional) of observing two tails is 1/4 since $P((T, T)) = N(A)/N(S) = 1/4$, where $A = \{(T, T)\}$. However, the probability of observing two tails must be zero if it is known that the outcome of the first coin toss was heads. We develop the notion of conditional probability ahead.

Suppose that we are analyzing an experiment with an associated probability space $\{S, \Upsilon, P\}$ and it is *given* that the outcome of the experiment is some element of a subset, $B$, of the sample space. How should the probability of an event, $A$, be defined given the additional information that event $B$ has occurred? By making a number of observations concerning properties that conditional probabilities should possess, we will be led to a definition of the conditional probability of event $A$ given event $B$.

First of all, since it is given that $B$ occurs, it is certain that $\bar{B}$ will not occur. In effect, the sample space has been reduced to the subset $B$, i.e., the outcomes in $S - B$ are no longer relevant, and $B$ can be interpreted as the **conditional**

**sample space**. Letting the symbol $P(A \mid B)$ represent **the conditional probability of event *A*, given event *B***, it follows that since the new sample space $B$ is an event that is now certain to occur, the *conditional* probability assigned to event $B$ should be 1, so that $P(B \mid B) = 1$. Note further that since $B$ will occur, it is clear that an event $A$ can also occur iff $A$ occurs concurrently with $B$, that is, iff $A \cap B$ occurs (see Figure 1.1). This suggests that conditional probability should be defined so that $P(A \mid B) = P(A \cap B \mid B)$ for any event $A$.

Now note that Theorem 1.4 implies the probability equation $P(B) = P(A \cap B) + P(\bar{A} \cap B)$, since $B$ can be partitioned into the two disjoint subsets $A \cap B$ and $\bar{A} \cap B$ because $B = (A \cap B) \cup (\bar{A} \cap B)$. Thus, the probability of event $B$ can be decomposed into the probabilities of the events $(A \cap B)$ and $(\bar{A} \cap B)$. Dividing both sides of the probability equation by $P(B)$ (assuming $P(B) \neq 0$) obtains a proportional decomposition of the probability of event $B$ as

$$1 = \frac{P(A \cap B)}{P(B)} + \frac{P(\bar{A} \cap B)}{P(B)},$$

where the proportion $k = P(A \cap B)/P(B) \in [0, 1]$ of event $B$'s probability is attributable to event $A \cap B$ with the remaining proportion, $1 - k$, attributable to event $\bar{A} \cap B$. Restricting the sample space to the set $B$ or, equivalently, conditioning on event $B$ does not eliminate any of the outcomes in the collection comprising either of the events $A \cap B$ or $\bar{A} \cap B$, nor does it change the fact that $B = (A \cap B) \cup (\bar{A} \cap B)$. Then, since event $A \cap B$ accounts for a proportion, $k$, of the probability that event $B$ will occur, and since $B$ is now certain to occur and so is assigned (conditional) probability $P(B \mid B) = 1$, then the proportion, $k$, of this unit probability should be attributable to event $A \cap B$. We thus assign $P(A \cap B \mid B) = k = P(A \cap B)/P(B)$, and since we also require that $P(A \mid B) = P(A \cap B \mid B)$ we are led to the following definition of conditional probability.

**Definition 1.10**
***Conditional probability***

Let $A$ and $B$ be any two events in a sample space $S$. If $P(B) \neq 0$, then the conditional probability of event $A$, given event $B$, is given by $P(A \mid B) = P(A \cap B)/P(B)$.

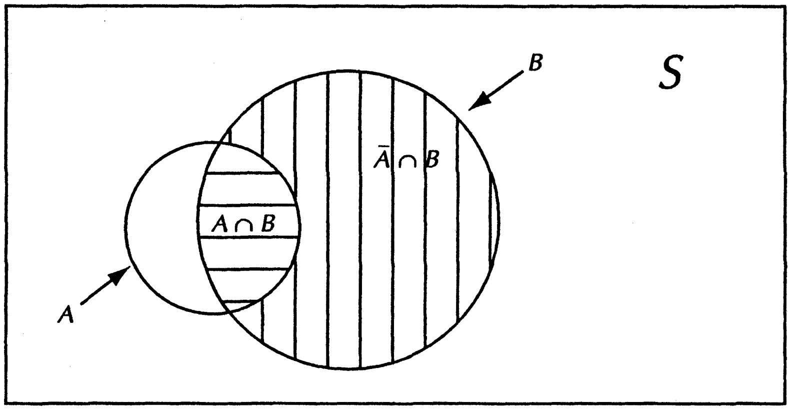


**Figure 1-1**
Conditional sample space $B$.

Our intuition concerning the meaning of conditional probability can be enhanced by examining the definition in light of the classical and relative frequency definitions of probability. In an experiment for which classical probability is applicable, the probability space consists of

1. a finite sample space,
2. an event space that consists of all subsets of the sample space, and
3. a probability set function that assigns probability to an event $A \subset S$ as $P(A) = N(A)/N(S)$.

Conditioning on an event $B \subset S$, the probability of an event $A$, by the definition of conditional probability, is given by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{N(A \cap B)/N(S)}{N(B)/N(S)} = \frac{N(A \cap B)}{N(B)}.$$

Since all outcomes are equally likely in this case, and since we are effectively restricting the sample space to the set $B$ by conditioning on event $B$, it stands to reason that the probability of observing $A$ is given by the number of outcomes in $B$ that result in $A$'s occurring relative to the total number of outcomes in $B$. This is, of course, consistent with the classical probability definition applied to the event $A$ in the context of the new sample space, $B$.

Regarding the relative frequency definition and conditional probability, recall that the probability set function assigns probabilities to events via $P(A) = \lim_{n\to\infty}(n_A/n)$. Conditioning on event $B \subset S$, the probability of event $A$, by the definition of conditional probability, is given by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\lim\limits_{n\to\infty}(n_{A\cap B}/n)}{\lim\limits_{n\to\infty}(n_B/n)} = \lim_{n\to\infty}(n_{A\cap B}/n_B).$$

Note the last equality follows from the fact that the limit of a ratio (here, the ratio is $(n_{A\cap B}/n)/(n_B/n) = n_{A\cap B}/n_B$) equals the ratio of the limits if all limits exist and if the limit in the denominator of the ratio of limits is not zero.[15] Restricting the sample space to the set $B$ by conditioning on event $B$, $P(A \mid B)$ is seen to equal the limiting fraction of the number of occurrences of event $B$ that also result in the occurrence of $A$. Consistent with the logic of the relative frequency definition of probability, $P(A \mid B)$ could then be interpreted as the limit of the frequency of observing event $A$ relative to the total number of outcomes generated from the conditional sample space $B$.

It is clear that conditional probabilities are values of a set function, since unique real numbers are being assigned to subsets of $S$. That these are values of a *probability* set function is established in the following theorem.

[15] See R. C. Buck, (1978), *Advanced Calculus*, 3rd ed., New York: McGraw-Hill, p. 44. We will discuss the concept of limits in more detail in Chapter 5. For now, a more intuitive understanding of limits is sufficient.

**Theorem 1.10** *Given a probability space $\{S,\Upsilon,P\}$ and an event $B$ for which $P(B) \neq 0$, $P(A \mid B) = P(A \cap B)/P(B)$ defines a probability set function with domain $\Upsilon$.*

**Proof** To prove the theorem, we need to establish that the set function defined by $P(A \mid B) = P(A \cap B)/P(B)$ adheres to the three axioms of probability on the domain $\Upsilon$. First, for any event $A$, it is clear that $P(A \mid B) \geq 0$, since $P(A \cap B) \geq 0$ and $P(B) > 0$. Second, $P(S \mid B) = 1$, since $P(S \mid B) = P(S \cap B)/P(B) = P(B)/P(B) = 1$. Finally, if $\{A_i : i \in I\}$ is a finite or countably infinite collection of disjoint events contained in $S$, then

$$\begin{aligned} P\left(\bigcup_{i\in I} A_i \mid B\right) &= P\left(\left(\bigcup_{i\in I} A_i\right) \cap B\right) / P(B) \\ &= P\left(\bigcup_{i\in I} (A_i \cap B)\right) / P(B) \quad \text{(since } \cap \text{ is distributive)} \\ &= \sum_{i\in I} P(A_i \cap B)/P(B) \quad \text{(since } (A_i \cap B) \cap (A_j \cap B) = \emptyset \text{ for } i \neq j) \\ &= \sum_{i\in I} P(A_i \mid B) \quad \text{(by definition of conditional probability),} \end{aligned}$$

and thus the third axiom of probability holds for the set function. ■

The conditional probability set function $P(\cdot \mid B)$ can be used to define two new probability spaces, $\{S, \Upsilon, P(\cdot \mid B)\}$ and $\{B, \Upsilon_B, P(\cdot \mid B)\}$, where $\Upsilon_B$ is the collection of all events contained in the set $B$. The validity of the first new probability space follows directly from Theorem 1.10. With minor changes, the reader can apply the proof of Theorem 1.10 to justify the second probability space. It follows that conditional *probabilities* can be legitimately assigned by $P(\cdot \mid B)$ to all of the events in the original sample space $S$ as well as to all of the events in the conditional sample space $B$.

Since $P(A \mid B)$ adheres to the probability axioms, all of the theorems that were proved for unconditional probabilities apply equally well to conditional probabilities.[16] This follows because the theorems are valid when applied to any set function that adheres to the probability axioms regardless of whether the set function is representing unconditional or conditional probabilities. Put another way, the proofs of all of the theorems would apply analogously to conditional probabilities by simply changing $P(\cdot)$ to $P(\cdot \mid D)$, say, in the proofs and recognizing that the probability axioms apply to the set function $P(\cdot \mid D)$, as demonstrated in the proof of Theorem 1.10. Note that we use the letter $D$ here to allow for the possibility that the event being conditioned on is different than the events, $A$ and $B$, that are referred to in some of the previous probability theorems. For convenience, we list the probability theorems below as they apply

[16] Note that, in a sense, all probabilities could be viewed as *conditional*, where $P(A \mid S)$ could be used to denote probabilities in previous sections. We will continue to use "unconditional" to refer to the case where the original sample space, $S$, has been left "unconditioned."

to conditional probabilities. It is assumed that the conditional probability is defined, i.e., $P(D) \neq 0$, and that $A$ and/or $B$ are events in either the conditional sample space $D$ or the original sample space $S$.

**Theorem 1.1$^c$** $P(A \mid D) = 1 - P(\bar{A} \mid D)$.

**Theorem 1.2$^c$** $P(\emptyset \mid D) = 0$.

**Theorem 1.3$^c$** *If* $A \subset B$, *then* $P(A \mid D) \leq P(B \mid D)$ *and* $P(B - A \mid D) = P(B \mid D) - P(A \mid D)$.

**Theorem 1.4$^c$** $P(A \mid D) = P(A \cap B \mid D) + P(A \cap \bar{B} \mid D)$.

**Theorem 1.5$^c$** $P(A \cup B \mid D) = P(A \mid D) + P(B \mid D) - P(A \cap B \mid D)$.

**Corollary 1.1$^c$** $P(A \cup B \mid D) \leq P(A \mid D) + P(B \mid D)$.

**Theorem 1.6$^c$** $P(A \mid D) \in [0,1]$.

**Theorem 1.7$^c$** $P(A \cap B \mid D) \geq 1 - P(\bar{A} \mid D) - P(\bar{B} \mid D)$.

**Theorem 1.8$^c$** $P(\cap_{i=1}^{n} A_i \mid D) \geq 1 - \sum_{i=1}^{n} P(\bar{A}_i \mid D)$.

The following examples illustrate the application of conditional probability.

**Example 1.10** Consider the experiments of tossing two coins in succession, and let $S = \{(H,H),(H,T),(T,H),(T,T)\}$, where $H$ =heads and $T$ =tails. Assume that all outcomes are equally likely.

**a.** What is the probability of obtaining two heads, given that the first coin toss came up heads?
**Answer**: $B = \{(H,H),(H,T)\}$ is the event that the first coin toss results in heads, which is our conditional sample space. $A = \{(H,H)\}$ is the event of obtaining two heads. Then $P(A \mid B) = P(A \cap B)/P(B) = (1/4)/(1/2) = 1/2$.

**b.** What is the probability of obtaining two heads, given that at least one of the coins comes up heads?
**Answer**: $C = \{(H,H),(H,T),(T,H)\}$ is the event that at least one of the coins came up heads, which is our conditional sample space. Then $P(A \mid C) = P(A \cap C)/P(C) = (1/4)/(3/4) = 1/3$.

**c.** What is the probability of obtaining one heads and one tails given that the first coin is tails?
**Answer**: $D = \{(T,H),(T,T)\}$ is the event that the first coin was tails. $E = \{(T,H),(H,T)\}$ is the event of obtaining one heads and one tails. Then $P(E \mid D) = P(E \cap D)/P(D) = (1/4)/(1/2) = 1/2$. □

**Example 1.11** A perplexed investor must choose an investment instrument from among 15 different stocks, 10 different bonds, and 5 different mutual funds. Allowing each instrument an equal probability of being chosen, the investor randomly

chooses an instrument. Given that the chosen instrument was not a bond, what is the probability that a stock was chosen?
**Answer**: We are conditioning on the event, $B$, that the instrument is either a stock or mutual fund (i.e., not a bond), and $P(B) = 2/3$. Let $A$ represent the event that the outcome is a stock, so that $P(A \cap B) = 1/2$. Then $P(A \mid B) = P(A \cap B)/P(B) = (1/2)/(2/3) = 3/4$ is the probability we seek. Note this makes sense from the standpoint that the conditional sample space is 20 investment instruments, of which 15 are stocks and 5 are mutual funds. The classical probability definition suggests that the probability of observing a stock in this sample space of 20 instruments is $15/20 = 3/4$. □

The definition of conditional probability can be transformed to obtain a result known as the **multiplication rule**. The multiplication rule allows one to calculate the probability of event $A \cap B$ from the knowledge of the conditional probability of event $A$, given event $B$, and the unconditional probability of $B$.

**Theorem 1.11** (Multiplication Rule) *Let $A$ and $B$ be any two events in the sample space for which $P(B) \neq 0$. Then $P(A \cap B) = P(A \mid B)P(B)$.*

**Proof** Multiply both sides of $P(A \mid B) = P(A \cap B)/P(B)$ in Definition 1.10 by $P(B)$. ■

The multiplication rule is especially useful in cases where an experiment can be interpreted as being conducted in two stages.

**Example 1.12.** What is the probability of drawing two aces in succession from a well-shuffled deck of poker cards? Assume cards drawn are not replaced in the deck.
**Answer**: Let $B$ be the event that the first card drawn is an ace. Since there are 4 aces in a poker deck, with a total of 52 cards in the deck, $P(B) = 4/52 = 1/13$. Now let $A$ be the event that the second card drawn is an ace. Given that the first card drawn is an ace (i.e., given event $B$), there are 3 aces remaining to be chosen from the remaining 51 cards, and thus the probability that the second draw is an ace, given that the first card drawn is an ace, i.e., $P(A \mid B)$, equals $3/51 = 1/17$. Then, by the multiplication rule, the probability that both draws result in aces is given by $P(A \cap B) = P(A \mid B)P(B) = (1/17)(1/13) = 1/221$. □

**Example 1.13** As part of its quality-control program, an apparel manufacturer has inspectors examine every garment the company produces. A garment is shipped to a retail outlet only if it passes inspection. The probability that a garment is defective is .02. The probability that an inspector assigns a "pass" to a defective garment is .05. What is the probability that a garment is defective and shipped to a retail outlet?
**Answer**: Let $D$ be the event that a garment is defective. Let $B$ be the event that the inspector assigns a "pass" to a garment. We know that $P(D) = .02$ and $P(B \mid D) = .05$. $B \cap D$ is the event that a garment is defective and is passed by the inspector. Then $P(B \cap D) = P(B \mid D)P(D) = (.05)(.02) = .001$. □

The multiplication rule can be extended to three or more events, in which case the probability of the intersection of all of the events can be represented as follows.

**Theorem 1.12** (Extended Multiplication Rule) *Let $A_1, A_2, \ldots, A_n$, $n \geq 2$, be events in the sample space. Then if all of the conditional probabilities exist,*

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1)\prod_{i=2}^{n} P\left(A_i \mid \bigcap_{j=1}^{i-1} A_j\right).$$

**Proof** We know from Theorem 1.11 that the result holds for $n = 2$ (note that $\cap_{j=1}^{1} A_j = A_1$ by definition). In an attempt to invoke the induction principle, *assume* that the result is true for $n = k$, where $k$ is some arbitrary positive integer $\geq 3$. We will show that the result is then also true for $n = k + 1$. Let $B = \cap_{i=1}^{k} A_i$. Then

$$\begin{aligned}
P\left(\bigcap_{i=1}^{k+1} A_i\right) &= P(A_{k+1} \cap B) = P(A_{k+1} \mid B)P(B) \quad \text{(Theorem 1.11)} \\
&= \underbrace{P(A_1)\prod_{i=2}^{k} P\left(A_i \mid \bigcap_{j=1}^{i-1} A_j\right)}_{P(B)} P(A_{k+1} \mid B) \quad \text{assuming result holds for } n = k \\
&= P(A_1)\prod_{i=2}^{k+1} P\left(A_i \mid \bigcap_{j=1}^{i-1} A_j\right) \quad \left(\text{substitution for } B = \bigcap_{i=1}^{k} A_i\right)
\end{aligned}$$

Thus, by mathematical induction, the theorem holds. (See Def. 1.12.) ■

Similar to the case of the multiplication rule for two events, the extended multiplication rule is especially useful in cases where an experiment can be interpreted as being conducted in $n$ stages.

**Example 1.14** What is the probability of drawing four aces in succession from a well-shuffled deck of poker cards? Assume cards drawn are not replaced in the deck.
**Answer:** Let $A_i$ be the event that the $i$th card drawn is an ace, $i = 1, 2, 3, 4$. Then, using Theorem 1.12,

$$\begin{aligned}
P\left(\bigcap_{i=1}^{4} A_i\right) &= P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2)P(A_4 \mid A_1 \cap A_2 \cap A_3) \\
&= \left(\frac{4}{52}\right)\left(\frac{3}{51}\right)\left(\frac{2}{50}\right)\left(\frac{1}{49}\right) = .3694 \times 10^{-5}.
\end{aligned}$$

□

**Example 1.15** Recall the garment inspection problem of Ex. 1.13. Suppose the retailers who market the garments of the apparel manufacturer also inspect each garment they purchase and place on sale only those for which they perceive no defects.

The probability that a retailer places a defective garment on sale is .10. What is the probability that a garment is defective, shipped to the retail outlet, and placed on sale by retailers?

**Answer**: Let $A$ be the event that the retailer places a garment on sale. $A \cap B \cap D$ is the event of interest, and by Theorem 1.12,

$$P(A \cap B \cap D) = P(D)P(B \mid D)P(A \mid B \cap D) = (.02)(.05)(.10) = .0001. \qquad \square$$

## 1.8 Independence

In everyday language, when one says that two events are *independent*, it is generally meant that the occurrence of one event does not affect the likelihood of an occurrence of the other event, and vice versa. There is a related notion of independence in the theory of probability. We begin with the technical definition of **independent events.**

**Definition 1.11**
***Independence of events (2-event case)***

Let $A$ and $B$ be two events in a sample space $S$. Then $A$ and $B$ are independent iff $P(A \cap B) = P(A)P(B)$. If $A$ and $B$ are not independent, $A$ and $B$ are said to be **dependent events.**

An intuitively appealing interpretation of the independence condition in Definition 1.11, which is closely allied to our layman's interpretation of independence, is available when $P(A) > 0$ and $P(B) > 0$. In this case, $P(A \cap B) = P(A)P(B)$ implies

$$P(A \mid B) = P(A \cap B)/P(B) = P(A)P(B)/P(B) = P(A),$$

$$P(B \mid A) = P(B \cap A)/P(A) = P(B)P(A)/P(A) = P(B).$$

Thus the probability of event $A$ occurring is unaffected by the occurrence of event $B$, and the probability of event $B$ occurring is unaffected by the occurrence of event $A$.

If event $A$ and/or event $B$ has probability zero, then *by definition*, events $A$ and $B$ are independent. This follows immediately from the fact that if either $P(A) = 0$ or $P(B) = 0$, then $P(A \cap B) = 0 = P(A)P(B)$ (since $(A \cap B) \subset A$ and $(A \cap B) \subset B$ imply both $P(A \cap B) \le P(A)$ and $P(A \cap B) \le P(B)$ by Theorem 1.3, and $P(A \cap B) \ge 0$ by Axiom 1.1, so that together the inequalities imply $P(A \cap B) = 0$), and thus the independence condition is fulfilled. However, in this case one or both conditional probabilities $P(A \mid B)$ and $P(B \mid A)$ are undefined, and the basis no longer exists for stating that "the independence of events $A$ and $B$ implies the probability of either event is unaffected by the occurrence of the other."

If $A$ and $B$ are independent, then $A$ and $\bar{B}$, $\bar{A}$ and $B$, and $\bar{A}$ and $\bar{B}$ are also independent. This result can be demonstrated by showing that the independence definition is satisfied for each of the preceding pairs of events if independence is satisfied for $A$ and $B$. We state the result as a theorem.

**Theorem 1.13** *If events $A$ and $B$ are independent, then events $A$ and $\bar{B}$, $\bar{A}$ and $B$, and $\bar{A}$ and $\bar{B}$ are also independent.*

**Proof**

$$\begin{aligned}
P(A \cap \bar{B}) &= P(A) - P(A \cap B) \quad \text{(Theorem 1.4)}\\
&= P(A) - P(A)P(B) \quad \text{(independence of } A \text{ and } B)\\
&= P(A)[1 - P(B)] \quad \text{(algebra)}\\
&= P(A)P(\bar{B}) \quad \text{(Theorem 1.1)}\\
P(\bar{A} \cap B) &= P(B) - P(A \cap B) \quad \text{(Theorem 1.4)}\\
&= P(B) - P(A)P(B) \quad \text{(independence of } A \text{ and } B)\\
&= P(B)[1 - P(A)] \quad \text{(algebra)}\\
&= P(\bar{A})P(B) \quad \text{(Theorem 1.1)}\\
P(\bar{A} \cap \bar{B}) &= P(\overline{A \cup B}) \quad \text{(DeMorgan's laws)}\\
&= 1 - P(A \cup B) \quad \text{(Theorem 1.1)}\\
&= 1 - (P(A) + P(B) - P(A \cap B)) \quad \text{(Theorem 1.5)}\\
&= 1 - P(A) - P(B) + P(A)P(B) \quad \text{(independence of } A \text{ and } B)\\
&= P(\bar{A}) - P(B)[1 - P(A)] \quad \text{(Theorem 1.1 and algebra)}\\
&= P(\bar{A})[1 - P(B)] \quad \text{(Theorem 1.1 and algebra)}\\
&= P(\bar{A})P(\bar{B}) \quad \text{(Theorem 1.1)}
\end{aligned}$$

■

The following example illustrates the concept of independence of events.

**Example 1.16** The work force of the Excelsior Corporation has the following distribution among type and gender of workers:

| | Type of Worker | | | |
|---|---|---|---|---|
| **Sex** | **Sales** | **Clerical** | **Production** | **Total** |
| Male | 825 | 675 | 750 | 2,250 |
| Female | 1,675 | 825 | 250 | 2,750 |
| Total | 2,500 | 1,500 | 1,000 | 5,000 |

In order to promote loyalty to the company, the company randomly chooses a worker to receive an all-expenses-paid vacation each month. Is the event of choosing a female independent of the event of choosing a clerical worker? **Answer**: Let $F$ =event of choosing a female and $C$ =event of choosing a clerical worker. From the data in the table, we know that $P(F) = .55$ and $P(C) = .30$. Also, $P(F \cap C) = .165$. Then $P(F \cap C) = P(F)P(C)$, and the events are *independent*.

Is the event of choosing a female independent of the event of choosing a production worker?

**Answer**: Let $A$ =event of choosing a production worker. Then $P(A) = .20$ and $P(A \cap F) = .05$, and thus $P(A \cap F) = .05 \neq .11 = P(A)P(F)$, so the events are *dependent*. □

The property that $A$ and $B$ are independent events is sometimes confused with the property that sets $A$ and $B$ are mutually exclusive or disjoint. It should be noted that the two properties are distinct, but related concepts. The mutually exclusive property is a property of sets, while the independence property is a property of the probability set function defined on a collection of sets. The following table presents the relationship between the two properties.

*Disjointness versus Independence*

| | $A \cap B = \emptyset$ | $A \cap B \neq \emptyset$ |
|---|---|---|
| $P(A) > 0$ and $P(B) > 0$ | Dependent | Independent if $P(A \cap B) = P(A)P(B)$ |
| $P(A)$ and/or $P(B) = 0$ | Independent | Independent |

Next we verify the three cases in which an immediate conclusion can be reached regarding the independence of events $A$ and $B$. We introduce the notation "$\Rightarrow$" which is to be read *implies that*.

**Theorem 1.14** (Independence and Disjointness)

1. *$P(A) > 0$, $P(B) > 0$, $A \cap B = \emptyset \Rightarrow A$ and $B$ are dependent.*
2. *$P(A)$ and/or $P(B) = 0$, $A \cap B = \emptyset \Rightarrow A$ and $B$ are independent.*
3. *$P(A)$ and/or $P(B) = 0$, $A \cap B \neq \emptyset \Rightarrow A$ and $B$ are independent.*

**Proof**

1. $P(A \cap B) = P(\emptyset) = 0$ by Theorem 1.2. Since $P(A \cap B) = 0 < P(A)P(B)$ because $P(A)$ and $P(B)$ are both positive, $A$ and $B$ cannot be independent events, and so they are dependent.
2. $P(A \cap B) = P(\emptyset) = 0$ by Theorem 1.2. Since $P(A)$ and/or $P(B) = 0$, then $P(A)P(B) = 0$, and thus $P(A \cap B) = 0 = P(A)P(B)$. Therefore, $A$ and $B$ are independent.
3. $P(A \cap B) \leq P(A)$ and $P(A \cap B) \leq P(B)$ by Theorem 1.3, since $(A \cap B) \subset A$ and $(A \cap B) \subset B$. If $P(A)$ and/or $P(B) = 0$, then $P(A \cap B) \leq 0$. $P(A \cap B) \geq 0$ by Axiom 1.1. Then $P(A \cap B) = 0 = P(A)P(B)$, and $A$ and $B$ are independent events. ■

The concept of independent events can be generalized to more than two events. The generalized statement utilizes product notation, which we review here.

**Definition 1.12**
***Product notation***

$$\prod_{i=1}^{n} a_i = a_1 a_2 \cdots a_n$$

$$\prod_{j \in J} a_j = \text{ product of all } a_j \text{ terms for which } j \in J.$$

**Definition 1.13**
***Independence of events (n-event case)***

Let $A_1, A_2, \ldots, A_n$ be events in the sample space $S$. The events $A_1, A_2, \ldots, A_n$ are independent iff

$$P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j),$$

for all subsets $J \subset \{1, 2, \ldots, n\}$ for which $N(J) \geq 2$. If the events $A_1, A_2, \ldots, A_n$ are not independent, they are said to be dependent events.

Note that some authors refer to the independence concept defined in Def. 1.13 as the **joint** or **mutual** or **complete independence** of the events $A_1, A_2, \ldots,$ $A_n$, when $n \geq 3$, to emphasize that additional conditions are required beyond the condition given in Def. 1.11 applicable to pairs of events. We will refrain from using these additional adjectives and simply refer to the *independence of events,* regardless of $n$. Furthermore, note that if the condition $P(A_i \cap A_j) = P(A_i)P(A_j)$ of Def. 1.11 applies to all pairs of events in $A_1, A_2, \ldots, A_n$, the events are referred to as being **pairwise independent** whether or not they are independent in the sense of Def. 1.13.

In the case of three events, independence requires that $P(A_1 \cap A_2) = P(A_1)P(A_2)$, $P(A_1 \cap A_3) = P(A_1)P(A_3)$, $P(A_2 \cap A_3) = P(A_2)P(A_3)$, and $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$. Note that if all three events have nonzero probability, then the reader can straightforwardly verify from the definition of conditional probability that independence implies $P(A_i \mid A_j) = P(A_i)\ \forall\ i \neq j$, $P(A_i \mid A_j \cap A_k) = P(A_i)\ \forall$ $i \neq j \neq k$, and $P(A_i \cap A_j \mid A_k) = P(A_i \cap A_j)\ \forall\ i \neq j \neq k$. It is not as straightforward to demonstrate that independence implies $P(A_i \mid A_j \cup A_k) = P(A_i)$ for $i \neq j \neq k$, and so we prove it here.

$$\begin{aligned}
P(A_i \mid A_j \cup A_k) &= \frac{P(A_i \cap (A_j \cup A_k))}{P(A_j \cup A_k)} \quad \text{(by definition)} \\
&= \frac{P((A_i \cap A_j) \cup (A_i \cap A_k))}{P(A_j \cup A_k)} \quad \text{(distributive law)} \\
&= \frac{P(A_i \cap A_j) + P(A_i \cap A_k) - P(A_i \cap A_j \cap A_k)}{P(A_j \cup A_k)} \quad \text{(Theorem 1.5)} \\
&= \frac{P(A_i)P(A_j) + P(A_i)P(A_k) - P(A_i)P(A_j)P(A_k)}{P(A_j \cup A_k)} \quad \text{(independence)}
\end{aligned}$$

$$= \frac{P(A_i)[P(A_j) + P(A_k) - P(A_j)P(A_k)]}{P(A_j \cup A_k)} \quad \text{(algebra)}$$

$$= \frac{P(A_i)P(A_j \cup A_k)}{P(A_j \cup A_k)} \quad \text{(Theorem 1.5 and independence)}$$

$$= P(A_i) \qquad \blacksquare$$

It is thus recognized that if events $A_1, A_2$, and $A_3$ are independent in the sense of Def. 1.13, and if each of the events occurs with nonzero probability, then the probability of any one of the events is unaffected by the occurrence of any of the remaining events and is also unaffected by the occurrence of the union or intersection of the remaining events. This interpretation extends in a straightforward way to cases involving four or more independent events, and the reader should attempt some of the possible extensions.

The reader may wonder whether the numerous conditions (for $n$ events, the number of conditions will be $2^n - n - 1$) cited in Def. 1.13 for independence of events are all necessary, i.e., wouldn't the one condition $P\left(\cap_{j=1}^n A_j\right) = \prod_{j=1}^n P(A_j)$ suffice and imply all of the others? Unfortunately, the answer is no—all of the conditions are required. The following example illustrates the point.

**Example 1.17** The Venn diagram in Figure 1.2 summarizes the probabilities assigned to some of the events in the sample space $S$. Note that $P(A \cap B \cap C) = .15 = (.5)(.6)(.5) = P(A)P(B)P(C)$. However,

$$P(A \cap B) = .25 \neq .3 = P(A)P(B),$$

$$P(A \cap C) = .20 \neq .25 = P(A)P(C),$$

$$P(B \cap C) = .40 \neq .3 = P(B)P(C),$$

and thus $P(A \cap B \cap C) = P(A)P(B)P(C)$ does *not* imply the pairwise independence conditions. □

The reader should construct an example illustrating that pairwise independence among $A$, $B$, and $C$ does *not* imply $P(A \cap B \cap C) = P(A)P(B)P(C)$.

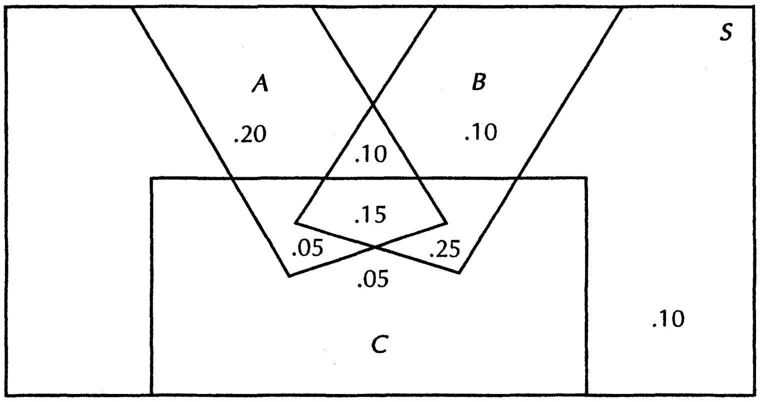


**Figure 1-2** Probability assignments in $S$.

## 1.9 Bayes's Rule

Bayes's rule, discovered by the 17th-century English clergyman and mathematician Thomas Bayes, provides an alternative representation of conditional probabilities. At first glance, the representation may appear somewhat convoluted in comparison to the representation of conditional probability given in Def. 1.10. However, the rule is well suited for providing conditional probabilities in certain experimental situations, which we will identify following the formal derivation of the rule.

Bayes's rule is actually a simple consequence of a result known as the theorem of total probability, which we state and prove next.

**Theorem 1.15** (Theorem of Total Probability) *Let the events $B_i$, $i \in I$, be a finite or countably infinite partition of the sample space, $S$, so that $B_j \cap B_k = \emptyset$ for $j \neq k$, and $\cup_{i \in I} B_i = S$. Let $P(B_i) > 0 \; \forall \; i$. Then $P(A) = \sum_{i \in I} P(A \mid B_i) P(B_i)$.*

**Proof** Since $\cup_{i \in I} B_i = S$, it follows that $A = A \cap (\cup_{i \in I} B_i) = \cup_{i \in I} (A \cap B_i)$, where we have used the fact that the intersection operation is distributive over the union operation. Now note that $(A \cap B_j) \cap (A \cap B_k) = \emptyset$ for $j \neq k$ since the $B_i$'s are disjoint. But then by Axiom 1.3, $P(A) = P(\cup_{i \in I}(A \cap B_i)) = \sum_{i \in I} P(A \cap B_i)$. The result of the theorem follows applying the multiplication rule to each term, $P(A \cap B_i)$, in the sum. ■

Regarding the title of Theorem 1.15, note that the "total probability" of event $A$ is represented as the sum of the portions of $A$'s probability distributed over the events in the partition of $A$ represented by the events $A \cap B_i$, $i \in I$. We state Bayes's rule as a corollary to Theorem 1.15.

**Corollary 1.2** (Bayes's Rule) *Let the events $B_i$, $i \in I$, be a finite or countably infinite partition of the sample space, $S$, so that $B_j \cap B_k = \emptyset$ for $j \neq k$ and $\cup_{i \in I} B_i = S$. Let $P(B_i) > 0$ $\forall \; i \in I$. Then, provided $P(A) \neq 0$,*

$$P(B_j \mid A) = \frac{P(A \mid B_j) P(B_j)}{\sum_{i \in I} P(A \mid B_i) P(B_i)}, \quad \forall \; j \in I.$$

**Proof** This follows directly the definition of the conditional probability $P(B_j \mid A) = P(A \cap B_j)/P(A)$, after substituting for the denominator using Theorem 1.15 and rewriting the numerator using the multiplication rule. ■

**Corollary 1.3** (Bayes's Rule (2-event case))

$$P(B \mid A) = \frac{P(A \mid B) P(B)}{P(A \mid B) P(B) + P(A \mid \bar{B}) P(\bar{B})}$$

**Proof** This is a direct consequence of Corollary 1.2 when $I = \{1, 2\}$. ■

In the next two examples we provide illustrations of the types of experimental situations for which Bayes's rule has useful applications.

**Example 1.18** Explorations, Inc., is in the oil well-drilling business. Let $B$ be the event that a well being drilled will produce oil, and let $A$ be an event representing geological well-site characteristics that are conducive to discovering oil at a site. Suppose further that, from past experience, it is known that the (unconditional) probability of a successful strike when drilling for oil is .06. Also suppose it is known that when oil is discovered, the probability is .85 that the geological characteristics are given by event $A$, whereas the probability is only .4 that geological characteristics represented by $A$ are present when no oil is discovered. If event $A$ occurs at a site, what is the probability of discovering oil at the site, i.e., what is $P(B \mid A)$?

**Answer**: It is known that $P(B) = .06$, $P(\bar{B}) = .94$, $P(A \mid B) = .85$, and $P(A \mid \bar{B}) = .40$. Bayes's rule applies here, so that

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid \bar{B})P(\bar{B})} = \frac{(.85)(.06)}{(.85)(.06) + (.40)(.94)} = .12.$$

Note that the occurrence of event $A$ increases considerably the probability of discovering oil at a particular site. □

**Example 1.19** A blood test developed by a pharmaceutical company for detecting a certain disease is 98 percent effective in detecting the disease *given* that the disease is, in fact, present in the individual being tested. The test yields a "false positive" result (meaning a person without the disease is incorrectly indicated as having the disease) for only one percent of the disease-free persons tested.

If an individual is randomly chosen from the population and tested for the disease, and given that .1 percent of the population actually has the disease, what is the probability that the person tested actually has the disease if the test result is positive (i.e., the disease is indicated as being present by the test)?

**Answer**: In this case, let $A$ be the event that the test result is positive, and let $B$ be the event that the individual actually has the disease. Then, from the preceding discussion concerning the characteristics of the test, it is known that $P(A \mid B) = .98$, $P(B) = .001$, and $P(A \mid \bar{B}) = .01$. Then, an application of Bayes' rule yields

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid \bar{B})P(\bar{B})} = \frac{(.98)(.001)}{(.98)(.001) + (.01)(.999)} = .089.$$

Thus, one has very little confidence that a positive test result implies that the disease is present. □

A common thread in the two examples, consistent with the statement of Bayes's rule itself, is that the sample space is partitioned into a collection of (disjoint) events $(B_i, i \in I)$ that are of interest and whose nonzero probabilities are known. Furthermore, an event occurs whose various conditional probabilities formed by conditioning on each of the events in the partition are known. Given this background information, Bayes's rule provides the means for reevaluating the probabilities of the various events in the partition of $S$, *given* the information that event $A$ occurs. The probabilities of the events in the partition are, in effect, "updated" in light of the new information provided by the occurrence of $A$. This interpretation of Bayes's rule has led to the use of the terms **prior probabilities** and **posterior probabilities** to refer to the $P(B_i)$'s and $P(B_i \mid A)$'s, respectively. That is, $P(B_i)$ is the probability of event $B_i$ in the partition of $S$ *prior* to the occurrence of event $A$, whereas $P(B_i \mid A)$ is the probability of $B_i$ *posterior* to, or after, event $A$ occurs.

Returning to the oil well-drilling example, note that each elementary event in the (implied) sample space is given by a pair of observations, one being whether or not the geological characteristics of the well site favor the discovery of oil, and the other being whether or not oil is actually discovered at the well site. The partition of the sample space that is of interest to the oil well-drilling company is the event "oil is discovered" versus the (complementary) event that "oil is not discovered" at the well site. The additional information used to update the (prior) probabilities concerning oil discovery is whether the geological characteristics of the site favor the discovery of oil. Bayes's rule can be applied to generate posterior probabilities of oil discovery because the conditional probabilities of favorable well-site characteristics being observed, *with* and *without* the condition of oil being discovered, are known. The reader should provide a characterization of the sample space, partition of interest, and additional information used to update probabilities in the case of the drug test example.

## Key Words, Phrases, and Symbols

Experiment
Outcome
Sample space
Element (or member)
Sample point
Discrete sample space
Continuous sample space
Event
Elementary event
The occurrence of event $A$
Mutually exclusive (or disjoint) events
Mathematical induction
Classical probability
Relative frequency probability

Subjective probability
Axiomatic approach
Event space, $\Upsilon$
Certain event, $S$
Partition of a set
Mathematical induction
Probability set function (or probability measure), $P$
Probability of event $A$
Probability space
Probability theorems
Bonferroni's inequality
Borel sets
$\sigma$-field
Conditional sample space

Conditional probability of event $A$, given event $B$
$P(A \mid B)$
Multiplication rule
Independence of events
$\Rightarrow$, implies that
Joint, mutual, or complete independence of events
Pairwise independence
Product notation
$\prod_{i=1}^{n}$ or $\prod_{j \in j}$
Bayes's rule
Prior probability
Posterior probability

## Problems

**1.** Define an appropriate sample space for each of the experiments described below:

a. At the close of business each day, the Acme Department Store's accountant counts the number of customer transactions that were made in cash. On a particular day, there were 100 customer transactions at the department store. The outcome of interest is the number of cash transactions made.

b. An Italian restaurant in the city of Spokane runs an ad in the city newspaper, *The Spokesman Review*, that contains a coupon that allows a customer to purchase two meals for the price of one for each newspaper coupon the customer has. The coupon is valid for 30 days after the ad is run. The outcome of interest is how many free meals the restaurant serves at the end of the 30-day period.

c. On a local 11 o'clock news broadcast for the town of College Station, the weather report includes the high and low temperatures, in Fahrenheit, for the preceding 24 hours. The outcome of interest is the pair of high and low temperatures on any given day.

d. A local gasoline jobber supplies a number of the area's independent gas stations with unleaded gasoline. The outcome of interest is the quantity of gasoline demanded from the jobber in any given week.

e. The mutual funds management company of Dewey, Cheatum, and Howe posts the daily closing net asset value of shares in its mutual fund on a readerboard outside of its headquarters. The outcome of interest is the posted net asset value of the shares at the end of a given day.

f. The office manager of a business specializing in copying services is counting the number of copies that a given copying machine produces before suffering a paper jam. The outcome of interest is the number of copies made before the machine suffers a paper jam.

**2.** For each of the sample spaces you have defined above, indicate whether the sample space is finite, countably infinite, or uncountably infinite. Justify your answers.

**3.** The sales team of a large car dealership in Seattle consists of the following individuals:

| Name | Sales Experience | Age | Education | Married |
|---|---|---|---|---|
| Tom | 4 years | 34 | High school | Yes |
| Karen | 12 years | 31 | < High school | No |
| Frank | 21 years | 56 | College graduate | Yes |
| Eric | 9 years | 42 | High school | Yes |
| Wendy | 3 years | 24 | College graduate | No |
| Brenda | 7 years | 29 | High school | No |
| Scott | 15 years | 44 | College graduate | Yes |
| Richard | 2 years | 25 | < High school | No |

A customer visiting the dealership randomly chooses one of the salespersons to discuss the purchase of a new vehicle. Define the set and assign the probability associated with each of the following events:

a. A woman is chosen.

b. A man less than 40 years of age is chosen.

c. An individual with at least 10 years of sales experience is chosen.

d. A married college graduate is chosen.

e. A married female with a high school education and at least five years of sales experience is chosen.

f. An individual with at least 2 years' experience and at least 21 years of age is chosen.

**4.** Assign probabilities to events a to f in the preceding question, but include the condition "*given that* the individual chosen $\geq$ 30 years old."

**5.** The manager of the cost accounting department of a large computer manufacturing firm always tells three jokes during her monthly report to the board of directors in an attempt to inject a bit of levity into an otherwise sobering presentation. She has an inventory of a dozen different jokes from which she chooses three to present for any given monthly report.

a. If she chooses the 3 jokes randomly from the inventory of 12 each month, what is the probability that, in any given month, at least 2 of the 3 jokes will be different from the jokes she told the month before?

b. If she chooses the 3 jokes randomly from the inventory of 12 each month, what is the probability that, in any given month, all 3 jokes will be different from the 3 she told the month before?

**6.** Schneider's Plumbing and Heating, located in Fargo, North Dakota, has 300 accounts receivable distributed as follows:

| Current | 1–30 days past due | 31–60 days past due | 61–90 days past due | Sent for collection |
|---|---|---|---|---|
| 140 | 80 | 40 | 25 | 15 |

An auditor is coming to inspect Schneider's financial records. Included in the auditor's analysis is a randomly chosen sample of four accounts from the company's collection of accounts receivable.

a. What is the probability that all of the accounts chosen by the auditor will be current accounts?

b. What is the probability that all of the accounts chosen by the auditor will be less than or equal to 60 days past due?

c. What is the probability that all of the accounts chosen by the auditor will be more than 60 days past due?

d. What is the probability that 2 of the accounts will be current, and 2 will be 1 to 30 days past due?

**7.** A computer manufacturing firm produces three product lines: (1) desktop computer systems, (2) notebook computers, and (3) subnotebook computers. The sales department has convened its monthly meeting in which the four staff members of the department provide the department manager with their indications of whether sales will increase for each of the product lines in the coming month. Let $A_i$ represent the event that sales for product line $i$ (= 1, 2, or 3) will increase in the coming month. The manager will consider the information of a given staff member to be usable if that information is *internally consistent*, where internally consistent in this context means consistent with the axioms and theorems of probability. Which of the staff members have provided the manager with usable information? Be sure to provide a convincing reason if you decide that a staff member's information needs to be discarded.

| Staff Member | Tom | Dick | Harry | Sally |
|---|---|---|---|---|
| $P(A_1)$ | .5 | .3 | .3 | .2 |
| $P(A_2)$ | .3 | .2 | .6 | .3 |
| $P(A_3)$ | .7 | .8 | −.4 | .5 |
| $P(A_1 \cap A_2)$ | .9 | .4 | .4 | .2 |
| $P(A_1 \cap A_3)$ | .6 | .15 | .2 | .3 |
| $P(A_2 \cap A_3)$ | .15 | .1 | .1 | .4 |
| $P(A_1 \cap A_2 \cap A_3)$ | .1 | 1.5 | .05 | .1 |

**8.** A large electronics firm is attempting to hire six new electrical engineers. It has been the firm's experience that 35 percent of the college graduates who are offered positions with the firm have turned down the offer of employment. After interviewing candidates for the positions, the firm offers employment contracts to seven college graduates. What is the probability that the firm will receive acceptances of employment from one too many engineers? You may assume that the decisions of the college graduates are independent.

**9.** A computer manufacturing firm accepts a shipment of CPU chips from its suppliers only if an inspection of 5 percent of the chips, randomly chosen from the shipment, does not contain any defective chips. If a shipment contains five defective chips and there are 1,000 chips in the shipment, what is the probability that the shipment will be accepted?

**10.** The probability that a stereo shop sells at least one amplifier on a given day is .75; the probability of selling at least one CD player is .6; and the probability of selling at least one amplifier *and* one CD player is .5.

a. What is the probability that the stereo shop will sell at least one of the two products on a given day?

b. What is the probability that the stereo shop will sell at least one CD player, given that the shop sells at least one amplifier?

c. What is the probability that the stereo shop will sell at least one amplifier, given that the shop sells at least one CD player?

d. What is the probability that the shop sells neither of the products on a given day?

**11.** Prove that the set function defined by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \text{ for } P(B) \neq 0$$

is a valid *probability* set function in the probability space $\{B, \Upsilon_B, P(\cdot \mid B)\}$, where $\Upsilon_B$ is the event space for the sample space $B$.

**12.** A large midwestern bank has devised a math aptitude test that it claims provides valuable input into the hiring decision for bank tellers. The bank's research indicates that 60 percent of all tellers hired by midwestern banks are classified as performing satisfactorily in the position at their initial six-month performance review, while the rest are rated as unsatisfactory. Of the tellers whose performance is rated as satisfactory, 90 percent had passed the math aptitude test. Of the tellers who were rated as unsatisfactory, only 20 percent had passed the math aptitude test.

a. What is the probability that a teller would be rated as satisfactory at her six-month performance review given that she passed the math aptitude test?

b. What is the probability that a teller would be rated as satisfactory at her six-month performance review given that she did not pass the math aptitude test?

c. Does the test seem to be an effective screening device to use in hiring tellers for the bank? Why or why not?

**13.** A large-scale firm specializing in providing temporary secretarial services to corporate clients has completed a study of the main reason why secretaries become dissatisfied with their work assignments and how likely it is that a dissatisfied secretary will quit her job. It was found that 20 percent of all secretaries were dissatisfied with some aspect of their job assignment. Of all dissatisfied secretaries, it was found that 55 percent were dissatisfied mainly because they disliked their supervisor; 30 percent were dissatisfied mainly because they felt they were not paid enough; 10 percent were dissatisfied mainly because they disliked the type of work; and 5 percent were dissatisfied mainly because they had conflicts with other employees. The probabilities that the dissatisfied secretaries would quit their jobs were respectively .20, .30, .90, and .05.

a. Given that a dissatisfied secretary quits her job, what is the most probable main reason why she was dissatisfied with her job assignment?

b. If a secretary were chosen at random, what is the probability that she would be dissatisfied, the main reason being her level of pay?

c. Given that a secretary is dissatisfied with her job assignment, what is the probability that she will quit?

**14.** A clerk is maintaining three different files containing job applications submitted for three different positions currently open in the firm at which the clerk is employed. One file contains two completed applications, one file contains one complete and one incomplete application, and the third file contains two incomplete applications. The clerk wishes to examine the files and chooses one of the files at random. She then chooses at random one of the applications contained in the chosen file. If the application chosen is complete, what is the probability that the remaining application in the file is also complete?

**15.** A company manages three different mutual funds. Let $A_i$ be the event that the $i$th mutual fund increases in value on a given day. Probabilities of various events relating to the mutual funds are given as follows:

$$P(A_1) = .55, \quad P(A_2) = .60, \quad P(A_3) = .45,$$

$$P(A_1 \cup A_2) = .82, \quad P(A_1 \cup A_3) = .7525,$$

$$P(A_2 \cup A_3) = .78, \quad P(A_2 \cap A_3 \mid A_1) = .20.$$

a. Are events $A_1, A_2$, and $A_3$ pairwise independent?

b. Are events $A_1, A_2$, and $A_3$ independent?

c. What is the probability that funds 1 and 2 both increase in value, given that fund 3 increases in value? Is this different from the unconditional probability that funds 1 and 2 both increase in value?

d. What is the probability that at least one mutual fund will increase in value on a given day?

**16.** Answer the following questions regarding the validity of probability assignments. If you answer false, explain why the statement is false.

a. If $P(A) = .2$, $P(B) = .3$, and $A \cap B = \phi$, then $P(A \cup B) = .06$. True or false?

b. If $A \cap B = \phi$ and $P(B) = .2$, then $P(A \mid B) = 0$. True or false?

c. If $P(B) = .05$, $P(A \mid B) = .80$, and $P(A \mid \bar{B}) = .5$, then $P(B \mid A) = .0777$ (to four digits of accuracy). True or false?

d. If $P(A) = .8$ and $P(B) = .7$, then $P(A \cap B) \geq .5$. True or false?

e. It is possible that $P(A) = .7$, $P(B) = .4$, and $A \cap B = \emptyset$. True or false?

**17.** The ZAP Electric Co. manufactures electric circuit breakers. The circuit breakers are produced on two different assembly lines in the company's Spokane plant. Assembly line I is highly automated and produces 85 percent of the plant's output. Assembly line II uses older technology that is more labor intensive, producing 15 percent of the plant's output. The probability that a circuit breaker manufactured on assembly line I is defective is .04, while the corresponding probability for assembly line II is .01.

As part of its quality-control program, ZAP uses a testing device for determining whether a circuit breaker is faulty. Some important characteristics of the testing device are as follows:

$$P(A \mid B) = P(\bar{A} \mid \bar{B}) = .985,$$

where $A$ is the event that the testing device *indicates* that a circuit breaker is faulty and $B$ is the event that the circuit breaker *really is* faulty.

a. If a circuit breaker is randomly chosen from a bin containing a day's production and the circuit breaker is actually defective, what is the probability that it was produced on assembly line II?

b. What is the probability that the testing device *indicates* that a circuit breaker is *not* faulty, given that the circuit breaker really is faulty?

c. If the testing device is applied to circuit breakers produced on assembly line I, what is the probability that a circuit breaker really is faulty, given that the testing device indicates that the circuit breaker is faulty? Would you say that this is a good testing device?

**18.** The ACME Computer Co. operates three plants that manufacture notebook computers. The plants are located in Seattle, Singapore, and New York. The plants produce 20, 30, and 50 percent of the company's output, respectively. ACME attaches the labels "Seattle", "SING", or "NY" to the underside of the computer in order to identify the plant in which a notebook computer was manufactured. The computers carry a two-year warranty, and if a customer requires repairs during the warranty period, he or she must send the computer back to the plant in which the computer was manufactured. There is also a stamp on the motherboard inside the computer which technicians at a plant can use as an additional way of identifying which plant manufactured the computer. The consumer is unable to examine this inside stamp, because if the consumer opens up the computer housing to look inside, a seal is broken, which voids the warranty.

Regarding quality control at the plants, the warranty-period failure rates of computers manufactured in the three plants are known to be .01, .05, and .02 for the Seattle, Singapore, and New York plants, respectively. You have bought an ACME computer, and it has failed during the warranty period. You need to send the computer back to the plant for repairs, but the label on the underside of the computer has been lost and so you don't know which plant manufactured your computer.

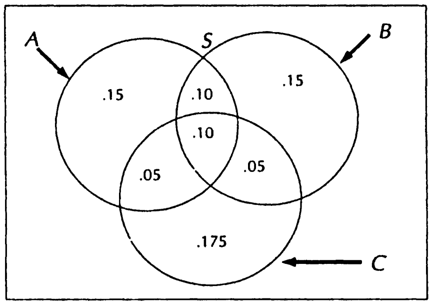
a. Which plant is the most probable plant to have manufactured your computer?

b. Which plant is the least probable plant to have manufactured your computer?

c. What is the probability that an ACME notebook computer will fail during the warranty period?

d. Given that an ACME computer does not fail during the warranty period, what is the probability that the computer was manufactured in New York?

**19.** The following diagram indicates how probabilities have been assigned to various subsets of the sample space $S$:



a. Are the three events $A$, $B$, and $C$ *pairwise-independent* events?

b. Are the three events $A$, $B$, and $C$ independent events?

c. What is the value of $P(A \cap B)$? What is the value of $P((A \cap B) \mid C)$?

d. Suppose event $D$ is such that $P(D) = .05$ and $D \cap (A \cup B \cup C) = \emptyset$. Are events $D$ and $A$ independent? Are events $D$ and $(A \cup B \cup C)$ independent?

e. What is the probability of event $C$, *given* $(A \cap B)$?

**20.** A large sack contains 1,000 flower seeds consisting of 300 carnations and 700 impatiens. Of the 300 carnation seeds, 200 will produce red flowers and 100 will produce white flowers. Of the 700 impatiens seeds, 400 will produce red flowers and 300 will produce white flowers.

a. If you randomly choose five seeds in succession (without replacing any seeds that have been chosen), what is the probability that these seeds will produce two impatiens with red flowers, two carnations with red flowers, and one carnation with white flowers?

b. If you randomly choose four seeds in succession (without replacing any seeds that have been chosen), what is the probability that these seeds will produce red- and white-flowered impatiens and red- and white-flowered carnations?

c. Given that four randomly chosen seeds all produce carnations, what is the probability that three are red flowered and one is white flowered?

d. Is the event of randomly choosing a carnation seed on the first draw independent of choosing an impatiens seed on the second draw?

**21.** For each case below, determine whether or not the real-valued set function $P(A)$ is in fact a *probability* set function:

a. sample space $S = \{1, 2, 3, 4, 5, 6, 7, 8\}$,
event space $\Upsilon = \{A : A \subset S\}$,
set function $P(A) = \sum_{x \in A}(x/36)$ for $A \in \Upsilon$.

b. sample space $S = [0, \infty)$,
event space $\Upsilon = \{A : A$ is an interval subset of $S$, or any set formed by unions, intersections, or complements of these interval subsets$\}$,
set function $P(A) = \int_{x \in A} e^{-x} dx$ for $A \in \Upsilon$.

c. sample space $S = \{x : x$ is a positive integer$\}$,
event space $\Upsilon = \{A : A \subset S\}$,
set function $P(A) = \sum_{x \in A}(x^2/10^5)$ for $A \in \Upsilon$.

d. sample space $S = (0, 1)$,
event space $\Upsilon = \{A : A$ is an interval subset of $S$, or any set formed by unions, intersections, or complements of these interval subsets$\}$,
set function $P(A) = \int_{x \in A} 12x(1-x)^2 dx$ for $A \in \Upsilon$.

**22.** The Smith Floor Wax Company manufactures and sells industrial-strength floor wax in the wholesale market for home care products. The factory produces 10,000 gallons of floor wax daily and currently has an inventory of 5,000 gallons of floor wax in its warehouse. If sales of floor wax exceed production, the company meets the excess demand by using inventory; if sales are less than production, the company adds this excess production to inventory. The company economist provides you with the following information concerning probabilities of daily sales events, where events are measured in gallons of wax sold.

$A = [0,\ 5{,}000],\ P(A) = .25$

$B = (5{,}000,\ 10{,}000],\ P(B) = .65$

$C = [2{,}500,\ 7{,}500],\ P(C) = .35$

$D = (5{,}000,\ 7{,}500],\ P(D) = .20$

a. What is the probability that inventory will have to be used to satisfy sales on a given day?

b. What is the probability that fewer than 2,500 gallons of wax will be sold on a given day?

c. What is the probability of the event $E = [0,\ 2{,}500) \cup (7{,}500,\ 10{,}000]$?

**23.** A box contains four different computer disks, labeled 1, 2, 3, and 4. Two disks are selected at random from the box "with replacement," meaning that after the first selection is made, the selected disk is returned to the box before the second selection is made. "At random" means that all disks in the box have an equal chance of being selected.

a. Define the sample space for this experiment.

b. Is the event of choosing disk 1 or 3 on the first selection independent of choosing disk 1 or 2 on the second selection? Why or why not?

c. Is the event of choosing disk 1 on the first selection independent of choosing disk 1 on the second selection? Why or why not?

d. Is the event of choosing disk 1 on both the first and second selections independent of the event that neither disk 3 nor 4 is chosen in the selection process?

**24.** The AJAX Microchip Company produces memory chips for personal computers. The company's entire production is generated from two assembly lines, labeled "I" and "II." Assembly line I uses more rapid assembly techniques and produces 80 percent of the company's output, while assembly line II produces 20 percent of the output. The probability that a memory chip produced on assembly line I is defective is .05, while the corresponding probability for assembly line II is .01.

A memory chip is chosen at random from a bin containing a day's production. Given that the chip is found to be defective, what is the probability that the chip was made on assembly line II? (*Hint*: Can you put this problem in a form for which Bayes's rule would be applicable?)

**25.** The management of the AJAX Microchip Company (mentioned in Problem 24) is interested in increasing quality control at the plant and is considering the purchase of a testing device that can determine when a memory chip is faulty. In particular, the specifications on the device are as follows:

$P(A \mid B) = P(\bar{A} \mid \bar{B}) = .98,$

where $A$ is the event that the testing device *indicates* that a memory chip is faulty, and $B$ is the event that the memory chip really *is* faulty.

a. What is the probability that the testing device indicates that a memory chip is not faulty, given that the memory chip really is faulty?

b. If the testing device is applied to the memory chips produced by AJAX's assembly line I, what is $P(B \mid A)$, i.e., the probability that a chip really is faulty, given that the testing device indicates the chip is faulty?

c. Suppose AJAX management wants $P(B \mid A)$ to be .95. What is the value of $r = P(A \mid B) = P(\bar{A} \mid \bar{B})$ that will ensure this testing accuracy if the test is applied to the chips produced on assembly line I?

**26.** Let $S = [0, 5]$ be a sample space containing all possible values of the daily quantity demanded of electric power for a large midwestern city in the summer months. The units of measurement are millions of megawatts, and the capacity of the power grid is 5 million megawatts.

Answer the following questions concerning probability assignments to events in the sample space, $S$, relating to the daily demand for electric power. Treat the information provided in the questions as cumulative. Justify your answers.

a. Given that $A = [0, 4]$, $B = [3, 5]$, $P(A) = .512$ and $P(B) = .784$, what is the probability that the power demand will be no greater than 4 million megawatts and no less than 3 million megawatts, i.e., what is the probability of $A \cap B$?

b. What is the probability of event $C = [0, 3)$?

c. Can $P(D) = .6$, given that $D = [0, 2.5]$?

d. Given that $P([0, 2]) = .064$, what is the probability of event $E = (2, 4]$?

**27.** SUPERCOMP, a retail computer store, sells personal computers and printers. The number of computers and printers sold on any given day varies, with the probabilities of the various possible sales outcomes being given by the following table:

| | | Number of Computers Sold | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | |
| Number of printers sold | 0 | .03 | .03 | .02 | .02 | .01 | Probabilities of elementary events |
| | 1 | .02 | .05 | .06 | .02 | .01 | |
| | 2 | .01 | .02 | .10 | .05 | .05 | |
| | 3 | .01 | .01 | .05 | .10 | .10 | |
| | 4 | .01 | .01 | .01 | .05 | .15 | |

a. Define an appropriate sample space for the experiment of observing how many computers and printers are sold on any given day.

b. Can the information provided in the table be used to define a probability set function for assigning probabilities to all events in the sample space? Explain (briefly, but clearly).

c. What is the probability that more than two computers will be sold on any given day? What is the probability that more than two printers will be sold on any given day? In each case, define the set of outcomes in $S$ that corresponds to the stated events.

d. What is the probability of selling more than two printers, *given* that more than two computers are sold? Show your calculation.

e. What is the probability of selling more than two printers *and* more than two computers? Show your calculation.

f. What is the probability that SUPERCOMP has no sales on a given day? *Given* that SUPERCOMP sells no computers, what is the probability that it sells no printers on a given day?

# 2 Random Variables, Densities, and Cumulative Distribution Functions

## 2.1 Introduction

The outcomes of many types of experiments are inherently in the form of real numbers. For example, measuring the height and weight of individuals, observing the market price and quantity demanded of a commodity, measuring the yield of a new type of wheat, or measuring the miles per gallon achievable by a new compact automobile all result in real-valued outcomes. The sample spaces associated with these types of experiments are subsets of the real line or, at least, subsets of n-dimensional real space, $R^n$.

There are also experiments whose outcomes are not inherently numbers and whose sample space is not inherently a subset of a real space. For example, observing whether a tossed coin results in heads or tails, observing whether an item selected from an assembly line is defective or nondefective, observing the type of weeds growing in a field of wheat, and observing which engine components caused an engine failure in an automobile are not experiments characterized by real-valued outcomes. It will prove to be both convenient and useful to convert these sample spaces into real-valued sample spaces by associating a real number to each outcome in the original sample space. Such a procedure might be looked upon as coding the outcomes of an experiment with real numbers.

Furthermore, the outcomes of an experiment may not be of direct interest in a given problem setting; instead, images of a real-valued function of the outcomes may be of prime importance. For example, in a game of craps, it is not the outcome of each die that is of primary importance, but rather the sum of the dots facing up determines whether a player has won or lost. As another example, if a firm is interested in calculating the profit associated with a given operation, it is the price of the product multiplied by the quantity sold, which gives the total revenue, that will be of primary importance in the profit calculation, and not price and quantity, per se.

All of the previous situations involve the concept of a random variable, which can be used to characterize the outcomes of an experiment as a set of real numbers. We now develop the concept of a random variable.

## 2.2 Univariate Random Variables and Density Functions

We begin with a definition of the term **random variable** appropriate for the univariate, or one-variable, case.

**Definition 2.1** *Univariate random variable*

Let $\{S, \Upsilon, P\}$ be a probability space. If $X : S \rightarrow R$ (or simply, $X$) is a real-valued function having as its domain the elements of $S$, then $X : S \rightarrow R$ (or $X$) is called a random variable.

A pictorial illustration of the random variable concept is given in Figure 2.1.

The reader might at least find it curious, and perhaps even consider it a misnomer, for the expression "random *variable*" to be used as a label for the concept just given. The expression *random-valued function* would seem more appropriate since it is, after all, a real-valued *function* that is at the heart of the

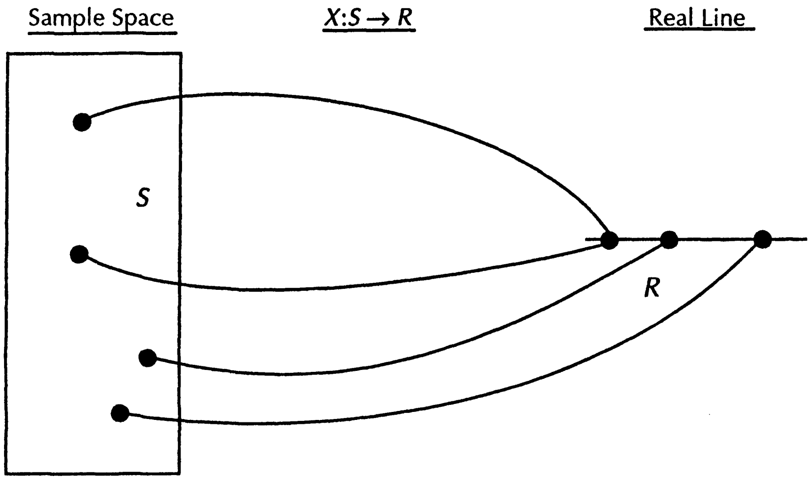


**Figure 2-1** Random variable $X$.

concept presented in the definition. Nonetheless, usage of "random variable" has become standard terminology, and we will use it also.

Henceforth, the symbol $X(w)$ will be used to represent the image of $w \in S$ generated by the random variable $X : S \rightarrow R$. We will use the lowercase $x$ to indicate a value of the function $X$, and $x$ can also be interpreted as a shorthand notation for an image value $x = X(w)$. As indicated in the definition, $X$ will be used as an abbreviated notation for representing a random variable whenever it is both convenient and appropriate. The letter $X$ that we use here is arbitrary, and any other symbol could be used to denote a random variable. For the most part, we will use letters in the latter part of the alphabet for representing random variables.[1] Letters at the beginning of the alphabet will be used to denote constants, and so the expression $x = a$ will mean that the value, $x$, of the random variable, $X$, equals the constant $a$. Similarly, $x \in A$ will mean that the value of $X$ is an element of the set $A$.[2]

If the outcomes of an experiment are real numbers to begin with, they are directly interpretable as values of a random variable since we can always represent the real-valued outcomes $w \in S$ as images of an identity function, e.g., $X(w) = w$. If the outcomes of an experiment are not initially in the form of real numbers, a random variable can be defined that associates a real number with each outcome, as $X(w) = x$, i.e., the random variable effectively codes the outcomes of a sample space with real numbers. Thus, through the use of the random-variable concept, all experiments with univariate outcomes can be interpreted as having sample spaces consisting of real-valued elements. In particular, the range of the random variable, $R(X)$, represents a real-valued sample space for the experiment. In terms of set notation, the range of the random variable, $X$, can be defined as $R(X) = \{x : x = X(w), w \in S\}$. When we refer to the **outcome of the random variable**, we mean the particular image element in $R(X)$ that occurs as a result of observing the outcome of a given experiment, i.e., if the outcome of an experiment is $w \in S$, then the outcome of the random variable is $x = X(w)$.

### Probability Space Induced by a Random Variable

Now that we have defined a real-valued sample space for a given experiment using the random-variable concept, we seek to embed the sample space within a probability space that can be used for assigning probabilities to events involving random-variable outcomes. To accomplish this, we must establish how probabilities are to be assigned to subsets of the real-valued sample space $R(X)$, i.e., we must define an appropriate *probability set function*. We also need to

[1] Teaching suggestion: When lecturing in statistics and econometrics classes, I have found it useful to underscore with a tilde, ~, the capital letter used to represent a random variable. This provides an unmistakable distinction between the function and a value of the function when writing expressions on the blackboard. Thus, $X$ and $\underset{\sim}{X}$ will mean precisely the same thing, with the tilde being used for emphasis to avoid misinterpreting handwritten capital letters as lowercase letters, and vice versa.

[2] Note that $X = a$ and $X \in A$ is alternative shorthand notation that is often used to represent that the value of $X$ is equal to $a$ or is in the set $A$. Our notation establishes a distinction between the function, $X$, and a value of the function, $x$.

identify the collection of subsets of $R(X)$ that will be assigned probability, i.e., define the *event space* or domain of the probability set function.

We begin with knowledge of the probability space, $\{S, \Upsilon, P\}$, so we are equipped to assign probabilities to events in $S$. What is the probability that an outcome of $X$ resides in the set $A \subset R(X)$? It stands to reason that if an event in $S$ can be found, say $B$, that occurs *iff* the event $A \subset R(X)$ occurs, then the two events must have the same probability of occurring, and we can state that $P_X(A) \equiv P(B)$, where $P_X(\cdot)$ denotes the probability set function for assigning probability to events for outcomes of $X$. Two events that are associated with different probability spaces but that occur only simultaneously are called **equivalent events**. Such an event $B$ can be defined as $B = \{w : X(w) \in A, w \in S\}$, which is the set of inverse images of the elements in $A$ based on the function $X$. By definition, $w \in B \Leftrightarrow x \in A$, and thus $A$ and $B$ are *equivalent events* (see Figure 2.2). We then have the following representation of probability assignments to events involving random-variable outcomes:

$$P_X(A) = P(B) \quad \text{for } B = \{w : X(w) \in A, w \in S\}.$$

Thus, probabilities assigned to events in $S$ are transferred to events in $R(X)$ through the functional relationship $x = X(w)$, which relates outcomes $w$ in $S$ and outcomes $x$ in $R(X)$.

What should the domain of $P_X(\cdot)$ be, i.e., what is the event space for random-variable outcomes? It is clear from the foregoing discussion that to be able to assign probabilities to a set $A \subset R(X)$ it must be the case that its associated set of inverse images in $S$, $B = \{w : X(w) \in A, w \in S\}$, is an *event* in $S$, i.e., it must be the case that $B \in \Upsilon$. If not, a problem exists in that there is no basis for assigning probability to either set $B$ or $A$ from knowledge of the probability space $\{S, \Upsilon, P\}$ since $B$ would not be in the domain of $P(\cdot)$. No difficulty will arise if $S$ is a finite or countably infinite sample space, since then the event space $\Upsilon$ equals the collection of *all* subsets of $S$, and *whatever* subset $B \subset S$ is associated with the subset $A \subset R(X)$, $B$ will be an event in $S$. Thus, *any* real-valued function defined on a discrete sample space will generate a real-valued sample space for which all subsets can be assigned probability. Henceforth, the event space, $\Upsilon_X$, for outcomes of random variables defined on finite or countably infinite sample spaces is defined to be the set of *all* subsets of $R(X)$.

In order to avoid the problem that occurs when $S$ is uncountably infinite, one simply restricts the types of real-valued functions that are used to define random variables to those for which the problem will not occur. To this effect,

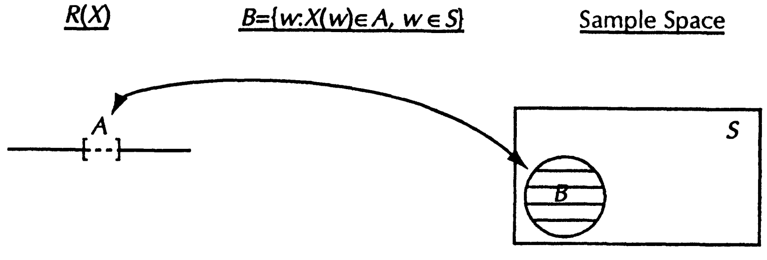


**Figure 2-2**
Subset of range elements, $A$, and associated domain elements, $B$, of $X$.

a proviso is generally added, either explicitly or implicitly, to the definition of a random variable $X$ requiring the real-valued function defined on $S$ to be such that for every Borel set, $A$, contained in $R(X)$ (recall Def. 1.9), the set $B = \{w\colon X(w) \in A, w \in S\}$ is an event in $S$, i.e., $B \in \Upsilon$. Then, since every Borel set $A \subset R(X)$ would be associated with an *event* $B \subset S$, every Borel set could be assigned a probability as $P_X(A) = P(B)$. Since the collection of Borel sets includes all intervals in $R(X)$ (and thus all points in $R(X)$), as well as all other sets that can be formed from the intervals by a countable number of union, intersection, and/or complement operations, the collection of Borel sets defines an event space sufficiently large for all practical purposes. Furthermore, we know from the discussion in Section 1.6 that an event space consisting of the Borel sets does not cause inherent conflicts with the probability axioms, and we will also see that such an event space facilitates the use of integrals for defining probabilities of events when $R(X)$ is uncountable.

In practice, it requires a great deal of ingenuity to define a random variable for which probability cannot be associated with each of the Borel sets in $R(X)$, and the types of functions that naturally arise when defining random variables in applications will generally satisfy the aforementioned proviso. Henceforth, we will assume that the event space, $\Upsilon_X$, for random-variable outcomes consists of all Borel sets in $R(X)$ if $R(X)$ is uncountable. We add that for all practical purposes, the reader need not unduly worry about the latter restriction to Borel sets, since any subset of an uncountable $R(X)$ that is of practical interest will be a Borel set.

In summary, a random variable induces an alternative probability space for the experiment. The **induced probability space** takes the form $\{R(X), \Upsilon_X, P_X\}$, where the range of the random variable $R(X)$ is the real-valued sample space, $\Upsilon_X$ is the event space for random-variable outcomes, and $P_X$ is a probability set function defined on the events in $\Upsilon_X$.

| | Random Variable | |
|---|---|---|
| Probability Space | $X : S \to R$ | Induced Probability Space |
| $\{S, \Upsilon, P(\cdot)\}$ | $x = X(w)$ | $\left\{\begin{array}{l} R(X) = \{x : x = X(w), w \in S\} \\ \Upsilon_X = \{A : A \text{ is an event in } R(X)\} \\ P_X(A) = P(B), B = \{w : X(w) \in A, w \in S\}, \forall A \in \Upsilon_X \end{array}\right\}$ |

**Example 2.1** Let $S = \{1, 2, 3, \ldots, 10\}$ represent the potential numbers of cars that a car salesperson sells in a given week, let the event space $\Upsilon$ be the set of all subsets of $S$, and let the probability set function be defined as $P(B) = (1/55)\sum_{w \in B} w$ for $B \in \Upsilon$. Suppose the salesperson's weekly pay consists of a base salary of \$100/week plus a \$100 commission for each car sold. We can represent the salesperson's weekly pay by the random variable $X(w) = 100 + 100w$, for $w \in S$. The *induced probability space* $\{R(X), \Upsilon_X, P_X\}$ is then characterized by $R(X) = \{200, 300, 400, \ldots, 1, 100\}$, $\Upsilon_X = \{A\colon A \subset R(X)\}$, and $P_X(A) = (1/55)\sum_{w \in B} w$ for $B = \{w\colon (100 + 100w) \in A, w \in S\}$ and $A \in \Upsilon_X$. Then, for example, the

event that the salesperson makes ≤ \$300/week, $A = \{200, 300\}$, has probability $P_X(A) = (1/55)\sum_{w\in\{1,2\}} w = (3/55)$. □

A major advantage in dealing with only real-valued sample spaces is that all of the mathematical tools developed for the real number system are available when analyzing the sample spaces. In practice, once the induced probability space has been identified; the underlying probability space $\{S, \Upsilon, P\}$ is generally ignored for purposes of defining random-variable events and their probabilities. In fact, we will most often choose to deal with the induced probability space $\{R(X), \Upsilon_X, P_X\}$ directly at the outset of an experiment, paying little attention to the underlying definition of the *function* having the range $R(X)$ or to the original probability space $\{S, \Upsilon, P\}$. However, we will sometimes return to the relationship between $\{S, \Upsilon, P\}$ and $\{R(X), \Upsilon_X, P_X\}$ to facilitate the proofs of certain propositions relating to random-variable properties.

Note for future reference that a real-valued function of a random variable is, itself, a random variable. This follows by definition, since a real-valued function of a random variable, say $Y$ defined by $y = Y(X(w))$ for $w \in S$, is a function of a function (i.e., a composition of functions) of the elements in a sample space $S$, which is then indirectly also a real-valued function of the elements in the sample space $S$ (recall Def. 2.1). One might refer to such a random variable as a **composite random variable.**

### Discrete Random Variables and Probability Density Functions

In practice, it is useful to have a representation of the probability set function, $P_X$, that is in the form of a well-defined algebraic formula and that does not require constant reference either to events in $S$ or to the probability set function defined on the events in $S$. A conceptually straightforward way of representing $P_X$ is available when the real-valued sample space $R(X)$ contains, at most, a countable number of elements. In this case, any subset of $R(X)$ can be represented as the union of the specific elements comprising the subset, i.e., if $A \subset R(X)$, then $A = \bigcup_{x\in A}\{x\}$. Since the elementary events in $A$ are clearly disjoint, we know from Axiom 1.3 that $P_X(A) = \sum_{x\in A} P_X(\{x\})$. It follows that once we know the probability of every elementary event in $R(X)$, we can assign probability to any other event in $R(X)$ by summing the probabilities of the elementary events contained in the event. This suggests that we define a *point* function $f : R(X) \rightarrow R$ as $f(x) =$ probability of $x = P_X(\{x\})\ \forall\ x \in R(X)$. Once $f$ is defined, then $P_X$ can be defined for all events as $P_X(A) = \sum_{x\in A} f(x)$. Furthermore, knowledge of $f(x)$ eliminates the need for any further reference to the probability space $\{S, \Upsilon, P\}$ for assigning probabilities to events in $R(X)$.

In the following example we illustrate the specification of the point function, $f$.

**Example 2.2** Examine the experiment of rolling a pair of dice and observing the number of dots facing up on each die. Assume the dice are fair. Letting $i$ and $j$ represent the number of dots facing up on each die, respectively, the sample space for the experiment is $S = \{(i, j)\colon\ i \text{ and } j \in \{1, 2, 3, 4, 5, 6\}\}$. Now define the random variable $x = X((i, j)) = i + j$, for $(i, j) \in S$. Then the following correspondence can be set up between outcomes of $X$, events in $S$, and the probability of outcomes of $X$ and events in $S$, where $w = (i, j)$:

| | $X(w) = x$ | $B_x = \{w\colon\ X(w) = x,\ w \in S\}$ | $f(x) = P(B_x)$ |
|---|---|---|---|
| $R(X)$ | 2 | $\{(1, 1)\}$ | 1/36 |
| | 3 | $\{(1, 2), (2, 1)\}$ | 2/36 |
| | 4 | $\{(1, 3), (2, 2), (3, 1)\}$ | 3/36 |
| | 5 | $\{(1, 4), (2, 3), (3, 2), (4, 1)\}$ | 4/36 |
| | 6 | $\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$ | 5/36 |
| | 7 | $\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$ | 6/36 |
| | 8 | $\{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$ | 5/36 |
| | 9 | $\{(3, 6), (4, 5), (5, 4), (6, 3)\}$ | 4/36 |
| | 10 | $\{(4, 6), (5, 5), (6, 4)\}$ | 3/36 |
| | 11 | $\{(5, 6), (6, 5)\}$ | 2/36 |
| | 12 | $\{(6, 6)\}$ | 1/36 |

The range of the random variable is $R(X) = \{2, 3, \ldots, 12\}$, which represents the collection of images of the points $(i, j) \in S$ generated by the function $x = X((i, j)) = i + j$. Probabilities of the various outcomes of $X$ are given by $f(x) = P(B_x)$, where $B_x$ is the collection of inverse images of $x$.

If we desired the probability of the event that $x \in A = \{7, 11\}$, then $P_X(A) = \sum_{x \in A} f(x) = f(7) + f(11) = 8/36$ (which, incidentally, is the probability of winning a game of craps on the first roll of the dice). If $A = \{2\}$, the singleton set representing "snake eyes," we find that $P_X(A) = \sum_{x \in A} f(x) = f(2) = 1/36$. □

In examining the outcomes of $X$ and their respective probabilities in Ex. 2.2, it is recognized that a compact algebraic specification can be suggested for $f(x)$, namely[3]

$$f(x) = \frac{6 - |x - 7|}{36} I_{\{2,3,\ldots,12\}}(x).$$

It is generally desirable to express the relationship between domain and image elements of a function in a compact algebraic formula whenever possible, as opposed to expressing the relationship in tabular form as in Ex. 2.2. This is especially true if the number of elements in $R(X)$ is large. Of course, if the number of elements in the domain is infinite, the relationship cannot be represented in tabular form and must be expressed algebraically. The reader is asked to define an appropriate point function $f$ for representing probabilities of the elementary events in the sample space $R(X)$ of Ex. 2.1.

[3] Notice that the algebraic specification faithfully represents the positive values of $f(x)$ in the preceding table of values and defines $f(x)$ to equal 0 $\forall\, x \notin \{2, 3, \ldots, 12\}$. Thus, the domain of $f$ is the entire real line. The reason for extending the domain of $f$ from $R(X)$ to $R$ will be discussed shortly. Note that assignments of probabilities to events as $P_X(A) = \sum_{x \in A} f(x)$ are unaffected by this domain extension.

We emphasize that if the outcomes of the random variable $X$ are the outcomes of interest in a given experimental situation, then given that a probability set function, $P_X(A) = \sum_{x \in A} f(x)$, has been defined on the events in $R(X)$, the original probability space $\{S, \Upsilon, P\}$ is no longer needed for defining probabilities of events in $R(X)$. Note that in Ex. 2.2, given $f(x)$, the probability set function $P_X(A) = \sum_{x \in A} f(x)$ can be used to define probabilities for all events $A \subset R(X)$ without reference to $S$ or $P(B_x)$, $B_x \subset S$.

The next example illustrates a case where an experiment is analyzed exclusively in terms of the probability space relating to random-variable outcomes.

**Example 2.3**

The Bippo Lighter Co. manufactures a budget-priced cigarette lighter that has a .90 probability of lighting on any given attempt to use the lighter. The probability that it lights on a given trial is independent of what occurs on any other trial. Define the probability space for the experiment of observing the number of trials required to obtain the first light. What is the probability that the lighter lights in 3 or fewer trials?

**Answer:** The range of the random variable, or equivalently the real-valued sample space, can be specified as $R(X) = \{1, 2, 3, \ldots\}$. Since $R(X)$ is countable, the event space $\Upsilon_X$ will be defined as the set of all subsets of $R(X)$. The probability that the lighter lights on the first attempt is clearly .90, and so $f(1) = .90$. Using independence of events, the probability it lights for the first time on the second trial is $(.10)(.90) = .09$, on the third trial is $(.10)^2(.90) = .009$, on the fourth trial is $(.10)^3(.90) = .0009$, and so on. In general, the probability that it takes $x$ trials to obtain the first light is $f(x) = (.10)^{x-1}(.90) I_{\{1,2,3,\ldots\}}(x)$. Then the probability set function is given by $P_X(A) = \sum_{x \in A}(.10)^{x-1}(.90) I_{\{1,2,3,\ldots\}}(x)$. The event that the lighter lights in three trials or less is represented by $A = \{1, 2, 3\}$. Then $P_X(A) = \sum_{x=1}^{3}(.10)^{x-1}(.90) = .999$. □

The preceding examples illustrate the concept of a **discrete random variable** and a **discrete probability density function**, which we formalize in the following definitions.

**Definition 2.2**
***Discrete random variable***

A random variable is called discrete if its range consists of a countable number of elements.

**Definition 2.3**
***Discrete probability density function***

The discrete probability density function, $f$, for a discrete random variable $X$ is defined as $f(x) =$ probability of $x$, $\forall\, x \in R(X)$, and $f(x) = 0$, $\forall\, x \notin R(X)$.

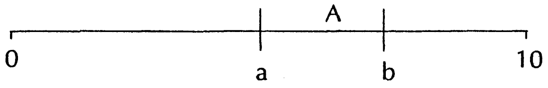
It should be noted that even though there is only a countable number of elements in the range of the discrete random variable $X$, the probability density function (PDF) defined here has the entire (uncountable) real line for its domain. The value of $f$ at a point $x$ in the range of the random variable is the probability of $x$, while the value of $f$ is zero at all other points on the real line. This definition is adopted for the sake of convenience—it standardizes the domain of all discrete density functions to be the real line while it has no effect on the assignment of probabilities to events via the set function $P_X(A) = \sum_{x \in A} f(x)$. This convention also provides a considerable simplification in the definition of marginal and conditional density functions, which we shall examine shortly.

In our previous examples, the probability space for the experiment was a priori deducible under the stated assumptions of the problems. It is most often the case in practice that the probability space is not a priori deducible, and an important problem in statistical inference is the identification of the appropriate density function, $f(x)$, to use in defining the probability set function component of the probability space.

### Continuous Random Variables and Probability Density Functions

So far, our discussion concerning the representation of $P_X$ in terms of the point function, $f(x)$, is applicable only to those random variables that have a countable number of possible outcomes. Can $P_X$ be similarly represented when the range of $X$ is uncountably infinite? Given that we can have an event $A$ defined as an uncountable subset of $R(X)$, it is clear that the summation operation over the elements of the set (i.e., $\sum_{x \in A}$) is not generally defined. Thus, defining a probability set function on the events in $R(X)$ as $P(A) = \sum_{x \in A} f(x)$ will not be possible. However, integration over uncountable sets is possible, suggesting that the probability set function might be defined as $P(A) = \int_{x \in A} f(x)dx$ when $R(X)$ is uncountably infinite. In this case the point function $f(x)$ would be defined so that $\int_{x \in A} f(x)dx$ defines the probability of event $A$. The following example illustrates the specification of such a point function $f(x)$ when $R(X)$ is uncountably infinite.

**Example 2.4** Suppose a trucking company has observed that accidents are equally likely to occur on a certain 10-mile stretch of highway, beginning at point 0 and ending at point 10. Let $R(X) = [0, 10]$ define the real-valued sample space of potential accident points.



It is clear that given that all points are equally likely, the probability set function should assign probabilities to intervals of the highway, say $A$, in such a way that the probability of an accident is equal to the proportion of the total highway length represented by the stretch of highway, $A$, as

$$P_X(A) = \frac{\text{length of } A}{10} = \frac{b-a}{10}, \quad \text{for } A = [a, b].$$

If we wish to assign these probabilities using $P_X(A) = \int_{x \in A} f(x)dx$, we require that

$$\int_a^b f(x)dx \equiv \frac{b-a}{10} \quad \text{for all } 0 \leq a \leq b \leq 10.$$

The following lemma will be useful in deriving the explicit functional form of $f(x)$.

**Lemma 2.1**
***Fundamental Theorem of Calculus***

Let $f(x)$ be a continuous function at $b$ and $a$, respectively.[4] Then

$$\frac{\partial \int_a^b f(x)dx}{\partial b} = f(b) \quad \text{and} \quad \frac{\partial \int_a^b f(x)dx}{\partial a} = -f(a).$$

Applying the lemma to the preceding integral identity yields

$$\frac{\partial \int_a^b f(x)dx}{\partial b} = f(b) \equiv \frac{\partial \left(\frac{b-a}{10}\right)}{\partial b} = \frac{1}{10}, \quad \forall\, b \in [0, 10],$$

which implies that the function defined by $f(x) = .1I_{[0,10]}(x)$ can be used to define the probability set function $P_X(A) = \int_{x \in A} .1dx$, for $A \in \Upsilon_X$.

For an example of the use of this representation, the probability that an accident occurs in the first half of the stretch of highway, i.e., the probability of event $A = [0, 5]$, is given by $P_X(A) = \int_0^5 .1dx = .5$. □

The preceding example illustrates the concept of a **continuous random variable** and a **continuous probability density function,** which we formalize in the next definition.

**Definition 2.4**
***Continuous random variables and continuous probability density functions***

A random variable is called continuous if its range is uncountably infinite, and if there exists a nonnegative-valued function $f(x)$, defined for all $x \in (-\infty, \infty)$, such that for any event $A \subset R(X)$, $P_X(A) = \int_{x \in A} f(x)dx$, and $f(x) = 0 \;\forall\, x \notin R(X)$. The function $f(x)$ is called a continuous probability density function.

Clarification of a number of important characteristics of continuous random variables is warranted. First of all, note that probability in the case of a continuous random variable is represented by the area under the graph of the density function $f$ and above the points in the set $A$, as illustrated in Figure 2.3.

Of course, the event in question need not be an interval, but given our convention regarding the event space $\Upsilon_X$, the event will be a Borel set for which an integral can be defined. A justification for the existence of the integral for

[4] See F.S. Woods (1954) *Advanced Calculus.* Boston: Ginn and Co., p. 141. Regarding **continuity** of $f(x)$, note that $f(x)$ is continuous at a point $d \in D(f)$ if, $\forall\, \epsilon > 0$, $\exists$ a number $\delta(\epsilon) > 0$ such that if $|x - d| < \delta(\epsilon)$, then $f(x) - f(d) < \epsilon$. The function $f$ is continuous if it is continuous at every point in its domain. Heuristically, a function will be continuous if there are no breaks in the graph of $y = f(x)$. Put another way, if the graph of $y = f(x)$ can be completely drawn without ever lifting a pencil from the graph paper, then $f$ is a continuous function.

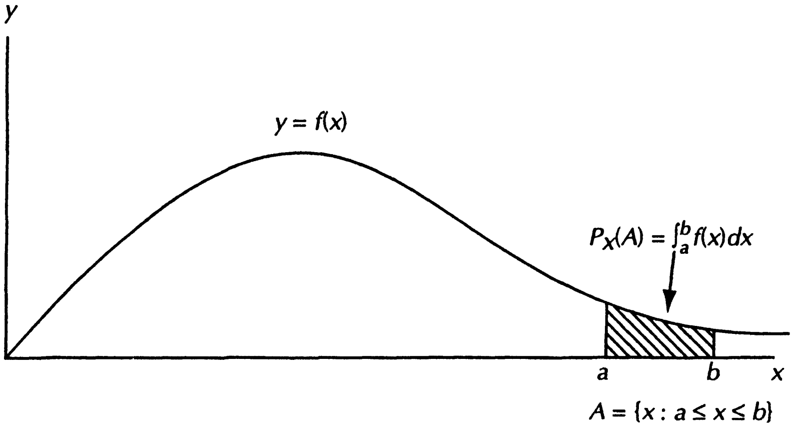


**Figure 2-3** Probability represented as area.

Borel sets is beyond the scope of this text, but implementation of the integration process in these cases is both natural and straightforward.[5] The next example illustrates the procedure of determining probabilities for events more complicated than a single interval.

**Example 2.5** Reexamine the highway accident example (Ex. 2.4), where $R(X) = [0, 10]$ and $f(x) = .1I_{[0,10]}(x)$.

**a.** What is the probability of $A = [1,2] \cup [7,9]$? The probability of $A$ is given by the area above the points in $A$ and below the graph of $f$, i.e.,

$$P_X(A) = \int_{x \in A} f(x)dx = \int_1^2 \left(\frac{1}{10}\right) dx + \int_7^9 \left(\frac{1}{10}\right) dx = .1 + .2 = .3.$$

**b.** Given $A$ defined above, what is the probability of $\bar{A} = [0, 1) \cup (2, 7) \cup (9, 10]$? The area representing the probability in question is calculated as

$$P_X(\bar{A}) = \int_{x \in \bar{A}} f(x)dx = \int_0^1 \left(\frac{1}{10}\right) dx + \int_2^7 \left(\frac{1}{10}\right) dx + \int_9^{10} \left(\frac{1}{10}\right) dx$$
$$= .1 + .5 + .1 = .7. \qquad \square$$

A consequence of the definition of the probability set function $P_X$ in Def. 2.4 is that, for a continuous random variable, the probability of any elementary event is zero, i.e., if $A = \{a\}$, then $P_X(A) = \int_a^a f(x)dx = 0$. Note that this certainly does not imply that every outcome of $X$ in $R(X)$ is impossible, since some elementary event in $R(X)$ will occur as a result of a given experiment. Instead, $P_X(\{x\}) = 0 \; \forall \; x \in R(X)$ suggests that zero probability is not synonymous with impossibility. In cases where an event, say $A$, can occur, but the probability set

[5] It can be shown that Borel sets are representable as the union of a collection of disjoint intervals, some of which may be single points. The collective area in question can then be defined as the sum of the areas lying above the various intervals and below the graph of $f$.

function assigns the event the value zero, we say **event $A$ occurs with probability zero**. The reader might intuitively interpret this to mean that **event $A$ is relatively impossible**, i.e., relative to the other outcomes that can occur $(R(X)-A)$, the likelihood that $A$ would occur is nil. We add that the notion of an event occurring with zero probability is not restricted to continuous random variables—it arises in discrete cases as well. In Section 2.9, we present examples where $R(X)$ is countably infinite and where $R(X)$ is finite, and we show that in both cases one can define events in $R(X)$ that occur with probability zero. Note that the above argument together with Theorem 1.1 then suggest that if $P_X(A) = 1$, it does *not* follow that event $A$ is *certain* to occur. In the spirit of our preceding discussion, if event $A$ is assigned a probability of 1, we say **event $A$ occurs with probability 1**, and if, in addition, $A \neq R(X)$, we might also state that **event $A$ is relatively certain**.

Note that an important implication of the preceding property for continuous random variables, which has already been utilized in Ex. 2.5b, is that the sets $[a, b]$, $(a, b]$, $[a, b)$, and $(a, b)$ are all assigned the same probability value $\int_a^b f(x)dx$, since adding or removing a finite number of elementary events to another event means adding or removing a collection of outcomes that occur with probability zero. That is, since $[a, b] = (a, b] \cup \{a\} = [a, b) \cup \{b\} = (a, b) \cup \{a\} \cup \{b\}$, and since $P_X(\{a\}) = P_X(\{b\}) = 0$, Axiom 1.3 implies that $P_X([a, b]) = P_X((a, b]) = P_X([a, b)) = P_X((a, b))$, so that the integral $\int_a^b f(x)dx$ suffices to assign the appropriate probability to all four interval events.

There is a fundamental difference in the interpretation of the image value $f(x)$ depending on whether $f$ is a discrete or continuous probability density function. In particular, while $f(x)$ is the probability of the outcome $x$ in the discrete case, $f(x)$ is *not* the probability of $x$ in the continuous case. To motivate this latter point, recognize that *if* $f(x)$ were the probability of outcome $x$ in the continuous case, then by our argument above, $f(x) = 0 \; \forall \; x \in R(X)$ since the probability of elementary events are zero. But this would imply that for every event $A$, including the certain event $R(X)$, $P_X(A) = \int_{x \in A} f(x)dx = \int_{x \in A} 0 \, dx = 0$, since having an integrand of 0 ensures that the integral has a zero value. The preceding property would contradict the interpretation of $P_X$ as a probability set function, and so $f(x)$ is clearly not interpretable as a probability. It *is* interpretable as a density function value, but nothing more—the continuous PDF must be integrated to define probabilities.

As in the discrete case, a continuous PDF has the entire real line for its domain. Again, this convention is adopted for the sake of convenience, as it standardizes the domain of all continuous density functions while leaving probabilities of events unaffected. It also simplifies the definition of marginal and conditional probability density functions, which we will examine shortly. The convention is generally straightforward to accommodate through the use of indicator functions. For example, in Ex. 2.5, the extension of the domain of $f(x)$ to the entire real line is accomplished simply by utilizing the indicator function in defining $f(x)$ as $f(x) = (1/10)I_{[0,10]}(x)$.

We now provide another example of a continuous random variable together with its density function, where the latter, we will assume, has been discovered by your personnel department.

**Example 2.6** Examine the experiment of observing the amount of time that passes between work-related injuries to employees working at a metal fabricating plant. Let $R(X) = \{x : x \geq 0\}$ represent the potential outcomes of the experiment measured in hours, and let the density of the continuous random variable be given by

$$f(x) = \frac{1}{100}e^{-x/100}I_{[0,\infty)}(x).$$

**a.** What is the probability of the event that 100 or more hours pass between work-related injuries? Letting $A = \{x\colon x \geq 100\}$ represent the event in question,

$$P(A) = \int_{100}^{\infty} \frac{1}{100}e^{-x/100}dx = -e^{-x/100}\ |_{100}^{\infty} = -e^{-\infty/100} + e^{-1} = e^{-1} = .37.$$

**b.** What is the probability that an injury occurs within 50 hours of the previous injury? Letting $B = \{x\colon 0 \leq x \leq 50\}$ represent the event in question,

$$P(B) = \int_{0}^{50} \frac{1}{100}e^{-x/100}\,dx = -e^{-x/100}\ |_{0}^{50} = -e^{-50/100} + e^{-0} = 1 - .61 = .39. \qquad \square$$

### Classes of Discrete and Continuous PDFs

In our later study of statistical inference, we will generally identify an appropriate range for a random variable based on the characteristics of a particular experiment being analyzed and have as an objective the identification of an appropriate $f(x)$ with which to complete the specification of the probability space. The fact that for all events $A \subset R(X)$ the values generated by $\sum_{x\in A} f(x)$ or $\int_{x\in A} f(x)dx$ must adhere to the probability axioms places some general restrictions on the types of functions that can be used as density functions, regardless of the specific characteristics of a given experiment. These general restrictions on the admissible choices of $f(x)$ are identified in the following definition.

**Definition 2.5** *The classes of discrete and continuous probability density functions—univariate case*

**a. Class of discrete density functions.** The function $f : R \rightarrow R$ is a member of the class of discrete density functions iff (1) the set $C = \{x\colon f(x) > 0, x \in R\}$ (i.e., the subset of points in $R$ having a positive image under $f$) is countable; (2) $f(x) = 0$ for $x \in \bar{C}$; and (3) $\sum_{x\in C} f(x) = 1$.

**b. Class of continuous density functions.** The function $f : R \rightarrow R$ is a member of the class of continuous density functions iff (1) $f(x) \geq 0$ for $x \in (-\infty, \infty)$; and (2) $\int_{-\infty}^{\infty} f(x)dx = 1$.

Some clarifying remarks concerning Def. 2.5 are warranted. First, it should be noted that the definition simply identifies the respective classes of function specifications that are to be considered as *candidates* for use as probability density functions. The *specific* functional form of the density function chosen to define the probability set function for a real-world experimental situation depends on the particular characteristics of the process generating the outcomes of the experiment.

A second observation concerns the fact that the definitions focus exclusively on real-valued functions having the entire real line for their domains. As we discussed earlier, this is a convention adopted as a matter of convenience. To ensure that subsets of points outside of the range of $X$ are properly assigned zero probability, all one needs to do is extend the domain of $f$ to the remaining points $R - R(X)$ on the real line by assigning to each point a zero density weighting, i.e., $f(x) = 0$ if $x \in \overline{R(X)}$.

A final remark concerns the rationale in support of the properties that are required for a function $f$ to be considered a probability density function. The properties are imposed on $f : R \rightarrow R$ to ensure that the set functions constructed from $f$, i.e.,

$$P_X(A) = \sum_{x \in A} f(x) \quad \text{(discrete case)}$$

or

$$P_X(A) = \int_{x \in A} f(x)dx \quad \text{(continuous case)},$$

are in fact *probability* set functions, which of course requires that the probability assignments adhere to the axioms of probability. To motivate the *sufficiency* of the conditions for $P_X$ to satisfy the probability axioms, first examine the discrete case. Since $f(x) \geq 0 \; \forall \; x$, $P_X(A) = \sum_{x \in A} f(x) \geq 0$ for any event $A$, and Axiom 1.1 is satisfied. Letting $R(X)$ equal the set $C$ defined in Def. 2.5.a, it follows that $P_X(R(X)) = \sum_{x \in R(X)} f(x) = 1$, satisfying Axiom 1.2. Finally, if $\bigcup_{i \in I} A_i$ is the union of a collection of disjoint events indexed by the index set $I$, then summing over all of the elementary events in $A = \bigcup_{i \in I} A_i$ obtains

$$P_X\left(\bigcup_{i \in I} A_i\right) = \sum_{x \in A} f(x) = \sum_{i \in I}\left(\sum_{x \in A_i} f(x)\right) = \sum_{i \in I} P_X(A_i).$$

Thus, the three probability axioms are satisfied, and $P_X$ is a probability set function.

To motivate *sufficiency* in the continuous case, first note that Axiom 1.1 is satisfied since if $f(x) \geq 0 \; \forall \; x$, then $P_X(A) = \int_{x \in A} f(x)dx \geq 0$, because integrating a nonnegative integrand over any interval (or Borel) set, $A$, results in a nonnegative number. Furthermore, since $\int_{-\infty}^{\infty} f(x)dx = 1$, there exists at least one event $A \subset (-\infty, \infty)$ such that $\int_{x \in A} f(x)dx = 1$ (the event can be $(-\infty, \infty)$ itself, or else there may be some other partition of $(-\infty, \infty)$ into $A \cup B$ such that $\int_{x \in A} f(x)dx = 1$ and $\int_{x \in B} f(x)dx = 0$). Letting $R(X) = A$, we have that $P_X(R(X)) = \int_{x \in R(X)} f(x)dx = 1$ and Axiom 1.2 is satisfied. Finally, if $D = \bigcup_{i \in I} A_i$ is the union of a collection of disjoint events indexed by the index set $I$, then by the additive property of integrals

$$P_X(D) = \int_{x \in D} f(x)dx = \sum_{i \in I}\left(\int_{x \in A_i} f(x)dx\right) = \sum_{i \in I} P_X(A_i).$$

Thus, the three probability axioms are satisfied, and $P_X$ is a probability set function.

It can also be shown that the function properties presented in Def. 2.5 are actually *necessary* for the discrete case and *practically necessary* in the continuous case. For the discrete case, first recall that $f(x)$ is directly interpretable as the probability of the outcome $x$, and this requires that $f(x) \geq 0 \; \forall \; x \in R$ (or else we would be assigning negative probabilities to some $x$'s). Second, the number of outcomes that can receive positive probability must be countable in the discrete case since $R(X)$ is countable, leading to the requirement that $C = \{x \colon f(x) > 0, x \in R\}$ is countable. Finally, $\sum_{x \in C} f(x) = 1$ is required if the probability assigned to the certain event is to equal 1.

In the continuous case, it is necessary that $\int_{-\infty}^{\infty} f(x)dx = 1$. To see this, first note that $R(X)$ is the certain event, implying $P(R(X)) = 1$. Now since $R(X)$ and $\overline{R(X)}$ are disjoint, we have that $P_X(R(X) \cup \overline{R(X)}) = P_X(R(X)) + P_X(\overline{R(X)}) = 1 + P_X(\overline{R(X)})$, which implies $P_X(\overline{R(X)}) = 0$ since probabilities cannot exceed 1. But, since $R(X) \cup \overline{R(X)} = R$ by definition, then $P(R(X) \cup \overline{R(X)}) = \int_{-\infty}^{\infty} f(x)dx = 1$. Regarding the requirement that $f(x) \geq 0$ for $x \in (-\infty, \infty)$, note that the condition is technically *not necessary*. It is known from the properties of integrals that the value of $\int_a^b f(x)dx$ is invariant to changes in the value of $f(x)$ at a finite number of points, and thus $f(x)$ could technically be negative for a finite number of $x$ values without affecting the values of the probability set function. As others do, we will ignore this technical anomaly since its practical significance in defining probability density functions is nil. We thus insist, as a practical matter, on the nonnegativity of $f(x)$. It also follows from the preceding invariance property of the integral that density functions are *not* unique for representing probability set functions in the continuous case.

**Example 2.7** For each case, determine whether the stated function can serve as a probability density function.

**a.** $f(x) = (1/2)I_{[0,2]}(x)$.
**Answer**: The function can serve as a continuous probability density function since $f(x) \geq 0 \; \forall \; x \in (-\infty, \infty)$ (note $f(x) = 1/2 > 0 \; \forall \; x \in [0, 2]$ and $f(x) = 0$ for $x \notin [0, 2]$), and $\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} (1/2)I_{[0,2]}(x)dx = \int_0^2 (1/2)dx = \frac{x}{2} \big|_0^2 = 1$.

**b.** $f(x) = (.3)^x(.7)^{1-x}I_{\{0,1\}}(x)$.
**Answer**: The function can serve as a discrete probability density function, since $f(x) > 0$ on the countable set $\{0, 1\}$, $\sum_{x=0}^{1} f(x) = 1$, and $f(x) = 0 \; \forall \; x \notin \{0, 1\}$.

**c.** $f(x) = (x^2 + 1)I_{[-1,1]}(x)$.
**Answer**: The function *cannot* serve as a probability density function. While $f(x) \geq 0 \; \forall \; x \in (-\infty, \infty)$, the function does not integrate to 1:

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^{\infty} (x^2 + 1)I_{[-1,1]}(x)dx \\ &= \int_{-1}^{1} (x^2 + 1)dx \\ &= \frac{x^3}{3} + x \, \Big|_{-1}^{1} = \frac{8}{3} \neq 1.\end{aligned}$$

**d.** $f(x) = (3/8)(x^2 + 1)I_{[-1,1]}(x)$.
**Answer**: The reader should demonstrate that this function *can* serve as a continuous probability density function. Note its relationship to the function in part c. □

### *Mixed Discrete-Continuous Random Variables

The categories of *discrete* and *continuous* random variables do not exhaust the possible types of random variables. There is a category of random variable called *mixed discrete-continuous* which exhibits the characteristics of a discrete random variable for some events and the characteristics of a continuous random variable for other events. In particular, a mixed discrete-continuous random variable is such that a countable subset of the elementary events is assigned positive probability, as in the case of a discrete random variable, *except* the sum of the probabilities over the countable set does *not* equal 1. The remaining probability is attributable to an uncountable collection of elementary events, each elementary event being assigned zero probability, as in the case of a continuous random variable. The following example illustrates the concept of a mixed discrete-continuous random variable.

**Example 2.8** Let $X$ be a random variable representing the length of time, measured in thousands of hours, that an active matrix color screen for a notebook computer operates properly. Assume the probability set function associated with the random variable is $P_X(A) = .25I_A(0) + .75\int_{x\in A-\{0\}} e^{-x}I_{(0,\infty)}(x)dx$ for every event $A$ (i.e., Borel set) contained in $R(X) = [0, \infty)$.

**a.** What is the probability that the color screen is defective, i.e., it does not function properly at the outset?
**Answer**: The event in question is $A = \{0\}$. Using $P_X$, we calculate the probability to be $P_X(\{0\}) = .25I_{\{0\}}(0) + .75\int_{x\in\emptyset} e^{-x}dx = .25$. (Note: By definition, $\int_{x\in\emptyset} f(x)dx = 0$.)

**b.** What is the probability that the color screen operates satisfactorily for less than 1,000 hours?
**Answer**: Here, $A = [0, 1)$. Using $P_X$, we calculate $P_X([0, 1)) = .25I_{[0,1)}(0) + .75\int_0^1 e^{-x}dx = .25 + .474 = .724$.

**c.** What is the probability that the color screen operates satisfactorily for at least 500 hours?
**Answer**: The event in question is $A = [.5, \infty)$. The probability assigned to this event is given by $P_X([.5, \infty)) = .25I_{[.5,\infty)}(0) + .75\int_{.5}^{\infty} e^{-x}dx = 0 + .6065 = .6065$. □

We formalize the concept of a mixed discrete-continuous random variable in the following definition.

**Definition 2.6** ***Mixed discrete-continuous random variables***

A random variable is called mixed discrete-continuous iff

**a.** its range is uncountably infinite;

**b.** there exists a countable set, $C$, of outcomes of $X$ such that $P_X(\{x\}) = f_d(x) > 0 \ \forall \ x \in C$, $f_d(x) = 0 \ \forall \ x \notin C$, and $\sum_{x \in C} f_d(x) < 1$, where the function $f_d$ is referred to as the **discrete density function component** of the probability set function of $X$;

**c.** there exists a nonnegative-valued function, $f_c$, defined for all $x \in (-\infty, \infty)$ such that for every event $B \subset R(X) - C$, $P_X(B) = \int_{x \in B} f_c(x)dx$, $f(x) = 0 \ \forall \ x \in R - R(X)$, and $\int_{-\infty}^{\infty} f_c(x)dx = 1 - \sum_{x \in C} f_d(x)$, where the function $f_c$ is referred to as the **continuous density function component** of the probability set function of $X$; and

**d.** the probability set function for $X$ is given by combining or *mixing* the discrete and continuous density function components in parts (b) and (c) above, as $P_X(A) = \sum_{x \in A \cap C} f_d(x) + \int_{x \in A - C} f_c(x)dx$ for every event $A$.

To see how the definition applies to a specific experimental situation, recall Ex. 2.8. If we substitute $f_d(x) = .25I_{\{0\}}(x)$, $C = \{0\}$, and $f_c(x) = .75e^{-x}I_{(0,\infty)}(x)$ into the definition of $P_X$ given in Def. 2.6.d, we obtain

$$\begin{aligned} P_X(A) &= \sum_{x \in A \cap \{0\}} (.25I_{\{0\}}(x)) + .75 \int_{x \in A - \{0\}} e^{-x} I_{(0,\infty)}(x)dx \\ &= .25I_A(0) + .75 \int_{x \in A - \{0\}} e^{-x} I_{(0,\infty)}(x)dx \end{aligned}$$

which is identical to the probability set function defined in Ex. 2.8.

As the reader may have concluded from examining the definition, the concept of a mixed discrete-continuous random variable is more complicated than either the discrete or continuous random-variable case, since there is no single PDF that can be either summed or integrated to define probabilities of events.[6] On the other hand, once the discrete and continuous random-variable concepts are understood, the notion of a mixed discrete-continuous random variable is a rather straightforward conceptual extension. Note that the definition of the probability set function in Def. 2.6.d essentially implies that to calculate the probability of an event $A$, one adds together the probabilities of the mutually exclusive events $A \cap C$ and $A - C$. Assigning probability to the event $A \cap C$ is done in a way that emulates the discrete random-variable case—a real-valued function (the discrete density component) is summed over the points in the event $A \cap C$. The probability of the event $A - C$ is calculated in a way that emulates

[6] We are somewhat hampered here by the scope of our study, where we are restricted to Riemann-type integrals. In a more advanced treatment of the subject, we could resort to the use of Stieltjes integrals, in which case a single *integral* could once again be used to define $P_X$. On Stieltjes integration, see R. G. Bartle (1976), *The Elements of Real Analysis*, 2nd ed. New York: John Wiley.

the continuous random-variable case—a real-valued function (the continuous density component) is integrated over the points in the event $A - C$. Adding together the results obtained for the "discrete" event $A \cap C$ and the "continuous" event $A - C$ defines the probability of the "mixed" event $A$.

Having introduced the concept of a mixed discrete-continuous random variable, we now note that there is an alternative definition for $P_X$ that can simplify the integral involving the continuous density component. The alternative definition can be used if the countable set $C$ is such that no more than a finite number of its elements is contained in any interval of finite length (which will be true in virtually all applications). In this case, $\int_{x \in A-C} f_c(x)dx = \int_{x \in A} f_c(x)dx$, since the integral over the countable points in $A \cap C$ will be zero, and we can define $P_X$ in the mixed discrete-continuous case alternatively as $P_X(A) = \sum_{x \in A \cap C} f_d(x) + \sum_{x \in A} f_c(x)dx$. We will incorporate this alternative definition in our discussion of cumulative distribution functions, which occurs in the next section.[7]

## 2.3 Univariate Cumulative Distribution Functions

Situations arise in practice that require finding the probability of the event that the outcome of a random variable is less than or equal to some real number, i.e., the event in question is $\{x : x \leq b, x \in R(X)\}$ for some real number $b$. These types of probabilities are given directly by the **cumulative distribution function** (CDF), which we introduce in this section.

Henceforth, we will eliminate the random variable subscript used heretofore in our probability set function notation; we will now write $P(A)$ rather than $P_X(A)$ wherever the context makes it clear to which probability space the event $A$ refers. Thus, the notation $P(A)$ may be used to represent the probability of either an event $A \subset S$ or an event $A \subset R(X)$. To economize on notation further, we introduce an *abbreviated set definition* for representing events.

**Definition 2.7**
*Abbreviated set definition for events*

Let an event be represented in the form $\{x$: set defining conditions, $\{x \in R(X)\}$ and the probability of the event be represented by $P(\{x$: set defining conditions, $x \in R(X)\})$ (alternatively, $S$ can be used in place of $R(X)$). The **abbreviated set definition** for the event, and the probability of the event expressed in terms of the abbreviated set definition, are respectively {set-defining conditions} and $P$(set-defining conditions), the condition $x \in R(X)$ always being tacitly assumed.

For an example of an abbreviated set definition that is particularly relevant to our current discussion of CDFs, note that $\{x \leq b\}$ will be used to represent

[7] There are still other types of random variables besides those we have examined, but they are rarely utilized in applied work. See T. S. Chow and H. Teicher (1978), *Probability Theory*, New York: Springer-Verlag, pp. 247–248.

$\{x\colon x \le b, x \in R(X)\}$, and $P(x \le b)$ will be used to represent $P(\{x\colon x \le b, x \in R(X)\})$.[8]

The formal definition of the cumulative distribution function, and its particular algebraic representations in the discrete, continuous, and mixed discrete-continuous cases, are given next.

**Definition 2.8** *Univariate cumulative distribution function*

The cumulative distribution function of a random variable $X$ is defined by $F(b) = P(x \le b)\ \forall\ b \in (-\infty, \infty)$. The functional representation of $F(b)$ in particular cases is as follows:

**a.** *Discrete X*:

$$F(b) = \sum_{\substack{x \le b \\ f(x)>0}} f(x), \quad \text{for } b \in (-\infty, \infty);$$

**b.** *Continuous X*:

$$F(b) = \int_{-\infty}^{b} f(x)dx, \quad \text{for } b \in (-\infty, \infty);$$

***c.** *Mixed discrete-continuous X*:

$$F(b) = \sum_{\substack{x \le b \\ f_d(x)>0}} f_d(x) + \int_{-\infty}^{b} f_c(x)dx, \quad \text{for } b \in (-\infty, \infty).$$

Examples of cumulative distribution functions now follow.

**Example 2.9** Reexamine Ex. 2.6, where the amount of time that passes between work-related injuries is observed. We can define the cumulative distribution function for $X$ as

$$\begin{aligned} F(b) &= \int_{-\infty}^{b} \frac{1}{100} e^{-x/100} I_{(0,\infty)}(x)dx \\ &= \left[\int_{0}^{b} \frac{1}{100} e^{-x/100} dx\right] I_{(0,\infty)}(b) = \left[1 - e^{-b/100}\right] I_{(0,\infty)}(b). \end{aligned}$$

If one were interested in the event that an injury occurs within 50 hours of the previous injury, the probability would be given by

$$F(50) = [1 - e^{-50/100}]I_{(0,\infty)}(50) = 1 - .61 = .39.$$

A graph of the cumulative distribution function is given in Figure 2.4. □

[8] Alternative shorthand notation that is often used in the literature is respectively $\{X \le b\}$ and $P(X \le b)$. Our notation establishes a distinction between the function $X$ and a value of the function $x$.

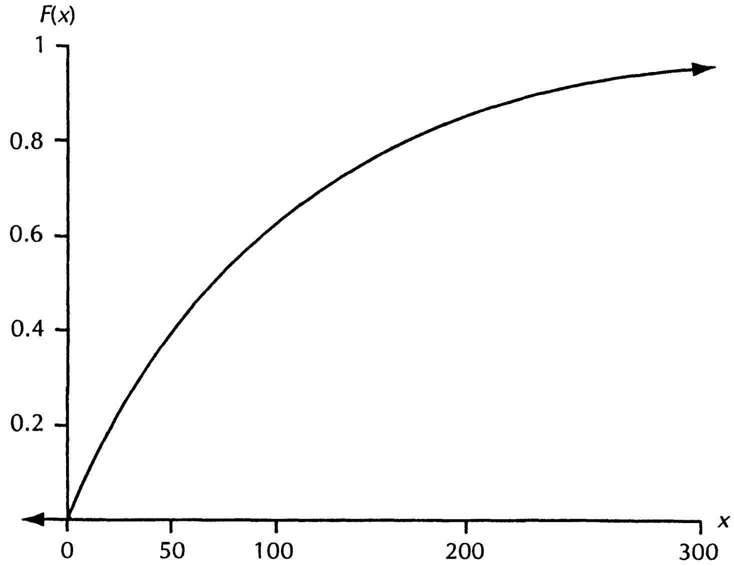


**Figure 2-4**
A CDF for a continuous $X$.

**Example 2.10** Examine the experiment of rolling a fair die and observing the number of dots facing up. Let the random variable $X$ represent the possible outcomes of the experiment, so that $R(X) = \{1, 2, 3, 4, 5, 6\}$ and $f(x) = (1/6)I_{\{1,2,3,4,5,6\}}(x)$. The cumulative distribution function for $X$ can be defined as

$$F(b) = \sum_{\substack{x \leq b \\ f(x)>0}} \tfrac{1}{6} I_{\{1,2,3,4,5,6\}}(x) = \tfrac{1}{6} \operatorname{trunc}(b) I_{[0,6)}(b) + I_{[6,\infty)}(b),$$

where trunc($b$) is the **truncation function** defined by assigning to any domain element $b$ the number resulting after truncating the decimal part of $b$. For example, trunc(5.97) = 5, or trunc(−2.12) = −2. If we were interested in the probability of tossing a 3 or less, the probability would be given by

$$F(3) = \tfrac{1}{6} \operatorname{trunc}(3) I_{[0,6)}(3) + I_{[6,\infty)}(3) = \tfrac{1}{2} + 0 = \tfrac{1}{2}.$$

A graph of the cumulative distribution function is given in Figure 2.5. □

***Example 2.11** Recall Ex. 2.8, where color screen lifetimes were represented by a mixed discrete-continuous random variable. The cumulative distribution for $X$ is given by

$$\begin{aligned} F(b) &= .25 I_{[0,\infty)}(b) + .75 \int_{-\infty}^{b} e^{-x} I_{(0,\infty)}(x) dx \\ &= .25 I_{[0,\infty)}(b) + .75[1 - e^{-b}] I_{(0,\infty)}(b). \end{aligned}$$

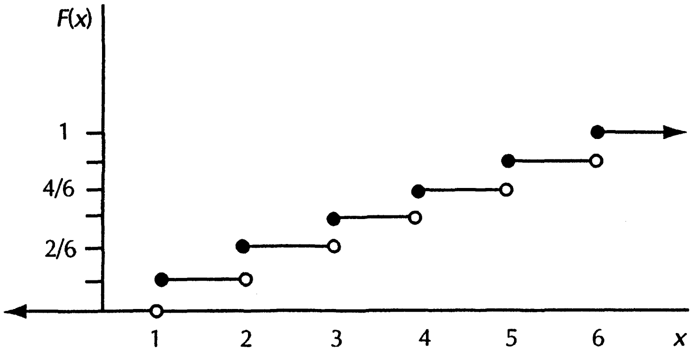


**Figure 2-5**
A CDF for a discrete $X$.

If one were interested in the probability that the color screen functioned for 1,000 hours or less, the probability would be given by

$$
\begin{aligned}
F(1) &= .25I_{[0,\infty)}(1) + .75[1 - e^{-1}]I_{(0,\infty)}(1) \\
&= .25 + .474 = .724.
\end{aligned}
$$

A graph of the cumulative distribution function is given in Figure 2.6. □

### CDF Properties

The graphs in the preceding examples actually illustrate some general properties of CDFs (the reader is asked to provide the appropriate justification for the properties that follow). First, CDFs have the entire real line for their domain, while their range is contained in the interval [0, 1]. Second, $\lim_{b \to -\infty} F(b) =$

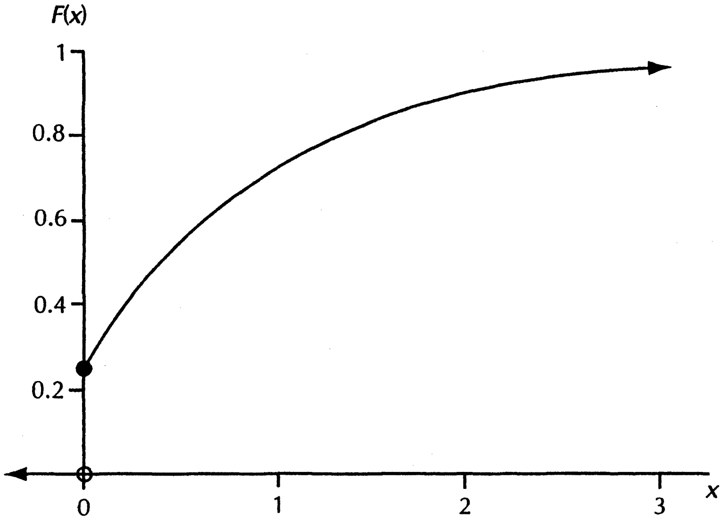


**Figure 2-6**
A CDF for a mixed discrete-continuous $X$.

$\lim_{b \to -\infty} P(x \leq b) = P(\emptyset) = 0$, and $\lim_{b \to \infty} F(b) = \lim_{b \to \infty} P(x \leq b) = P(R(X)) = 1$.[9] It is also true that if $a < b$, then necessarily $F(a) = P(x \leq a) \leq P(x \leq b) = F(b)$, which is the defining property of what is called a **nondecreasing function.**

The CDFs of discrete, continuous, and mixed discrete-continuous random variables can be distinguished by their continuity properties and by the behavior of $F(b)$ on sets of domain elements for which $F$ is continuous. The CDF of a continuous random variable must be a continuous function on the entire real line, as illustrated in Figure 2.4, for suppose there existed a discontinuous "jumping up" point at a point $d$. Then $P(x = d) = \lim_{b \to d^-} P(b < x \leq d) = F(d) - \lim_{b \to d^-} F(b) > 0$ because of the discontinuity (see Figure 2.7), contradicting that $P(x = d) = 0 \ \forall \ d$ if $X$ is continuous.[10]

The CDFs for both discrete and mixed discrete-continuous random variables exhibit a countable number of discontinuities at "jumping up" points representing the assignments of positive probabilities to a countable number of elementary events (recall Figures 2.5 and 2.6). The discrete case is distinguished from the mixed case by the property that the CDF in the former case is a constant function on all intervals for which $F$ is continuous. The mixed case will have a CDF that is an increasing function of x on one or more interval subsets of the real line.[11]

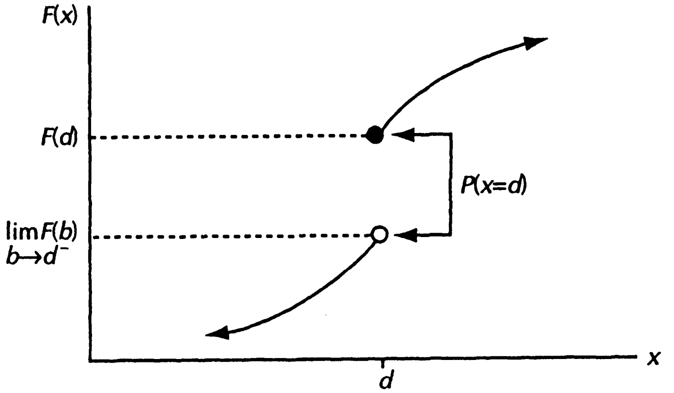


**Figure 2-7**
Discontinuity in a CDF.

[9] For those readers whose recollection of the limit concept from calculus courses is not clear, it suffices here to appeal to intuition and interpret the limit of $F(b)$ as "the real number to which $F(b)$ becomes and remains infinitesimally close as $b$ increases without bound (or as $b$ decreases without bound)." We will examine the limit concept in more detail in Chapter 5.

[10] $\lim_{b \to d^-}$ indicates that we are examining the limit as $b$ approaches $d$ from below (also called a left-hand limit). $\lim_{b \to d^+}$ would indicate the limit as $b$ approached $d$ from above (also called a right-hand limit). For now, it will suffice for the reader to appeal to intuition and interpret $\lim_{b \to d^-} F(b)$ as "the real number to which $F(b)$ becomes and remains infinitesimally close as $b$ increases and becomes infinitesimally close to $d$."

[11] $F$ is an **increasing function** on an interval of points $[a, b]$ if $\forall \ x_i$ and $x_j \in [a, b]$ for which $x_i < x_j$, $F(x_i) < F(x_j)$.

### Duality Between CDFs and PDFs

A CDF can be used to derive a probability density function as well as discrete and continuous density components in the mixed discrete-continuous random-variable case.

**Theorem 2.1** (Discrete PDFs from CDFs) *Let $x_1 < x_2 < x_3 < \cdots$ be the countable collection of outcomes in the range of the discrete random variable X. Then the discrete probability density function for X can be defined as*

$$f(x_1) = F(x_1),$$
$$f(x_i) = F(x_i) - F(x_{i-1}), \quad i = 2, 3, \ldots,$$
$$f(x) = 0 \quad \textit{for } x \notin R(X).$$

**Proof** The proof follows directly from the definition of the CDF and is left to the reader. ■

**Theorem 2.2** (Continuous PDFs from CDFs) *Let $f(x)$ and $F(x)$ represent the probability density function and CDF, respectively, for the continuous random variable X. The density function for X can be defined as $f(x) = dF(x)/dx$ wherever $f(x)$ is continuous, and $f(x) = 0$ (or any nonnegative number) elsewhere.*

**Proof** By the fundamental theorem of calculus (recall Lemma 2.1), it follows that

$$\frac{dF(x)}{dx} = \frac{d\int_{-\infty}^{x} f(t)dt}{dx} = f(x)$$

wherever $f(x)$ is continuous, and the first part of the theorem is demonstrated. Now, since $X$ is a continuous random variable, then $P(x \leq b) = F(b) = \int_{-\infty}^{b} f(x)dx$ exists ∀ $b$ by definition. Changing the value of the nonnegative integrand at points of discontinuity will have no effect on the value of $F(b) = \int_{-\infty}^{b} f(x)dx$,[12] so that $f(x)$ can be defined arbitrarily at the points of discontinuity. ■

***Theorem 2.3** (Density Components of a Mixed Discrete-Continuous Random Variable from CDFs) *Let X be a mixed discrete-continuous random variable with a CDF, F. Let $x_1 < x_2 < x_3 < \cdots$ be the countable collection of outcomes of X for which $F(x)$ is discontinuous. Then the discrete density component of X can be defined as $f_d(x_i) = F(x_i) - \lim_{b \to x_i^-} F(b)$ for $i = 1, 2, 3, \ldots$; and $f_d(x) = 0$ elsewhere.*

*The continuous density component of X can be defined as $f_c(x) = dF(x)/dx$ wherever $f(x)$ is continuous, and $f_c(x) = 0$ (or any arbitrary nonnegative number) elsewhere.*

[12] This can be rigorously justified by the fact that under the conditions stated: (1) the (improper) Riemann integral is equivalent to a Lebesque integral; (2) the largest set of points for which $f(x)$ can be discontinuous and still have the integral $\int_{-\infty}^{b} f(x)dx$ defined ∀ $b$ has "measure zero"; and (3) the values of the integrals are unaffected by changing the values of the integrand on a set of points having "measure zero." This result applies to multivariate integrals as well. See C. W. Burill (1972), *Measure, Integration, and Probability*. New York: McGraw-Hill, pp. 106–109, for further details.

**Proof** The proof is a combination of the arguments used in the proofs of the preceding two theorems and is left to the reader. ■

Given Theorems 2.1 to 2.3, it follows that there is a **duality between CDFs and probability density functions** whereby either function can be derived from the other. We illustrate Theorems 2.1 to 2.3 in the following examples.

**Example 2.12** Recall Ex. 2.10, where the outcome of rolling a fair die is observed. We can define the discrete density function for $X$ using the CDF for $X$ as follows:

$$f(1) = F(1) = \tfrac{1}{6},$$

$$f(x) = \begin{cases} F(x) - F(x-1) = \dfrac{x}{6} - \dfrac{x-1}{6} = \dfrac{1}{6} & \text{for } x = 2, 3, 4, 5, 6, \\ 0 \quad \text{elsewhere.} \end{cases}$$

A more compact representation of $f(x)$ can be given as

$$f(x) = \tfrac{1}{6} I_{\{1,2,3,4,5,6\}}(x),$$

which we know to be the appropriate discrete density function for the case at hand. □

**Example 2.13** Recall Ex. 2.9, where the time that passes between work-related injuries is observed. We can define the continuous density function for $X$ using the stated CDF for $X$ as follows:

$$f(x) = \begin{cases} \dfrac{dF(x)}{dx} = \dfrac{d(1 - e^{-x/100})I_{(0,\infty)}(x)}{dx} = \dfrac{1}{100} e^{-x/100} & \text{for } x \in (0, \infty), \\ 0 \quad \text{for } x \in (-\infty, 0). \end{cases}$$

The derivative of $F(x)$ does not exist at the point $x = 0$ (recall Figure 2.4), a reflection of the fact that $f(x)$ is discontinuous at $x = 0$. We arbitrarily assign $f(x) = 0$ when $x = 0$ so that the density function of x is ultimately defined by $f(x) = (1/100)e^{-x/100} I_{(0,\infty)}(x)$, which we know to be an appropriate continuous density function for the case at hand. □

***Example 2.14** Recall Ex. 2.11, where the operating lives of notebook color screens are observed. The CDF of the mixed discrete-continuous random variable $X$ is discontinuous only at the point $x = 0$ (recall Figure 2.6). Then the discrete density component of $X$ is given by

$$f_d(0) = F(0) - \lim_{b \to 0^-} F(b) = .25 - 0 = .25$$

and

$$f_d(x) = 0, \qquad x \neq 0,$$

or alternatively,

$$f_d(x) = .25 I_{\{0\}}(x),$$

which we know to be the appropriate discrete density function component in this case.

The continuous density function component can be defined as

$$f_c = \begin{cases} \dfrac{dF(x)}{dx} = .75e^{-x} & \text{for } x \in (0, \infty), \\ 0 \quad \text{for } x \in (-\infty, 0), \end{cases}$$

but the derivative of $F(x)$ does not exist at the point $x = 0$ (recall Figure 2.6). We arbitrarily assign $f_c(x) = 0$ when $x = 0$ so that the continuous density function component of $X$ is finally representable as $f_c(x) = .75e^{-x} I_{(0,\infty)}(x)$, which we know to be an appropriate continuous density function component in this case. □

## 2.4 Multivariate Random Variables, PDFs, and CDFs

In the preceding sections of this chapter, we have examined the concept of a univariate random variable, where only one real-valued function was defined on the elements of a sample space. The concept of a multivariate random variable is an extension of the univariate case, where two or more real-valued functions are concurrently defined on the elements of a given sample space. Underlying the concept of a multivariate random variable is the notion of a **real-valued vector function**, which we'll now define.

**Definition 2.9**
***Real-valued vector function***

Let $f_i\colon A \rightarrow R$, $i = 1, \ldots, n$, be a collection of $n$ real-valued functions, each function being defined on the domain $A$. Then the function $f\colon A \rightarrow R^n$ defined by

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix} = f(x), \quad \text{for } x \in A,$$

is called an **(*n*-dimensional) real-valued vector function**. The real-valued functions $f_1, \ldots, f_n$ are called **coordinate functions** of the vector function $f$.

Note that the real-valued vector function $f\colon A \rightarrow R^n$ is distinguished from the scalar function $f\colon A \rightarrow R$ by the fact that its range elements are *n-dimensional vectors* of real numbers as opposed to scalar real numbers. The range of the real-valued vector function is given by $R(f) = \{(y_1, \ldots, y_n)\colon y_i = f_i(x), i = 1, \ldots, n; x \in A\}$.

We now provide a formal definition of the notion of a multivariate random variable.

**Definition 2.10**
***Multivariate (n-variate) random variable***

Let $\{S, \Upsilon, P\}$ be a probability space. If $X\colon S \rightarrow R^n$ (or simply $X$) is a real-valued vector function having as its domain the elements of $S$, then $X\colon S \rightarrow R^n$ (or $X$) is called a **multivariate (*n*-variate) random variable**.

Since the multivariate random variable is defined by

$$\underset{(n\times 1)}{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} X_1(w) \\ X_2(w) \\ \vdots \\ X_n(w) \end{bmatrix} = \underset{(n\times 1)}{\mathbf{X}(w)} \quad \text{for } w \in S,$$

it is admissible to interpret $X$ as a collection of $n$ univariate random variables, each defined on the same probability space $\{S, \Upsilon, P\}$. The range of the $n$-variate random variable is given by $R(X) = \{(x_1, \ldots, x_n) : x_i = X_i(w), i = 1, \ldots, n; w \in S\}$.

The multivariate random-variable concept applies quite naturally and generally to any real-world experiment in which more than one characteristic is observed for each outcome of the experiment. For example, upon making an observation concerning a futures trade on the Chicago Mercantile Exchange, one could record the price, quantity, delivery date, and commodity grade associated with the trade. Upon conducting a poll of registered voters, one could record various political preferences and a myriad of sociodemographic data associated with each randomly chosen interviewee. Upon making a sale, a car dealership will record the price, model, year, color, and the selections from the options list that were made by the buyer. Applications abound in business and economic settings, and the reader is asked to provide additional examples of the multivariate random-variable concept.

Definitions for the concept of *discrete* and *continuous* multivariate random variables and their associated density functions are as follows:

**Definition 2.11**
***Discrete multivariate random variables and discrete joint probability density functions***

A multivariate random variable is called *discrete* if its range consists of a countable number of elements. The **discrete joint probability density function,** $f$, for a discrete multivariate random variable $X = (X_1, \ldots, X_n)$ is defined as $f(x_1, \ldots, x_n) =$ probability of $(x_1, \ldots, x_n)$ if $(x_1, \ldots, x_n) \in R(X)$, and 0 otherwise.

**Definition 2.12**
***Continuous multivariate random variables and continuous joint probability density functions***

A multivariate random variable is called *continuous* if its range is uncountably infinite and if there exists a nonnegative-valued function $f(x_1, \ldots, x_n)$, defined for all $(x_1, \ldots, x_n) \in R^n$, such that for any event $A \subset R(X)$,

$$P(A) = \underset{(x_1,\ldots,x_n)\in A}{\int \cdots \int} f(x_1, \ldots, x_n) dx_1 \cdots dx_n$$

and

$$f(x_1, \ldots, x_n) = 0 \quad \forall\, (x_1, \ldots, x_n) \notin R(X).$$

The function $f(x_1, \ldots, x_n)$ is called a **continuous joint probability density function.**

### Multivariate Random Variable Properties and Classes of PDFs

A number of properties of discrete and continuous multivariate random variables, and their joint probability densities, can be identified through analogy with the univariate case. In particular, the multivariate random variable induces a new probability space, $\{R(X), \Upsilon_X, P_X\}$, for the experiment, where the event space $\Upsilon_X$ consists of all subsets of $R(X)$ in the discrete case and all Borel subsets of $R(X)$ in the continuous case. The rationale underlying the transition from the probability space $\{S, \Upsilon, P\}$ to the induced probability space $\{R(X), \Upsilon_X, P_X\}$ is precisely the same as in the univariate case except for the increased dimensionality of the elements in $R(X)$ in the multivariate case. The probability set function defined on the events in the event space is represented in terms of multiple summation of a PDF in the discrete case and multiple integration of a PDF in the continuous case. In the discrete case, $f(x_1, \ldots, x_n)$ is directly interpretable as the probability of the outcome $(x_1, \ldots, x_n)$; in the continuous case, the probability of each elementary event is zero and $f(x_1, \ldots, x_n)$ is not interpretable as a probability. As a matter of convenience, both density functions are defined to have the entire $n$-dimensional real space for their domain, so that $f(x_1, \ldots, x_n) = 0 \; \forall \, x \notin R(X)$.

Regarding the classes of functions that can be used as discrete or continuous joint density functions, we provide the following generalization of Def. 2.5:

**Definition 2.13**
***The classes of discrete and continuous joint probability density functions***

a. **Class of discrete joint density functions**. A function $f\colon R^n \rightarrow R$ is a member of the class of discrete joint density functions iff:
   1. the set $C = \{(x_1, \ldots, x_n)\colon f(x_1, \ldots, x_n) > 0, (x_1, \ldots, x_n) \in R^n\}$ is countable,
   2. $f(x_1, \ldots, x_n) = 0$ for $x \in \bar{C}$, and
   3. $\sum \cdots \sum_{(x_1, \ldots, x_n) \in C} f(x_1, \ldots, x_n) = 1$.
b. **Class of continuous joint density functions**. A function $f\colon R^n \rightarrow R$ is a member of the class of continuous joint density functions iff:
   1. $f(x_1, \ldots, x_n) \geq 0 \; \forall \, (x_1, \ldots, x_n) \in R^n$ and
   2. $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) dx_1 \cdots dx_n = 1$.

The reader can generalize the arguments used in the univariate case to demonstrate that the properties stated in Def. 2.13 are sufficient, as well as necessary in the discrete case and "almost necessary" in the continuous case, for set functions defined as

$$P(A) = \begin{cases} \displaystyle\sum_{(x_1, \ldots, x_n) \in A} \cdots \sum f(x_1, \ldots, x_n) & \text{(discrete case),} \\ \displaystyle\int_{(x_1, \ldots, x_n) \in A} \cdots \int f(x_1, \ldots, x_n) dx_1 \cdots dx_n & \text{(continuous case)} \end{cases}$$

to satisfy the probability axioms $\forall\, A \in \Upsilon_X$.

The following is an example of the specification of the probability space for a bivariate discrete random variable.

**Example 2.15** For the experiment of rolling a pair of dice in Ex. 2.2, distinguish the two die by letting the first die be "red" and second "green." Thus an outcome $(i, j)$ refers to $i$ dots on the red die and $j$ dots on the green die. Define the following two random variables:

$$x_1 = X_1(w) = i \quad \text{and} \quad x_2 = X_2(w) = i + j.$$

The range of the bivariate random variable $(X_1, X_2)$ is given by $R(X) = \{(x_1, x_2) : x_1 = i, x_2 = i + j, i \text{ and } j \in \{1, \ldots, 6\}\}$. The event space is $\Upsilon_X = \{A : A \subset R(X)\}$.

The correspondence between elementary events in $R(X)$ and elementary events in $S$ is displayed as follows:

| | | | | $x_1$ | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| | 2 | (1, 1) | ∅ | ∅ | ∅ | ∅ | ∅ |
| | 3 | (1, 2) | (2, 1) | ∅ | ∅ | ∅ | ∅ |
| | 4 | (1, 3) | (2, 2) | (3, 1) | ∅ | ∅ | ∅ |
| | 5 | (1, 4) | (2, 3) | (3, 2) | (4, 1) | ∅ | ∅ |
| | 6 | (1, 5) | (2, 4) | (3, 3) | (4, 2) | (5, 1) | ∅ |
| $x_2$ | 7 | (1, 6) | (2, 5) | (3, 4) | (4, 3) | (5, 2) | (6, 1) |
| | 8 | ∅ | (2, 6) | (3, 5) | (4, 4) | (5, 3) | (6, 2) |
| | 9 | ∅ | ∅ | (3, 6) | (4, 5) | (5, 4) | (6, 3) |
| | 10 | ∅ | ∅ | ∅ | (4, 6) | (5, 5) | (6, 4) |
| | 11 | ∅ | ∅ | ∅ | ∅ | (5, 6) | (6, 5) |
| | 12 | ∅ | ∅ | ∅ | ∅ | ∅ | (6, 6) |

Elementary Events in $S$

It follows immediately from the correspondence with the probability space $\{S, \Upsilon, P\}$ that the discrete density function for the bivariate random variable $(X_1, X_2)$ can be represented as

$$f(x_1, x_2) = \tfrac{1}{36} I_{\{1,\ldots,6\}}(x_1) I_{\{1,\ldots,6\}}(x_2 - x_1),$$

and the probability set function defined on the events in $R(X)$ is then

$$P(A) = \sum_{(x_1,x_2)\in A}\sum f(x_1, x_2) \quad \text{for } A \in \Upsilon_X.$$

Let $A = \{(x_1, x_2) : 1 \le x_1 \le 2, 2 \le x_2 \le 5, (x_1, x_2) \in R(X)\}$, which is the event of rolling 2 or less on the red die and a total of 5 or less on the pair of dice. Then the probability of this event is given by

$$P(A) = \sum_{(x_1,x_2)\in A}\sum f(x_1, x_2) = \sum_{x_1=1}^{2} \sum_{x_2=x_1+1}^{5} f(x_1, x_2) = \frac{7}{36}. \qquad \square$$

The preceding example illustrates two general characteristics of the multivariate random-variable concept that should be noted. First of all, even though a multivariate random variable can be viewed as a collection of univariate random variables, it is *not* necessarily the case that the range of the multivariate random variable $X$ equals the Cartesian product of the ranges of the univariate random variable defining $X$. Depending on the definition of the $X_i$'s, either $R(X) \neq \times_{i=1}^{n} R(X_i)$ and $R(X) \subset \times_{i=1}^{n} R(X_i)$, or $R(X) = \times_{i=1}^{n} R(X_i)$ is possible. Example 2.15 is an example of the former case, where a number of scalar outcomes that are *individually* possible for the univariate random variables $X_1$ and $X_2$ are not *simultaneously* possible as outcomes for the bivariate random variable $(X_1, X_2)$. Second, note that our convention of defining $f(x_1, x_2) = 0 \;\forall\; (x_1, x_2) \notin R(X)$ allows an alternative summation expression for defining the probability of event $A$ in Ex. 2.15:

$$P(A) = \sum_{x_1=1}^{2} \sum_{x_2=2}^{5} f(x_1, x_2) = \frac{7}{36}.$$

We have included the point $(2, 2)$ in the summation above, which is an impossible event—we cannot roll a 2 on the red die and a total of 2 on the pair of dice, so that $(2, 2) \notin R(X)$. Nonetheless, the probability assigned to $A$ is correct since $f(2, 2) = 0$, by definition. In general, when defining the probability of an event $A$ for an $n$-dimensional discrete random variable $X$, $f(x_1, \ldots, x_n)$ can be summed over the points identified in the set-defining conditions for $A$ *without* regard for the condition that $x \in R(X)$, since any $x \notin R(X)$ will be such that $f(x) = 0$, and the value of the summation will be left unaltered. This approach can be especially convenient if set $A$ is defined by individual, independent set-defining conditions applied to each $X_i$ in an $n$-dimensional random variable $(X_1, \ldots, X_n)$, as in the preceding example. An analogous argument applies to the continuous case, with integration replacing summation.

We now present an example of the specification of the probability space for a bivariate continuous random variable.

**Example 2.16** Your company manufactures big-screen television sets. The screens are 3 feet high by 4 feet wide rectangles that must be coated with a metallic reflective coating (see Figure 2.8). The machine that is coating the screens begins to randomly produce a coating flaw at a point on the screen surface, where all points on the screen are equally likely to be the point of the flaw. Letting $(0, 0)$ be the center of the screen, we represent the collection of potential flaw points as

$$R(X) = \{(x_1, x_2) : x_1 \in [-2, 2], x_2 \in [-1.5, 1.5]\}.$$

Clearly, the total area of the screen is $3 \cdot 4 = 12$ square feet, and any closed rectangle on the screen having width $W$ and height $H$ contains the proportion $WH/12$ of the total area of the screen. Since all of the points are equally likely, the probability set function defined on the events in $R(X)$ should assign to each closed rectangle of points a probability equal to $WH/12$ where $W$ and $H$ are,

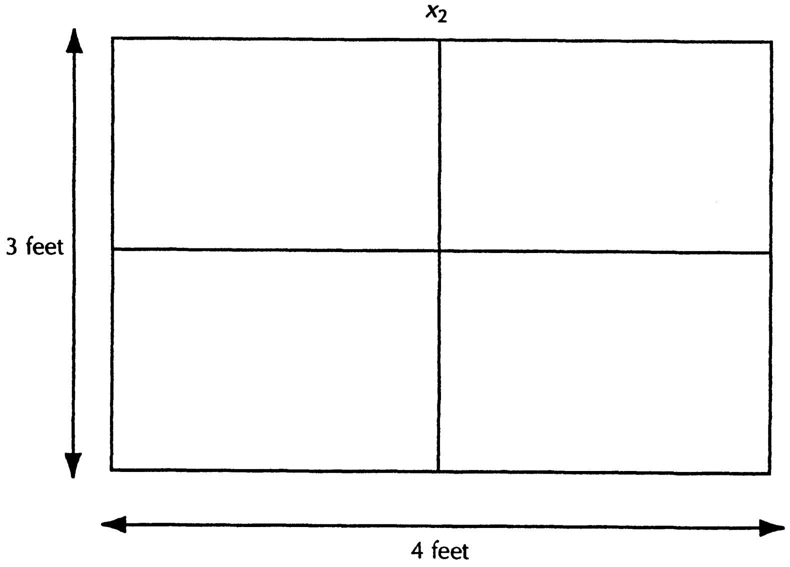


**Figure 2-8**
Television screen.

respectively, the width and height of the rectangular event. We thus seek a function $f(x_1, x_2)$ such that

$$\int_c^d \int_a^b f(x_1, x_2) dx_1\, dx_2 \equiv \frac{(b-a)(d-c)}{12}$$

$\forall\, a, b, c,$ and $d$ such that $-2 \leq a \leq b \leq 2$ and $-1.5 \leq c \leq d \leq 1.5$. Differentiating the iterated integral above, first with respect to $d$ and then with respect to $b$, yields $f(b, d) = 1/12 \,\forall\, b \in [-2, 2]$ and $\forall\, d \in [-1.5, 1.5]$.[13] The form of the continuous joint density function is then given by $f(x_1, x_2) = \frac{1}{12} I_{[-2,2]}(x_1) I_{[-1.5,1.5]}(x_2)$. The probability set function is thus defined as $P(A) = \int \int_{(x_1,x_2)\in A} \frac{1}{12} dx_1\, dx_2$. Then, for example, the probability that the flaw occurs in the upper left quarter of the screen is given by

$$P(-2 \leq x_1 \leq 0, 0 \leq x_2 \leq 1.5) = \int_0^{1.5} \int_{-2}^{0} \frac{1}{12} dx_1\, dx_2 = \int_0^{1.5} \frac{1}{6} dx_2 = .25. \qquad \square$$

### Multivariate CDFs and Duality with PDFs

The CDF concept can be generalized to the multivariate case as follows:

[13] The differentiation is accomplished by applying Lemma 2.1 twice: once to the integral $\int_c^d \left[\int_a^b f(x_1, x_2) dx_1\right] dx_2$, differentiating with respect to $d$ to yield $\int_a^b f(x_1, d) dx_1$, and then differentiating the latter integral with respect to $b$ to obtain $f(b, d)$. In summary, $(\partial^2/\partial b \partial d) \int_c^d \int_a^b f(x_1, x_2) dx_1\, dx_2 = f(b, d)$.

**Definition 2.14**
***Joint cumulative distribution function***

The joint cumulative distribution function of an $n$-dimensional random variable $X$ is defined by $F(b_1, \ldots, b_n) = P(x_i \leq b_i, i = 1, \ldots, n) \; \forall \; (b_1, \ldots, b_n) \in R^n$. The algebraic representation of $F(b_1, \ldots, b_n)$ in the discrete and continuous cases can be given as follows:

**a. Discrete $X$:**

$$F(b_1, \ldots, b_n) = \sum_{x_1 \leq b_1} \cdots \sum_{\substack{x_n \leq b_n \\ f(x_1, \ldots, x_n) > 0}} f(x_1, \ldots, x_n) \quad \text{for } (b_1, \ldots, b_n) \in R^n.$$

**b. Continuous $X$:**

$$F(b_1, \ldots, b_n) = \int_{-\infty}^{b_n} \cdots \int_{-\infty}^{b_1} f(x_1, \ldots, x_n) dx_1 \cdots dx_n \quad \text{for } (b_1, \ldots, b_n) \in R^n.$$

Some general properties of the joint cumulative distribution function include the following three:

1. $\lim_{b_i \to -\infty} F(b_1, \ldots, b_n) = P(\emptyset) = 0, i = 1, \ldots, n;$
2. $\lim_{b_i \to \infty, \forall i} F(b_1, \ldots, b_n) = P(R(X)) = 1;$
3. $F(\mathbf{a}) \leq F(\mathbf{b})$ for $\mathbf{a} < \mathbf{b}$ where

$$\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} < \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = \mathbf{b}$$

and the *vector* inequality is taken in the usual sense to mean $a_i \leq b_i \; \forall \; i$, and $a_i < b_i$ for at least one $i$. The reader should convince herself that these properties follow directly from the definition of the joint cumulative distribution function.

Similar to the univariate case, the joint CDF can be used to derive joint discrete and continuous probability densities. For the discrete case, we state the result for bivariate random variables only. For multivariate random variables of three dimensions or higher, the large number of terms required in the density-defining procedure makes its use somewhat cumbersome.

**Theorem 2.4** (Discrete Bivariate PDFs from Joint CDFs) *Let $(X,Y)$ be a discrete bivariate random variable with joint cumulative distribution function $F(x,y)$, and let $x_1 < x_2 < x_3 < \cdots$ and $y_1 < y_2 < y_3 < \cdots$ represent the possible outcomes of $X$ and $Y$. Then*

$$f(x_1, y_1) = F(x_1, y_1),$$

$$f(x_1, y_j) = F(x_1, y_j) - F(x_1, y_{j-1}), \qquad j \geq 2,$$

$$f(x_i, y_1) = F(x_i, y_1) - F(x_{i-1}, y_1), \qquad i \geq 2,$$

$$f(x_i, y_j) = F(x_i, y_j) - F(x_i, y_{j-1}) - F(x_{i-1}, y_j) + F(x_{i-1}, y_{j-1}), \qquad i \text{ and } j \geq 2.$$

**Proof** The proof is left to the reader. ■

**Theorem 2.5** (Continuous Joint PDFs from Joint CDFs) *Let $F(x_1,\ldots,x_n)$ and $f(x_1,\ldots,x_n)$ represent the joint CDF and PDF for the continuous multivariate random variable $X = (X_1,\ldots,X_n)$. The joint PDF of $X$ can be defined as*

$$f(x_1,\ldots,x_n) = \begin{cases} \dfrac{\partial^n F(x_1,\ldots,x_n)}{\partial x_1 \cdots \partial x_n} & \textit{where } f(\cdot) \textit{ is continuous} \\ 0 \ \textit{(or any nonnegative numbers) elsewhere.} \end{cases}$$

**Proof** The first part of the definition follows directly from an $n$-fold application of Lemma 2.1 for differentiating the iterated integral defining the joint CDF. In particular,

$$\frac{\partial^n F(x_1,\ldots,x_n)}{\partial x_1 \cdots \partial x_n} = \frac{\partial^n \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(t_1,\ldots,t_n)dt_1 \cdots dt_n}{\partial x_1 \cdots \partial x_n} = f(x_1,\ldots,x_n)$$

wherever $f(\cdot)$ is continuous.

Regarding the second part of the definition, as long as the integral exists, arbitrarily changing the values of the nonnegative integrand at the points of discontinuity will not affect the value of

$$F(b_1,\ldots,b_n) = \int_{-\infty}^{b_n} \cdots \int_{-\infty}^{b_1} f(x_1,\ldots,x_n)dx_1 \cdots dx_n$$

(recall Footnote 12). ■

**Example 2.17** Examine the experiment of tossing two fair coins independently and observing whether heads $(H)$ or tails $(T)$ occurs on each toss, so that $S = \{(H,H),(H,T),(T,H),(T,T)\}$ with all elementary events in $S$ being equally likely. Define a bivariate random variable on the elements of $S$ by letting $x$ represent the total number of heads and $y$ represent the total number of tails resulting from the two tosses. The joint density function for the bivariate random variable $(X, Y)$ is then defined by

$$f(x,y) = \tfrac{1}{4} I_{\{(0,2),(2,0)\}}(x,y) + \tfrac{1}{2} I_{\{(1,1)\}}(x,y).$$

It follows from Def. 2.14 that the joint CDF for $(X, Y)$ can be represented as

$$\begin{aligned} F(b_1,b_2) = {} & \tfrac{1}{4} I_{[2,\infty)}(b_1) I_{(-\infty,1)}(b_2) + \tfrac{1}{4} I_{(-\infty,1)}(b_1) I_{[2,\infty)}(b_2) \\ & + \tfrac{1}{2} I_{[1,2)}(b_1) I_{[1,2)}(b_2) + \tfrac{3}{4} I_{[2,\infty)}(b_1) I_{[1,2)}(b_2) \\ & + \tfrac{3}{4} I_{[1,2)}(b_1) I_{[2,\infty)}(b_2) + I_{[2,\infty)}(b_1) I_{[2,\infty)}(b_2). \end{aligned}$$

The CDF no doubt appears to be somewhat "cobbled" together, making the definition of $F$ a rather complicated expression. Unfortunately, such function definitions often arise when specifying joint CDFs in the discrete case, even for seemingly simple experiments such as the one at hand. To understand more clearly the underlying rationale for the preceding definition of $F$, it is useful to partition $R^2$ into subsets that correspond to the events in $S$. In particular, we are interested in defining the collection of elements $w \in S$ for which $X(w) \leq b_1$ and $Y(w) \leq b_2$ is true for the various values of $(b_1, b_2) \in R^2$. Examine the following table:

| $b_1$ | $b_2$ | $A = \{w: X(w) \leq b_1, Y(w) \leq b_2, w \in S\}$ | $P(A)$ |
|---|---|---|---|
| $b_1 < 1$ | $b_2 < 1$ | $\emptyset$ | 0 |
| $1 \leq b_1 < 2$ | $b_2 < 1$ | $\emptyset$ | 0 |
| $b_1 < 1$ | $1 \leq b_2 < 2$ | $\emptyset$ | 0 |
| $b_1 \geq 2$ | $b_2 < 1$ | $\{(H, H)\}$ | 1/4 |
| $b_1 < 1$ | $b_2 \geq 2$ | $\{(T, T)\}$ | 1/4 |
| $1 \leq b_1 < 2$ | $1 \leq b_2 < 2$ | $\{(H, T), (T, H)\}$ | 1/2 |
| $b_1 \geq 2$ | $1 \leq b_2 < 2$ | $\{(H, T), (T, H), (H, H)\}$ | 3/4 |
| $1 \leq b_1 < 2$ | $b_2 \geq 2$ | $\{(H, T), (T, H), (T, T)\}$ | 3/4 |
| $b_1 \geq 2$ | $b_2 \geq 2$ | $S$ | 1 |

The reader should convince herself using a graphical representation of $R^2$ that the conditions defined on $(b_1, b_2)$ can be used to define nine disjoint subsets of $R^2$ that exhaustively partition $R^2$ (i.e., the union of the disjoint sets $= R^2$). The reader will notice that the indicator functions used in the definition of $F$ were based on the latter six sets of conditions on $(b_1, b_2)$ exhibited in the preceding table. If one were interested in the probability $P(x \leq 1, y \leq 1) = F(1, 1)$, for example, the joint CDF indicates that 1/2 is the number we seek. □

**Example 2.18** Reexamine the projection television screen example Ex. 2.16. The joint CDF for the bivariate random variable $(X_1, X_2)$, whose outcome represents the location of the flaw point, is given by

$$\begin{aligned} F(b_1, b_2) &= \int_{-\infty}^{b_2} \int_{-\infty}^{b_1} \frac{1}{12} I_{[-2,2]}(x_1) I_{[-1.5,1.5]}(x_2) dx_1 dx_2 \\ &= \frac{(b_1 + 2)(b_2 + 1.5)}{12} I_{[-2,2]}(b_1) I_{[-1.5,1.5]}(b_2) \\ &+ \frac{4(b_2 + 1.5)}{12} I_{(2,\infty)}(b_1) I_{[-1.5,1.5]}(b_2) \\ &+ \frac{3(b_1 + 2)}{12} I_{[-2,2]}(b_1) I_{(1.5,\infty)}(b_2) \\ &+ I_{(2,\infty)}(b_1) I_{(1.5,\infty)}(b_2). \end{aligned}$$

It is seen that "cobbled-together" definitions of joint CDFs occur in the continuous case as well. To understand the rationale for the piecewise definition, first note that if $b_1 < -2$ and/or $b_2 < -1.5$, then we are integrating over a set of $(x_1, x_2)$ points $\{(x_1, x_2): x_1 < b_1, x_2 < b_2\}$ for which the integrand has a zero value, resulting in a zero value for the definite integral. Thus, $F(b_1, b_2) = 0$ if $b_1 < -2$ and/or $b_2 < -1.5$. If $b_1 \in [-2, 2]$ and $b_2 \in [-1.5, 1.5]$, then taking the effect of the indicator functions into account, the integral defining $F$ can be represented as

$$F(b_1, b_2) = \int_{-1.5}^{b_2} \int_{-2}^{b_1} \frac{1}{12} dx_1 \, dx_2 = \frac{(b_1 + 2)(b_2 + 1.5)}{12},$$

which is represented by the first term in the preceding definition of $F$. If $b_1 > 2$, but $b_2 \in [-1.5, 1.5]$, then since the integrand is zero for all values of $x_1 > 2$, we can represent the integral defining $F$ as

$$F(b_1, b_2) = \int_{-1.5}^{b_2} \int_{-2}^{2} \frac{1}{12} dx_1 \, dx_2 = \frac{4(b_2 + 1.5)}{12},$$

which is represented by the second term in our definition of $F$. If $b_2 > 1.5$ but $b_1 \in [-2, 2]$, then since the integrand is zero for all values of $x_2 > 1.5$, we have

$$F(b_1, b_2) = \int_{-1.5}^{1.5} \int_{-2}^{b_1} \frac{1}{12} dx_1 \, dx_2 = \frac{3(b_1 + 2)}{12},$$

which is represented by the third term in our definition of $F$. Finally, if both $b_1 > 2$ and $b_2 > 1.5$, then since the integrand is zero for all values of $x_1 > 2$ and/or $x_2 > 1.5$, the integral defining $F$ can be written as

$$F(b_1, b_2) = \int_{-1.5}^{1.5} \int_{-2}^{2} \frac{1}{12} dx_1 \, dx_2 = 1,$$

which justifies the final term in our definition of $F$. The reader should convince himself that the preceding conditions on $(b_1, b_2)$ collectively exhaust the possible values of $(b_1, b_2) \in R^2$.

If one were interested in the probability $P(x_1 \leq 1, x_2 \leq 1)$, the "relevant piece" in the definition of $F$ would be the first term, and thus $F(1, 1) = \frac{(3)(2.5)}{12} = .625$. Alternatively, the probability $P(x_1 \leq 1, x_2 \leq 10)$ would be assigned using the third term in the definition of $F$, yielding $F(1, 10) = .75$. □

### Multivariate Mixed Discrete-Continuous and Composite Random Variables

A discussion of multivariate random variables in the mixed discrete-continuous case could be presented here. However, we choose not to do so. In fact, we will not examine the mixed case any further in this text. We are content with having introduced the mixed case in the univariate context. The problem is that in the multivariate case, representations of the relevant probability set functions—especially when dealing with the concepts of marginal and conditional densities, which will be discussed subsequently—become extremely tedious and cumbersome unless one allows a more general notion of integration than that of Riemann, which would then require us to venture beyond the intended scope of this text. We thus leave further study of mixed discrete-continuous random variables to a more advanced course. Note, however, that since elements of both the discrete and continuous random-variable concepts are involved in the mixed case, our continued study of the discrete and continuous cases will provide the necessary foundation on which to base further study of the mixed case.

As a final remark concerning our general discussion of multivariate random variables, note that a function (or vector function) of a multivariate random variable is also a random variable (or multivariate random variable). This follows from the same composition of functions argument that was noted in the univariate case. That is,

$$y = Y(X(w)), \quad \text{or} \quad y = Y(X_1(w), \ldots, X_n(w)), \quad \text{or}$$

$$\underset{m \times 1}{\mathbf{y}} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} Y_1(X_1(w), \ldots, X_n(w)) \\ \vdots \\ Y_m(X_1(w), \ldots, X_n(w)) \end{bmatrix} = \underset{m \times 1}{\mathbf{Y}}(\mathbf{X}(w))$$

are all in the context of "functions of functions," so that ultimately $Y$ is a function of the elements $w \in S$ and is therefore a random variable.[14] One might refer to such as a **composite random variable**.

## 2.5 Marginal Probability Density Functions and CDFs

Suppose that we have knowledge of the probability space corresponding to an experiment involving outcomes of the $n$-dimensional random variable $X_{(n)} = (X_1, \ldots, X_m, X_{m+1}, \ldots, X_n)$ but our real interest lies in assigning probabilities to events involving only the $m$-dimensional random variable $X_{(m)} = (X_1, \ldots, X_m)$, $m < n$. In practical terms, this relates to an experiment in which $n$ different characteristics were recorded for each outcome but we are specifically interested in analyzing only a subset of the characteristics. We will now examine the concept of a **marginal probability density function (MPDF)** for $X_{(m)}$, which will be derived from knowledge of the joint density function for $X_{(n)}$. Once defined, the MPDF can be used to identify the appropriate probability space only for the portion of the experiment characterized by the outcomes of $(X_1, \ldots, X_m)$, and we will be able to use the MPDF in the usual way (summation in the discrete case, integration in the continuous case) to assign probabilities to events concerning $(X_1, \ldots, X_m)$.

The key to understanding the definition of a marginal probability density is to establish the equivalence between events of the form $(x_1, \ldots, x_m) \in B$ in the probability space for $(X_1, \ldots, X_m)$ and events of the form $(x_1, \ldots, x_n) \in A$ in the probability space for $(X_1, \ldots, X_n)$ since it is the latter events to which we can assign probabilities knowing $f(x_1, \ldots, x_n)$.

### Bivariate Case

Let $f(x_1, x_2)$ be the joint density function and $R(X)$ be the range of the bivariate random variable $(X_1, X_2)$. Suppose we want to assign a probability to the event

[14] The reader is reminded that we are suppressing the technical requirement that for every Borel set of $y$ values, the associated collection of $w$ values in $S$ must constitute an *event* in $S$ for the function $Y$ to be called a random variable. As we have remarked previously, this technical difficulty does not cause a problem in applied work.

$x_1 \in B$. Which event for the *bivariate* random variable is equivalent to event $B$ occurring for the *univariate* random variable $X_1$? By definition, this event is given by $A = \{(x_1, x_2): x_1 \in B, (x_1, x_2) \in R(X)\}$, i.e., the event $B$ occurs for $x_1$ iff the outcome of $(X_1, X_2)$ is in $A$ so that $x_1 \in B$. Then since $B$ and $A$ are *equivalent events*, the probability that we will observe $x_1 \in B$ is identically equal to the probability that we will observe $(x_1, x_2) \in A$ (recall the discussion of equivalent events in Section 2.2).

For the discrete case, the foregoing probability correspondence implies that

$$P_{X_1}(B) = P(x_1 \in B) = P(A) = \sum_{(x_1,x_2)\in A}\sum f(x_1, x_2).$$

Our convention of defining $f(x_1, x_2) = 0 \; \forall \; (x_1, x_2) \notin R(X)$ allows the following alternative representation of $P_{X_1}(B)$:

$$P_{X_1}(B) = \sum_{x_1 \in B} \sum_{x_2 \in R(X_2)} f(x_1, x_2).$$

The equivalence of the two representations of $P_{X_1}(B)$ follows from the fact that the set of elementary events being summed over in the latter case, $C = \{(x_1, x_2): x_1 \in B, x_2 \in R(X_2)\}$, is such that $A \subset C$, and $f(x_1, x_2) = 0 \; \forall \; (x_1, x_2) \in C - A$. The latter representation of $P_{X_1}(B)$ leads to the following definition of the **marginal probability density** of $X_1$:

$$f_1(x_1) = \sum_{x_2 \in R(X_2)} f(x_1, x_2).$$

This function, when summed over the points comprising the event $x_1 \in B$, yields the probability that $x_1 \in B$, i.e.,

$$P_{X_1}(B) = \sum_{x_1 \in B} f_1(x_1) = \sum_{x_1 \in B} \sum_{x_2 \in R(X_2)} f(x_1, x_2).$$

Heuristically, one can think of the marginal density of $X_1$ as having been defined by "summing out" the values of $x_2$ in the bivariate PDF for $(X_1, X_2)$. Having defined $f_1(x_1)$, the probability space for the portion of the experiment involving only $X_1$ can then be defined as $\{R(X_1), \Upsilon_{X_1}, P_{X_1}\}$ where $P_{X_1}(B) = \sum_{x_1 \in B} f_1(x_1)$ for $B \in \Upsilon_{X_1}$. Note that the order in which the random variables are originally listed is immaterial to the approach taken above, and the marginal density function and probability space for $X_2$ could be defined in an analogous manner by simply reversing the roles of $X_1$ and $X_2$ in the preceding arguments. The MPDF for $X_2$ would be defined as

$$f_2(x_2) = \sum_{x_1 \in R(X_1)} f(x_1, x_2),$$

with the probability space for $X_2$ defined accordingly.

**Example 2.19** Reexamine Ex. 1.16, in which an individual was to be drawn randomly from the work force of the Excelsior Corporation to receive a monthly "loyalty award."

Define the bivariate random variable $(X_1, X_2)$ as

$$x_1 = \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} \text{ if } \begin{Bmatrix} \text{male} \\ \text{female} \end{Bmatrix} \text{ is drawn,}$$

$$x_2 = \begin{Bmatrix} 0 \\ 1 \\ 2 \end{Bmatrix} \text{ if } \begin{Bmatrix} \text{sales} \\ \text{clerical} \\ \text{production} \end{Bmatrix} \text{ worker is drawn,}$$

so that the bivariate random variable is measuring two characteristics of the outcome of the experiment: gender and type of worker. The joint density of the bivariate random variable is represented in tabular form below, where the nonzero values of $f(x_1, x_2)$ are given in the cells formed by intersecting an $x_1$-row with a $x_2$-column.

| | | $R(X_2)$ | | | |
|---|---|---|---|---|---|
| | | 0 | 1 | 2 | $f_1(x_1)$ |
| $R(X_1)$ | 0 | .165 | .135 | .150 | .450 |
| | 1 | .335 | .165 | .050 | .550 |
| | $f_2(x_2)$ | .500 | .300 | .200 | |

The nonzero values of the marginal density of $X_2$ are given in the bottom *margin* of the table, the definition of the marginal density being

$$f_2(x_2) = \sum_{x_1 \in R(X_1)} f(x_1, x_2) = \sum_{x_1=0}^{1} f(x_1, x_2) = .5I_{\{0\}}(x_2) + .3I_{\{1\}}(x_2) + .2I_{\{2\}}(x_2).$$

The probability space for $X_2$ is thus $\{R(X_2), \Upsilon_{X_2}, P_{X_2}\}$, with $\Upsilon_{X_2} = \{A : A \subset R(X_2)\}$ and $P_{X_2}(A) = \sum_{x_2 \in A} f_2(x_2)$. If one were interested in the probability that the individual chosen was a sales or clerical worker, i.e., the event $A = \{0, 1\}$, then $P_{X_2}(A) = \sum_{x_2=0}^{1} f_{X_2}(x_2) = .5 + .3 = .8$.

The nonzero values of the marginal density for $X_1$ are given in the right-hand *margin* of the table, the definition of the density being

$$f_1(x_1) = \sum_{x_2 \in R(X_2)} f(x_1, x_2) = \sum_{x_2=0}^{2} f(x_1, x_2) = .45I_{\{0\}}(x_1) + .55I_{\{1\}}(x_1).$$

The probability space for $X_1$ is thus $\{R(X_1), \Upsilon_{X_1}, P_{X_1}\}$, with $\Upsilon_{X_1} = \{A : A \subset R(X_1)\}$ and $P_{X_1}(A) = \sum_{x_1 \in A} f_1(x_1)$. If one were interested in the probability that the individual chosen was male, i.e., the event $A = \{0\}$, then $P_{X_1}(A) = \sum_{x_1=0}^{0} f_{X_1}(x_1) =$ .45. □

The preceding example provides a heuristic justification for the term *marginal* in the bivariate case and reflects the historical basis for the name *marginal density function.* In particular, by summing across the rows or columns of a tabular representation of the joint PDF $f(x_1, x_2)$, one can calculate the marginal densities of $X_1$ and $X_2$ in the *margins* of the table.

We now examine the marginal density function concept for continuous random variables. Recall that the probability of event $B$ occurring for the univariate random variable $X_1$ is identically the probability that the event $A = \{(x_1, x_2): x_1 \in B, (x_1, x_2) \in R(X)\}$ occurs for the bivariate random variable $X = (X_1, X_2)$. Then

$$P_{X_1}(B) = P(x_1 \in B) = P(A) = \iint_{(x_1,x_2)\in A} f(x_1, x_2)dx_1\, dx_2.$$

Our convention of defining $f(x_1, x_2) = 0 \;\forall\; (x_1, x_2) \notin R(X)$ allows an alternative representation of $P_{X_1}(B)$ to be given by

$$P_{X_1}(B) = \int_{x_1\in B}\int_{-\infty}^{\infty} f(x_1, x_2)dx_2\, dx_1.$$

The equivalence of the two representations follows from the fact that the set of elementary events being integrated over in the latter case, $C = \{(x_1, x_2): x_1 \in B, x_2 \in (-\infty, \infty)\}$, is such that $A \subset C$, and $f(x_1, x_2) = 0 \;\forall\; (x_1, x_2) \in C - A$. The latter representation of $P_{X_1}(B)$ leads to the definition of the marginal density of $X_1$ as

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2.$$

This function, when integrated over the elementary events comprising the event $x_1 \in B$, yields the probability that $x_1 \in B$, i.e.,

$$P_{X_1}(B) = \int_{x_1\in B} f_1(x_1)dx_1 = \int_{x_1\in B}\int_{-\infty}^{\infty} f(x_1, x_2)dx_2\, dx_1.$$

Heuristically, one might think of the marginal density of $X_1$ as having been defined by "integrating out" the values of $X_2$ in the bivariate density function for $(X_1, X_2)$. Having defined $f_1(x_1)$, the probability space for the portion of the experiment involving only $X_1$ can then be defined as $\{R(X_1), \Upsilon_{X_1}, P_{X_1}\}$, where $P_{X_1}(A) = \int_{x_1\in A} f_1(x_1)dx_1$ for $A \in \Upsilon_{X_1}$. Since the order in which the random variables were originally listed is immaterial, the marginal density function and probability space for $X_2$ can be defined in an analogous manner by simply reversing the roles of $X_1$ and $X_2$ in the preceding arguments. The MPDF for $X_2$ would be defined as

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_1,$$

with the probability space for $X_2$ defined accordingly.

**Example 2.20** The Seafresh Fish Processing Company operates two fish processing plants. The proportion of processing capacity at which each of the plants operates on any given day is the outcome of a bivariate random variable having joint density function $f(x_1, x_2) = (x_1 + x_2)I_{[0,1]}(x_1)I_{[0,1]}(x_2)$. The marginal density function for the proportion of processing capacity at which plant 1 operates can be defined by integrating out $x_2$ from $f(x_1, x_2)$ as

$$\begin{aligned} f_1(x_1) &= \int_{-\infty}^{\infty} f(x_1, x_2)dx_2 = \int_{-\infty}^{\infty} (x_1 + x_2)I_{[0,1]}(x_1)I_{[0,1]}(x_2)dx_2 \\ &= \int_0^1 (x_1 + x_2)I_{[0,1]}(x_1)dx_2 = \left(x_1x_2 + \frac{x_2^2}{2}\right) I_{[0,1]}(x_1) \Big|_0^1 \\ &= (x_1 + \tfrac{1}{2})I_{[0,1]}(x_1). \end{aligned}$$

The probability space for plant 1 outcomes is given by $\{R(X_1), \Upsilon_{X_1}, P_{X_1}\}$, where $R(X_1) = [0, 1]$, $\Upsilon_{X_1} = \{A\colon A \text{ is a Borel set} \subset R(X_1)\}$, and $P_{X_1}(A) = \int_{x_1 \in A} f_1(x_1)dx_1$, $\forall\, A \in \Upsilon_{X_1}$. If one were interested in the probability that plant 1 will operate at less than half of capacity on a given day, i.e., the event $A = [0, .5]$, then

$$P_{X_1}(x_1 \leq .5) = \int_0^{.5} \left(x_1 + \frac{1}{2}\right) I_{[0,1]}(x_1)dx_1 = \frac{x_1^2}{2} + \frac{x_1}{2} \Big|_0^{.5} = .375. \qquad \square$$

Regarding other properties of marginal density functions, note that the significance of the term *marginal* is only to indicate the context in which the density was derived, i.e., the marginal density of $X_1$ is deduced from the joint density for $(X_1, X_2)$. Otherwise, the MPDF has no special properties that differ from the basic properties of any other probability density function.

### *N*-Variate Case

The concept of a discrete MPDF can be straightforwardly generalized to the $n$-variate case, in which case the marginal densities may themselves be joint density functions. For example, if we have the density function $f(x_1, x_2, x_3)$ for the trivariate random variable $(X_1, X_2, X_3)$, then we may conceive of six marginal density functions: $f_1(x_1)$, $f_2(x_2)$, $f_3(x_3)$, $f_{12}(x_1, x_2)$, $f_{13}(x_1, x_3)$, and $f_{23}(x_2, x_3)$. In general, for an $n$-variate random variable, there are $(2^n - 2)$ possible MPDFs that can be defined from knowledge of $f(x_1, \ldots, x_n)$. We present the $n$-variate generalization in the following definition. We use the notation $f_{j_1 \ldots j_m}(x_{j_1}, \ldots, x_{j_m})$ to represent the MPDF of the $m$-variate random variable $(X_{j_1}, \ldots, X_{j_m})$ with the $j_i$'s being the indices that identify the particular random vector of interest. The motivation for the definition is analogous to the argument in the bivariate case upon identifying the equivalent events $(x_{j_1}, \ldots, x_{j_m}) \in B$ and $A = \{x\colon (x_{j_1}, \ldots, x_{j_m}) \in B, x \in R(X)\}$ and is left to the reader as an exercise.

**Definition 2.15** ***Discrete marginal probability density functions***

Let $f(x_1, \ldots, x_n)$ be the joint discrete probability density function for the $n$-dimensional random variable $(X_1, \ldots, X_n)$. Let $J = \{j_1, j_2, \ldots, j_m\}$, $1 \leq m < n$, be a set of indices selected from the index set $I = \{1, 2, \ldots, n\}$. Then the marginal density function for the $m$-dimensional discrete random variable $(X_{j_1}, \ldots, X_{j_m})$ is given by

$$f_{j_1 \ldots j_m}(x_{j_1}, \ldots, x_{j_m}) = \underset{(x_i \in R(X_i), i \in I-J)}{\sum \cdots \sum} f(x_1, \ldots, x_n).$$

In other words, to define a MPDF in the general discrete case, we simply "sum out" the variables that are not of interest in the joint density function. We are left with the marginal density function for the random variable in which we are interested. For example, if $n = 3$, so that $I = \{1, 2, 3\}$, and if $J = \{j_1, j_2\} = \{1, 3\}$ so that $I - J = \{2\}$, then Def. 2.15 indicates that the MPDF of the random variable $(X_1, X_3)$ is given by

$$f_{13}(x_1, x_3) = \sum_{x_2 \in R(X_2)} f(x_1, x_2, x_3).$$

Similarly, the marginal density for $x_1$ would be defined by

$$f_1(x_1) = \sum_{x_2 \in R(X_2)} \sum_{x_3 \in R(X_3)} f(x_1, x_2, x_3).$$

The concept of a continuous MPDF can be generalized to the $n$-variate case as follows:

**Definition 2.16** ***Continuous marginal probability density functions***

Let $f(x_1, \ldots, x_n)$ be the joint continuous probability density function for the $n$-variate random variable $(X_1, \ldots, X_n)$. Let $J = \{j_1, j_2, \ldots, j_m\}$, $1 \leq m < n$, be a set of indices selected from the index set $I = \{1, 2, \ldots, n\}$. Then the marginal density function for the $m$-variate continuous random variable $(X_{j_1}, \ldots, X_{j_m})$ is given by

$$f_{j_1 \ldots j_m}(x_{j_1}, \ldots, x_{j_m}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) \prod_{i \in I-J} dx_i.$$

In other words, to define a MPDF function in the general continuous case, we simply "integrate out" the variables in the joint density function that are not of interest. We are left with the marginal density function for the random variables in which we are interested. An example of marginal densities in the context of a trivariate random variable will be presented in Section 2.8.

### Marginal Cumulative Distribution Functions (MCDFs)

*Marginal* CDFs are simply CDFs that have been derived for a subset of the random variables in $X = (X_1, \ldots, X_n)$ from initial knowledge of the joint PDF or joint CDF of $X$. For example, ordering the elements of a continuous random variable $(X_1, \ldots, X_n)$ so that the first $m < n$ random variables are of primary interest, the MCDF of $(X_1, \ldots, X_m)$ can be defined as

$$\begin{aligned}
F_{1\ldots m}(b_1, \ldots, b_m) &= P_{X_1\ldots X_m}(x_i \leq b_i, i = 1, \ldots, m) \quad \text{(Def. of CDF)} \\
&= P(x_i \leq b_i, i = 1, \ldots, m; x_i < \infty, i = m+1, \ldots, n) \quad \text{(equivalent events)} \\
&= F(b_1, \ldots, b_m, \infty, \ldots, \infty) \quad \text{(Def. in terms of joint CDF)} \\
&= \int_{-\infty}^{b_1} \cdots \int_{-\infty}^{b_m} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) dx_n \cdots dx_1 \quad \text{(Def. in terms of joint PDF)} \\
&= \int_{-\infty}^{b_1} \cdots \int_{-\infty}^{b_m} f_{1\ldots m}(x_1, \ldots, x_m) dx_m \ldots dx_1 \quad \text{(Def. in terms of marginal PDF).}
\end{aligned}$$

In the case of an arbitrary subset $(X_{j_1}, \ldots, X_{j_m})$, $m < n$, of the random variables $(X_1, \ldots, X_n)$, the MCDF in terms of the joint CDF or marginal PDF can be represented as

$$F_{j_1\ldots j_m}(b_{j_1}, \ldots, b_{j_m}) = F(b) = \int_{-\infty}^{b_{j_1}} \cdots \int_{-\infty}^{b_{j_m}} f_{j_1\ldots j_m}(x_{j_1}, \ldots, x_{j_m}) dx_{j_m} \cdots dx_{j_1},$$

where $b_{j_i}$ is the $j_i$th entry in $b$ and $b_i = \infty$ if $i \notin \{j_1, \ldots, j_m\}$.

Examples of marginal CDFs in the trivariate case are presented in Section 2.8. The discrete case is analogous, with summation replacing integration.

## 2.6 Conditional Density Functions

Suppose that we have knowledge of the probability space corresponding to an experiment involving outcomes of the $n$-dimensional random variable $X_{(n)} = (X_1, \ldots, X_m, X_{m+1}, \ldots, X_n)$ and we are interested in assigning probabilities to the event $(x_1, \ldots, x_m) \in C$ *given that* $(x_{m+1}, \ldots, x_n) \in D$. In practical terms, this relates to an experiment in which $n$ different characteristics were recorded for each outcome and we are specifically interested in analyzing a subset of these characteristics *given that* a fixed set of possibilities will occur with certainty for the remaining characteristics. Note that this is different from asking for the *probability* of observing the event $(x_1, \ldots, x_m) \in C$ *and* $(x_{m+1}, \ldots, x_n) \in D$, for we are saying that $(x_{m+1}, \ldots, x_n) \in D$ *will happen with certainty*. In other words, we are asking for the probability that $(x_1, \ldots, x_m) \in C$, *conditional* on the event that $(x_{m+1}, \ldots, x_n) \in D$. How do we assign the appropriate probability in this case? Questions of this type can be addressed through the use of **conditional probability density functions**, which can be derived from knowledge of the joint density function $f(x_1, \ldots, x_n)$.

The key to the definition of a conditional probability density function is to establish the equivalence between events for the $m$-dimensional random variable $(X_1, \ldots, X_m)$ and $(n-m)$-dimensional random variable $(X_{m+1}, \ldots, X_n)$ with events for the $n$-dimensional random variable $(X_1, \ldots, X_n)$. Then conditional probabilities in the probability space for $(X_1, \ldots, X_n)$ can be used to define a conditional probability density function.

### Bivariate Case

Let $f(x_1, x_2)$ be the joint density function and $R(X)$ be the range of the bivariate random variable $(X_1, X_2)$. The event for the bivariate random variable that is equivalent to the event $C$ occurring for the random variable $X_1$ is given by

$$A = \{(x_1, x_2) : x_1 \in C, (x_1, x_2) \in R(X)\}.$$

Similarly, the event for the bivariate random variable that is equivalent to the event $D$ occurring for the random variable $X_2$ is given by

$$B = \{(x_1, x_2) : x_2 \in D, (x_1, x_2) \in R(X)\}.$$

Then the probability that $x_1 \in C$ *given that* $x_2 \in D$ can be defined by the conditional probability

$$P_{X_1|X_2}(C \mid D) = P(x_1 \in C \mid x_2 \in D) = P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \text{for } P(B) \neq 0,$$

where

$$A \cap B = \{(x_1, x_2) : x_1 \in C, x_2 \in D, (x_1, x_2) \in R(X)\}.$$

In the case of a discrete random variable, the foregoing conditional probability is represented by

$$P_{X_1|X_2}(C \mid D) = P(A \mid B) = \frac{\sum \sum_{(x_1,x_2) \in A \cap B} f(x_1, x_2)}{\sum \sum_{(x_1,x_2) \in B} f(x_1, x_2)}.$$

Given our convention that $f(x_1, x_2) = 0$ whenever $(x_1, x_2) \notin R(X)$, we can ignore the set-defining condition $(x_1, x_2) \in R(X)$ in both the sets $A \cap B$ and $B$ and represent the conditional probability as

$$P_{X_1|X_2}(C \mid D) = \frac{\sum_{x_1 \in C} \sum_{x_2 \in D} f(x_1, x_2)}{\sum_{x_1 \in R(X_1)} \sum_{x_2 \in D} f(x_1, x_2)} = \sum_{x_1 \in C} \left[ \frac{\sum_{x_2 \in D} f(x_1, x_2)}{\sum_{x_2 \in D} f_2(x_2)} \right],$$

where we have used the fact that $f_2(x_2) = \sum_{x_1 \in R(X_1)} f(x_1, x_2)$. The expression in brackets is the conditional density function we seek, since it is the function that would be summed over the elements in $C$ to assign probability to the event $x_1 \in C$, given $x_2 \in D$, for any event $C$. We will denote the conditional density of $X_1$, given $x_2 \in D$, by the notation $f(x_1 \mid x_2 \in D)$. If $D$ is a singleton set $\{d\}$, we will also represent the conditional density function as $f(x_1 \mid x_2 = d)$.

In the case of a continuous bivariate random variable, the probability that $x_1 \in C$ given that $x_2 \in D$ would be given by (assuming $P_{X_2}(D) = P(B) \neq 0$)

$$P_{X_1|X_2}(C \mid D) = P(x \in C \mid x_2 \in D) = P(A \mid B) = \frac{\iint_{(x_1,x_2) \in A \cap B} f(x_1, x_2) dx_1\, dx_2}{\iint_{(x_1,x_2) \in B} f(x_1, x_2) dx_1\, dx_2}$$

Using our convention that $f(x_1, x_2) = 0 \;\forall\; (x_1, x_2) \notin R(X)$, we can also represent the conditional probability as

$$P_{X_1|X_2}(C \mid D) = \frac{\int_{x_1 \in C} \int_{x_2 \in D} f(x_1, x_2) dx_2\, dx_1}{\int_{-\infty}^{\infty} \int_{x_2 \in D} f(x_1, x_2) dx_2\, dx_1} = \int_{x_1 \in C} \left[ \frac{\int_{x_2 \in D} f(x, x_2) dx_2}{\int_{x_2 \in D} f_2(x_2) dx_2} \right] dx_1,$$

where we have used the fact that $f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1$. The expression in brackets is the conditional density function we seek, since it is the function that would be integrated over the elements in $C$ to assign probability to the event $x_1 \in C$, given $x_2 \in D$, for any event $C$. As in the discrete case, we will use the notation $f(x_1 \mid x_2 \in D)$ or $f(x_1 \mid x_2 = d)$ to represent the conditional density function. In both the discrete and continuous cases, we will eliminate the random variable subscripts on $P_{X_1|X_2}(\cdot)$ when the random variable context of the probability set function is clear.

Once derived, a conditional probability density function exhibits all of the standard properties of a PDF. The significance of the term *conditional* PDF is to indicate that the density of $X_1$ was derived from the joint density for $(X_1, X_2)$ conditional on a specific event for $X_2$. Otherwise, there are no special general properties of a conditional PDF that distinguishes it from any other PDF.

We provide examples of the derivation and use of discrete and continuous conditional PDFs in the following examples.

**Example 2.21** Recall the dice example, Ex. 2.15, where $f(x_1, x_2) = (1/36) I_{\{1,\ldots,6\}}(x_1) I_{\{1,\ldots,6\}}(x_2 - x_1)$. The conditional density function for $X_1$, given that $x_2 = 5$, is given by

$$f(x_1 \mid x_2 = 5) = \frac{f(x_1, 5)}{f_2(5)} = \frac{\frac{1}{36} I_{\{1,\ldots,6\}}(x_1) I_{\{1,\ldots,6\}}(5 - x_1)}{\frac{6-|5-7|}{36} I_{\{2,\ldots,12\}}(5)} = \frac{1}{4} I_{\{1,\ldots,4\}}(x_1).$$

The probability of rolling a 3 or less on the red die, given that the total of the two dice will be 5, is then

$$P(x_1 \leq 3 \mid x_2 = 5) = \sum_{x_1=1}^{3} f(x_1 \mid x_2 = 5) = \frac{3}{4}.$$

Note that the *unconditional* probability that $x_1 \leq 3$ is equal to 1/2.

The conditional density function for $X_1$, given that $x_2 \in D = \{7, 11\}$, is given by

$$f(x_1 \mid x_2 \in D) = \frac{\sum_{x_2 \in D} f(x_1, x_2)}{\sum_{x_2 \in D} f_2(x_2)} = \frac{\frac{1}{36} I_{\{1,\ldots,6\}}(x_1) \left[ I_{\{1,\ldots,6\}}(7 - x_1) + I_{\{1,\ldots,6\}}(11 - x_1) \right]}{\frac{8}{36}}$$

$$= \tfrac{1}{8} I_{\{1,\ldots,4\}}(x_1) + \tfrac{1}{4} I_{\{5,6\}}(x_1).$$

The probability of rolling a 3 or less on the red die, given that the total of the two dice will be either a 7 or 11, is then

$$P(x_1 \leq 3 \mid x_2 \in D) = \sum_{x_1=1}^{3} f(x_1 \mid x_2 \in D) = \frac{3}{8}. \qquad \square$$

**Example 2.22** Recall Ex. 2.20 regarding the proportion of daily capacity at which two fish processing plants operate. The conditional density function of plant 1's capacity, given that plant 2 operates at less than half of capacity, is given by

$$\begin{aligned} f(x_1 \mid x_2 \leq .5) &= \frac{\int_{-\infty}^{.5} f(x_1, x_2) dx_2}{\int_{-\infty}^{.5} f_2(x_2) dx_2} = \frac{\int_0^{.5} (x_1 + x_2) I_{[0,1]}(x_1) dx_2}{\int_0^{.5} (x_2 + \frac{1}{2}) dx_2} \\ &= \frac{.5x_1 + .125}{.375} I_{[0,1]}(x_1) = \left(\frac{4}{3}x_1 + \frac{1}{3}\right) I_{[0,1]}(x_1). \end{aligned}$$

The probability that $x_1 \leq .5$, given that $x_2 \leq .5$, is given by

$$P(x_1 \leq .5 \mid x_2 \leq .5) = \int_0^{.5} \left(\frac{4}{3}x_1 + \frac{1}{3}\right) dx_1 = \frac{1}{3}.$$

Recall that the *unconditional* probability that $x_1 \leq .5$ was .375. □

*Conditioning on Elementary Events in Continuous Cases* A problem arises in the continuous case when defining a conditional PDF for $X_1$, conditional on an *elementary* event occurring for $X_2$. Namely, since all elementary events are assigned probability zero in the continuous case, and since, more generally, the integral over a singleton set is zero, our definition of the conditional density, as presented earlier, yields

$$f(x_1 \mid x_2 = b) = \frac{\int_b^b f(x_1, x_2) dx_2}{\int_b^b f_2(x_2) dx_2} = \frac{0}{0},$$

which is an indeterminate form. Thus $f(x_1 \mid x_2 = b)$ is undefined, so that $P(x_1 \in A \mid x_2 = b)$ is undefined as well. This is different than the discrete case, where

$$f(x_1 \mid x_2 = b) = \frac{f(x_1, b)}{f_2(b)}$$

is well-defined, provided $f_2(b) \neq 0$.

The problem is circumvented by redefining the conditional probability, $P(x_1 \in A \mid x_2 = b)$, in the continuous case in terms of a limit, as

$$\begin{aligned} P(x_1 \in A \mid x_2 = b) &= \lim_{\epsilon \to 0^+} P(x_1 \in A \mid x_2 \in [b - \epsilon, b + \epsilon]) \\ &= \lim_{\epsilon \to 0^+} \left[ \frac{\int_{x_1 \in A} \int_{b-\epsilon}^{b+\epsilon} f(x_1, x_2) dx_2 \, dx_1}{\int_{b-\epsilon}^{b+\epsilon} f_2(x_2) dx_2} \right], \end{aligned}$$

where $\lim_{\epsilon \to 0^+}$ means we are examining a limit for a sequence of $\epsilon$ values that approach zero from positive values (i.e., $\epsilon > 0$). The idea is to examine the limiting value of a sequence of probabilities that are conditioned on a corresponding sequence of events, $[b - \epsilon, b + \epsilon]$ for $\epsilon \to 0^+$, that converge to the elementary event $\{b\}$. The following lemma will facilitate the identification of the limit.

**Lemma 2.2** ***Mean Value Theorem for Integrals***

If $g(x)$ is continuous $\forall\, x \in [c_1, c_2]$, then $\exists\, x_0 \in [c_1, c_2]$ such that $\int_{c_1}^{c_2} g(x)dx = g(x_0)(c_2 - c_1)$.[15]

To use the mean value theorem, and to ensure that the limit of the conditional probabilities exists, we assume that there exists a choice of $\epsilon > 0$ such that $f_2(x_2)$ and $f(x_1, x_2)$ are continuous in $x_2$, $\forall\, x_2 \in [b-\epsilon, b+\epsilon]$, and that $f_2(b) > 0$. Then, by the mean value theorem,

$$P(x_1 \in A \mid x_2 = b) = \lim_{\epsilon \to 0^+} \left[ \frac{2\epsilon \int_{x_1 \in A} f(x_1, x_2^0) dx_1}{2\epsilon f_2(x_2^*)} \right],$$

where both $x_2^0$ and $x_2^* \in [b - \epsilon, b + \epsilon]$, and $x_2^0$ will generally depend on the value of $x_1$.[16] The $2\epsilon$'s in the numerator and denominator cancel each other, and as $\epsilon \to 0^+$, the interval $[b - \epsilon, b + \epsilon]$ reduces to $[b, b] = b$, so that in the limit, both $x_2^0$ and $x_2^* = b$. The limiting value of the conditional probability is then

$$P(x_1 \in A \mid x_2 = b) = \int_{x_1 \in A} \frac{f(x_1, b)}{f_2(b)} dx_1.$$

Since the choice of event $A$ is arbitrary, it follows that the appropriate conditional probability density in this case is

$$f(x_1 \mid x_2 = b) = \frac{f(x_1, b)}{f_2(b)},$$

which is precisely of the same form as the discrete case. Thus, the definition of conditional density functions, when conditioning on elementary events, will be identical for continuous and discrete random variables.

**Example 2.23** Recall Ex. 2.22. The conditional PDF for plant 1's proportion of capacity $X_1$, given that plant 2's capacity proportion is $x_2 = .75$, can be defined as

$$f(x_1 \mid x_2 = .75) = \frac{f(x_1, .75)}{f_2(.75)} = \frac{(x_1 + .75) I_{[0,1]}(x_1)}{1.25} = \left(\frac{4}{5} x_1 + \frac{3}{5}\right) I_{[0,1]}(x_1).$$

[15] R. Courant and F. John (1965), *Introduction to Calculus and Analysis*, New York: John Wiley-Interscience, p. 143.

[16] In applying the mean value theorem to the numerator, we treat $f(x_1, x_2)$ as a function of the single variable $x_2$, fixing the value of $x_1$ for each application.

The probability that $x_1 \leq .5$, given that $x_2 = .75$, is then given by

$$P(x_1 \leq .5 \mid x_2 = .75) = \int_0^{.5} \left(\frac{4}{5}x_1 + \frac{3}{5}\right) dx_1 = .4. \qquad \square$$

### N-Variate Case

The preceding concepts of discrete and continuous conditional probability density functions in the bivariate case can be generalized to the $n$-variate case, as indicated in the following definition.

**Definition 2.17**
***Conditional probability density functions***

Let $f(x_1, \ldots, x_n)$ be the joint density function for the $n$-dimensional random variable $(X_1, \ldots, X_n)$. Let $J_1 = \{j_1, \ldots, j_m\}$ and $J_2 = \{j_{m+1}, \ldots, j_n\}$ be two mutually exclusive index sets whose union is equal to the index set $\{1, 2, \ldots, n\}$. Then the conditional density function for the $m$-dimensional random variable $(X_{j_1}, \ldots, X_{j_m})$, given that $(X_{j_{m+1}}, \ldots, X_{j_n}) \in D$ and $P_{X_{j_{m+1}} \cdots X_{j_n}}(D) > 0$, is as follows:

*Discrete Case*:

$$f(x_{j_1}, \ldots, x_{j_m} \mid x_{j_{m+1}}, \ldots, x_{j_n}) \in D) = \frac{\sum \cdots \sum_{(x_{j_{m+1}}, \ldots, x_{j_n}) \in D} f(x_1, \ldots, x_n)}{\sum \cdots \sum_{(x_{j_{m+1}}, \ldots, x_{j_n}) \in D} f_{j_{m+1} \cdots j_n}(x_{j_{m+1}}, \ldots, x_{j_n})}$$

*Continuous Case*:

$$f(x_{j_1}, \ldots, x_{j_m} \mid (x_{j_{m+1}}, \ldots, x_{j_n}) \in D)$$
$$= \frac{\int \cdots \int_{(x_{j_{m+1}}, \ldots, x_{j_n}) \in D} f(x_1, \ldots, x_n) dx_{j_{m+1}} \ldots dx_{j_n}}{\int \cdots \int_{(x_{j_{m+1}}, \ldots, x_{j_n}) \in D} f_{j_{m+1} \cdots j_n}(x_{j_{m+1}}, \ldots, x_{j_n}) dx_{j_{m+1}} \ldots dx_{j_n}}$$

If $D$ is equal to the elementary event $(d_{m+1}, \ldots, d_n)$, then the definition of the conditional density in both the discrete and continuous cases can be represented as

$$f(x_{j_1}, \ldots, x_{j_m} \mid x_{j_i} = d_i, i = m+1, \ldots, n) = \frac{f(x_1, \ldots, x_n)}{f_{j_{m+1} \cdots j_n}(d_{m+1}, \ldots, d_n)}$$

where $x_{j_i} = d_i$ if $j_i \in J_2$, when the marginal density in the denominator is positive valued.[17]

[17] In the continuous case, it is also presumed that $f$ and $f_{j_{m+1} \cdots j_n}$ are continuous in $(x_{j_{m+1}}, \ldots, x_{j_n})$ within some neighborhood of points around the point where the conditional density is evaluated in order to justify the conditional density definition via a limiting argument analogous to the bivariate case. Motivation for the conditional density expression when conditioning on an elementary event in the continuous case can then be provided by extending the mean value theorem argument used in the bivariate case. See R. G. Bartle, *Real Analysis*, p. 429, for a statement of the general mean value theorem for integrals.

For example, if $n = 3$, $J_1 = \{1, 3\}$, and $J_2 = \{2\}$, then the conditional density function of $(X_1, X_3)$, given that $x_2 \in D$, would be defined as

$$f(x_1, x_3 \mid x_2 \in D) = \frac{\sum_{x_2 \in D} f(x_1, x_2, x_3)}{\sum_{x_2 \in D} f_2(x_2)}$$

in the discrete case, with integration replacing summation in the continuous case. If $D = d_2$, then for both the discrete and continuous cases,

$$f(x_1, x_3 \mid x_2 = d_2) = \frac{f(x_1, d_2, x_3)}{f_2(d_2)}.$$

Similarly, if $J_1 = \{1\}$ and $J_2 = \{2, 3\}$, then the conditional density for $X_1$, given $(x_2, x_3) \in D$, would be defined as

$$f(x_1 \mid (x_2, x_3) \in D) = \frac{\sum \sum_{(x_2, x_3) \in D} f(x_1, x_2, x_3)}{\sum \sum_{(x_2, x_3) \in D} f_{23}(x_2, x_3)}$$

in the discrete case, with integration replacing summation in the continuous case. If $D = \{(d_2, d_3)\}$, then for both the discrete and continuous cases,

$$f(x_1 \mid x_2 = d_2, x_3 = d_3) = \frac{f(x_1, d_2, d_3)}{f_{23}(d_2, d_3)}.$$

An example of conditional PDFs in the trivariate case will be presented in Section 2.8.

In summary, if we begin with the joint density function appropriate for assigning probabilities to events involving the $n$-dimensional random variable $(X_1, \ldots, X_n)$, we can derive a conditional probability density function that is the PDF appropriate for assigning probabilities to events for an $m$-dimensional subset of the random variables in $(X_1, \ldots, X_n)$, *given* (or *conditional*) on an event for the remaining $n - m$ random variables. The construction of the conditional density involves both the joint density of $(X_1, \ldots, X_n)$ and the marginal density of the $(n-m)$-dimensional random variable on which we are conditioning. In the special case where we are conditioning on an elementary event, the conditional density function simply becomes the ratio of the joint density function to the marginal density function, replacing the arguments of these functions with their conditioned values for those arguments corresponding to random variables on which we are conditioning (which represents *all* of the arguments of the marginal density, and a subset of the arguments of the joint density).

### Conditional CDFs

We can define the concept of a **conditional CDF** by simply using a conditional density function in the definition of the CDF. For example, for the bivariate random variable $(X_1, X_2)$, we can define

$$F(b_1 \mid x_2 \in D) = P(x_1 \leq b_1 \mid x_2 \in D) = \int_{-\infty}^{b_1} f(x_1 \mid x_2 \in D) dx_1$$

as one such conditional CDF, representing the CDF of $X_1$, conditional on $x_2 \in D$. Once defined, the conditional CDF possesses no special properties that distinguish it in concept from any other CDF. The reader is asked to contemplate the various conditional CDFs that can be defined for the $n$-dimensional random variable $(X_1, \ldots, X_n)$.

## 2.7 Independence of Random Variables

From our previous discussion of independence of events, we know that $A$ and $B$ are independent iff $P(A \cap B) = P(A)P(B)$. This concept can be applied directly to determine whether two events for the $n$-dimensional random variable $(X_1, \ldots, X_n)$ are independent. The general definition of independence of events (Def. 1.13) can be used directly to examine the independence of $k$ events for the random variable $(X_1, \ldots, X_n)$. The concept of independence of events will now be extended further to the idea of **independence of random variables**, which is related to the question of whether the $n$ events (recall the abbreviated set definition notation of Def. 2.7) $\{x_i \in A_i\} = \{(x_1, \ldots, x_n): x_i \in A_i, (x_1, \ldots, x_n) \in R(X)\}$, $i = 1, \ldots, n$, are independent for *all* possible choices of the events $A_1, \ldots, A_n$. If so, the *n random variables* are said to be independent. In effect, the concept is one of **global independence of events** for random variables—we define an event $A_i$ for each of the $n$ random variables in $(X_1, \ldots, X_n)$ and, *no matter how we define the events* (which is the meaning of the term "global" here), the events $\{x_i \in A_i\}$, $i = 1, \ldots, n$, are independent. Among other things, this implies that the probability assigned to *any* event $A_i$ for *any* random variable $X_i$ in $(X_1, \ldots, X_n)$ would be unaffected by conditioning on *any* event $B$ for the remaining random variables (assuming $P(B) > 0$ for the existence of the conditional probability).

### Bivariate Case

We seek to establish a condition that will ensure that the events $\{x_1 \in A_1\}$ and $\{x_2 \in A_2\}$ are independent for *all* possible choices of the events $A_1$ and $A_2$. This can be accomplished by applying independence conditions to events in the probability space, $\{R(X), \Upsilon, P\}$ for the bivariate random variable $X = (X_1, X_2)$. The events $x_1 \in A_1$ and $x_2 \in A_2$ are equivalent, respectively, to the following events for the bivariate random variable:

$$B_1 = \{(x_1, x_2): x_1 \in A_1, (x_1, x_2) \in R(X)\},$$

$$B_2 = \{(x_1, x_2): x_2 \in A_2, (x_1, x_2) \in R(X)\}.$$

The two events $B_1$ and $B_2$ are independent iff $P(B_1 \cap B_2) = P(B_1)P(B_2)$, which can also be represented as $P(x_1 \in A_1, x_2 \in A_2) = P(x_1 \in A_1)P(x_2 \in A_2)$. Requiring the independence condition to hold for *all* choices of the events $A_1$ and $A_2$ leads to the definition of the independence condition for random variables.

**Definition 2.18** ***Independence of random variables***

> The random variables $X_1$ and $X_2$ are said to be independent iff $P(x_1 \in A_1, x_2 \in A_2) = P(x_1 \in A_1)P(x_2 \in A_2)$ for all events $A_1, A_2$.

There is an equivalent characterization of independence of random variables in terms of PDFs that can be useful in practice and that also further facilitates the investigation of the implications of random-variable independence.

**Theorem 2.6** (Joint Density Factorization for Independence of Random Variables) *The random variables $X_1$ and $X_2$ with joint probability density function $f(x_1,x_2)$ and marginal probability density functions $f_i(x_i)$, $i = 1,2$, are independent iff the joint density factors into the product of the marginal densities as $f(x_1,x_2) = f_1(x_1)f_2(x_2)$ $\forall$ $(x_1,x_2)$ except, possibly, at points of discontinuity for the joint density function of a continuous random variable.*

**Proof** (Discrete case) Let $A_1$ and $A_2$ be any two events for $X_1$ and $X_2$, respectively. Then if the joint density function $f(x_1, x_2)$ factors,

$$P(x_1 \in A_1, x_2 \in A_2) = \sum_{x_1 \in A_1} \sum_{x_2 \in A_2} f(x_1, x_2) = \sum_{x_1 \in A_1} f_1(x_1) \sum_{x_2 \in A_2} f_2(x_2) = P(x_1 \in A_1)P(x_2 \in A_2),$$

and so $X_1$ and $X_2$ are independent. Thus, factorization is *sufficient* for independence. Now assume $(X_1, X_2)$ *are* independent random variables. Let $A_1 = \{a_1\}$ and $A_2 = \{a_2\}$ for any choice of elementary events, $a_i \in R(X_i)$, corresponding to the random variable $X_i$, $i = 1, 2$, respectively. Then, by independence,

$$P(x_1 = a_1, x_2 = a_2) = f(a_1, a_2) = P(x_1 = a_1)P(x_2 = a_2) = f_1(a_1)f_2(a_2).$$

If $a_i \notin R(X_i)$, then $f_i(a_i) = 0$ and $f(a_1, a_2) = 0$ for $i = 1, 2$, and thus factorization will automatically hold. Thus, factorization is *necessary* for independence.

(Continuous case) Let $A_1$ and $A_2$ be any two events for $X_1$ and $X_2$, respectively. Then if the joint density function $f(x_1, x_2)$ factors,[18]

$$\begin{aligned} P(x_1 \in A_1, x_2 \in A_2) &= \int_{x_2 \in A_2} \int_{x_1 \in A_1} f(x_1, x_2) dx_1\, dx_2 \\ &= \int_{x_1 \in A_1} f_1(x_1) dx_1 \int_{x_2 \in A_2} f_2(x_2) dx_2 \\ &= P(x_1 \in A_1)P(x_2 \in A_2), \end{aligned}$$

so that $X_1$ and $X_2$ are independent. Thus, factorization is *sufficient* for independence. Now assume $(X_1, X_2)$ *are* independent random variables. Let $A_i =$

[18] Any points of discontinuity can be ignored in the definitions of the probability integrals without affecting the probability assignments. Recall Footnote 12.

$\{x_i : x_i \leq a_i\}$ for arbitrary choice of $a_i$, $i = 1, 2$. Then by independence,

$$P(x_1 \leq a_1, x_2 \leq a_2) = \int_{-\infty}^{a_2}\int_{-\infty}^{a_1} f(x_1, x_2)dx_1\, dx_2$$
$$= P(x_1 \leq a_1)P(x_2 \leq a_2) = \int_{-\infty}^{a_1} f_1(x_1)dx_1 \int_{-\infty}^{a_2} f_2(x_2)dx_2.$$

Differentiating the integrals with respect to $a_1$ and $a_2$ yields $f(a_1, a_2) = f_1(a_1)f_2(a_2)$ wherever the joint density function is continuous. Thus, the factorization condition stated in the theorem is *necessary* for independence. ■

In other words, two random variables are independent iff their joint PDF can be expressed equivalently as the product of their respective marginal PDFs (the condition not being required to hold at points of discontinuity in the continuous case). An important implication of the pairwise independence of $X_1$ and $X_2$ is that the conditional and marginal PDFs of the respective random variables are identical.[19] For example, assuming independence,

$$f(x_1 \mid x_2 = b) = \frac{f(x_1, b)}{f_2(b)} = \frac{f_1(x_1)f_2(b)}{f_2(b)} = f_1(x_1)$$

and

$$f(x_1 \mid x_2 \in B) = \frac{\int_{x_2 \in B} f(x_1, x_2)dx_2}{\int_{x_2 \in B} f_2(x_2)dx_2} = \frac{f_1(x_1)\int_{x_2 \in B} f_2(x_2)dx_2}{\int_{x_2 \in B} f_2(x_2)dx_2} = f_1(x_1)$$

(in the discrete case, replace integration by summation). The fact that conditional and marginal PDFs are identical implies that the probability of $x_1 \in A$, for any event $A$, is unaffected by the occurrence or nonoccurrence of event $B$ (or $b$) for $X_2$, e.g., in the continuous case,

$$P(x_1 \in A \mid x_2 \in B) = \int_{x_1 \in A} f(x_1 \mid x_2 \in B)dx_1 = \int_{x_1 \in A} f_1(x_1)dx_1 = P(x_1 \in A)$$

(replace integration by summation in the discrete case). The result holds for events involving $X_2$ for which the conditional density function is defined. The roles of $X_1$ and $X_2$ can be reversed in the preceding discussion.

**Example 2.24** Recall Ex. 2.16 concerning coating flaws in the manufacture of television screens. The horizontal and vertical coordinates of the coating flaw was the outcome of a bivariate random variable with joint density function

$$f(x_1, x_2) = \tfrac{1}{12} I_{[-2,2]}(x_1) I_{[-1.5,1.5]}(x_2).$$

Are the random variables independent?

[19] We will henceforth suppress constant reference to the fact that factorization might not hold for some points of discontinuity in the continuous case—it will be tacitly understood that results we derive based on the factorization of $f(x_1, x_2)$ may bè violated at some isolated points. For example, for the case at hand, marginal and conditional densities may not be equal at some isolated points. Assignments of probability will be unaffected by this technical anomaly.

**Answer**: The marginal densities of $X_1$ and $X_2$ are given by

$$f_1(x_1) = I_{[-2,2]}(x_1)\int_{-1.5}^{1.5}\frac{1}{12}dx_2 = .25I_{[-2,2]}(x_1),$$

$$f_2(x_2) = I_{[-1.5,1.5]}(x_2)\int_{-2}^{2}\frac{1}{12}dx_1 = \frac{1}{3}I_{[-1.5,1.5]}(x_2).$$

It follows that $f(x_1,x_2) = f_1(x_1)f_2(x_2)\ \forall\ (x_1,x_2)$, and the random variables are independent. Therefore, knowledge that an event for $X_2$ has occurred has no effect on the probability assigned to events for $X_1$, and vice versa. □

**Example 2.25** Recall the dice example, Ex. 2.15. Are $X_1$ and $X_2$ independent random variables?
**Answer**: Examine the validity of the independence condition:

$$f(x_1,x_2) \overset{?}{=} f_1(x_1)f_2(x_2) \quad \forall\ (x_1,x_2),$$

or, specifically,

$$\frac{1}{36}I_{\{1,2,\ldots,6\}}(x_1)I_{\{1,2,\ldots,6\}}(x_2-x_1) \overset{?}{=} \frac{1}{6}I_{\{1,2,\ldots,6\}}(x_1)\left(\frac{6-|x_2-7|}{36}\right)I_{\{2,\ldots,12\}}(x_2) \quad \forall\ (x_1,x_2).$$

The random variables $X_1$ and $X_2$ are *not* independent, since, for example, letting $x_1 = 2$ and $x_2 = 4$ results in $1/36 \neq 1/72$. Therefore, knowledge that an event for $X_2$ has occurred *can* affect the probability assigned to events for $X_1$, and vice versa. □

### *N*-Variate

The independence concept can be extended beyond the bivariate case to the case of **independence of random variables** $X_1,\ldots,X_n$. The formal definition of independence in the $n$-variate case is as follows:

**Definition 2.19**
***Independence of random variables (n-variate)***

The random variables $X_1, X_2, \ldots, X_n$ are said to be independent iff $P(x_i \in A_i, i = 1,\ldots,n) = \prod_{i=1}^{n} P(x_i \in A_i)$ for all choices of the events $A_1, \ldots, A_n$.

The motivation for the definition is similar to the argument used in the bivariate case. For $B_i = \{(x_1,\ldots,x_n)\colon x_i \in A_i, (x_1,\ldots,x_n) \in R(X)\}, i = 1,\ldots,n$, to be independent events, we require (recall Def. 1.13)

$$P\left(\bigcap_{j\in J} B_j\right) = \prod_{j\in J} P(B_j)\ \forall\ J \quad \subset \{1,2,\ldots,n\}, \quad \text{with } N(J) \geq 2.$$

If we further require this condition to hold for *all* possible choices of the events $(B_1,\ldots,B_n)$, then the totality of the conditions can be represented as

$$P(x_i \in A_i, i = 1,\ldots,n) = P\left(\bigcap_{i=1}^{n} B_i\right) = \prod_{i=1}^{n} P(B_i) = \prod_{i=1}^{n} P(x_i \in A_i)$$

for all choices of the events $A_1, \ldots, A_n$ (or, equivalently, for corresponding choices of $B_1, \ldots, B_n$). Any of the other conditions required for independence of events, i.e.,

$$P\left(\bigcap_{j\in J} B_j\right) = \prod_{j\in J} P(B_j) \quad \text{with } J \subset \{1,2,\ldots,n\} \text{ and } N(J) < n,$$

are implied by the preceding condition upon letting $A_j = R(X_j)$ (or, equivalently, $B_j = R(X))$ for $j \in \bar{J}$.

The generalization of the joint density factorization theorem is given as Theorem 2.7. The proof is a direct extension of the arguments used in proving Theorem 2.6 and is left to the reader.

**Theorem 2.7** (Joint Density Factorization for Independence of Random Variables ($n$-variate case)) *The random variables $X_1, X_2, \ldots, X_n$ with joint probability density function $f(x_1,\ldots,x_n)$ and marginal probability density functions $f_i(x_i)$, $i = 1, \ldots, n$, are independent iff the joint density can be factored into the product of the marginal densities as*

$$f(x_1,\ldots,x_n) = \prod_{i=1}^{n} f_i(x_i) \quad \forall\ (x_1,\ldots,x_n)$$

*except, possibly, at points of discontinuity for the joint density function of a continuous random variable.*

An example of the application of Theorem 2.7 is given in Section 2.8.

If $(X_1,\ldots,X_n)$ are independent random variables, then knowing the marginal densities $f_i(x_i)$, $i = 1, \ldots, n$, is equivalent to knowing the joint density function for $(X_1,\ldots,X_n)$, since then $f(x_1,\ldots,x_n) = \prod_{i=1}^{n} f_i(x_i)$. However, if the random variables in the collection $(X_1,\ldots,X_n)$ are *not* independent, then knowing each of the marginal densities of the $X_i$'s is *not* sufficient to determine the joint density function for $(X_1,\ldots,X_n)$. In fact, it can be shown that an uncountably infinite family of different joint density functions can give rise to the same collection of marginal density functions.[20] We provide the following counterexample in the bivariate case to the proposition that knowledge of the marginal PDFs is sufficient for determining the joint PDF.

**Example 2.26** Examine the function

$$f_\alpha(x_1,x_2) = [1 + \alpha(2x_1 - 1)(2x_2 - 1)]I_{[0,1]}(x_1)I_{[0,1]}(x_2).$$

The reader should verify that $f_\alpha(x_1,x_2)$ is a probability density function $\forall\ \alpha \in [-1, 1]$. For any choice of $\alpha \in [-1, 1]$, the marginal density function for $X_1$ is given by

$$f_1(x_1) = \int_{-\infty}^{\infty} f_\alpha(x_1,x_2)dx_2 = I_{[0,1]}(x_1).$$

[20] E. J. Gumbel (1958), Distributions à plusieurs variables dont les marges sont données, *C.R. Acad. Sci., Paris*, 246, pp. 2717–2720.

Similarly, the marginal density of $X_2$, for any choice of $\alpha \in [-1, 1]$, is given by

$$f_2(x_2) = \int_{-\infty}^{\infty} f_\alpha(x_1, x_2)dx_1 = I_{[0,1]}(x_2).$$

Since the *same* marginal density functions are associated with each of an uncountably infinite collection of joint density functions, it is clear that knowledge of $f_1(x_1)$ and $f_2(x_2)$ is insufficient to determine which is the appropriate joint density function for $(X_1, X_2)$. If we knew the marginal densities of $X_1$ and $X_2$, as stated, and if $X_1$ and $X_2$ are *independent* random variables, then we would know that $f(x_1, x_2) = I_{[0,1]}(x_1)I_{[0,1]}(x_2)$. □

### Independence Between Random Vectors and Between Functions of Random Vectors

The independence concepts can be extended so that they apply to independence among two or more random *vectors*. Essentially, all that is required is to interpret the $X_i$'s as *multivariate* random variables in the appropriate definitions and theorems presented heretofore, and the statements are valid. Motivation for the validity of the extensions can be provided using arguments that are analogous to those used previously. For example, to extend the previous bivariate result to two random *vectors*, let $X_1 = (X_{11}, \ldots, X_{1m})$ be an $m$-dimensional random variable and $X_2 = (X_{21}, \ldots, X_{2n})$ be an $n$-dimensional random variable. Then $X_1$ and $X_2$ are independent iff

$$\begin{aligned} P(x_1 \in A_1, x_2 \in A_2) &= P((x_{11}, \ldots, x_{1m}) \in A_1, (x_{21}, \ldots, x_{2n}) \in A_2) \\ &= P((x_{11}, \ldots, x_{1m}) \in A_1)P((x_{21}, \ldots, x_{2n}) \in A_2) = P(x_1 \in A_1)P(x_2 \in A_2), \end{aligned}$$

for *all* event pairs $A_1, A_2$. Furthermore, in terms of joint density factorization, $X_1$ and $X_2$ are independent iff

$$f(x_1, x_2) = f(x_{11}, \ldots, x_{1m}, x_{21}, \ldots, x_{2n}) = f_1(x_{11}, \ldots, x_{1m})f_2(x_{21}, \ldots, x_{2n}) = f_1(x_1)f_2(x_2) \quad \forall\ (x_1, x_2)$$

holds except, perhaps, at points of discontinuity for $f(x_1, x_2)$ for continuous random variables. The reader can contemplate the myriad of other independence conditions that can be constructed for discrete and continuous random *vectors*.

Implications of the extended independence definitions and theorems are qualitatively similar to the implications identified previously for the case where the $X_i$'s were interpreted as scalars. For example, if $X_1 = (X_{11}, \ldots, X_{1m})$ and $X_2 = (X_{21}, \ldots, X_{2n})$ are independent random variables, then

$$P((x_{11}, \ldots, x_{1m}) \in A_1 \mid (x_{21}, \ldots, x_{2n} \in A_2)) = P((x_{11}, \ldots, x_{1m}) \in A_1),$$

i.e., conditional and unconditional probability of events for the random variable $X_1$ are identical (and similarly for $X_2$) for all choices of $A_1$ and $A_2$ for which the conditional probability is defined.

It is also useful to note some results concerning the independence of random variables which are defined as functions of other independent random variables. We begin with the simplest case of two independent random variables $X_1$ and $X_2$.

**Theorem 2.8** *If $X_1$ and $X_2$ are independent random variables, and if the random variables $Y_1$ and $Y_2$ are defined by $y_1 = Y_1(x_1)$ and $y_2 = Y_2(x_2)$, then $Y_1$ and $Y_2$ are independent random variables.*

**Proof** The event involving outcomes of $X_i$ that is equivalent to the event $y_i \in A_i$ is given by $B_i = \{x_i\colon\ Y_i(x_i) \in A_i, x_i \in R(X_i)\}$ for $i = 1,2$. Then

$$\begin{aligned} P(y_1 \in A_1, y_2 \in A_2) &= P(x_1 \in B_1, x_2 \in B_2) \\ &= P(x_1 \in B_1)P(x_2 \in B_2) \quad \text{(by independence of } x_1, x_2\text{)} \\ &= P(y_1 \in A_1)P(y_2 \in A_2), \end{aligned}$$

and since this holds for every event pair $A_1, A_2$, the random variables $Y_1$ and $Y_2$ are independent. ■

**Example 2.27** A large service station sells unleaded and premium-grade gasoline. The quantities sold of each type of fuel on a given day is the outcome of a bivariate random variable with density function[21]

$$f(x_1, x_2) = \tfrac{1}{20} e^{-(.1x_1+.5x_2)} I_{(0,\infty)}(x_1) I_{(0,\infty)}(x_2),$$

where the $x_i$'s are measured in thousands of gallons. The marginal densities are given by (reader, please verify)

$$f_1(x_1) = \tfrac{1}{10} e^{-.1x_1} I_{(0,\infty)}(x_1) \quad \text{and} \quad f_2(x_2) = \tfrac{1}{2} e^{-.5x_2} I_{(0,\infty)}(x_2),$$

and so the random variables are independent. The prices of unleaded and premium gasoline are \$1.25 and \$1.45 per gallon, respectively. The wholesale cost of gasoline plus federal state and local taxes amounts to \$1.00 and \$1.10 per gallon, respectively. Other daily variable costs in selling the two products amount to $C_i(x_i) = 20x_i^2$, $i = 1,2$. Are daily profits above variable costs for the two products independent random variables?

**Answer**: Yes. Note that the profit levels in the two cases are $\Pi_1 = 250x_1 - 20x_1^2$ and $\Pi_2 = 350x_2 - 20x_2^2$, respectively. Since $\Pi_1$ is only a function of $x_1$, $\Pi_2$ is only a function of $x_2$, and $X_1$ and $X_2$ are independent, then $\Pi_1$ and $\Pi_2$ are independent by Theorem 2.8. □

A more general theorem explicitly involving random *vectors* is stated as follows:

**Theorem 2.9** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a collection of n independent random vectors, and let the random vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be defined by $\mathbf{y}_i = \mathbf{Y}_i(\mathbf{x}_i)$, $i = 1, \ldots, n$. Then the random vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are independent.*

[21] This must be an approximation—why?

**Proof** The event involving outcomes of the random vector $\mathbf{X}_i$ that is equivalent to the event $A_i$ for the random vector $\mathbf{Y}_i$ is given by $B_i = \{\mathbf{x}_i: \mathbf{Y}_i(\mathbf{x}_i) \in A_i, \mathbf{x}_i \in R(\mathbf{X}_i)\}$, $i = 1, \ldots, n$. Then

$$\begin{aligned} P(\mathbf{y}_i \in A_i, i = 1, \ldots, n) = P(\mathbf{x}_i \in B_i, i = 1, \ldots, n) &= \prod_{i=1}^{n} P(\mathbf{x}_i \in B_i) \quad \text{(by independence of random vectors)} \\ &= \prod_{i=1}^{n} P(\mathbf{y}_i \in A_i), \end{aligned}$$

and since this holds for every collection of events $A_1, \ldots, A_n$, the random vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are independent by a vector interpretation of the random variables in Def. 2.19. ■

**Example 2.28** Examine the experiment of independently tossing two fair coins and rolling three fair dice. Let $X_1$ and $X_2$ represent whether heads $(x_i = 1)$ or tails $(x_i = 0)$ appears on the first and second coins, respectively, and let $X_3$, $X_4$, and $X_5$ represent the number of dots facing up on each of the three dice, respectively. Since the random variables are independent, the joint density of $X_1, \ldots, X_5$ can be written as

$$f(x_1, \ldots, x_5) = \prod_{i=1}^{2} \frac{1}{2} I_{\{0,1\}}(x_i) \prod_{i=3}^{5} \frac{1}{6} I_{\{1,\ldots,6\}}(x_i).$$

Define two new random *vectors* $\mathbf{Y}_1$ and $\mathbf{Y}_2$ using the vector functions

$$\underset{(2\times1)}{\mathbf{y}_1} = \begin{bmatrix} y_{11} \\ y_{12} \end{bmatrix} = \begin{bmatrix} x_1 + x_2 \\ x_1 x_2 \end{bmatrix} = \underset{(2\times1)}{\mathbf{Y}_1}(x_1, x_2),$$

$$\underset{(2\times1)}{\mathbf{y}_2} = \begin{bmatrix} y_{21} \\ y_{22} \end{bmatrix} = \begin{bmatrix} x_3 + x_4 + x_5 \\ x_3 x_4 / x_5 \end{bmatrix} = \underset{(2\times1)}{\mathbf{Y}_2}(x_3, x_4, x_5).$$

Then since the vector $\mathbf{y}_1$ is a function of $(x_1, x_2)$, $\mathbf{y}_2$ is a function of $(x_3, x_4, x_5)$, and since the random vectors $(X_1, X_2)$ and $(X_3, X_4, X_5)$ are independent (why?), Theorem 2.9 indicates that the random vectors $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent. This is clearly consistent with intuition, since outcomes of the vector $\mathbf{Y}_1$ obviously have nothing to do with outcomes of the vector $\mathbf{Y}_2$. The reader should note that within vectors, the random variables are *not* independent, i.e., $Y_{11}$ and $Y_{12}$ are not independent, and neither are $Y_{21}$ and $Y_{22}$. □

## 2.8 Extended Example of Multivariate Concepts in the Continuous Case

We now further illustrate some of the concepts of this chapter with an extended example involving a trivariate continuous random variable. Let $(X_1, X_2, X_3)$ be a continuous trivariate random variable with joint density function

$f(x_1, x_2, x_3) = \frac{3}{16}x_1x_2^2e^{-x_3}I_{[0,2]}(x_1)I_{[0,2]}(x_2)I_{[0,\infty)}(x_3)$.

**a.** What is the marginal density of $X_1$? of $X_2$? of $X_3$?
**Answer**:

$$\begin{aligned}f_1(x_1) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3)dx_2\, dx_3\\ &= \tfrac{3}{16}x_1I_{[0,2]}(x_1)\int_{-\infty}^{\infty} x_2^2I_{[0,2]}(x_2)dx_2\int_{-\infty}^{\infty} e^{-x_3}I_{[0,\infty)}(x_3)dx_3\\ &= \tfrac{3}{16}x_1I_{[0,2]}(x_1)(\tfrac{8}{3})(1) = \tfrac{1}{2}x_1I_{[0,2]}(x_1).\end{aligned}$$

Similarly,

$$f_2(x_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3)dx_1\, dx_3 = \frac{3}{8}x_2^2I_{[0,2]}(x_2)$$

$$f_3(x_3) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3)dx_1\, dx_2 = e^{-x_3}I_{[0,\infty)}(x_3).$$

**b.** What is the probability that $x_1 \geq 1$?
**Answer**:

$$P(x_1 \geq 1) = \int_1^{\infty} f_1(x_1)dx_1 = \int_1^2 \frac{1}{2}x_1\, dx_1 = \frac{x_1^2}{4}\Big|_1^2 = .75.$$

**c.** Are the three random variables independent?
**Answer**: Yes. Since we have derived the marginal densities of $X_1$, $X_2$, and $X_3$, it is clear that

$$f(x_1, x_2, x_3) = f_1(x_1)f_2(x_2)f_3(x_3) \quad \forall\ (x_1, x_2, x_3).$$

**d.** What is the marginal cumulative distribution function for $x_1$? for $x_3$?
**Answer**: By definition,

$$\begin{aligned}F_1(b) &= \int_{-\infty}^{b} f_1(x_1)dx_1 = \int_{-\infty}^{b} \frac{1}{2}x_1I_{[0,2]}(x_1)dx_1\\ &= \frac{1}{2}\frac{x_1^2}{2}\Big|_0^b I_{[0,2]}(b) + I_{(2,\infty)}(b) = \frac{b^2}{4}I_{[0,2]}(b) + I_{(2,\infty)}(b),\end{aligned}$$

$$\begin{aligned}F_3(b) &= \int_{-\infty}^{b} f_3(x_3)dx_3 = \int_{-\infty}^{b} e^{-x_3}I_{[0,\infty)}(x_3)dx_3\\ &= -e^{-x_3}\big|_0^b\, I_{[0,\infty)}(b) = (1 - e^{-b})I_{[0,\infty)}(b).\end{aligned}$$

**e.** What is the probability that $x_1 \leq 1$? that $x_3 > 1$?
**Answer**: $P(x_1 \leq 1) = F_1(1) = .25$. $P(x_3 > 1) = 1 - F_3(1) = e^{-1} = .3679$.

**f.** What is the joint cumulative distribution function for $X_1$, $X_2$, $X_3$?

**Answer**: By definition,

$$\begin{aligned}F(b_1, b_2, b_3) &= \int_{-\infty}^{b_1}\int_{-\infty}^{b_2}\int_{-\infty}^{b_3} f(x_1, x_2, x_3)dx_3\,dx_2\,dx_1\\ &= \int_{-\infty}^{b_1} \frac{1}{2}x_1 I_{[0,2]}(x_1)dx_1 \int_{-\infty}^{b_2} \frac{3}{8}x_2^2 I_{[0,2]}(x_2)dx_2 \int_{-\infty}^{b_3} e^{-x_3} I_{[0,\infty)}(x_3)dx_3\\ &= \left[\frac{b_1^2}{4}I_{[0,2]}(b_1) + I_{(2,\infty)}(b_1)\right]\left[\frac{3b_2^3}{24}I_{[0,2]}(b_2) + I_{(2,\infty)}(b_2)\right]\left[(1 - e^{-b_3})I_{[0,\infty)}(b_3)\right].\end{aligned}$$

**g.** What is the probability that $x_1 \leq 1$, $x_2 \leq 1$, $x_3 \leq 10$?
**Answer**: $F(1, 1, 10) = (1/4)(3/24)(1 - e^{-10}) = .031$.

**h.** What is the conditional PDF of $X_1$, given that $x_2 = 1$ and $x_3 = 0$?
**Answer**: By definition,

$$f(x_1 \mid x_2 = 1, x_3 = 0) = \frac{f(x_1, 1, 0)}{f_{23}(1, 0)}.$$

Also,

$$f_{23}(x_2, x_3) = \int_{-\infty}^{\infty} f(x_1, x_2, x_3)dx_1 = \frac{3}{8}x_2^2 I_{[0,2]}(x_2)e^{-x_3}I_{[0,\infty)}(x_3).$$

Thus,

$$f(x_1 \mid x_2 = 1, x_3 = 0) = \frac{\frac{3}{16}x_1 I_{[0,2]}(x_1)}{\frac{3}{8}} = \frac{1}{2}x_1 I_{[0,2]}(x_1)$$

**i.** What is the probability that $x_1 \in [0, 1/2]$, given that $x_2 = 1$ and $x_3 = 0$?
**Answer**:

$$\begin{aligned}P(x_1 \in [0, \tfrac{1}{2}] \mid x_2 = 1, x_3 = 0) &= \int_0^{1/2} f(x_1 \mid x_2 = 1, x_3 = 0)dx_1\\ &= \int_0^{1/2} \frac{1}{2}x_1 I_{[0,2]}(x_1)dx_1\\ &= \frac{x_1^2}{4}\bigg|_0^{1/2} = \frac{1}{16}\end{aligned}$$

**j.** Let the two random variables $Y_1$ and $Y_2$ be defined by $y_1 = Y_1(x_1, x_2) = x_1^2 x_2$ and $y_2 = Y_2(x_3) = x_3/2$. Are the random variables $Y_1$ and $Y_2$ independent?
**Answer**: Yes, they are independent. The bivariate random variable $(X_1, X_2)$ is independent of the random variable $X_3$ since $f(x_1, x_2, x_3) = f_{12}(x_1, x_2)f_3(x_3)$, i.e., the joint density function factors into the product of the marginal density of $(X_1, X_2)$ and the marginal density of $X_3$. Then, since $y_1$ is a function of only $(x_1, x_2)$ and $y_2$ is a function of only $x_3$, $Y_1$ and $Y_2$ are independent random variables, by Theorem 2.9.

## 2.9 Events Occurring with Probability Zero

We present examples of events that occur with probability zero in discrete sample spaces. The examples illustrate that the concept of an event occurring with probability zero is not isolated to the case of uncountably infinite sample spaces, but is also applicable to countable and even finite sample spaces.

**Example 2.29** The annual break-even level of electricity sales for a midwestern electric utility is calculated by the accountants of the utility to be one billion kilowatt hours. Assume the annual level of electricity sales is the outcome of some *continuous* random variable with density function $f(x)$. Define the random variable $Y$ such that

$$y = \left\{ \begin{matrix} -1 \\ 0 \\ 1 \end{matrix} \right\} \quad \text{if } x \left\{ \begin{matrix} < 1 \text{ billion} \\ = 1 \text{ billion} \\ > 1 \text{ billion} \end{matrix} \right\} \text{ kilowatt hours.}$$

Then $R(Y)$ is finite, with the outcome $y = 0$ occurring with probability zero. . □

**Example 2.30** Let the discrete random variable $X$ have a countably infinite range, $R(X) = \{x_1, x_2, x_3, \ldots\}$, and density function $f(x)$. Suppose there exists an elementary event in $R(X)$ that is less likely to occur than any other elementary event. Let this elementary event be $x_1$. Then

$$P(x_1) = f(x_1) < f(x_i) = P(x_i) \quad \forall\, i > 1,$$

and it follows that

$$\sum_{i=1}^{n} f(x_1) < \sum_{i=1}^{n} f(x_i) \quad \forall\, n > 1,$$

or

$$f(x_1) < \frac{1}{n} \sum_{i=1}^{n} f(x_i) \quad \forall\, n > 1.$$

Since $\sum_{i=1}^{n} f(x_i) \leq 1 \ \forall\, n$, it follows that the only value for $f(x_1)$ that can satisfy the above strict inequality $\forall\, n$ is $f(x_1) = 0$. Thus, the event $x_1$ occurs with probability zero.

Define another random variable using an indicator function as

$$y = I_{\{R(X) - x_1\}}(x)$$

so that $y = 1$ when $x_1$ does not occur and $y = 0$ when $x_1$ does occur. The range of $Y$ is finite, equal to $R(Y) = \{0, 1\}$, and letting $h$ be the discrete probability density function for $Y$, we have $h(y) = I_{\{1\}}(y)$, so that $y = 0$ occurs with probability zero, while $y = 1$ occurs with probability one. □

## Key Words, Phrases, and Symbols

random variable, $X$
$X: S \rightarrow R$
$X(w)$
$R(X)$
outcome of the random variable, $x$
induced probability space, $\{R(X), \Upsilon_X, P_X\}$
equivalent events
composite random variable
discrete random variable
discrete probability density function
PDF
continuous random variable
continuous probability density function
event $A$ occurs with probability zero
event $A$ is relatively impossible
event $A$ occurs with probability one
event $A$ is relatively certain
classes of discrete and continuous density functions
mixed discrete-continuous random variables
discrete density component
continuous density component
abbreviated set notation
$P(x \leq b)$
cumulative distribution function
CDF
truncation function
$F(b)$
nondecreasing function
increasing function
duality between CDFs and PDFs
real-valued vector function
multivariate random variable
discrete joint probability density function
continuous joint probability density function
$f(x_1, \ldots, x_n)$
joint cumulative distribution function
marginal probability density function
MPDF
$f_{1\ldots m}(x_1, \ldots, x_m)$
marginal cumulative distribution function
MCDF
conditional density function
$f(x_1, \ldots, x_m \mid (x_{m+1}, \ldots, x_n) \in B)$
conditional cumulative distribution function
marginal cumulative distribution function
independence of random variables
joint density factorization for independence

## Problems

**1.** Which of the following are valid probability density functions? Justify your answer.

a. $f(x) = (.2)^x(.6)^{1-x} I_{\{0,1\}}(x)$

b. $f(x) = (.3)(.7)^x I_{\{0,1,2,\ldots\}}(x)$

c. $f(x) = .6e^{-x/4} I_{[0,\infty)}(x)$

d. $f(x) = x^{-1} I_{[1,e]}(x)$

**2.** Graph each of the *density* functions in Problem 1.

**3.** Sparkle Cola, Inc., manufactures a cola drink. The cola is sold in 12 oz. bottles. The probability distribution associated with the random variable whose outcome represents the actual quantity of soda placed in a bottle of Sparkle Cola by the soda bottling line is specified to be

$$f(x) = 50[e^{-100(12-x)} I_{(-\infty,12]}(x) + e^{-100(x-12)} I_{(12,\infty)}(x)].$$

In order to be considered full, a bottle must contain within .25 oz. of 12 oz. of soda.

a. Define a random variable whose outcome indicates whether or not a bottle is considered full.

b. What is the range of this random variable?

c. Define a probability density function for the random variable. Use it to assign probability to the event that a bottle is "considered full."

d. The PDF $f(x)$ is only an approximation. Why?

**4.** A health maintenance organization (HMO) is currently treating 10 patients with a deadly bacterial infection. The best-known antibiotic treatment is being used in these cases, and this treatment is effective 95 percent of the time. If the treatment is not effective, the patient expires.

a. Define a random variable whose outcome represents the number of patients being treated by the HMO that survive the deadly bacterial infection. What is the range of this random variable? What is the event space for outcomes of this random variable?

b. Define the appropriate probability density function for the random variable you defined in (a). Define the probability set function appropriate for assigning probabilities to events regarding the outcome of the random variable.

c. Using the probability space you defined in (a) and (b), what is the probability that all 10 of the patients survive the infection?

d. What is the probability that no more than two patients expire?

e. If 50 percent of the patients were to expire, the government would require that the HMO suspend operations, and an investigation into the medical practices of the HMO would be conducted. Provide an argument in defense of the government's actions in this case.

**5.** Star Enterprises is a small firm that produces a product that is simple to manufacture, involving only one variable input. The relationship between input and output levels is given by $q = x^{.5}$, where $q$ is the quantity of product produced and $x$ is the quantity of variable input used. For any given output and input prices, Star Enterprises operates at a level of production that maximizes its profit over variable cost. The possible prices facing the firm on a given day is represented by a random variable $V$ with $R(V) = \{10, 20, 30\}$ and probability density function

$$f(v) = .2I_{\{10\}}(v) + .5I_{\{20\}}(v) + .3I_{\{30\}}(v).$$

Input prices vary independently of output prices, and input price on a given day is the outcome of $W$ with $R(W) = \{1, 2, 3\}$ and probability density function

$$g(w) = .4I_{\{1\}}(w) + .3I_{\{2\}}(w) + .3I_{\{3\}}(w).$$

a. Define a random variable whose outcome represents Star's profit over variable cost on a given day. What is the range of the random variable? What is the event space?

b. Define the appropriate probability density function for profit over variable cost. Define a probability set function appropriate for assigning probability to events relating to profit over variable cost.

c. What is the probability that the firm makes at least \$100 profit over variable cost?

d. What is the probability that the firm makes a positive profit on a given day? Is making a positive profit a certain event? Why or why not?

e. *Given* that the firm makes at least \$100 profit over variable cost, what is the probability that it makes at least \$200 profit over variable cost?

**6.** The ACME Freight Co. has containerized a large quantity of 4-megabyte memory chips that are to be shipped to a personal computer manufacturer in California. The shipment contains 1,000 boxes of memory chips, with each box containing a dozen chips. The chip manufacturer calls and says that due to an error in manufacturing, each box contains exactly one defective chip. The defect can be detected through an easily administered, nondestructive continuity test using an ohmmeter. The chip maker requests that ACME break open the container, find the defective chip in each box, discard them, and then reassemble the container for shipment. The testing of each chip requires one minute to accomplish.

a. Define a random variable representing the amount of testing time required to find the defective chip in a box of chips. What is the range of the random variable? What is the event space?

b. Define a probability density function for the random variable you have defined in (a). Define a probability set function appropriate for assigning probabilities to events relating to testing time required to find the defective chip in a box of chips.

c. What is the probability that it will take longer than five minutes to find the defective chip in a box of chips?

d. If ACME uses two eight-hour-shift workers for one shift each to perform the testing, what is the probability that testing of all of the boxes in the container will be completed?

**7.** Intelligent Electronics, Inc., manufactures monochrome liquid crystal display (LCD) notebook computer screens. The number of hours an LCD screen functions until failure is represented by the outcome of a random variable $X$ having range $R(X) = [0, \infty)$ and probability density function

$$f(x) = .01 \exp\left(-\frac{x}{100}\right) I_{[0,\infty)}(x).$$

The value of $x$ is measured in thousands of hours. The company has a one-year warranty on its LCD screen, during which time the LCD screen will be replaced free of charge if it fails to function.

a. Assuming that the LCD screen is used for 10,000 hours per year, what is the probability that the firm will have to perform warranty service on an LCD screen?

b. What is the probability that the screen functions for at least 50,000 hours? *Given* that the screen has already functioned for 50,000 hours, what is the probability that it will function for at least *another* 50,000 hours?

**8.** People Power, Inc., is a firm that specializes in providing temporary help to various businesses. Job applicants are administered an aptitude test that evaluates mathematics, writing, and manual dexterity skills. After the firm analyzed thousands of job applicants who took the test, it found that the scores on the three tests could be viewed as outcomes of random variables with the following joint density function (the tests are graded on a 0–1 scale, with 0 the lowest score and 1 the highest):

$$f(x_1, x_2, x_3) = .80(2x_1 + 3x_2)x_3 \prod_{i=1}^{3} I_{[0,1]}(x_i).$$

a. A job opening has occurred for an office manager. People Power, Inc., requires scores of > .75 on both the mathematics and writing tests for a job applicant to be offered the position. Define the marginal density function for the mathematics and writing scores. Use it to define a probability space in which probability questions concerning events for the mathematics and writing scores can be answered. What is the probability that a job applicant who has just entered the office to take the test will qualify for the office manager position?

b. A job opening has occurred for a warehouse worker. People Power, Inc., requires a score of > .80 on the manual dexterity test for a job applicant to be offered the position. Define the marginal density function for the dexterity score. Use it to define a probability space in which probability questions concerning events for the dexterity score can be answered. What is the probability that a job applicant who has just entered the office to take the test will qualify for the warehouse worker position?

c. Find the conditional density of the writing test score, given that the job applicant achieves a score of > .75 on the mathematics test. Given that the job applicant scores > .75 on the mathematics test, what is the probability that she scores > .75 on the writing test? Are the two test scores independent random variables?

d. Is the manual dexterity score independent of the writing and mathematics scores? Why or why not?

**9.** The weekly average price (in dollars/foot) and total quantity sold (measured in thousands of feet) of copper wire manufactured by the Colton Cable Co. can be viewed as the outcome of the bivariate random variable $(P, Q)$ having the joint density function

$$f(p, q) = 5pe^{-pq} I_{[.1,.3]}(p) I_{(0,\infty)}(q).$$

a. What is the probability that total dollar sales in a week will be less than \$2,000?

b. Find the marginal density of price. What is the probability that price will exceed \$.25/foot?

c. Find the conditional density of quantity, given price = .20. What is the probability that > 5,000 feet of cable will be sold in a given week?

d. Find the conditional density of quantity, given price = .10. What is the probability that > 5,000 feet of cable will be sold in a given week? Compare this result to your answer in (c). Does this make economic sense? Explain.

**10.** A personal computer manufacturer produces both desktop computers and notebook computers. The monthly proportions of customer orders received for desktop and notebook computers that are shipped within one week's time can be viewed as the outcome of a bivariate random variable $(X, Y)$ with joint probability density

$$f(x, y) = (2 - x - y) I_{[0,1]}(x) I_{[0,1]}(y)$$

a. In a given month, what is the probability that more than 75 percent of notebook computers and 75 percent of desktop computers are shipped within one week of ordering?

b. Assuming that an equal number of desktop and notebook computers are ordered in a given month, what is the probability that more than 75 percent of all orders received will be shipped within one week?

c. Are the random variables independent?

d. Define the conditional probability that less than 50 percent of the notebook orders are shipped within one week, given that $x$ proportion of the desktop orders are shipped within one week (the probability will be a function of the proportion $x$). How does this probability change as $x$ increases?

**11.** A small nursery has seven employees, three of whom are salespersons, and four of whom are gardeners who tend to the growing and caring of the nursery stock. With such a small staff, employee absenteeism can be critical. The number of salespersons and gardeners absent on any given day is the outcome of a bivariate random variable $(X, Y)$. The nonzero values of the joint density function are given in tabular form as

| | | | | Y | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| | 0 | .75 | .025 | .01 | .01 | .03 |
| X | 1 | .06 | .03 | .01 | .01 | .003 |
| | 2 | .025 | .01 | .005 | .005 | .002 |
| | 3 | .005 | .004 | .003 | .002 | .001 |

a. What is the probability that more than two employees will be absent on any given day?

b. Find the marginal density function of the number of gardeners that are absent. What is the probability that more than two gardeners will be absent on any given day?

c. Are the number of gardener absences and the number of salesperson absences independent random variables?

d. Find the conditional density function for the number of salespersons who are absent, given that there are no gardeners absent. What is the probability that there are no salespersons absent, given that there are no gardeners absent? Is the conditional probability higher or lower given that there is at least one gardener absent?

**12.** The joint density of the bivariate random variable $(X, Y)$ is given by

$$f(x, y) = xy I_{[0,1]}(x) I_{[0,2]}(y).$$

a. Find the joint cumulative distribution function of $(X, Y)$. Use it to find the probability that $x \leq .5$ and $y \leq 1$.

b. Find the marginal cumulative distribution function of $X$. What is the probability that $x \leq .5$?

c. Find the marginal density of $X$ from the marginal cumulative distribution of $X$.

**13.** The joint cumulative distribution function for $(X, Y)$ is given by

$$F(x, y) = (1 - e^{-x/10} - e^{-y/2} + e^{-(x+5y)/10}) I_{(0,\infty)}(x) I_{(0,\infty)}(y).$$

a. Find the joint density function of $(X, Y)$.

b. Find the marginal density function of $X$.

c. Find the marginal cumulative distribution function of $X$.

**14.** The cumulative distribution of the random variable $X$ is given by

$$F(x) = (1 - p^{x+1}) I_{\{0,1,2,\ldots\}}(x),$$

for some choice of $p \in (0, 1)$.

a. Find the density function of the random variable $X$.

b. What is the probability that $x \leq 8$ if $p = .75$?

c. What is the probability that $x \leq 1$ given that $x \leq 8$?

**15.** The federal mint uses a stamping machine to make coins. Each stamping produces 10 coins. The number of the stamping at which the machine breaks down and begins to produce defective coins can be viewed as the outcome of a random variable, $X$, having a probability density function with general functional form

$$f(x) = \alpha(1 - \beta)^{x-1} I_{\{1,2,3,\ldots\}}(x), \quad \text{where } \beta \in (0, 1).$$

a. Are there any constraints on the choice of $\alpha$ for $f(x)$ to be a probability density function? If so, precisely what are they?

b. Is the random variable $X$ a discrete or a continuous random variable? Why?

c. It is known that the probability the machine will break down on the first stamping is equal to .05. What is the specific functional form of the probability density function $f(x)$? What is the probability that the machine will break down on the tenth stamping?

d. Derive a functional representation for the cumulative distribution function corresponding to the random variable $X$. Use it to assign the appropriate probability to the event that the machine does not break down for at least 10 stampings.

e. What is the probability that the machine does not break down for at least 20 stampings, *given* that the machine does not break down for at least 10 stampings?

**16.** The daily quantity demanded of unleaded gasoline in a regional market can be represented as

$$Q = 100 - 10p + E, \quad \text{where } p \in [0, 8],$$

and $E$ is a random variable having a probability density given by

$$f(e) = .025 I_{[-20,20]}(e)$$

Quantity demanded, $Q$, is measured in thousands of gallons, and price, $p$, is measured in dollars.

a. What is the probability of the quantity demanded being greater than 70,000 gallons if price is equal to \$4? if price is equal to \$3?

b. If the average variable cost of supplying $Q$ amount of unleaded gasoline is given by $C(Q) = Q^{.5}/2$, define a random variable that can be used to represent the daily profit above variable cost from the sale of unleaded gasoline.

c. If price is set equal to \$4, what is the probability that there will be a positive profit above variable cost on a given day? What if price is set to \$3? to \$5?

d. If price is set to \$6, what is the probability that quantity demanded will equal 40,000 gallons?

**17.** For each of the cumulative distribution functions listed below, find the associated probability density functions. For each CDF, calculate $P(x \leq 6)$.

a. $F(b) = (1 - e^{-b/6})I_{(0,\infty)}(b)$

b. $F(b) = (5/3)(.6 - .6^{\text{trunc}(b)+1})I_{(0,\infty)}(b)$

**18.** An economics class has a total of 20 students with the following age distribution:

| # of students | age |
|---|---|
| 10 | 19 |
| 4 | 20 |
| 4 | 21 |
| 1 | 24 |
| 1 | 29 |

Two students are to be selected randomly, without replacement, from the class to give a team report on the state of the economy. Define a random variable whose outcome represents the average age of the two students selected. Also, define a discrete probability density function for the random variable. Finally, what is the probability space for this experiment?

**19.** Let $X$ be a random variable representing the *minimum* of the two numbers of dots that are facing up after a pair of fair dice is rolled. Define the appropriate probability density for $X$. What is the probability space for the experiment of rolling the fair dice and observing the minimum of the two numbers of dots?

**20.** A package of a half-dozen light bulbs contains two defective bulbs. Two bulbs are randomly selected from the package and are to be used in the same light fixture. Let the random variable $X$ represent the number of light bulbs selected that function properly (i.e., that are not defective). Define the appropriate probability density function for $X$. What is the probability space for the experiment?

**21.** A committee of three students will be randomly selected from a senior-level political science class to present an assessment of the impacts of an antitax initiative to some visiting state legislators. The class consists of five economists, eight political science majors, four business majors, and three art majors. Referring to the experiment of drawing three students randomly from the class, let the bivariate random variable $(X, Y)$ be defined by $x$ = number of economists on the committee, and $y$ = number of business majors on the committee.

a. What is the range of the bivariate random variable $(X, Y)$? What is the probability density function, $f(x, y)$, for this bivariate random variable? What is the probability space?

b. What is the probability that the committee will contain at least one economist and at least one business major?

c. What is the probability that the committee will consist of only political science and art majors?

d. *On the basis of the probability space you defined in (a)*, is it possible for you to assign probability to the event that the committee consists entirely of art majors? Why or why not? If you answer yes, calculate this probability using $f(x, y)$ from (a).

e. Calculate the marginal density function for the random variable $X$. What is the probability that the committee contains three economists?

f. Define the conditional density function for the number of business majors on the committee, *given* that the committee contains two economists. What is the probability that the committee contains less than one business major, *given* that the committee contains two economists?

g. Define the conditional density function for the number of business majors on the committee, *given* that the committee contains at least two economists. What is the probability that the committee contains less than one business major, *given* that the committee contains at least two economists?

h. Are the random variables $X$ and $Y$ independent? Justify your answer.

**22.** The Imperial Electric Co. makes high-quality portable compact disc players for sale in international and domestic markets. The company operates two plants in the United States, where one plant is located in the Pacific Northwest and one is located in the South. At either plant, once a disc player is assembled, it is subjected to a stringent quality-control inspection, at which time the disc player is either approved for shipment or else sent back for adjustment before it is shipped. On any given day, the proportion of the units produced at each plant that require adjustment before shipping, and the total production of disc players at the company's two plants, are outcomes of a trivariate random variable, with the following joint probability density function:

$$f(x,y,z) = \tfrac{2}{3}(x+y)e^{-x}I_{(0,\infty)}(x)I_{(0,1)}(y)I_{(0,1)}(z),$$

where

$x =$ total production of disc players at the two plants, measured in thousands of units,

$y =$ proportion of the units produced at the Pacific Northwest plant that are shipped without adjustment, and

$z =$ proportion of the units produced in the southern plant that are shipped without adjustment.

a. In this application, the use of a *continuous* trivariate random variable to represent proportions and total production values must be viewed as only an *approximation* to the underlying real-world situation. Why? In the remaining parts, assume the approximation is acceptably accurate, and use the approximation to answer questions where appropriate.

b. What is the probability that less than 50 percent of the disc players produced in each plant will be shipped without adjustment and that production will be less than 1,000 units on a given day?

c. Derive the marginal probability density function for the total production of disc players at the two plants. What is the probability that less than 1,000 units will be produced on a given day?

d. Derive the marginal probability density function for the bivariate random variable $(Y,Z)$. What is the probability that more than 75 percent of the disc players will be shipped without adjustment from each plant?

e. Derive the conditional density function for $X$, *given* that 50 percent of the disc players are shipped from the Pacific Northwest plant without adjustment. What is the probability that 1,500 disc players will be produced by the Imperial Electric Co. on a day for which 50 percent of the disc players are shipped from the Pacific Northwest plant without adjustment?

f. Answer (e) for the case where 90 percent of the disc players are shipped from the Pacific Northwest plant without adjustment.

g. Are the random variables $(X,Y,Z)$ independent random variables?

h. Are the random variables $(Y,Z)$ independent random variables?

**23.** ACE Rentals, a car-rental company, rents three types of cars: compacts, mid-size sedans, and large luxury cars. Let $(x_1,x_2,x_3)$ represent the number of compacts, mid-size sedans, and luxury cars, respectively, that ACE rents per day. Let the sample space for the possible outcomes of $(X_1,X_2,X_3)$ be given by

$$S = \{(x_1,x_2,x_3):\ x_1,x_2,\text{ and } x_3 \in \{0,1,2,3\}\}$$

(ACE has an inventory of nine cars, evenly distributed among the three types of cars).

The discrete probability density function associated with $(X_1,X_2,X_3)$ is given by

$$f(x_1,x_2,x_3) = \left[\frac{.004(3+2x_1+x_2)}{(1+x_3)}\right]\prod_{i=1}^{3} I_{\{0,1,2,3\}}(x_i).$$

The compact car rents for \$20/day, the mid-size sedan rents for \$30/day, and the luxury car rents for \$60/day.

a. Derive the marginal density function for $X_3$. What is the probability that all three luxury cars are rented on a given day?

b. Derive the marginal density function for $(X_1,X_2)$. What is the probability of more than one compact and more than one mid-size sedan being rented on a given day?

c. Derive the conditional density function for $X_1$, given $x_2 \geq 2$. What is the probability of renting no more than one compact car, given that two or more mid-size sedans are rented?

d. Are $X_1$, $X_2$, and $X_3$ jointly independent random variables? Why or why not? Is $(X_1,X_2)$ independent of $X_3$?

e. Derive the conditional density function for $(X_1,X_2)$, given that $x_3 = 0$. What is the probability of renting more than one compact and more than one mid-size sedan given that no luxury cars are rented?

f. If it costs \$150/day to operate ACE Rentals, define a random variable that represents the daily profit made by the company. Define an appropriate density function for this random variable. What is the probability that ACE Rentals makes a positive daily profit on a given day?

**24.** If $(X_1, X_2)$ and $(X_3, X_4)$ are independent bivariate random variables, are $X_2$ and $X_3$ independent random variables? Why or why not?

**25.** The joint density function of the discrete trivariate random variable $(X_1, X_2, X_3)$ is given by

$$f(x_1, x_2, x_3) = .20 I_{\{0,1\}}(x_1) I_{\{0,1\}}(x_2) I_{\{|x_1 - x_2|\}}(x_3) + .05 I_{\{0,1\}}(x_1) I_{\{0,1\}}(x_2) I_{\{1-|x_1 - x_2|\}}(x_3).$$

a. Are $(X_1, X_2)$, $(X_1, X_3)$, and $(X_2, X_3)$ *each* pairwise independent random variables?

b. Are $X_1$, $X_2$, $X_3$ jointly independent random variables?

**26.** SUPERCOMP, a retail computer store, sells personal computers and printers. The number of computers and printers sold on any given day varies, with the probabilities of the various possible sales outcomes being given by the following table:

| | | Number of Computers Sold | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | |
| | 0 | .03 | .03 | .02 | .02 | .01 | Probabilities of elementary events |
| Number | 1 | .02 | .05 | .06 | .02 | .01 | |
| of | 2 | .01 | .02 | .10 | .05 | .05 | |
| printers | 3 | .01 | .01 | .05 | .10 | .10 | |
| | 4 | .01 | .01 | .01 | .05 | .15 | |

a. If SUPERCOMP has a profit margin (product sales price – product unit cost) of \$100 per computer sold and \$50 per printer sold, define a random variable representing aggregate profit margin from the sale of computers and printers on a given day. What is the range of this random variable?

b. Define a discrete density function appropriate for use in calculating probabilities of all events concerning aggregate profit margin outcomes on a given day.

c. What is the probability that the aggregate profit margin $\geq$ \$300 on a given day?

d. The daily variable cost of running the store is \$200/day. What is the probability that SUPERCOMP's aggregate profit margin on computer and printer sales will equal or exceed variable costs on a given day?

e. Assuming that events involving the number of computers and printers sold are independent from day to day, what is the probability that for any given six-day business week, aggregate profit margins equal or exceed variable cost all six days?

# 3 Mathematical Expectation and Moments

## 3.1 Expectation of a Random Variable

The definition of the expectation of a random variable can be motivated both by the concept of a weighted average and through the use of the physics concept of the center of gravity, or the balancing point of a distribution of weights. We first examine the case of a discrete random variable and look at a problem involving the balancing-point concept.[1]

**Example 3.1** Suppose that a weightless rod is placed on a fulcrum, a weight of 10 lbs. is placed on the rod exactly four feet to the right of the fulcrum, and a weight of 5 lbs. is placed on the rod exactly eight feet to the left of the fulcrum, as shown in Figure 3.1.

Assume that $\delta = 0$ is the point at which the fulcrum is placed, so that the 10 lb. weight is at the point $x_1 = 4$, and the 5 lb. weight is at the point $x_2 = -8$. Let mass$(x)$ denote the mass placed at point $x$. The *moment* of any mass placed at a point $x$ is defined to be the product of the mass times its signed distance from the fulcrum, $[\text{mass}(x)](x - \delta)$, where $\delta$ is the point at which the fulcrum

[1] Readers who recollect earlier days spent on a seesaw should possess ample intuition regarding the placement of weights appropriate distances from a fulcrum so as to achieve a "balanced seesaw."

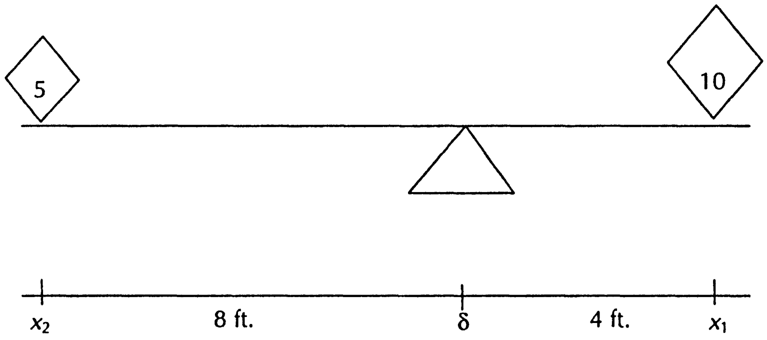


**Figure 3-1** Weights on a weightless rod.

is placed. Thus, the moment of the 10 lb. weight is $10(4-0) = 40$, while the moment of the 5 lb. weight is $5(-8-0) = -40$. A system of weights with fulcrum placed at $\delta$ will balance if the sum of the moments $\sum_{i=1}^{n}[\text{mass}(x_i)](x_i - \delta)$, called the *total moment of the system*, is equal to zero. Our system balances, since $40 + (-40) = 0$. □

The moments concept illustrated in Ex. 3.1 can be used to identify the point at which a probability density "balances." Recall the dice example of Ex. 2.2. We place the probability "weights" on a weightless rod at the points corresponding to the outcomes with which the probabilities are associated, as shown in Figure 3.2.

At what point should the fulcrum be placed so that the distribution of weights balances? We require that the total moment of the system be zero.

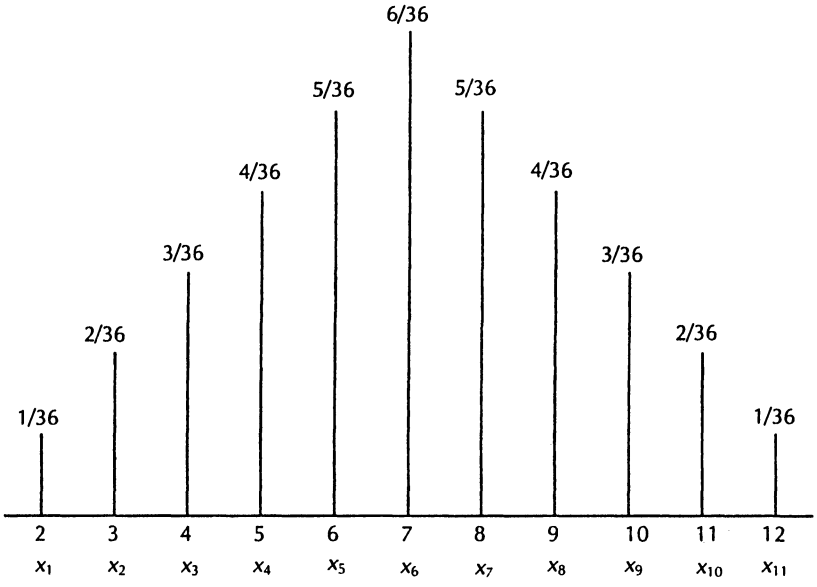


**Figure 3-2** Density function "weights" on a weightless rod.

Thus, we require that

$$\sum_{i=1}^{11}[\text{mass}(x_i)](x_i - \delta) = \sum_{i=1}^{11} f(x_i)(x_i - \delta) = 0,$$

which implies that

$$\sum_{i=1}^{11} f(x_i)x_i = \delta\left[\sum_{i=1}^{11} f(x_i)\right] = \delta,$$

where the sum in brackets equals 1 because the density function $f(x)$ is being summed over the entire range of the random variable $X$. Substituting the appropriate values of $x_i$ and $f(x_i)$ in the expression obtains the result $\delta = 7$. Thus, if the fulcrum were placed at the point 7, the system of weights would balance. The quantity $\delta$ is precisely what we mean by the **expected value** of the discrete random variable $X$ with density function $f(x)$. Thus, the expected value of a discrete random variable is a measure of the center of gravity of its density function.

**Definition 3.1**
***Expectation of a random variable; discrete case***

> The expected value of a discrete random variable exists, and is defined by $\mathrm{E}X = \sum_{x\in R(X)} xf(x)$, iff $\sum_{x\in R(X)} |x|f(x) < \infty$.

The significance of the existence condition stated in the definition is to ensure that the sum defining the expectation is *absolutely convergent*.[2] If all outcomes of $X$ are finite and $R(X)$ is finite, then the existence condition is met, and the expectation of $X$ will necessarily exist, since $f(x) \in [0, 1]\ \forall\ x$, so that $\sum_{x\in R(X)} |x|f(x)$ is a finite sum of finite numbers.

There is no guarantee that $\sum_{x\in R(X)} xf(x)$ converges to any finite number when $R(X)$ is countably infinite, even if all of the outcomes of $X$ are finite. Furthermore, even if the infinite sum converges to some real number for a given ordering of the terms in the sum, it is an unfortunate fact of infinite sums involving both positive and negative terms that, unless the sum is *absolutely convergent*, an appropriate reordering of terms will result in the infinite sum converging to other real numbers.[3] This is hardly consistent with the notion of a balancing point of the density $f(x)$. Moreover, the nonuniqueness of values to which the infinite sum can converge makes any particular convergence point arbitrary and meaningless as the expectation of $X$. Thus, to ensure the finiteness and *uniqueness* of the converged value in the countably infinite case, E$X$ is said to exist iff $\sum_{x\in R(X)} |x|f(x) < \infty$, which is to say, iff $\sum_{x\in R(X)} xf(x)$ is *absolutely convergent*. For virtually any problem of practical interest, if the sum used

[2] The sum $\sum_{x\in A} g(x)$ is absolutely convergent iff $\sum_{x\in A} |g(x)| < \infty$. In our case, let $g(x) = xf(x)$, and since $f(x) \geq 0$, $|g(x)| = |xf(x)| = |x|f(x)$. (See R. G. Bartle (1976), *The Elements of Real Analysis*, 2nd ed. New York: John Wiley, p. 289. Some authors define E$X = \infty$ if $\sum_{\substack{x\in R(X)\\ x>0}} xf(x) = \infty$ and $\sum_{\substack{x\in R(X)\\ x<0}} xf(x) > -\infty$, and E$X = -\infty$ if the first and second sums $< \infty$ and $= -\infty$ respectively.

[3] Bartle, Real Analysis, p. 292.

in the definition of the expectation is finite, the expectation can be said to exist. It should also be noted that in many applications, random variables are nonnegative valued, in which case if the sum is convergent, it is necessarily *absolutely* convergent.

Since $f(x) \geq 0 \ \forall \ x \in R(X)$, and $\sum_{x \in R(X)} f(x) = 1$, the expected value of a discrete random variable can also be straightforwardly interpreted as a weighted average of the possible outcomes (or range elements) of the random variable. In this context the weight assigned to a particular outcome of the random variable is equal to the probability that the outcome occurs (as given by $f(x)$).

**Example 3.2** A life insurance company offers a 50-year-old male a \$1,000 face value, one-year term life insurance policy for a premium of \$14. Standard mortality tables indicate that the probability a male in this age category will die within the year is .006. What is the insurance company's expected gain from issuing this policy?

**Answer**: Define a random variable $X$ having range $R(X) = \{14, -986\}$, the outcomes corresponding, respectively, to the premium of \$14 collected by the company if the person lives, or the net payment of \$986 (\$1,000 minus the premium collected) to the person's estate if he dies. The probabilities of the two elementary events are .994 and .006, respectively. Then, $\mathrm{E}X = (14)(.994) - (986)(.006) = 8$. □

Note that the expected value of $X$ need not be a value in the range of $X$ as the previous and following examples illustrate.

**Example 3.3** Examine the experiment of rolling a die, and recall that the density function associated with the dots facing up on the die is $f(x) = (1/6)I_{\{1,\ldots,6\}}(x)$. In this case $R(X) = \{1, 2, 3, 4, 5, 6\}$. The expected value of $X$ equals $\mathrm{E}X = \sum_{x \in R(X)} x f(x) = \sum_{x=1}^{6} (x/6) I_{\{1,\ldots,6\}}(x) = 3.5$, and thus $\mathrm{E}X \notin R(X)$. □

In the continuous case, the physics problem of balancing mass on a weightless rod can no longer be conceptualized as having weights applied to specific points on the rod. Instead, the mass is interpreted as being continuously spread out along the rod, exerting downward force along a continuum of points on the rod. The mass function, mass$(x)$, is now a density of the mass at point $x$, $\int_{-\infty}^{\infty}$ mass$(x)dx$ equals the total mass placed on the rod, and $\int_a^b$ mass$(x)dx$ equals the mass lying between the points $a$ and $b$[4] (see Figure 3.3). The mass is balanced on the rod when the fulcrum is placed at the point $\delta$ such that the total moment of the mass, $\int_{-\infty}^{\infty}$ mass$(x)(x - \delta)dx$, equals zero.

Viewing our density function as a (probability) mass, the continuous density "balances" with a fulcrum placed at the point $\delta$ if

$$\int_{-\infty}^{\infty} f(x)(x - \delta)dx = 0$$

[4] The reader might notice that the mass function would exhibit properties similar to a probability density function, except the integral over the real line would not necessarily = 1, but rather would equal the number reflecting the total mass placed on the rod.

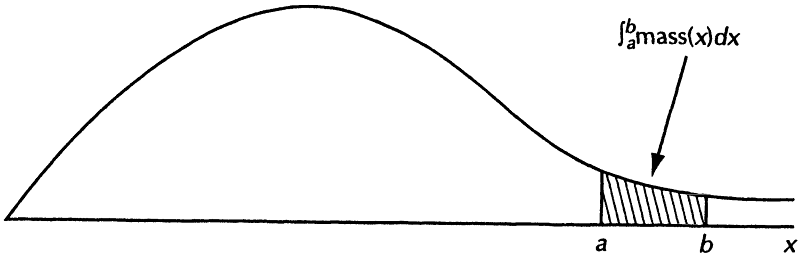


**Figure 3-3** Continuous mass on a weightless rod.

which implies that

$$\int_{-\infty}^{\infty} xf(x)dx = \delta \underbrace{\int_{-\infty}^{\infty} f(x)dx}_{1} = \delta.$$

Again, it is this balancing point or center of gravity, $\delta$, of the density that represents the expectation of the continuous random variable $X$ having density $f(x)$.

**Definition 3.2** ***Expected value of a random variable; continuous case***

The expected value of the continuous random variable $X$ exists, and is defined by $\mathrm{E}X = \int_{-\infty}^{\infty} xf(x)dx$, iff $\int_{-\infty}^{\infty} |x|f(x)dx < \infty$.

The significance of the existence condition is to ensure that the limit operations inherent in the definition of the improper Riemann integral $\int_{-\infty}^{\infty} xf(x)dx$ exhibit *absolute convergence*, which is in fact required for the existence of the improper Riemann integral.[5] Henceforth, it will be sufficient to state that the expectation exists if the integral exists.

The expected value of $X$ in the continuous case can also be viewed, in a limit sense, as a weighted average of the possible outcomes (or range elements) of the random variable. This interpretation follows fundamentally from the definition of the definite integral as the limit of a *Riemann sum*.[6] For the sake of exposition, we assume that the positive values of $f(x)$ all occur for x within the interval $x \in [a, b]$ for finite $a$ and $b$, although a similar argument holds when $a$ and/or $b$ are infinite.

Let $x_0 = a$, $x_n = b$, $x_0 < x_1 < x_2 < \cdots < x_n$, $\Delta x_i = x_i - x_{i-1}$, and examine the Riemann sum $\sum_{i=1}^{n} X_i^0 f(X_i^0)\Delta x_i$, where $X_i^0$ is a value chosen such that $X_i^0 \in [x_{i-1}, x_i]$.

This situation can be represented by the diagram in Figure 3.4. Thus, each $X_i^0$ is weighted by the value $f(X_i^0)\Delta x_i$, an area indicated in the diagram by a shaded rectangle, and a summation is taken over all the chosen values of the $X_i^0$. We have effectively divided the interval $[a, b]$ into a collection of

[5] See J. E. Marsden (1974), *Elementary Classical Analysis*, San Francisco: W. H. Freeman and Co., pp. 267–271. Some authors define $\mathrm{E}X = \infty$ if $\int_0^{\infty} xf(x)dx = \infty$ and $\int_{-\infty}^{0} xf(x) > -\infty$, and $\mathrm{E}X = -\infty$ if the first and second integrals are $< \infty$ and $= -\infty$, respectively.

[6] Bartle, Real Analysis, pp. 213-214.

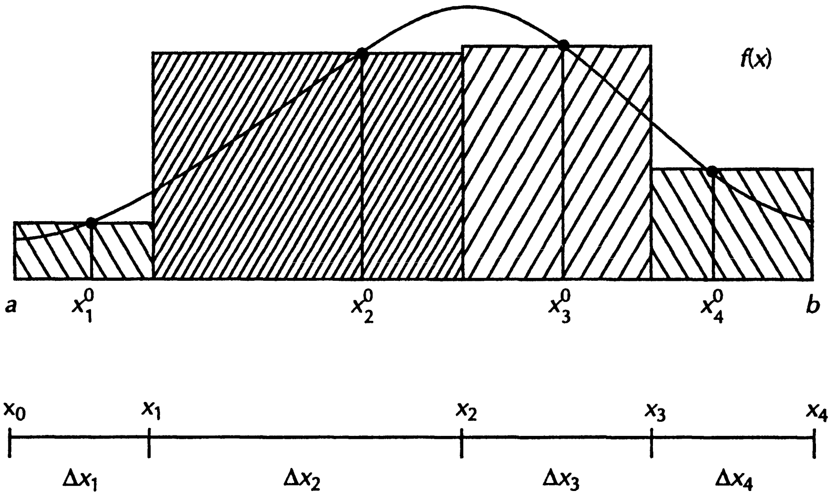


**Figure 3-4** Approximation of area under $f(x)$.

subintervals of various widths $\Delta x_i$, $i = 1, \ldots, n$. The subinterval of maximum width is referred to as the *mesh* (or sometimes, the norm) of the collection of subintervals, i.e., mesh $= \max(\Delta x_1, \ldots, \Delta x_n)$. If $f(x)$ is continuous,[7] then as we increase without bound the number of subintervals, and in so doing decrease the mesh to zero, we have

$$\lim_{\substack{n\to\infty \\ \text{mesh}\to 0}} \sum_{i=1}^{n} X_i^0 f\left(X_i^0\right) \Delta x_i = \int_a^b x f(x) dx = \int_{-\infty}^{\infty} x f(x) dx$$

and

$$\lim_{\substack{n\to\infty \\ \text{mesh}\to 0}} \sum_{i=1}^{n} f\left(X_i^0\right) \Delta x_i = \int_a^b f(x) dx = \int_{-\infty}^{\infty} f(x) dx = 1$$

(presuming the first limit, and hence first integral, exists—the second limit, and integral, necessarily exists since $f(x)$ is a density function). Therefore, under the assumptions of our argument, the expected value of $X$ can be viewed, in a limiting sense, as a weighted average of an infinite number of possible outcomes of the random variable.

**Example 3.4** A large domestic automobile manufacturer mails out quarterly customer satisfaction surveys to owners who have purchased new automobiles within the last three years. The proportion of surveys returned in any given quarter is the outcome of a random variable $X$ having density function $f(x) = 3x^2 I_{[0,1]}(x)$. What is the expected proportion of surveys returned in any given quarter?

[7] The argument can still be applied to cases where there is a finite number of discontinuities.

**Answer**: By definition,

$$\begin{aligned} \mathrm{E}X &= \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} x\left(3x^2 I_{[0,1]}(x)\right)dx \\ &= \int_0^1 3x^3 dx = \frac{3x^4}{4}\Big|_0^1 = .75. \end{aligned}$$

This is represented diagrammatically in Figure 3.5. □

In applications, the following result can often be a useful sufficient condition for the existence of the expected value in either the discrete or continuous case.

**Theorem 3.1** (Existence of E$X$ for Bounded $R(X)$) *If* $|x| < c \; \forall \; x \in R(X)$, *for some choice of* $c \in (0,\infty)$, *then* E$X$ *exists.*

**Proof** By the assumption of the theorem, $f(x) = 0 \; \forall \; x$ such that $|x| \geq c$.

Discrete:

$$\sum_{x \in R(X)} |x| f(x) < \sum_{x \in R(X)} cf(x) = c \sum_{x \in R(X)} f(x) = c < \infty,$$

so that E$X$ exists.

Continuous:

$$\int_{-\infty}^{\infty} |x| f(x) dx < \int_{-\infty}^{\infty} cf(x) dx = c \int_{-\infty}^{\infty} f(x) dx = c < \infty,$$

so that E$X$ exists. ■

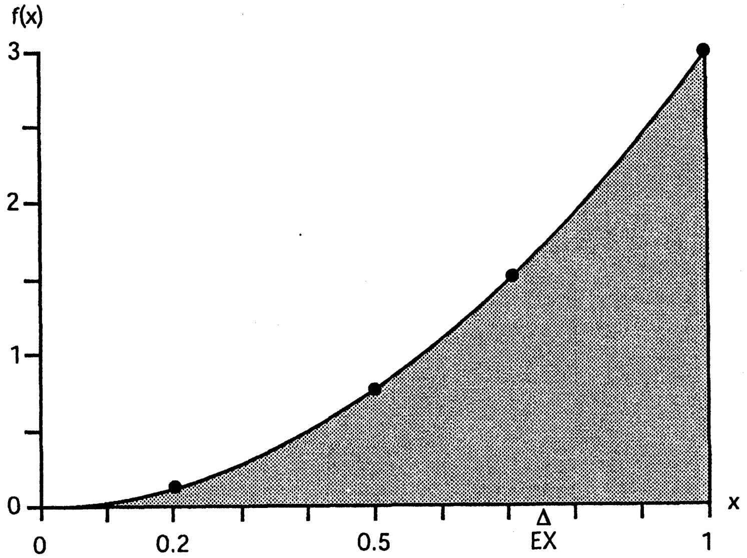


**Figure 3-5**
$f(x) = 3x^2 I_{[0,1]}(x)$.

The theorem indicates that expectations exist for any random variable whose outcomes are bounded in absolute value.

In order to economize on notation, we henceforth assume that whenever a discrete density function is involved in a summation expression, it is tacitly understood that summation is only over points for which $f(x) > 0$. Thus, we will generally abbreviate expressions such as $\sum_{\substack{x\in A\\ f(x)>0}} f(x)$ to simply $\sum_{x\in A} f(x)$, the condition $f(x) > 0$ always being understood.

## 3.2 Expectation of a Function of Random Variables

Many cases arise in practice where we are interested in the expectation of a function of a random variable rather than the expectation of a random variable itself. For example, the profit on a stock investment will be a *function* of the difference between the per share buying and selling prices of the stock, and the net return on an advertising campaign is a *function* of consumer buying response to the campaign—both the stock selling price and consumer buying response can be viewed as random variables. How might E$Y$ be determined when $y = g(x)$, $x \in R(X)$, and $X$ has density function $f(x)$? By definition, if we know the density of $Y$, $h(y)$, then

$$\text{(discrete)} \qquad \mathrm{E}Y = \sum_{y\in R(Y)} yh(y),$$

$$\text{(continuous)} \qquad \mathrm{E}Y = \int_{-\infty}^{\infty} yh(y)dy.$$

To use the expectation definition directly, we would need to establish the density of $Y$. This can be done in principle by exploiting the functional relationship between $y$ and $x$, given knowledge of the density $f(x)$, but often, finding $h(y)$ can be quite difficult (we will examine methods for deriving such densities in Chapter 6). Fortunately, we need not derive the density function of $y$ to obtain E$Y$.

Since $Y$ is defined via a composition of the functions $g$ and $X$, and since the domain of $X$ is conceptually the sample space $S$, then $Y$ is defined on the elements of $S$ via the composition, i.e., an outcome of $y$ can be viewed as being given by $y = g(X(w))$ for $w \in S$, so that $y : S \rightarrow R$. This implies that the range of $Y$ and probabilities of events for $Y$ can be represented alternatively as

$$R(Y) = \{y : y = g(x), x \in R(X)\} = \{y : y = g(X(w)), w \in S\}$$

and

$$P_Y(A) = P_X(\{x : g(x) \in A, x \in R(X)\}) = P(\{w : g(X(w)) \in A, w \in S\}).$$

Therefore, we can concentrate our attention on the $g$ function component of the composition, which has a real-valued domain $R(X)$, and conceptualize the outcomes of $Y$ as being generated by $y = g(x)$ for $x \in R(X)$, where $y : R(X) \rightarrow R$. In so doing, we lose no information concerning the possible outcomes of $Y$ or the probabilities of events for $Y$,[8] and we gain the convenience of being able to ignore the original probability space $\{S, \Upsilon, P\}$ and deal exclusively with a real-valued domain for the function $Y$. We will generally focus on this latter interpretation of the function $Y$ in our subsequent study, unless we make explicit reference to the domain of $Y$ as being $S$.

We now present a theorem identifying a straightforward approach for obtaining the expectation of $Y = g(X)$ by using density weightings applied to the outcomes of $X$.

**Theorem 3.2** *Let $X$ be a random variable having density function $f(x)$. Then the expectation of $Y = g(X)$ is given by*[9]

$$\text{(discrete)} \qquad \mathrm{E}g(X) = \sum_{x \in R(X)} g(x)f(x),$$

$$\text{(continuous)} \qquad \mathrm{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

**Proof** (Discrete case) Let $Y = g(X)$. The density function for the random variable $Y$ can be represented by

$$h(y) = P_Y(y) = P_X(\{x : g(x) = y, x \in R(X)\}) = \sum_{\{x:g(x)=y\}} f(x).$$

That is, the probability of the outcome $y$ is equal to the probability of the equivalent event $\{x : g(x) = y\}$, which is the inverse image of $y$. Then

$$\mathrm{E}g(X) = \mathrm{E}Y = \sum_{y \in R(Y)} yh(y) = \sum_{y \in R(Y)} y \sum_{\{x:g(x)=y\}} f(x)$$

$$= \sum_{y \in R(Y)} \sum_{\{x:g(x)=y\}} g(x)f(x) = \sum_{x \in R(X)} g(x)f(x),$$

where the next to last expression is true, since $g(x) = y$ for all $x \in \{x : g(x) = y\}$, and the last expression is true since $\sum_{y \in R(Y)} \sum_{\{x:g(x)=y\}} g(x)f(x)$ is equivalent to summing over all $x \in R(X)$ because the collection of all $y \in R(Y)$ (the outer sum) is the set $R(Y) = \{y : y = g(x), x \in R(X)\}$.

(Continuous case) To prove the theorem for the continuous case, we first need to establish the following lemma.

**Lemma 3.1**

For any continuous random variable $Y$, the expectation of $Y$, if it exists, can be written as

$$\mathrm{E}Y = \int_0^{\infty} P(y > z)dz - \int_0^{\infty} P(y \leq -z)dz.$$

[8] This presumes that the probability space $\{R(X), \Upsilon_X, P_X\}$ is known.

[9] It is tacitly assumed that the sum and integral are absolutely convergent for the expectation to exist.

**Proof of Lemma** Let $h(y)$ be the density function of $Y$. Then $P(y > z) = \int_z^\infty h(y)dy$, so that

$$\int_0^\infty P(y > z)dz = \int_0^\infty \int_z^\infty h(y)dydz = \int_0^\infty \left[\int_0^y dz\right] h(y)dy = \int_0^\infty yh(y)dy,$$

where the second equality was simply the result of changing the order of integration (note that the inner range of integration is a function of the outer range of integration, and the *same* set of $(y, z)$ points are being integrated over).

Similarly, $P(y \leq -z) = \int_{-\infty}^{-z} h(y)dy$, so that

$$\int_0^\infty P(y \leq -z)dz = \int_0^\infty \int_{-\infty}^{-z} h(y)dydz = \int_{-\infty}^0 \left[\int_0^{-y} dz\right] h(y)dy = -\int_{-\infty}^0 yh(y)dy.$$

Therefore,

$$\int_0^\infty P(y > z)dz - \int_0^\infty P(y \leq -z)dz = \int_0^\infty yh(y)dy + \int_{-\infty}^0 yh(y)dy = \mathrm{E}Y. \qquad \blacksquare$$

Note that the lemma (integrals and all) also applies to discrete random variables.[10]

Using the lemma, we have

$$\begin{aligned}
\mathrm{E}(g(X)) &= \int_0^\infty P(g(x) > z)dz - \int_0^\infty P(g(x) \leq -z)dz \\
&= \int_0^\infty \int_{\{x:g(x)>z\}} f(x)dxdz - \int_0^\infty \int_{\{x:g(x)\leq -z\}} f(x)dxdz \\
&= \int_{\{x:g(x)>0\}} \left[\int_0^{g(x)} dz\right] f(x)dx - \int_{\{x:g(x)\leq 0\}} \left[\int_0^{-g(x)} dz\right] f(x)dx \\
&= \int_{\{x:g(x)>0\}} g(x)f(x)dx + \int_{\{x:g(x)\leq 0\}} g(x)f(x)dx. \\
&= \int_{-\infty}^\infty g(x)f(x)dx. \qquad \blacksquare
\end{aligned}$$

**Example 3.5** Let the daily profit function of a firm be given by $\Pi(X) = pq(X) - rX$, where $X$ is a random variable whose outcome represents the daily quantity of a highly perishable agricultural commodity delivered to the firm for processing, measured in hundredweights (100 lb. units), $p = 5$ is the price of the processed product per pound, $r = 2$ is the cost of the raw agricultural commodity per pound, and $q(x) = x^{.9}$ is the production function indicating the relationship between raw and finished product measured in hundredweights. Let the density function of $X$ be $f(x) = \frac{1+2x}{110} I_{[0,10]}(x)$. What is the expected value of daily profit?

[10] See P. Billingsley (1986), *Probability and Measure,* 2nd ed. New York: John Wiley, pp. 73–74, for the method of proof in the discrete case.

**Answer**: A direct application of Theorem 3.2 yields

$$\mathrm{E}\Pi(X) = \int_{-\infty}^{\infty} \Pi(x)f(x)dx = \int_{0}^{10} (5x^{.9} - 2x)\left(\frac{1+2x}{110}\right) dx = 13.77.$$

Since quantities are measured in hundredweights, this means that the expected profit is \$1,377 per day. □

**Example 3.6** Your company manufactures a special 1/4-inch hexagonal bolt for the Defense Department. For the bolt to be usable in its intended application, the bolt must be manufactured within a 1 percent tolerance of the 1/4-inch specification. As part of your quality assurance program, each bolt is inspected by a laser measuring device that is 100 percent effective in detecting bolts that are not within the 1 percent tolerance. Bolts not meeting the tolerance are discarded. The actual size of a bolt manufactured on your assembly line is represented by a random variable, $X$, having a probability density $f(x) = (.006)^{-1}I_{[.247,.253]}(x)$, where $x$ is measured in inches. If your profit per bolt sold is \$.01, and if a discarded bolt costs your company \$.03, what is your expected profit per bolt manufactured?

**Answer**: We define a discrete random variable whose outcome represents whether a bolt provides the company with a \$.01 profit or a \$.03 loss. Specifically,

$$Y = g(X) = .01(I_{[.2475,.2525]}(X)) - .03(1 - I_{[.2475,.2525]}(X))$$

is the function of $X$ that we seek, where $y = .01$ if $x \in [.2475, .2525]$ (i.e., the bolt is within tolerance) and $y = -.03$ otherwise. Then

$$\begin{aligned} \mathrm{E}Y = \mathrm{E}g(X) &= \int_{.247}^{.253} g(x)f(x)dx \\ &= .01P(.2475 \leq x \leq .2525) - .03[1 - P(.2475 \leq x \leq .2525)] \\ &= .01(.83\underline{3}) - .03(.16\underline{6}) = .0033. \end{aligned}$$

□

The reader should note that in the preceding example, while $X$ was a *continuous* random variable, $Y = g(X)$ is a *discrete* random variable. Whether Y=g(X) is discrete or continuous depends on the nature of the function $g$ and whether $X$ is discrete or continuous. The reader should convince herself that if $X$ is discrete, then $Y$ must be discrete, but if $X$ is continuous, then $Y$ can be continuous or discrete (or mixed discrete-continuous).

Upon close examination of Ex. 3.6, the reader may have noticed that the expectation of an indicator function equals the probability of the set being indicated. It follows that probabilities can be represented as expectations.

**Theorem 3.3** *Let $X$ be a random variable with density function $f(x)$, and suppose $A$ is an event for $X$. Then $\mathrm{E}(I_A(X)) = P_X(A)$.*

**Proof** (Discrete) By definition,

$$\mathrm{E}(I_A(X)) = \sum_{x \in R(X)} I_A(x)f(x) = \sum_{x \in A} f(x) = P_X(A).$$

(Continuous) By definition,

$$\mathrm{E}I_A(X) = \int_{-\infty}^{\infty} I_A(x)f(x)dx = \int_{x \in A} f(x)dx = P_X(A).$$ ■

It should be noted that the existence of E$X$ does *not* imply that E$g(X)$ exists, as the following example illustrates.

**Example 3.7** Let $X$ be a random variable with density function $f(x) = (1/2)I_{\{0,1\}}(x)$. Then E$X = 0 \cdot 1/2 + 1 \cdot 1/2 = 1/2$. Define a new random variable $Y = g(X) = X^{-1}$. Since $|1/0|(1/2) + |1/1|(1/2) \not< \infty$, E$g(X)$ does not exist. □

The preceding example also illustrates that, in general, E$g(X) \neq g($E$X)$. Referring to Ex. 3.7, note that since E$X = 1/2$, $g($E$X) = ($E$X)^{-1} = 2$, which does *not* equal E$g(X)$ because E$g(X)$ does not exist! In the special case where the function $g$ is either concave or convex,[11] there is a definite relationship between E$g(X)$ and $g($E$X)$, as indicated by the following theorem.

**Theorem 3.4** (Jensen's Inequality) *Let $X$ be a random variable with expectation* E$X$, *and let $g$ be a continuous function on an open interval $I$ containing $R(X)$. Then*

**a.** E$g(X) \geq g($E$X)$ *if $g$ is convex on $I$, and* E$g(X) > g($E$X)$ *if $g$ is strictly convex on $I$ and $X$ is not degenerate;*[12]

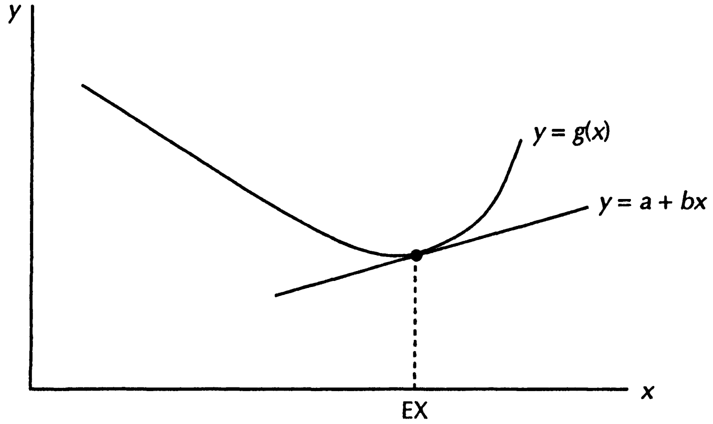


**Figure 3-6**
Convex function $g$.

[11] A continuous function, $g$, defined on a set $D$ is called concave if $\forall\, x \in D$, $\exists$ a line going through the point $(x, g(x))$ that lies on or above the graph of $g$. The function is convex if $\forall\, x$, $\exists$ a line going through the point $(x, g(x))$ that lies on or below the graph of $g$ (see Figure 3.6). The function is strictly convex or concave if the aforementioned line has only the point $(x, g(x))$ in common with the graph of $g$.

[12] A degenerate random variable is a random variable that has one outcome that is assigned a probability of 1. More will be said

**b.** $\mathrm{E}g(X) \leq g(\mathrm{E}X)$ *if g is concave on I, and* $\mathrm{E}g(X) < g(\mathrm{E}X)$ *if g is strictly concave on I and X is not degenerate.*

**Proof** We prove the result for the convex case. The proof of the concave case is analogous.

If $g$ is a convex function for $x \in I$, then there exists a line going through the point $\mathrm{E}X$, say $\ell(x) = a + bx$, such that $g(x) \geq \ell(x) = a + bx \; \forall \; x \in I$ and $g(\mathrm{E}X) = a + b\mathrm{E}X$ (see Figure 3.6). Now note that

$$\text{(discrete)} \qquad \mathrm{E}g(X) = \sum_{x \in R(X)} g(x)f(x) \geq \sum_{x \in R(X)} (a + bx)f(x) = a + b\mathrm{E}X = g(\mathrm{E}X),$$

$$\text{(continuous)} \qquad \mathrm{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)dx \geq \int_{-\infty}^{\infty} (a + bx)f(x)dx = a + b\mathrm{E}X = g(\mathrm{E}X),$$

since $g(x) \geq a + bx \; \forall \; x \in I$,[13] so that $\mathrm{E}g(X) \geq g(\mathrm{E}X)$.

If $g$ is strictly convex, then there exists a line going through the point $\mathrm{E}X$, say $\ell(x) = a + bx$, such that $g(x) > \ell(x) = a + bx \; \forall \; x \in I$ for which $x \neq \mathrm{E}X$, and $g(\mathrm{E}X) = a + b\mathrm{E}X$. Then, assuming that no element in $R(X)$ is assigned probability one (i.e., $X$ is not degenerate), the previous inequality results become strict, implying $\mathrm{E}g(X) > g(\mathrm{E}X)$ in either the discrete or continuous case. ■

**Example 3.8** Suppose that the yield per acre of a given agricultural crop under standard cultivation practices is represented by $Y = 5X - .1X^2$, where outcomes of $Y$ are measured in bushels, and $X$ represents the total rainfall during the growing season, measured in inches. If $\mathrm{E}X = 20$, can you place an upper bound on the expected yield per acre for this crop?

**Answer**: Yes. Note that $Y = 5X - .1X^2$ is a *concave* function, so that Jensen's inequality applies. Then $\mathrm{E}Y = \mathrm{E}g(X) \leq g(\mathrm{E}X) = 5\mathrm{E}X - .1(\mathrm{E}X)^2 = 60$ is an upper bound to the expected yield. In fact, the function is strictly concave, and so the inequality can be made strict (it is reasonable to assume that rainfall is not a degenerate random variable). □

### Expectation Properties

There are a number of properties of the expectation operation that follow directly from its definition. We prove the validity of the propositions using the case of a continuous random variable. The reader is asked to provide analogous demonstrations in the discrete cases.

**Theorem 3.5** *If c is a constant, then* $\mathrm{E}(c) = c$.

about degenerate random variables in Section 3.6.

[13] Recall the integral inequality that if $h(x) \geq t(x) \; \forall \; x \in (a, b)$, then $\int_a^b h(x)dx \geq \int_a^b t(x)dx$. Strict inequality holds if $h(x) > t(x) \; \forall \; x \in (a, b)$. The result also holds for $a = -\infty$ and/or $b = \infty$.

**Proof** Let $g(X) = c$. Then, by Theorem 3.2,

$$\mathrm{E}(c) = \int_{-\infty}^{\infty} cf(x)dx = c\underbrace{\int_{-\infty}^{\infty} f(x)dx}_{1} = c. \qquad \blacksquare$$

In words, "the expected value of a constant is the constant itself."

**Theorem 3.6** *If c is a constant, then* $\mathrm{E}(cX) = c\mathrm{E}X$.

**Proof** Let $g(X) = cX$. Then, by Theorem 3.2,

$$\mathrm{E}(cX) = \int_{-\infty}^{\infty} cxf(x)dx = c\underbrace{\int_{-\infty}^{\infty} xf(x)dx}_{\mathrm{E}X} = c\mathrm{E}X. \qquad \blacksquare$$

In words, "the expected value of a constant times a random variable is the constant times the expected value of the random variable."

**Theorem 3.7** $\mathrm{E}\sum_{i=1}^{k} g_i(X) = \sum_{i=1}^{k} \mathrm{E}g_i(X)$.

**Proof** Let $g(X) = \sum_{i=1}^{k} g_i(X)$. Then, by Theorem 3.2,

$$\mathrm{E}\sum_{i=1}^{k} g_i(X) = \int_{-\infty}^{\infty}\left[\sum_{i=1}^{k} g_i(x)\right]f(x)dx = \sum_{i=1}^{k}\int_{-\infty}^{\infty} g_i(x)f(x)dx = \sum_{i=1}^{k}\mathrm{E}g_i(X). \qquad \blacksquare$$

In words, "the expectation of a sum is the sum of the expectations" regarding $k$ functions of the random variable $X$. A useful corollary to Theorem 3.7 concerns the expectation of a linear function of $X$.

**Corollary 3.1** *Let* $Y = a + bX$ *for real constants a and b, and let* $\mathrm{E}X$ *exist. Then* $\mathrm{E}Y = a + b\mathrm{E}X$.

**Proof** This follows directly from Theorem 3.7 by defining $g_1(X) = a$, $g_2(X) = bX$, and then applying Theorems 3.5 and 3.6. ■

### Multivariate Extensions

The concept of an expectation of a function of a random variable is generalizable to a function of a multivariate random variable as indicated in the following theorem. The proof is based on an extension of the proof of Theorem 3.2 and is omitted.[14]

[14] See Steven F. Arnold (1990), *Mathematical Statistics*. Englewood Cliffs, NJ: Prentice Hall, pp. 92, 98.

**Theorem 3.8** (Expectation of a Function of a Multivariate Random Variable) *Let $(X_1,\ldots,X_n)$ be a multivariate random variable with joint density function $f(x_1,\ldots,x_n)$. Then the expectation of $Y = g(X_1,\ldots,X_n)$ is given by*[15]

$$\text{(discrete)} \qquad \mathrm{E}Y = \sum_{(x_1,\ldots,x_n)\in R(X)} \cdots \sum g(x_1,\ldots,x_n)f(x_1,\ldots,x_n),$$

$$\text{(continuous)} \qquad \mathrm{E}Y = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1,\ldots,x_n)f(x_1,\ldots,x_n)dx_1\cdots dx_n.$$

We remind the reader that since $f(x_1,\ldots,x_n) = 0 \,\forall\, (x_1,\ldots,x_n) \notin R(X)$, one could also sum over the points $(x_1,\ldots,x_n) \in \times_{i=1}^{n} R(X_i)$ to define $\mathrm{E}Y$ in the discrete case.

**Example 3.9** Let the bivariate random variable $(X_1, X_2)$ represent the proportions of operating capacity at which two electricity generating plants operate on a given spring day in an East Coast power grid. Assume the joint density of $(X_1, X_2)$ is given by

$$f(x_1, x_2) = 6x_1x_2^2 I_{[0,1]}(x_1)I_{[0,1]}(x_2).$$

What is the expected average proportion of operating capacity at which the two plants operate?

**Answer**: Define the average proportion of operating capacity via the function $g(X_1,X_2) = .5(X_1 + X_2)$. By Theorem 3.8,

$$\begin{aligned}\mathrm{E}g(X_1, X_2) &= \int_0^1\int_0^1 3x_1^2x_2^2dx_1dx_2 + \int_0^1\int_0^1 3x_1x_2^3dx_1dx_2 \\ &= \frac{1}{3} + \frac{3}{8} = \frac{17}{24} = .7083.\end{aligned}$$

□

The expectation property in Theorem 3.7 concerning the sum of functions of a random variable $X$ can also be extended to the sum of functions of a multivariate random variable, as the following theorem indicates.

**Theorem 3.9** $\mathrm{E}\sum_{i=1}^{k} g_i(X_1,\ldots,X_n) = \sum_{i=1}^{k} \mathrm{E}g_i(X_1,\ldots,X_n)$.

**Proof** (Continuous) Let $g(X_1,\ldots,X_n) = \sum_{i=1}^{k} g_i(X_1,\ldots,X_n)$. Then by Theorem 3.8, $\mathrm{E}g(X_1,\ldots,X_n)$ is given by

$$\begin{aligned}\mathrm{E}\sum_{i=1}^{k} g_i(X_1,\ldots,X_n) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sum_{i=1}^{k} g_i(x_1,\ldots,x_n)\right)f(x_1,\ldots,x_n)dx_1\ldots dx_n \\ &= \sum_{i=1}^{k}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g_i(x_1,\ldots,x_n)f(x_1,\ldots,x_n)dx_1\ldots dx_n\end{aligned}$$

[15] It is tacitly assumed that the sum and integral are absolutely convergent for the expectation to exist.

$$= \sum_{i=1}^{k} \mathrm{E}g_i(X_1, \ldots, X_n).$$

The proof in the discrete case is left to the reader. ■

A useful corollary to Theorem 3.9 involving the sum of random variables themselves is given as follows:

**Corollary 3.2** $\mathrm{E}\sum_{i=1}^{k} X_i = \sum_{i=1}^{k} \mathrm{E}X_i$.

**Proof** This is an application of Theorem 3.9 with $g_i(X_1, \ldots, X_n) = X_i, i = 1, \ldots, n$. ■

In words, "the expectation of a sum is equal to the sum of the expectations" regarding the $n$ random variables $X_1, \ldots, X_n$.

In the case where the random variables $(X_1, \ldots, X_n)$ are *independent*, we can prove that "the expectation of a product is the product of the expectations."

**Theorem 3.10** *Let $(X_1, \ldots, X_n)$ be independent random variables. Then*

$$\mathrm{E}\prod_{i=1}^{n} X_i = \prod_{i=1}^{n} \mathrm{E}X_i.$$

**Proof** (Continuous case) Letting $g(X_1, \ldots, X_n) = \prod_{i=1}^{n} X_i$ in Theorem 3.8, we have

$$\mathrm{E}\prod_{i=1}^{n} X_i = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{n} x_i f(x_1, \ldots, x_n) dx_1 \cdots dx_n$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{n} x_i \prod_{j=1}^{n} f_j(x_j) dx_1 \ldots dx_n \qquad \text{(by independence)}$$

$$= \prod_{i=1}^{n} \int_{-\infty}^{\infty} x_i f_i(x_i) dx_i = \prod_{i=1}^{n} \mathrm{E}X_i.$$

The discrete case is analogous and is left to the reader. ■

Later in our study we will find it necessary to take expectations of a vector or matrix of random variables. The following definition describes what is involved in such an operation.

**Definition 3.3** ***Expectation of a matrix of random variables***

Let **W** be an $n \times k$ matrix of random variables whose $(i, j)$th element is $W_{ij}$. Then E**W**, the expectation of the matrix **W**, is the matrix of expectations of the elements of **W**, where the $(i, j)$th element of E**W** is equal to $\mathrm{E}W_{ij}$.

If we let $k = 1$ in the above definition, we have that the expectation of a vector is the vector of expectations, i.e.,

$$\mathrm{E}\mathbf{W} = E\begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_n \end{bmatrix} = \begin{bmatrix} \mathrm{E}W_1 \\ \mathrm{E}W_2 \\ \vdots \\ \mathrm{E}W_n \end{bmatrix}.$$

In general,

$$\underset{(\mathbf{n}\times\mathbf{k})}{\mathrm{E}\mathbf{W}} = \begin{bmatrix} \mathrm{E}W_{11} & \cdots & \mathrm{E}W_{1k} \\ \vdots & \ddots & \vdots \\ \mathrm{E}W_{n1} & \cdots & \mathrm{E}W_{nk} \end{bmatrix},$$

i.e., "the expectation of a matrix is the matrix of expectations."

Having introduced the concept of the expectation of a vector, we note that a **multivariate Jensen's inequality** (Theorem 3.4) holds true for multivariate random variables. In fact, the appropriate extension is made by letting $X$ denote an $n \times 1$ random vector and $I$ represent an open rectangle in the statement of Theorem 3.4. The reader is asked to prove the multivariate version of Theorem 3.4 (replace the line $\ell(x) = a + bx$ with the hyperplane $\ell(x) = a + \sum_{i=1}^{n} b_i x_i$ in the proof).

## 3.3 Conditional Expectation

Up to this point, expectations of random variables and functions of random variables have been taken in the unconditional sense, i.e., we have not assumed that information was available relating to the occurrence of an event for a subset of the random variables $(X_1, \ldots, X_n)$. When information is given concerning the occurrence of events for a subset of the random variables $(X_1, \ldots, X_n)$, the *conditional* expectation of a random variable becomes a relevant concept.

There is a myriad of situations that arise in business and economic settings in which the concept of conditional expectation is relevant. For example, the expected number of housing starts calculated for planning purposes by building supply manufacturers would depend on the given level of mortgage interest assumed, or the expected sales tax revenue accruing to state government would be a function of whatever reduced level of employment was assumed due to the downsizing of a major industry in the state. More generally, we will see that conditional expectation is at the heart of *regression analysis*, whereby one attempts to explain the expected value of one random variable as a function of the values of other related random variables, e.g., the expected yield/acre of an agricultural crop is conditional on the level of rainfall, temperature, sunshine, and degree of weed and pest infestation.

The difference between unconditional and conditional expectation is that the unconditional density function is used to weight outcomes in the former case, while the conditional density function supplies the weights in the latter case. The conditional expectation of a function of a random variable is defined in the bivariate case as follows.

**Definition 3.4** ***Conditional expectation; bivariate.***

Let $X$ and $Y$ be random variables with joint density function $f(x, y)$. Let the conditional density of $Y$, given $x \in B$, be $f(y \mid x \in B)$. Let $g(Y)$ be a real-valued function of $Y$. Then the conditional expectation of $g(Y)$, given $x \in B$, is defined as

$$\text{(discrete)} \quad \mathrm{E}(g(Y) \mid x \in B) = \sum_{y \in R(Y)} g(y) f(y \mid x \in B),$$

$$\text{(continuous)} \quad \mathrm{E}(g(Y) \mid x \in B) = \int_{-\infty}^{\infty} g(y) f(y \mid x \in B) dy.$$

Note in the special case where $g(Y) = Y$, we have by Definition 3.4 that $E(Y \mid x \in B) = \int_{-\infty}^{\infty} y f(y \mid x \in B) dy$ in the continuous case, and $E(Y \mid x \in B) = \sum_{y \in R(Y)} y f(y \mid x \in B)$ in the discrete case.

**Example 3.10** Let the bivariate random variable $(X, Y)$ represent the per dollar return on two investment projects. Let the joint density of $(X, Y)$ be

$$f(x, y) = \tfrac{1}{96}(x^2 + 2xy + 2y^2) I_{[0,4]}(x) I_{[0,2]}(y).$$

What is the conditional expectation of the per dollar return on the second project, given that the per dollar return on the first project is $x = 1$?

**Answer**: To answer the question, we first need to establish the conditional density $f(y \mid x = 1)$. This in turn requires knowledge of the marginal density of $X$, which we find as

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \frac{1}{96} \int_0^2 (x^2 + 2xy + 2y^2) I_{[0,4]}(x) dy$$
$$= \left(\tfrac{1}{48}x^2 + \tfrac{1}{24}x + \tfrac{1}{18}\right) I_{[0,4]}(x).$$

Then

$$f(y \mid x = 1) = \frac{f(1, y)}{f_X(1)} = \frac{\frac{1}{96}(1 + 2y + 2y^2) I_{[0,2]}(y)}{\frac{17}{144}} = [.088235 + .176471(y + y^2)] I_{[0,2]}(y).$$

Finally, by Definition 3.4,

$$\mathrm{E}(Y \mid x = 1) = \int_{-\infty}^{\infty} y f(y \mid x = 1) dy = \int_0^2 (.088235y + .176471(y^2 + y^3)) \, dy$$
$$= \frac{.088235y^2}{2} + .176471 \left[\frac{y^3}{3} + \frac{y^4}{4}\right] \Bigg|_0^2 = 1.3529. \qquad \square$$

It is important to note that *all* of the properties of expectations derived in the previous two sections apply equally well to conditional expectations. This follows from the fact that the operations of taking an expectation and a conditional expectation are precisely the same once a PDF has been derived, and the genesis of the PDF is irrelevant to the expectation properties derived heretofore (note that the origin of a PDF in Sections 3.1 and 3.2 was never an issue).

Rather than specifying a particular elementary event for the outcome of $X$ when defining a conditional expectation of $g(Y)$, we might conceptualize leaving the elementary event for $X$ *unspecified*, and express the conditional expectation of $g(Y)$ as a *function of* $x$. Let $\eta(x) = E(g(Y) \mid x)$ denote the function of $x$ whose value when $x = b$ is $E(g(Y) \mid x = b)$. Then, by definition, we can interpret $\eta(X) = E(g(Y) \mid X)$ as a random variable. If we take $\mathrm{E}\eta(X) = \mathrm{E}(\mathrm{E}(g(Y) \mid X))$, we obtain the *unconditional* expectation of $g(Y)$.

**Theorem 3.11** (Double Expectation Theorem) $\mathrm{E}(\mathrm{E}(g(Y) \mid X)) = \mathrm{E}g(Y)$.

**Proof** (Continuous case) Let $f_X(x)$ be the marginal density of $X$ and $f(x, y)$ be the joint density of $X$ and $Y$. Then

$$\eta(x) = \mathrm{E}(g(Y) \mid x) = \int_{-\infty}^{\infty} g(y)\frac{f(x,y)}{f_X(x)}\,dy$$

and

$$\mathrm{E}(\eta(X)) = \mathrm{E}(\mathrm{E}(g(Y) \mid X)) = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} g(y)\frac{f(x,y)}{f_X(x)}dy\right] f_X(x)dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y)f(x,y)dydx = \mathrm{E}g(Y).$$

The proof in the discrete case is left to the reader. ■

**Example 3.11** Suppose that the expectation of market supply for some commodity, given price $p$, is represented by $\mathrm{E}(Q \mid p) = 3p+7$ and $\mathrm{E}P = 2$. Then by the double expectation theorem, the unconditional expectation of market supply is given by

$$\mathrm{E}(\mathrm{E}Q \mid P)) = \mathrm{E}(3P + 7) = 3\mathrm{E}P + \mathrm{E}7 = 13. \qquad \square$$

In cases where one is conditioning on an elementary event $x = b$, there are useful generalizations of Definition 3.4 and Theorem 3.11, which are referred to as the *substitution theorem* and the *generalized double expectation theorem*, respectively.

**Theorem 3.12** (Substitution Theorem) $\mathrm{E}(g(X,Y) \mid x = b) = \mathrm{E}(g(b,Y) \mid x = b)$.

**Proof** (Discrete case) Let $z = g(x, y)$, and note that the PDF of $Z$, conditional on $x = b$, can be defined as

$$h(z \mid x = b) = P(g(x,y) = z \mid x = b) = P(g(x,y) = z, x = b)/P(x = b)$$

$$= \sum_{\{y:g(b,y)=z\}} f(b,y)/f_X(b) = \sum_{\{y:g(b,y)=z\}} f(y \mid x = b).$$

It is evident that the set of $z$ values for which $h(z \mid x = b) > 0$ is given by $\Xi = \{z : z = g(b, y), y \text{ is such that } f(y \mid x = b) > 0\}$. Then

$$\begin{aligned} \mathrm{E}(g(X, Y) \mid x = b) = \mathrm{E}(z \mid x = b) &= \sum_{z\in\Xi} z h(z \mid x = b) \\ &= \sum_{z\in\Xi} z \sum_{\{y: g(b,y)=z\}} f(y \mid x = b) \\ &= \sum_{y\in R(Y)} g(b, y) f(y \mid x = b) = \mathrm{E}(g(b, Y) \mid x = b). \end{aligned}$$

(Continuous case) See A. F. Karr (1993), *Probability*. New York: Springer-Verlag, p. 230. ■

The *substitution* theorem indicates that when taking the expectation of $g(X, Y)$ conditional on $x = b$, one can substitute the constant $b$ for $X$ as $g(b, Y)$ and then take the conditional expectation with respect to the random variable $Y$. The random variable $X$ essentially acts as a constant in $g(X, Y)$ under the condition $x = b$.

**Theorem 3.13** (Generalized Double Expectation Theorem) $\mathrm{EE}(g(X,Y) \mid X) = \mathrm{E}g(X,Y)$.

**Proof** Using the substitution theorem but leaving the elementary event for $X$ unspecified in order to express the conditional expectation as a function of the elementary event $x$ obtains (continuous case and discrete case are analogous)

$$\begin{aligned} \eta(x) &= \mathrm{E}(g(X, Y) \mid x) = \mathrm{E}(g(x, Y) \mid x) \\ &= \int_{-\infty}^{\infty} g(x, y) f(y \mid x) dy. \end{aligned}$$

Then

$$\begin{aligned} \mathrm{E}(\eta(X)) &= \mathrm{EE}(g(X, Y) \mid X) \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} g(x, y) f(y \mid x) dy\right] f_X(x) dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dy\, dx = \mathrm{E}g(X, Y). \end{aligned}$$ ■

Thus, taking the expectation of the conditional expectation of $g(X, Y)$ given $x$ yields the unconditional expectation of $g(X, Y)$.

**Example 3.12** Let $(X, Y)$ represent the per dollar return on the two investment projects of Ex. 3.10. Assume \$1,000 is invested in each project. What is the expected return on the portfolio, given that the per dollar return on the first project is $x = 1$?

**Answer**: The return on the portfolio can be represented as $z = g(x, y) = 1{,}000x + 1{,}000y$. The substitution theorem allows the conditional expectation to be defined as

$$\begin{aligned}\mathrm{E}(Z \mid x=1) &= \mathrm{E}(1000X + 1000Y \mid x=1) = \mathrm{E}(1000 + 1000Y \mid x=1)\\ &= \int_0^2 (1000 + 1000y)f(y \mid x=1)dy\\ &= 1000 + 1000\int_0^2 yf(y \mid x=1)dy = 2352.9. \qquad \square\end{aligned}$$

**Example 3.13** Given the representation of expected market supply in Ex. 3.11, we know by the substitution theorem that the expected dollar sales of the commodity, expressed as a function of price, is $\mathrm{E}(pQ \mid p) = 3p^2 + 7p$. Suppose $\mathrm{E}P^2 = 8$. Then, using Theorem 3.13, the (unconditional) expectation of dollar sales is given by $\mathrm{EE}(PQ \mid P) = \mathrm{E}(3P^2 + 7P) = 3\mathrm{E}P^2 + 7\mathrm{E}P = 38$. □

### Regression Function

In the special case where $g(Y) = Y$, the conditional expectation of $Y$ expressed as a function of x, i.e., $\mathrm{E}(Y \mid x)$, is called **the regression function of $Y$ on $X$**. The regression function depicts the functional relationship between the conditional expectation of $Y$ and the potential values of $X$ on which the expectation might be conditioned. In the continuous case, the graph of the function is generally a curve in the plane, in which case $E(Y \mid x)$ is often referred to as **the regression curve of $Y$ on $X$**.

**Example 3.14** Refer to the investment return example, Ex. 3.10, and rather than calculate $E(Y \mid x = 1)$, we calculate $E(Y \mid x)$, the regression function of $Y$ on $X$. Letting $f(y \mid x)$ denote the conditional density function of $y$ expressed as a function of x, we have $f(y \mid x) = f(x, y)/f_X(x)$. Then

$$\begin{aligned}\mathrm{E}(Y \mid x) &= \int_{-\infty}^{\infty} y\frac{f(x,y)}{f_X(x)}dy = \int_0^2 \frac{y(x^2 + 2xy + 2y^2)I_{[0,4]}(x)}{\left(2x^2 + 4x + \frac{16}{3}\right)I_{[0,4]}(x)}dy\\ &= \left[\frac{2x^2 + \frac{16}{3}x + 8}{2x^2 + 4x + \frac{16}{3}}\right] \quad \text{for } x \in [0,4],\end{aligned}$$

and the regression function is undefined for $x \notin [0, 4]$. The regression function represents the expected per dollar return on project 2 as a function of the various potential conditioning values for the per dollar return on project 1. Note that the regression function is a nonlinear function of x. The reader can verify that at the point $x = 1$, $\mathrm{E}(Y \mid x) = 1.3529$ is the value of the regression function, as it should be given the answer to Ex. 3.10. The reader is encouraged to sketch the graph of the regression function over its domain of definition. □

The regression function has an important interpretation in terms of approximating one random variable by a function of other random variables. Examine the problem of choosing a function of $X$, say $h(X)$, whose outcomes are the minimum expected squared distance[16] from the outcome of $Y$. Assuming that $(X, Y)$ is a continuous bivariate random variable (the discrete case is analogous), we thus seek an $h(X)$ that minimizes

$$\begin{aligned} \mathrm{E}(Y - h(X))^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y - h(x))^2 f(x, y)dxdy \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} (y - h(x))^2 f(y \mid x)dy\right] f_X(x)dx. \end{aligned}$$

If $h(x)$ could be chosen so as to minimize the bracketed integral for *each* possible $x$, then it would follow that the double integral, and thus the expected squared distance between outcomes of $Y$ and $h(X)$, would be minimized.

The optimal choice of approximating function is given by $h(x) = \mathrm{E}(Y \mid x)$. To see why, note that the substitution theorem allows the preceding bracketed expression to be written as

$$\begin{aligned} \mathrm{E}([Y - h(x)]^2 \mid x) &= \mathrm{E}([Y - \mathrm{E}(Y \mid x) + \mathrm{E}(Y \mid x) - h(x)]^2 \mid x) \\ &= \mathrm{E}([Y - \mathrm{E}(Y \mid x)]^2 \mid x) + [\mathrm{E}(Y \mid x) - h(x)]^2, \end{aligned}$$

where the cross-product term is zero and has been eliminated because

$$\begin{aligned} &\mathrm{E}([Y - \mathrm{E}(Y \mid x)][\mathrm{E}(Y \mid x) - h(x)] \mid x) \\ &\quad = [\mathrm{E}(Y \mid x) - h(x)]\mathrm{E}(Y - \mathrm{E}(Y \mid x) \mid x) \qquad \text{(by the substitution theorem)} \\ &\quad = [\mathrm{E}(Y \mid x) - h(x)][\mathrm{E}(Y \mid x) - \mathrm{E}(Y \mid x)] = 0. \end{aligned}$$

It follows that the choice of $h(x)$ that minimizes $\mathrm{E}([Y - h(x)]^2 \mid x)$ is given by $h(x) = \mathrm{E}(Y \mid x)$, since any other choice results in $[\mathrm{E}(Y \mid x) - h(x)]^2 > 0$.

The preceding result suggests that if one is attempting to explain or predict the outcome of one random variable from knowledge of the outcome of another random variable, and if expected squared distance (also called mean square error—to be discussed in Chapter 7) is used as the measure of closeness between actual and predicted outcomes, then the best (closest) prediction is given by values of the regression function, or equivalently by the conditional expectation of the random variable of interest. For example, in Ex. 3.14, if one were attempting to predict the expected dollar return on project 2 in terms of the dollar return on project 1, the regression function $E(Y \mid x)$ presented in the example provides the predictions that minimize expected squared distance between outcomes of $Y$ and outcomes of $h(X)$. If $x = 1$, then the best prediction of $Y$'s outcome would be 1.3529.

### Conditional Expectation and Regression in the Multivariate Case

The definition of conditional expectation (Definition 3.4) and the theorems involving conditional expectation extend to the case where $Y$ and/or $X$ is mul-

[16] Recall that the distance between the points $a$ and $b$ is defined by $d(a, b) = |b - a|$, and thus the squared distance would be given by $d^2(a, b) = (b - a)^2$.

tivariate, in which case the reader can interpret $Y$ and $X$ as referring to random *vectors* and introduce multiple summation or integration notation appropriately when reading the definition and the theorems. Also, the notion of the *regression function of* $Y$ *on* $X$ extends straightforwardly to the case where $X$ is multivariate, in which case $\mathrm{E}(Y \mid x)$ is interpreted as a function of the *vector* $\mathbf{x}$ and would be defined by $\mathrm{E}(Y \mid x) = \int_{-\infty}^{\infty} yf(y \mid x_1,\ldots,x_n)dy$ in the continuous case, with summation replacing integration in the discrete case. An argument analogous to the bivariate case can be used to prove that $h(X) = E(Y \mid X)$ is the function of the multivariate $X$ that is the minimum expected squared distance from $Y$. Thus, the best approximation or prediction of $Y$ outcomes via a function of the outcome of the multivariate $X$ is provided by values of the regression function.

For convenience, we list below a number of general expectation results applied specifically to conditional expectations involving multivariate random variables.

**Definition 3.5** ***Conditional expectation (general)***

Let $(X_1,\ldots,X_n)$ and $(Y_1,\ldots,Y_m)$ be random vectors having a joint density function $f(x_1,\ldots,x_n,y_1,\ldots,y_m)$. Let $g(Y_1,\ldots,Y_m)$ be a real-valued function of $(Y_1,\ldots,Y_m)$. Then the conditional expectation of $g(Y_1,\ldots,Y_m)$, given $(x_1,\ldots,x_n) \in B$, is defined as

(discrete) $\mathrm{E}(g(Y_1,\ldots,Y_m) \mid (x_1,\ldots,x_n) \in B)$

$$= \sum_{(y_1,\ldots,y_m)\in R(Y)} \cdots \sum g(y_1,\ldots,y_m)f(y_1,\ldots,y_m \mid (x_1,\ldots,x_n) \in B),^{17}$$

(continuous) $\mathrm{E}(g(Y_1,\ldots,Y_m) \mid (x_1,\ldots,x_n) \in B)$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(y_1,\ldots,y_m)f(y_1,\ldots,y_m \mid (x_1,\ldots,x_n) \in B)dy_1 \cdots dy_m.$$

**Theorem 3.14** (Substitution Theorem: Multivariate)

$$\mathrm{E}(g(X_1,\ldots,X_n,Y_1,\ldots,Y_m) \mid \mathbf{x} = \mathbf{b}) = \mathrm{E}(g(b_1,\ldots,b_n,Y_1,\ldots,Y_m) \mid \mathbf{x} = \mathbf{b})$$

**Theorem 3.15** (Double Expectation Theorems: Multivariate)

$$\mathrm{E}(\mathrm{E}(g(Y_1,\ldots,Y_m) \mid X_1,\ldots,X_n)) = \mathrm{E}g(Y_1,\ldots,Y_m)$$

$$\mathrm{E}(\mathrm{E}(g(X_1,\ldots,X_n,Y_1,\ldots,Y_m) \mid X_1,\ldots,X_n)) = \mathrm{E}g(X_1,\ldots,X_n,Y_1,\ldots,Y_m)$$

**Theorem 3.16** $E(c \mid (x_1,\ldots,x_n) \in B) = c.$

**Theorem 3.17** $\mathrm{E}(cY \mid (x_1,\ldots x_n) \in B) = c\mathrm{E}(Y \mid (x_1,\ldots,x_n) \in B)$

[17] One can equivalently sum over the points $(y_1,\ldots,y_m) \in \times_{i=1}^{m} R(Y_i)$ in defining the expectation in the discrete case.

**Theorem 3.18**

$$E\left(\sum_{i=1}^{k} g_i(Y_1,\ldots,Y_m) \mid (x_1,\ldots,x_n) \in B\right) = \sum_{i=1}^{k} \mathrm{E}\left(g_i(Y_1,\ldots,Y_m) \mid (x_1,\ldots,x_n) \in B\right).$$

## 3.4 Moments of a Random Variable

The expectations of certain power functions of a random variable have uses as measures of central tendency, spread or dispersion, and skewness of the density function of the random variable, and also are important components of statistical inference procedures that we will study in later chapters. These special expectations are called **moments** of the random variable (or of the density function). There are two types of moments that we will be concerned with—**moments about the origin** and **moments about the mean.**

**Definition 3.6**
***r*th moment about the origin**

Let $X$ be a random variable with density function $f(x)$. Then the $r$th **moment of $X$ about the origin**, denoted by $\mu_r'$, is defined for integers $r \geq 0$ as

$$\text{(discrete)} \qquad \mu_r' = \mathrm{E}(X^r) = \sum_{x \in R(X)} x^r f(x),$$

$$\text{(continuous)} \qquad \mu_r' = \mathrm{E}(X^r) = \int_{-\infty}^{\infty} x^r f(x) dx.$$

Note that $\mu_0' = 1$ for any discrete or continuous random variable, since $\mu_0' = \mathrm{E}(X^0) = \mathrm{E}(1) = 1$. The first moment about the origin is simply the expectation of the random variable $X$, i.e., $\mu_1' = \mathrm{E}(X^1) = \mathrm{E}X$, a quantity that we examined at the beginning of our discussion of mathematical expectation. This *balancing point* of a density function, or the weighted average of the elements in the range of the random variable, will now be given a special name and symbol.

**Definition 3.7**
***Mean of a random variable (or mean of a density function)***

The first moment about the origin of a random variable, $X$, is called the **mean of the random variable $X$** (or mean of the density function of $X$) and will be denoted by the symbol $\mu$.

Thus, the first moment about the origin characterizes the central tendency of a density function. Measures of spread and skewness of a density function are given by certain moments about the mean.

**Definition 3.8**
***r*th central moment (or *r*th moment about the mean)**

Let $X$ be a random variable with density function $f(x)$. Then the $r$th **central moment of $X$** (or the $r$th moment of $X$ about the mean), denoted by $\mu_r$, is defined for integers $r \geq 0$ as

$$\text{(discrete)} \qquad \mu_r = \mathrm{E}(X - \mu)^r = \sum_{x \in R(X)} (x - \mu)^r f(x),$$

$$\text{(continuous)} \qquad \mu_r = \mathrm{E}(X - \mu)^r = \int_{-\infty}^{\infty} (x - \mu)^r f(x) dx.$$

Note that $\mu_0 = 1$ for any discrete or continuous random variable, since $\mu_0 = \mathrm{E}(X - \mu)^0 = \mathrm{E}(1) = 1$. Furthermore, $\mu_1 = 0$ for any discrete or continuous random variable for which $\mathrm{E}X$ exists, since $\mu_1 = \mathrm{E}(X - \mu)^1 = \mathrm{E}X - \mathrm{E}\mu = \mu - \mu = 0$. The second central moment is given a special name and symbol.

**Definition 3.9**
***Variance of a random variable (or variance of a density function)***

The second central moment, $\mathrm{E}(X - \mu)^2$, of a random variable, $X$, is called the **variance of the random variable *X*** (or the variance of the density function of $X$) and will be denoted by the symbol $\sigma^2$, or by $\mathrm{var}(X)$.

We will also have use for the following function of the variance of a random variable.

**Definition 3.10**
***Standard deviation of a random variable (or standard deviation of a density function)***

The nonnegative square root of the variance of a random variable, $X$, (i.e., $\sqrt{\sigma^2}$), is called the **standard deviation of the random variable *X*** (or standard deviation of the density function of $X$) and will be denoted by the symbol $\sigma$, or by $\mathrm{std}(X)$.

The variance (and thus also the standard deviation) of $X$ is a measure of dispersion or spread of the density function $f(x)$ around its balancing point (the mean of $X$). The larger the variance, the greater the spread or dispersion of the density about its mean. In the extreme case where the entire density is concentrated at the mean of $X$ and thus has no spread or dispersion, i.e., $f(x) = I_{\{\mu\}}(x)$, then $\mathrm{E}(X - \mu)^2 = 0$ and the variance (and standard deviation) is zero.

In order to examine the relationship between the spread of a density and the magnitude of the variance in more detail, we first present **Markov's inequality** (named after the Russian mathematician A. Markov), whence we introduce **Chebyshev's inequality** (named after the Russian mathematician P. L. Chebyshev) as a corollary.

**Theorem 3.19** (Markov's Inequality) *Let $X$ be a random variable with density function $f(x)$, and let $g$ be a nonnegative-valued function of $X$. Then $P(g(x) \geq a) \leq \mathrm{E}g(X)/a$ for any value $a > 0$.*

**Proof** (Continuous case and discrete case are analogous. The discrete case is left to reader.)

$$\begin{aligned}
\mathrm{E}g(X) &= \int_{-\infty}^{\infty} g(x)f(x)dx \\
&= \int_{\{x:g(x)\geq a\}} g(x)f(x)dx \\
&\quad + \underbrace{\int_{\{x:g(x)<a\}} g(x)f(x)dx}_{\geq 0} \\
&\geq \int_{\{x:g(x)\geq a\}} g(x)f(x)dx \\
&\geq \int_{\{x:g(x)\geq a\}} af(x)dx \qquad (\text{since } g(x) \geq a \text{ for all } x \in \{x : g(x) \geq a\} \\
&\geq a\int_{\{x:g(x)\geq a\}} f(x)dx = aP(g(x) \geq a),
\end{aligned}$$

and thus, $\frac{\mathrm{E}g(X)}{a} \geq P(g(x) \geq a)$. ■

**Corollary 3.3** (Chebyshev's Inequality) $P(|x - \mu| \geq k\sigma) \leq 1/k^2$ *for* $k > 0$.

**Proof** This follows by letting $g(X) = (X - \mu)^2$ and $a = k^2\sigma^2$ in Markov's inequality and realizing that $(x - \mu)^2 \geq k^2\sigma^2$ is equivalent to $(x - \mu) \leq -k\sigma$ or $(x - \mu) \geq k\sigma$, which is in turn equivalent to $|x - \mu| \geq k\sigma$. ■

In words, Markov's inequality states that we can *always* place an upper bound on the probability that $g(x) \geq a$ so long as $g(x)$ is nonnegative valued and $\mathrm{E}g(X)$ exists. Chebyshev's inequality implies that if $\mu$ and $\sigma$ are, respectively, the mean and standard deviation of the density function of $X$, then for any positive constant $k$, the probability that $X$ will have an outcome that is $k$ or more standard deviations from its mean, i.e., outside the interval $(\mu-k\sigma, \mu+k\sigma)$, is less than or equal to $1/k^2$. Note that we are able to make these probability statements *without* knowledge of the algebraic form of the density function.

Chebyshev's inequality is sometimes stated in terms of an event that is the complement of the event in Corollary 3.3.

**Corollary 3.4** (Chebyshev's Inequality) $P(|x - \mu| < k\sigma) \geq 1 - 1/k^2$ *for* $k > 0$.

**Proof** This follows directly from Corollary 3.3, noting that $P(|x-\mu| \geq k\sigma) = 1-P(|x-\mu| < k\sigma) \leq 1/k^2$. ■

Markov's inequality and Chebyshev's inequalities are interesting in their own right, but at this point we will use the concepts only to further clarify our interpretation of the variance as a measure of the spread or dispersion of a density function. In Corollary 3.4, let $k\sigma = c$, where $c$ is any arbitrarily small positive number. Then

$$P(|x - \mu| < c) \geq 1 - \frac{\sigma^2}{c^2},$$

where we have substituted for $k$ the value $c/\sigma$. Then note that as $\sigma^2 \to 0$, the probability inequality approaches

$$P(\mu - c < x < \mu + c) \geq 1,$$

which implies that as $\sigma^2 \to 0$, the density concentrates in the interval $(\mu - c, \mu + c)$ for *any* arbitrarily small positive $c$. Diagrammatically, this can be illustrated as, say, Figure 3.7.

As a concrete example, let $\mu = 0$, and examine the event $B = \{x : |x| < 10\}$, where we are letting $c = 10$ in the preceding argument. Then using $P(|x - \mu| < k\sigma) \geq 1 - 1/k^2$ with $k\sigma = c = 10$, we have

$$\text{if} \begin{bmatrix} \sigma = 5 \\ \sigma = 2 \\ \sigma = 1 \end{bmatrix}, \quad \text{then } P(B) \geq \begin{bmatrix} 1 - 1/2^2 = .75 \\ 1 - 1/5^2 = .96 \\ 1 - 1/10^2 = .99 \end{bmatrix},$$

and thus the smaller is $\sigma$ (and thus the smaller the variance), the larger is the lower bound on the probability that the outcome of $X$ occurs in the interval $(-10, 10)$.

For an alternative argument in support of interpreting the variance as a measure of the spread of a density function, note that the variance of $X$ can be interpreted as the expected squared distance of the random variable $X$ from its mean. To see this, first recall that the distance between two points, $x$ and $y$, on the real line is defined as $d(x, y) = |x - y|$. Then $d^2(x, y) = (x - y)^2$, and letting $y = \mu$, we have $\mathrm{E}d^2(X, \mu) = \mathrm{E}(X - \mu)^2 = \sigma^2$. Therefore, the smaller is $\sigma^2$, the smaller is the expected squared distance of $X$ from its mean.

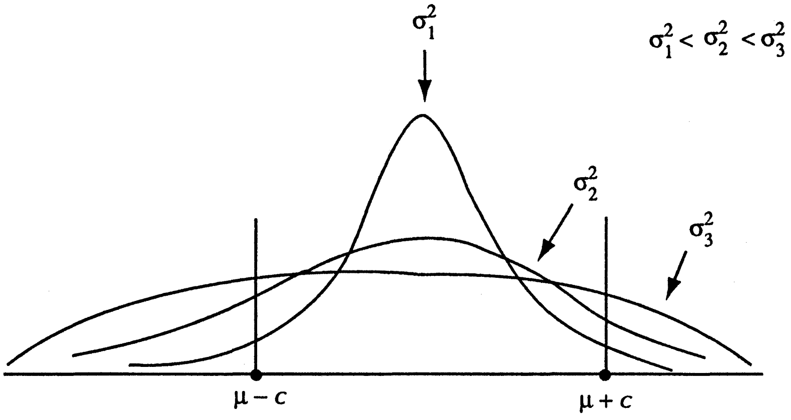


**Figure 3-7** Density functions and variances.

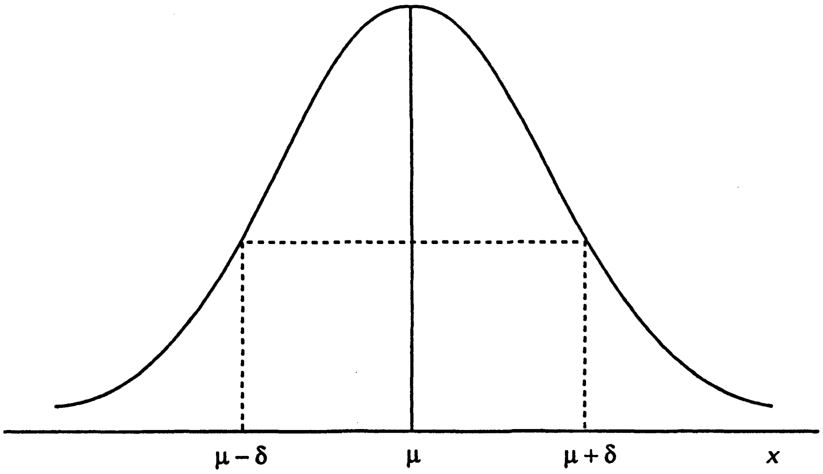


**Figure 3-8**
Density symmetric about $\mu$.

The third central moment is sometimes referred to as a measure of *skewness* of the density of $X$. A density is said to be **symmetric about** $\mu$ when $f(\mu + \delta) = f(\mu - \delta)\ \forall\ \delta > 0$ (see Figure 3.8).

If the density is not symmetric, then it is said to be a **skewed density.** A necessary condition for a density to be symmetric about $\mu$ is that $\mu_3 = \mathrm{E}(X - \mu)^3 = 0$. To see this, examine the continuous case, and note that

$$\mu_3 = \int_{-\infty}^{\infty} (x - \mu)^3 f(x)dx = \int_{\mu}^{\infty} (x - \mu)^3 f(x)dx + \int_{-\infty}^{\mu} (x - \mu)^3 f(x)dx.$$

By making the substitution $z = x - \mu$ in the first integral and $z = -x + \mu$ in the second integral, we obtain

$$\mu_3 = \int_0^{\infty} z^3 f(\mu + z)dz - \int_{\infty}^{0} (-z)^3 f(\mu - z)dz = \int_0^{\infty} z^3 f(\mu + z)dz - \int_0^{\infty} z^3 f(\mu - z)dz = 0,$$

since $f(\mu + z) = f(\mu - z)\ \forall\ z$ by the symmetry of $f$ about $\mu$. The reader can provide an analogous demonstration in the discrete case. Thus, if $\mu_3 \neq 0$, then the density function in question is skewed. The reader should note, however, that $\mu_3 = 0$ is not *sufficient* for symmetry, i.e., a density can be skewed and still have $\mu_3 = 0$, as the following example illustrates.

**Example 3.15** Let the random variable $X$ have the density function $f(x) = .22I_{\{1\}}(x) + .77I_{\{2\}}(x) + .01I_{\{4\}}(x)$ (see Figure 3.9). Note that $\mu = 1.8$, and it is clear that $f(x)$ is not symmetric about $\mu$. Nonetheless, $\mu_3 = \mathrm{E}(X - \mu)^3 = 0$. □

The sign of $\mu_3$ is sometimes interpreted as indicating the direction of the skew in the density function. In particular, density functions having long "tails" to the right are called **skewed to the right**, and these densities tend to have $\mu_3 > 0$, whereas density functions with long left-hand tails are called **skewed to the left**, and these densities tend to have $\mu_3 < 0$ (see Figure 3.10). Unfortunately, there are exceptions to these "tendencies," and the nature of

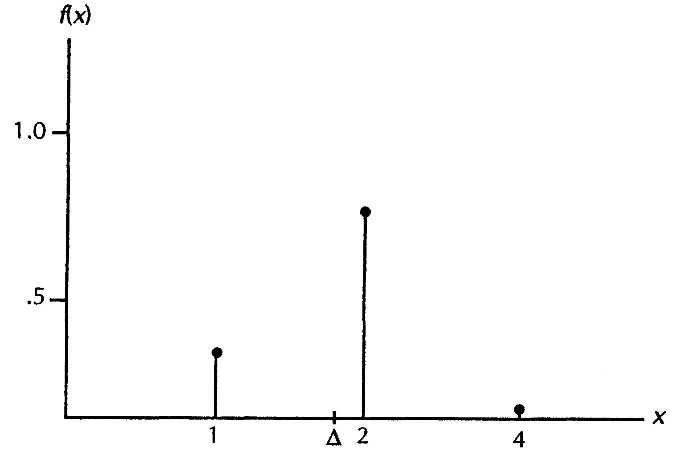


**Figure 3-9** A nonsymmetric density function.

the skewness is best determined by examining the graph of the density itself if the functional form of the density is known.

### Relationship Between Moments About the Origin and Mean

Integral moments about the origin and about the mean are functionally related. Central moments can be expressed solely in terms of moments about the origin, while moments about the origin can be expressed in terms of the mean and moments about the mean. The functional relationship is the direct result of the binomial theorem, which we review for the reader in the following lemma.

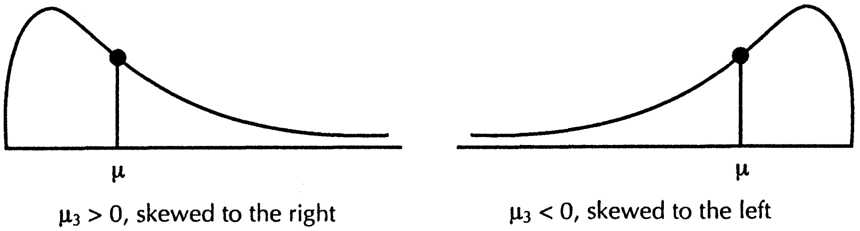
**Lemma 3.2** ***Binomial Theorem***

Let $a$ and $b$ be real numbers. Then $(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j}.$

**Theorem 3.20** (Central moments as functions of moments about the origin) *If $\mu_r$ exists and $r$ is a positive integer, then*

$$\mu_r = \sum_{j=0}^{r} (-1)^j \binom{r}{j} \mu'_{r-j} \mu^j.$$

**Proof** By definition, $\mu_r = \mathrm{E}(X-\mu)^r$. Substituting $(-\mu)$ for $a$, $X$ for $b$, and $r$ for $n$ in the binomial theorem (Lemma 3.2) and taking an expectation yield $\mu_r =$



$\mu_3 > 0$, skewed to the right

$\mu_3 < 0$, skewed to the left

**Figure 3-10** Skewed density functions.

$\mathrm{E}\sum_{j=0}^{r}\binom{r}{j}(-\mu)^j X^{r-j}$. An application of Theorems 3.6 and 3.7 results in $\mu_r = \sum_{j=0}^{r}(-1)^j\binom{r}{j}\mu^j \mathrm{E}X^{r-j}$. ■

**Theorem 3.21** (Moments about the origin as functions of central moments) *If $\mu_r'$ exists and $r$ is a positive integer, then $\mu_r' = \sum_{j=0}^{r}\binom{r}{j}\mu_{r-j}\mu^j$.*

**Proof** By definition, $\mu_r' = \mathrm{E}X^r = \mathrm{E}(X-\mu+\mu)^r$. Substituting $\mu$ for $a$, $(X-\mu)$ for $b$, and $r$ for $n$ in the binomial theorem (Lemma 3.2) and taking an expectation yield $\mu_r' = \mathrm{E}\sum_{j=0}^{r}\binom{r}{j}\mu^j(X-\mu)^{r-j}$. An application of Theorems 3.6 and 3.7 results in $\mu_r' = \sum_{j=0}^{r}\binom{r}{j}\mu^j \mathrm{E}(X-\mu)^{r-j}$. ■

A special case of Theorem 3.20, which we will use repeatedly in later chapters, is the case where $r = 2$, which provides a representation of the variance of a random variable in terms of moments about the origin. In particular, from Theorem 3.20,

$$\mathrm{var}(X) = \mu_2 = \mu_2' - \mu^2 = \mathrm{E}X^2 - (\mathrm{E}X)^2.$$

**Example 3.16** Let the random variable $X$ have density function $f(x) = I_{[0,1]}(x)$. Since $\mathrm{E}X = 1/2$, and since $\mathrm{E}X^2 = \int_{-\infty}^{\infty} x^2 I_{[0,1]}(x)dx = \frac{x^3}{3}\Big|_0^1 = 1/3$, then $\mathrm{var}(X) = \mathrm{E}X^2 - (\mathrm{E}X)^2 = 1/3 - (1/2)^2 = 1/12$. □

### Existence of Moments

Regarding the existence of moments, the following theorem is often quite useful.

**Theorem 3.22** *If $\mathrm{E}X^t$ exists for a given integral $r > 0$, then $\mathrm{E}X^s$ exists for all integral $s \in [0, r]$.*

**Proof** (Continuous) Define $A_{<1} = \{x : |x|^s < 1\}$ and $A_{\geq 1} = \{x : |x|^s \geq 1\}$. Note that

$$\int_{-\infty}^{\infty} |x|^s f(x)dx = \int_{x\in A_{<1}} |x|^s f(x)dx + \int_{x\in A_{\geq 1}} |x|^s f(x)dx.$$

Since $f(x) \geq |x|^s f(x)\ \forall\ x \in A_{<1}$,

$$P(|x|^s < 1) = \int_{x\in A_{<1}} f(x)dx \geq \int_{x\in A_{<1}} |x|^s f(x)dx.$$

Now let $r > s$, and note that $|x|^s \leq |x|^r\ \forall\ x \in A_{\geq 1}$, whence it follows that

$$\int_{x\in A_{\geq 1}} |x|^r f(x)dx \geq \int_{x\in A_{\geq 1}} |x|^s f(x)dx.$$

Finally, since $\int_{x\in A_{<1}} |x|^r f(x)dx \geq 0$,

$$\int_{-\infty}^{\infty} |x|^s f(x)dx \leq P(|x|^s < 1) + \int_{-\infty}^{\infty} |x|^r f(x)dx < \infty,$$

where the rightmost inequality is due to the fact that $P(|x|^s < 1) \in [0, 1]$ and $\mathrm{E}X^r$ exists, implying the absolute convergence of the improper integral defining the expectation. It follows from $\int_{-\infty}^{\infty} |x|^s f(x)dx < \infty$ that $\mathrm{E}X^s$ exists.

(Discrete) The proof in the discrete case is analogous to the preceding proof with summation replacing integration and is left to the reader. ■

The theorem implies that if the existence of the $r$th-order moment about the origin can be demonstrated, then lower-order moments about the origin are known to exist. Theorem 3.22 can also be used to demonstrate the nonexistence of moments, since if $\mathrm{E}X^r$ does *not* exist, then necessarily $\mathrm{E}X^s$ cannot exist for $s > r$ or else Theorem 3.22 would be contradicted.

**Example 3.17** Let the random variable $X$ have the density function $f(x) = 2(x + 1)^{-3} I_{[0,\infty)}(x)$. Examine $\mathrm{E}X^\alpha = \int_0^\infty x^\alpha 2(x+1)^{-3}dx$. To simplify the integral, make the substitution $y = x+1$, so that $x = y - 1$ and $dy = dx$, to yield $\mathrm{E}X^\alpha = 2\int_1^\infty (y-1)^\alpha y^{-3}dy$. Note that if $\alpha = 2$, then $\mathrm{E}X^2 = 2\int_1^\infty (y^{-1} - 2y^{-2} + y^{-3})dy = \lim_{y\to\infty} 2(\ln y + 2y^{-1} - (1/2)y^{-2})$, and since the limit diverges, $\mathrm{E}X^2$ does not exist (i.e., note that $\ln y \to \infty$ as $y \to \infty$). This implies by Theorem 3.22 that moments of order 2 *or greater* do *not* exist for $X$. The reader can verify that $\mathrm{E}X$ exists and is equal to 1. □

Existence results analogous to Theorem 3.22 can be stated for moments about the mean.

**Theorem 3.23** *If* $\mathrm{E}(Y-\mu)^r$ *exists for a given integral* $r > 0$, *then* $\mathrm{E}(Y-\mu)^s$ *exists for all integral* $s \in [0, r]$.

**Proof** This follows directly from Theorem 3.22 upon defining $X = Y - \mu$. ■

One can also infer the existence of moments about the mean from moments about the origin, and vice versa.

**Theorem 3.24** *If* $\mathrm{E}X^r$ (*or* $\mathrm{E}(X - \mu)^r$) *exists for a given integral* $r > 0$, *then* $\mathrm{E}(X - \mu)^s$ (*or* $\mathrm{E}X^s$) *exists* ∀ *integral* $s \in [0, r]$.

**Proof** This follows directly from Theorems 3.20 to 3.23. Details are left to the reader. ■

### Nonmoment Measures of Probability Density Characteristics

Note that whether or not moments exist for $X$, there are other measures of probability distribution characteristics that are also of interest in applications. An alternative measure of the central tendency of a density is the median, defined as follows.

**Definition 3.11** ***Median of X***

Any number, $b$, satisfying $P(x \leq b) \geq 1/2$ and $P(x \geq b) \geq 1/2$ is called a **median** of $X$ and is denoted by $\mathrm{med}(X)$.

The median is a measure of central tendency in the sense that $\geq 1/2$ of the probability mass of a density is both to the right and to the left of the median. In the continuous case, the probability inequalities in Definition 3.11 can be met with strict equalities, so that the median is a point at which half the probability mass is to the left and half is to the right, as $\int_{-\infty}^{\text{med}(X)} f(x)dx = \int_{\text{med}(X)}^{\infty} f(x)dx = .5$.

Depending on how the density is defined, the median may not be unique even in the continuous case. However, if $b = \text{med}(X)$, and if in the neighborhood of the point $b$ the CDF of $X$ is continuous and strictly increasing, then med$(X)$ is unique (why?).

**Example 3.18** Let the random variable $X$ have density function $f(x) = (1/6)I_{\{1,2,\ldots,6\}}(x)$. Then the median of $X$ is not unique and can be any number in the interval [3,4], since $P(x \leq b) \geq 1/2$ and $P(x \geq b) \geq 1/2 \;\forall\; b \in [3,4]$. □

**Example 3.19** The central processing unit (CPU) used by a company that manufactures personal computers has an operating life until failure that is given by the outcome of a random variable $X$ having density function $f(x) = (1/50)e^{-x/50}I_{(0,\infty)}(x)$, where $x$ is measured in thousands of hours. What is the median operating life of the CPU?

**Answer**: We must solve the following equation for med$(X)$:

$$\int_{-\infty}^{\text{med}(X)} \frac{1}{50}e^{-x/50}I_{(0,\infty)}(x)dx = .5 \quad \text{or} \quad 1 - e^{-\text{med}(X)/50} = .5,$$

so that med$(X)$=34.657. It is thus equally probable that the CPU will operate more or less than 34,657 hours until failure. □

Another probability density characteristic, which subsumes the median as a special case, is called a quantile.

**Definition 3.12** *Quantile of X*

Any number $b$ satisfying $P(x \leq b) \geq p$ and $P(x \geq b) \geq 1 - p$ for $p \in (0, 1)$ is called a **quantile** of $X$ of order $p$ (or the $(100p)$th percentile of the distribution of $X$).

Note the median is then simply the quantile of $X$ of order .5, or the 50th percentile of the distribution of $X$. As in the case of the median, the quantile of order $p$ may not be unique for a given random variable $X$. In Ex. 3.18, any $b \in [4, 5]$ would be a quantile of $X$ of order 2/3, while in Ex. 3.19, the quantile of order 2/3 would be $b = 54.931$.

One additional characteristic of a probability distribution that will be especially useful when we study the maximum likelihood procedure of statistical inference is a **mode of the distribution** of $X$.

**Definition 3.13**
*Mode of f(x)*

Let $X$ be a random variable with density function $f(x)$. Then any point $b$ at which $f(x)$ exhibits a maximum is called a **mode** of $X$, or a **mode of the distribution** of $X$, and is denoted by mode$(X)$.

Some density functions may not have a unique mode. Those that do are called **unimodal**. Note that the density function in Ex. 3.18 exhibits six modes corresponding to the points $x = 1, 2, 3, 4, 5, 6$. The density function in Ex. 3.17 has one mode, at the point $x = 0$. The density in Ex. 3.19 has no mode (why not? How might the problem be altered so that mode$(X)$ exists?).

## 3.5 Moment- and Cumulant-Generating Functions

The expectation of $e^{tX}$ results in a function of $t$ that, when differentiated with respect to the argument $t$ and then evaluated at $t = 0$, generates moments of $X$ about the origin. The function is aptly called the **moment-generating function** of $X$.

**Definition 3.14**
*Moment generating function (MGF)*

The expected value of $e^{tX}$ is defined to be the moment-generating function of $X$ if the expected value exists for every value of $t$ in some interval $t \in (-h, h)$, $h > 0$. The moment-generating function of $X$ will be denoted by $M_X(t)$. Thus,

$$\text{(discrete)} \qquad M_X(t) = \mathrm{E}e^{tX} = \sum_{x \in R(x)} e^{tx} f(x),$$

$$\text{(continuous)} \qquad M_X(t) = \mathrm{E}e^{tX} = \int_{-\infty}^{\infty} e^{tx} f(x) dx.$$

The reader should note that $M_X(0) = \mathrm{E}e^0 = E1 = 1$ is *always* defined, and from this property it is clear that a function of $t$ *cannot* be a MGF unless the value of the function at $t = 0$ is 1. The condition that $M_X(t)$ be defined $\forall$ $t \in (-h, h)$ is a technical condition ensuring that $M_X(t)$ is differentiable at the point zero, a property whose importance will become evident shortly.

We now indicate how the MGF can be used to generate moments about the origin. In the following theorem, we use the notation $d^r g(a)/dx^r$ to indicate the $r$th derivative of $g(x)$ with respect to $x$ evaluated at $x = a$.

**Theorem 3.25** *Let $X$ be a random variable for which the MGF, $M_X(t)$, exists. Then*

$$\mu'_r = \mathrm{E}X^r = \frac{d^r M_X(0)}{dt^r}.$$

**Proof** The proof is facilitated by the following lemma from advanced calculus.

**Lemma 3.3** If the function $g(t)$ defined by $g(t) = \sum_{x \in R(X)} e^{tx} f(x)$ or $\int_{-\infty}^{\infty} e^{tx} f(x) dx$ converges for $t \in (-h, h)$, $h > 0$, then $d^r g(t)/dt^r$ exists $\forall\ t \in (-h, h)$ and for all positive integers $r$, and the derivative can be found by *differentiating under the summation sign* or *differentiating under the integral sign*, respectively, as

$$\frac{d^r g(t)}{dt^r} = \sum_{x \in R(x)} \frac{d^r e^{tx}}{dt^r} f(x) \quad \text{or} \quad \int_{-\infty}^{\infty} \frac{d^r e^{tx}}{dt^r} f(x) dx$$

(see D. V. Widder (1961), *Advanced Calculus*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, pp. 442–447).

If the moment generating function

$$M_X(t) = \mathrm{E}e^{Xt} = \sum_{x \in R(X)} e^{tx} f(x) \quad \text{or} \quad \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

exists (converges) for $t \in (-h, h)$, $h > 0$, then from Lemma 3.3,

$$\frac{d^r M_X(t)}{dt^r} = \sum_{x \in R(X)} \frac{d^r e^{tx}}{dt^r} f(x) \quad \text{or} \quad \int_{-\infty}^{\infty} \frac{d^r e^{tx}}{dt^r} f(x) dx$$

$$= \sum_{x \in R(X)} x^r e^{tx} f(x) \quad \text{or} \quad \int_{-\infty}^{\infty} x^r e^{tx} f(x) dx.$$

Evaluating the $r$th derivative at $t = 0$ yields

$$\frac{d^r M_X(0)}{dt^r} = \sum_{x \in R(X)} x^r f(x) \quad \text{or} \quad \int_{-\infty}^{\infty} x^r f(x) dx = \mathrm{E}X^r. \qquad \blacksquare$$

**Example 3.20** The random variable $X$ has the density function $f(x) = e^{-x} I_{[0,\infty)}(x)$. Find the MGF, and use it to define the mean and variance of $X$.

**Answer**:

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} e^{-x} I_{[0,\infty)}(x) dx = \int_0^{\infty} e^{x(t-1)} dx = \left. \frac{e^{x(t-1)}}{t-1} \right|_0^{\infty}$$

$$= 0 - \frac{1}{t-1} = (1-t)^{-1} \quad \text{(provided } t < 1\text{)}$$

The mean is defined as $\mu = dM_X(0)/dt = (1-0)^{-2} = 1$. For the variance, recall from Theorem 3.20 that $\sigma^2 = \mu_2' - \mu^2$. Then, $\mu_2' = d^2 M_X(0)/dt^2 = 2(1-0)^{-3} = 2$, and thus, $\sigma^2 = 1$. □

There are a number of elementary results relating to moment-generating functions that can be quite useful in applications. We present these results in the next theorem, leaving the proof as an exercise for the reader.

**Theorem 3.26** (Properties of MGFs) *Let $(X_1, \ldots, X_n)$ be independent random variables having respective MGFs $M_{X_i}(t)$, $i = 1, \ldots, n$.*

**a.** *If* $Y_i = aX_i + b$, *then* $M_{Y_i}(t) = e^{bt}M_{X_i}(at)$.
**b.** *If* $Y = \sum_{i=1}^n X_i$, *then* $M_Y(t) = \prod_{i=1}^n M_{X_i}(t)$.
**c.** *If* $Y = \sum_{i=1}^n a_iX_i + b$, *then* $M_Y(t) = e^{bt}\prod_{i=1}^n M_{X_i}(a_it)$.

It is useful to note that since $X$ could be defined as a random variable that is itself a function of other random variables, a more general conceptualization of the MGF is $M_{g(X)}(t) = E\exp(g(X)t)$, where $g$ is a function of $X$.[18] The MGF $M_{g(X)}(t)$ could then be used to define moments about the origin for the random variable defined by $g(X)$. Note that moments about the *origin* for $g(X) = X - \mu$ coincide with moments about the *mean* for $X$, and thus the generalized MGF can be defined appropriately to generate moments about the mean directly.

**Example 3.21** Let $f(x) = e^{-x}I_{(0,\infty)}(x)$, as in Ex. 3.20. Recall that $\mu = 1$ in this case. We find the moment-generating function of the random variable $Y = g(X) = X - 1$, and use it to define the variance of $X$. First of all, note that Theorem 3.26. a is applicable with $a = 1$ and $b = -1$. Since $M_X(t) = (1-t)^{-1}$ for $t < 1$ from Ex. 3.20, it follows that

$$M_{(X-1)}(t) = e^{-t}(1-t)^{-1} \quad \text{for } t < 1.$$

To find var($X$), we need to evaluate $(d^2M_{(X-1)}(0))/dt^2 = \mathrm{E}(X-1)^2$:

$$\frac{dM_{(X-1)}(t)}{dt} = e^{-t}(1-t)^{-2} - (1-t)^{-1}e^{-t}$$

and

$$\frac{d^2M_{(X-1)}(t)}{dt^2} = 2e^{-t}(1-t)^{-3} - e^{-t}(1-t)^{-2} + (1-t)^{-1}e^{-t} - e^{-t}(1-t)^{-2},$$

so that

$$\frac{d^2M_{(X-1)}(0)}{dt^2} = 1 = \mathrm{var}(X).$$ □

### Uniqueness and Inversion of MGFs

Apart from generating moments, the MGF can be useful for identifying the density function of a given random variable. This is due to a uniqueness property possessed by MGFs that essentially establishes a one-to-one correspondence between density functions and MGFs. A formal statement of the uniqueness property is given in the following theorem. The proof of the theorem relies on the fact that the MGF is a bilateral Laplace transform of the function $f(x)$, and there is a unique association between a Laplace transform and the function being transformed. These concepts are beyond the scope of our study, and the proof will be omitted. The interested reader can refer to Widder, Advanced Calculus, p. 459–460 and D. A. S. Fraser (1976), *Probability and Statistics.* North Scituate, Duxbury Press, p. 544–546, for details.

[18] Recall that $\exp(a) \equiv e^a$.

**Theorem 3.27** (MGF Uniqueness Theorem) *If a moment-generating function exists for a random variable $X$ having density function $f(x)$, then the moment generating function is unique. Conversely, the moment generating function determines the density function of $X$ uniquely, at least up to a set of points having probability zero.*

In other words, a density function has one and only one MGF associated with it, if an MGF exists at all. Furthermore, if more than one density function is associated with a given MGF, the density functions differ only on a set of points *that are immaterial* for the purposes of assigning probabilities to events for $X$, i.e., they differ on a set of points having probability zero. Thus, for all practical purposes, if one knows the MGF for a given random variable $X$, and if one also knows of a density function that produces this MGF in the manner prescribed by Def. 3.14, then the density function can be treated as the density function of the random variable $X$. The following example illustrates the logic followed in applying the uniqueness theorem.

**Example 3.22** Examine the density function $f(x) = (b - a)^{-1} I_{[a,b]}(x)$, for $a < b$. The MGF associated with this density can be identified as follows:

$$M_X(t) = \mathrm{E}e^{Xt} = \int_{-\infty}^{\infty} e^{xt}(b-a)^{-1} I_{[a,b]}(x)dx = (b-a)^{-1}\int_a^b e^{xt}dx = (b-a)^{-1}\frac{e^{xt}}{t}\Big|_a^b$$

$$= \begin{cases} \dfrac{e^{bt}-e^{at}}{t(b-a)} & \text{for } t \neq 0, \\ 1 \quad \text{for } t = 0. \end{cases}$$

Now suppose a random variable $Z$ has an MGF defined by $M_Z(t) = (e^{bt} - e^{at})/t(b-a)$ for $t \neq 0$. Then by the uniqueness theorem, since $f(x)$ above is associated with this same MGF, the density function of $Z$ can be specified as $f(z) = (b-a)^{-1} I_{[a,b]}(z)$. □

When it exists, an MGF can be thought of as a "fingerprint" of a given density function. In Chapter 4, we will examine a collection of density functions that have been found to be useful in applications, and we will assemble a file of their MGFs. Later on we will examine a number of important functions of random variables that will be used for statistical inference purposes, and in a notable number of cases, we will be able to identify the probability densities of these functions by matching their MGFs to the appropriate MGFs in the file we will have assembled.

The "recognition" of an MGF as a known fingerprint of some probability density function is not the only way an MGF can be used to identify the probability distribution of a random variable. There is an **inversion relationship** that allows one to integrate a function involving the MGF to identify the CDF of a random variable, from which the density function can be deduced. Unfortunately, the technique involves transform theory and generally complicated integration of expressions involving complex numbers and is beyond the scope of our study. Nonetheless, without providing the formal details, we will provide the reader with the general idea of what is involved. In particular, if we replace the argument $t$ in an MGF by $(it)$, $i$ being the imaginary number $i = \sqrt{-1}$, then the CDF associated with the MGF can be found as

$$F(b) = \lim_{a \to -\infty} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} M_X(it)dt$$

$\forall$ $b$ at which $F$ is continuous. Then the density function can be determined from the CDF using the methods described in Chapter 2. Alternatively, if it can be shown that $\int_{-\infty}^{\infty} | M_X(it) | \, dt < \infty$, then the density function for a continuous random variable can be derived directly as

$$f(x) = \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-itx} M_X(it)dt.$$

For further information concerning this inversion property, the reader can consult M. Kendall and A. Stuart (1977), *The Advanced Theory of Statistics, Vol. 1* New York: Macmillan, Chapter 4.

In cases where the MGF does not exist, there is an alternative function that always exists, called the **characteristic function**, which serves the same purpose as the MGF. In particular, there is a unique relationship between characteristic functions and density functions, analogous to the result stated in Theorem 3.27. The characteristic function can be inverted to obtain the density function, and the characteristic function can be used to generate any moments that exist for a random variable by differentiating the characteristic function an appropriate number of times, evaluating the derivative at the point zero, and then dividing the result by $(i)^k$, $k$ being the order of the moment sought (equivalently, $k$ is the order of the derivative). The characteristic function is defined as $\phi_X(t) = \mathrm{E}e^{itX}$, and so complex numbers are involved in the definition of the characteristic function. When the MGF exists, the characteristic function is $\phi_X(t) \equiv M_X(it)$, i.e., the characteristic function is identically the MGF evaluated at $(it)$ rather than at $t$. For example, in Ex. 3.22, the characteristic function of $X$ would be $(e^{bit} - e^{ait})/it(b-a)$ for $t \neq 0$. Despite the advantage that $\phi_X(t)$ always exists, we will not pursue the study of characteristic functions any further, in order to avoid the use of complex numbers. Interested readers can examine Kendall and Stuart, Advanced Statistics, Chapter 4, for further details.

### Cumulant-Generating Function

The natural logarithm of the moment-generating function defines a function called the **cumulant-generating function**, which, when differentiated $r$ times with respect to $t$ and then evaluated at $t = 0$, defines the $r$th **cumulant** of a random variable.

**Definition 3.15**
***Cumulant-generating function and cumulants***

The cumulant-generating function of $X$ is defined as $\psi(t) = \ln(M_X(t))$. The $r$th cumulant of $X$ is given by $\kappa_r = (d^r\psi(0))/dt^r$. The first four cumulants are related to moments as follows: $\kappa_1 = \mu'_1$; $\kappa_2 = \sigma^2$; $\kappa_3 = \mu_3$; and $\kappa_4 = \mu_4 - 3\sigma^4$.

The cumulant-generating function can be used directly to generate the mean, variance, and third moment about the mean via differentiation of the function to the first, second, or third order, respectively. If $X_1, \ldots, X_n$ are independent random variables, it follows from Def. 3.15 and Theorem 3.26.b that the cumulant-generating function of $Y = \sum_{i=1}^n X_i$ equals the sum of the cumulant-generating functions of the $X_i$'s. It then also follows that the cumulant of the sum is the sum of the cumulants, which is the genesis of the name "cumulant." Often, the derivatives of the cumulant-generating function are easier to calculate than the derivatives of the MGF.

**Example 3.23** Recall Ex. 3.20 and 3.21, where $M_X(t) = (1 - t)^{-1}$ for $t < 1$. The cumulant-generating function of $X$ is given by $\psi_X(t) = \ln M_X(t) = -\ln(1 - t)$ for $t < 1$. Then

$$\mu = d\psi_X(0)/dt = (1 - t)^{-1} \,\big|_{t=0} = 1,$$

$$\sigma^2 = d^2\psi_X(0)/dt^2 = (1 - t)^{-2} \,\big|_{t=0} = 1,$$

$$\mu_3 = d^3\psi_X(0)/dt^3 = 2(1 - t)^{-3} \,\big|_{t=0} = 2.$$

□

### Multivariate Extensions

The MGF and cumulant-generating function can be extended to the case of a multivariate random variable $X = (X_1, \ldots, X_n)$.

**Definition 3.16**
***MGF and cumulant generating function; multivariate***

The expected value of $\exp(\sum_{j=1}^n t_j X_j)$ is defined to be the MGF of the $n$-variate random variable $X = (X_1, \ldots, X_n)$ if the expected value exists for all $t_i \in (-h, h)$, for some $h > 0$, $i = 1, \ldots, n$. The MGF will be denoted by $M_X(t)$, where $t = (t_1, \ldots, t_n)$. Thus,

$$\text{(discrete)} \qquad M_X(t) = \sum_{(x_1,\ldots,x_n)\in R(X)} \cdots \sum \exp\left(\sum_{j=1}^n t_j x_j\right) f(x_1, \ldots, x_n),$$

$$\text{(continuous)} \qquad M_X(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(\sum_{j=1}^n t_j x_j\right) f(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

The cumulant generating function of $X$ is defined as

$\psi_X(t) = \ln M_X(t)$.

Letting $\mu'_r(X_i)$ denote the $r$th moment of $X_i$ about the origin, it can be shown that

$$\mu'_r(X_i) = \mathrm{E}X_i^r = \frac{\partial^r M_X([0])}{\partial t_i^r}.$$

Thus, the $r$th order *partial* derivative of $M_X(t)$ with respect to $t_i$, evaluated at $t = [0]$ (i.e., the *vector* $t$ equal to the zero vector), equals the $r$th order moment about zero for $X_i$. Similarly, the $r$th *partial* derivative of $\psi_X(t)$ with respect to $t_i$, evaluated at $t = [0]$, equals the $r$th cumulant of the random variable $X_i$, which then allows means, variances, and third moments about the origin to be calculated directly for each of the $X_i$'s.

An analog of the **MFG uniqueness theorem** applies to the multivariate MGF. In fact, interpreting $X$ as a *vector* in the statement of Theorem 3.27 produces the appropriate **multivariate MGF uniqueness theorem.**

If the MGF for an $n$-variate random variable $(X_1,\ldots,X_n)$ is known, the **marginal MGF** for a subset of $m < n$ of the random variables is easily found by setting the $t_i$'s associated with the remaining $n - m$ random variables to zero.

**Theorem 3.28** (Marginal MGFs from Multivariate MGFs) *Let $X = (X_1,\ldots,X_n)$ have MGF $M_X(t)$, and let $X_{(m)} = (X_j, j \in J)$ be any m-element subset of the random variables in X, where $J \subset \{1,2,\ldots,n\}$, $N(J) = m < n$. Define $t_{(m)} = (t_j, j \in J)$. Then the MGF of $X_{(m)}$, referred to as the* ***marginal MGF of*** $X_{(m)}$, *can be represented as $M_{X_{(m)}}(t_{(m)}) = M_X(t^*)$, where the elements in $t^*$ are defined by $t_j^* = t_j I_J(j)$.*

**Proof**

$$M_X(t^*) = \mathrm{E}\left(\exp\left(\sum_{j=1}^{n} t_j^* X_j\right)\right) = \mathrm{E}\left(\exp\left(\sum_{j\in J} t_j X_j\right)\right) \qquad (\text{since } t_j^* = 0 \text{ if } j \notin J)$$

$$= M_{X_{(m)}}(t_{(m)}) \qquad (\text{by definition}).$$ ■

A **marginal cumulant-generating function** can be defined as the natural logarithm of a marginal MGF.

If $X$ has MGF $M_X(t)$, it can be shown that $X_1,\ldots,X_n$ are independent iff $M_X(t) = \prod_{i=1}^n M_{X_i}(t_i)$, or equivalently, iff $\psi_X(t) = \sum_{i=1}^n \psi_{X_i}(t)$ (see S. F. Arnold (1990), *Mathematical Statistics*. Englewood Cliffs, NJ: Prentice-Hall, pp. 118–119).

**Example 3.24** Suppose the joint MGF of the bivariate random variable $(X_1, X_2)$ is given by $M_X(t) = \exp(\sum_{i=1}^2 \mu_i t_i + (1/2)\sum_{i=1}^2 \sum_{j=1}^2 \sigma_{ij} t_i t_j)$ (we will see in Chapter 4 that this is the MGF associated with a bivariate "normal" density function). Then the marginal MGF of $X_1$ can be defined by setting $t_2 = 0$ in $M_X(t)$ to obtain $M_{X_1}(t_1) = \exp(\mu_1 t_1 + (1/2)\sigma_{11} t_1^2)$ (which is the MGF associated with a univariate "normal" density function). The marginal cumulant-generating function is given by $\psi_{X_1}(t) = \ln M_{X_1}(t) = \mu_1 t_1 + (1/2)\sigma_{11} t_1^2$. The two random variables will be independent if $\sigma_{12} = \sigma_{21} = 0$. □

If we take *cross* partial derivatives of the joint MGF of $X$, and evaluate the derivative at $t = [0]$, we obtain

$$\frac{\partial^{r+s} M_X([0])}{\partial t_i^r \partial t_j^s} = \mathrm{E} X_i^r X_j^s.$$

This expectation is an example of a *joint moment*. The cross partial derivative of the cumulant generating function given by $(\partial^2 \psi_X([0]))/(\partial t_i \partial t_j) = \mathrm{E}(X_i - \mathrm{E}X_i)(X_j - \mathrm{E}X_j)$ is the *covariance* between $X_i$ and $X_j$. These concepts are discussed further in the next section.

## 3.6 Joint Moments, Covariance, and Correlation

In the case of multivariate random variables, the concept of **joint moments** becomes relevant. The formal definitions of joint moments about the origin and about the mean are as follows:

**Definition 3.17** ***Joint moment about the origin***

Let $X$ and $Y$ be two random variables having joint density function $f(x, y)$. Then the $(r, s)$th joint moment of $(X, Y)$ (or of $f(x, y)$) about the origin is defined by

$$\text{(discrete)} \qquad \mu'_{r,s} = \sum_{x \in R(X)} \sum_{y \in R(Y)} x^r y^s f(x, y),$$

$$\text{(continuous)} \qquad \mu'_{r,s} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^r y^s f(x, y) dx dy.$$

**Definition 3.18** ***Joint moments about the mean (or central joint moment)***

Let $X$ and $Y$ be two random variables having joint density function $f(x, y)$. Then the $(r, s)$th joint moment of $(X, Y)$ (or of $f(x, y)$) about the mean is defined by

$$\text{(discrete)} \qquad \mu_{r,s} = \sum_{x \in R(X)} \sum_{y \in R(Y)} (x - \mathrm{E}X)^r (y - \mathrm{E}Y)^s f(x, y),$$

$$\text{(continuous)} \qquad \mu_{r,s} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mathrm{E}X)^r (y - \mathrm{E}Y)^s f(x, y) dx dy.$$

### Covariance and Correlation

Our immediate interest will focus on a particular joint moment about the mean, $\mu_{1,1}$, and the relationship between this moment and moments about the origin. The central moment $\mu_{1,1}$ is given a special name and symbol, and we will see that $\mu_{1,1}$ is useful as a measure of "linear association" between $X$ and $Y$.

**Definition 3.19** *Covariance*

The central joint moment $\mu_{1,1} = \mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y)$ is called the **covariance between $X$ and $Y$** and is denoted by the symbol $\sigma_{XY}$, or by $\mathrm{cov}(X, Y)$.

Note that there is a simple relationship between $\sigma_{XY}$ and moments about the origin that can be used for the calculation of the covariance.

**Theorem 3.29** $\sigma_{XY} = \mathrm{E}XY - \mathrm{E}X\mathrm{E}Y$.

**Proof** This result follows directly from the properties of the expectation operation. In particular, by definition,

$$\sigma_{XY} = \mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y) = \mathrm{E}[XY - (\mathrm{E}X)Y - X\mathrm{E}Y + \mathrm{E}X\mathrm{E}Y] = \mathrm{E}XY - \mathrm{E}X\mathrm{E}Y. \quad \blacksquare$$

**Example 3.25** Let the bivariate random variable $(X, Y)$ have a joint density function $f(x, y) = (x + y)I_{[0,1]}(x)I_{[0,1]}(y)$. Find $\mathrm{cov}(X, Y)$.
**Answer:** Note that

$$\mathrm{E}XY = \int_0^1 \int_0^1 xy(x + y)dxdy = \int_0^1 \int_0^1 [x^2y + xy^2]dxdy = \frac{1}{3},$$

$$\mathrm{E}X = \int_0^1 \int_0^1 x(x + y)dxdy = \int_0^1 \int_0^1 (x^2 + xy)dxdy = \frac{7}{12},$$

$$\mathrm{E}Y = \int_0^1 \int_0^1 y(x + y)dxdy = \int_0^1 \int_0^1 (yx + y^2)dxdy = \frac{7}{12}.$$

Then, by Theorem 3.29, $\mathrm{cov}(X, Y) = 1/3 - (7/12)(7/12) = -1/144$. □

A useful corollary to Theorem 3.29 is that *the expectation of a product is the product of the expectations iff* $\sigma_{XY} = 0$, formally stated as follows.

**Corollary 3.5** $\mathrm{E}XY = \mathrm{E}X\mathrm{E}Y$ *iff* $\sigma_{XY} = 0$.

**Proof** This follows directly from Theorem 3.29 upon setting $\sigma_{XY}$ to zero (sufficiency) or setting $\mathrm{E}XY$ equal to $\mathrm{E}X\mathrm{E}Y$ (necessity). ■

What does $\sigma_{XY}$ measure? The covariance is a measure of the *linear association* between two random variables, the precise meaning of linear association to be made clear shortly. Our discussion will be facilitated by observing that the value of $\sigma_{XY}$ exhibits a definite upper bound in absolute value which is expressible as a function of the variances of the two random variables involved. The bound on $\sigma_{XY}$ follows from the following inequality.

**Theorem 3.30** (Cauchy–Schwarz Inequality) $(\mathrm{E}WZ)^2 \leq \mathrm{E}W^2\mathrm{E}Z^2$.

**Proof** The quantity $\mathrm{E}(\lambda_1 W + \lambda_2 Z)^2$ must be greater than or equal to $0 \ \forall \ (\lambda_1, \lambda_2)$ since $(\lambda_1 W + \lambda_2 Z)^2$ is a random variable having only non-negative outcomes. Thus

$$\lambda_1^2\mathrm{E}W^2 + \lambda_2^2\mathrm{E}Z^2 + 2\lambda_1\lambda_2\mathrm{E}WZ \geq 0 \ \forall \ (\lambda_1, \lambda_2),$$

which in matrix terms can be represented as

$$[\lambda_1\lambda_2]\begin{bmatrix} \mathrm{E}W^2 & \mathrm{E}WZ \\ \mathrm{E}WZ & \mathrm{E}Z^2 \end{bmatrix}\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \geq 0 \quad \forall\,(\lambda_1, \lambda_2).$$

The last inequality is precisely the defining property of positive semidefiniteness for the (2 × 2) matrix in brackets,[19] and the matrix in brackets will be positive semidefinite iff $\mathrm{E}W^2\mathrm{E}Z^2 - (\mathrm{E}WZ)^2 \geq 0$ (see Section 3.8). ■

The covariance bound we seek is stated in the following theorem.

**Theorem 3.31** (Covariance Bound) $|\,\sigma_{XY}\,| \leq \sigma_X\sigma_Y$.

**Proof** Let $W = (X - \mathrm{E}X)$ and $Z = (Y - \mathrm{E}Y)$ in the Cauchy–Schwarz inequality. Then $(\mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y))^2 \leq \mathrm{E}(X - \mathrm{E}X)^2\mathrm{E}(Y - \mathrm{E}Y)^2$ or $\sigma_{XY}^2 \leq \sigma_X^2\sigma_Y^2$, which holds iff $|\,\sigma_{XY}\,| \leq \sigma_X\sigma_Y$. ■

Thus, the covariance between $X$ and $Y$ is upper-bounded in absolute value by the product of the standard deviations of $X$ and $Y$. Using this bound, we can define a useful scaled version of the covariance, called the **correlation** between $X$ and $Y$, as follows.

**Definition 3.20** ***Correlation***

The correlation between two random variables $X$ and $Y$ is defined by

$$\mathrm{corr}(X, Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

**Example 3.26** Refer to Ex. 3.25. Note that

$$\mathrm{E}X^2 = \int_0^1\int_0^1 x^2(x + y)dxdy = \frac{5}{12}$$

$$\mathrm{E}Y^2 = \int_0^1\int_0^1 y^2(x + y)dxdy = \frac{5}{12},$$

so that $\sigma_X^2 = \mathrm{E}X^2 - (\mathrm{E}X)^2 = 5/12 - (7/12)^2 = 11/144$, and $\sigma_Y^2 = \mathrm{E}Y^2 - (\mathrm{E}Y)^2 = 5/12 - (7/12)^2 = 11/144$. Then the correlation between $X$ and $Y$ is given by

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} = \frac{-1/144}{(11/144)^{1/2}(11/144)^{1/2}} = \frac{-1}{11}. \qquad \square$$

Bounds on the correlation between $X$ and $Y$ follow directly from the bounds on the covariance between $X$ and $Y$.

**Theorem 3.32** (Correlation Bound) $-1 \leq \rho_{XY} \leq 1$.

[19] Recall that a matrix **A** is positive semidefinite iff $\mathbf{t}'\mathbf{A}\mathbf{t} \geq 0\ \forall\ \mathbf{t}$, and **A** is positive definite iff $\mathbf{t}'\mathbf{A}\mathbf{t} > 0\ \forall\ \mathbf{t} \neq [\mathbf{0}]$.

**Proof** This follows directly from Theorem 3.31 via division by $\sigma_X\sigma_Y$. ■

The covariance equals the upper bound of $\sigma_X\sigma_Y$ iff the correlation equals its upper bound of 1, while the covariance equals its lower bound of $-\sigma_X\sigma_Y$ iff the correlation equals its lower bound of $-1$.

Assuming that the covariance exists, a *necessary* condition for the independence of $X$ and $Y$ is that $\sigma_{XY} = 0$ (or equivalently, that $\rho_{XY} = 0$ if $\sigma_X\sigma_Y \neq 0$).

**Theorem 3.33** *If $X$ and $Y$ are independent, then $\sigma_{XY} = 0$ (assuming the covariance exists).*

**Proof** If $X$ and $Y$ are independent, then $f(x, y) = f_X(x)f_Y(y)$. It follows that

$$\text{(discrete)}\quad \sigma_{XY} = \sum_{x\in R(X)}\sum_{y\in R(Y)} (x - \mathrm{E}X)(y - \mathrm{E}Y)f_X(x)f_Y(y)$$
$$= \sum_{x\in R(X)} (x - \mathrm{E}X)f_X(x) \sum_{y\in R(Y)} (y - \mathrm{E}Y)f_Y(y)$$
$$= (\mathrm{E}X - \mathrm{E}X)(\mathrm{E}Y - \mathrm{E}Y) = 0 \cdot 0 = 0.$$

(Continuous) Replace summation with integration above. ■

The converse of Theorem 3.33 is not true—there can be dependence between $X$ and $Y$, even *functional* dependence, and the covariance between $X$ and $Y$ could nonetheless be zero, as the following example illustrates.

**Example 3.27** Let $X$ and $Y$ be two random variables having a joint density function given by $f(x, y) = 1.5I_{[-1,1]}(x)I_{[0,x^2]}(y)$. Note this density implies that $(x, y)$ points are equally likely to occur on and below the parabola represented by the graph of $y = x^2$ (see Figure 3.11). There is a direct functional dependence between $X$ and the range of $Y$, so that $f(y \mid x)$ will change as $x$ changes, and thus $X$ and $Y$ must be dependent random variables. Nonetheless, $\sigma_{XY} = 0$. To see this, note that

$$\mathrm{E}XY = 1.5\int_{-1}^{1}\int_{0}^{x^2} xydydx = 1.5\int_{-1}^{1}(1/2)x^5dx = .75\frac{x^6}{6}\Bigg|_{-1}^{1} = 0,$$

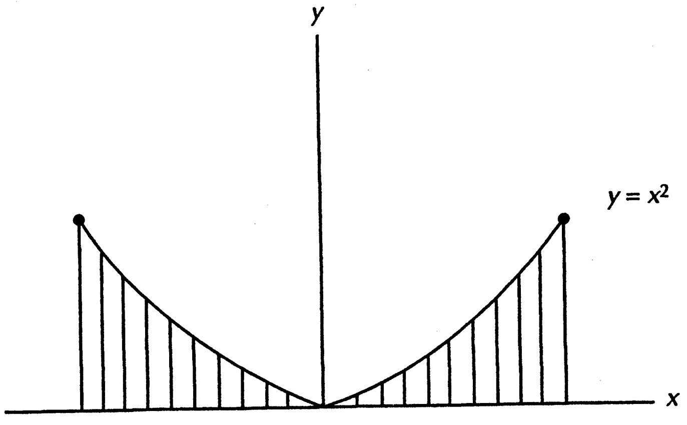


**Figure 3-11**
$y \in [0, x^2]$ for $x \in [-1, 1]$.

$$\mathrm{E}X = 1.5\int_{-1}^{1}\int_{0}^{x^2} xdydx = 1.5\int_{-1}^{1} x^3dx = 1.5\frac{x^4}{4}\Big|_{-1}^{1} = 0,$$

$$\mathrm{E}Y = 1.5\int_{-1}^{1}\int_{0}^{x^2} ydydx = 1.5\int_{-1}^{1}(1/2)x^4dx = .75\frac{x^5}{5}\Big|_{-1}^{1} = .3.$$

Therefore, $\sigma_{XY} = \mathrm{E}XY - \mathrm{E}X\mathrm{E}Y = 0 - 0(.3) = 0.$ □

### Correlation, Linear Association, and Degeneracy

We will now demonstrate that when the covariance takes its maximum absolute value, and thus $\rho_{XY} = +1$ or $-1$, then there is a perfect positive $(\rho_{XY} = +1)$ or negative $(\rho_{XY} = -1)$ *linear* relationship between $X$ and $Y$ that holds with probability one (i.e., $P(y = a + bx) = 1$ or $P(y = a - bx) = 1$). The demonstration is facilitated by the following useful result.

**Theorem 3.34** *Let $Z$ be a random variable for which $\sigma_Z^2 = 0$. Then $P(z = \mathrm{E}Z) = 1$.*

**Proof** Let $g(Z) = (Z - \mathrm{E}Z)^2$. Then

$$P(\mathrm{E}Z - a < z < \mathrm{E}Z + a) = P\left((z - \mathrm{E}Z)^2 < a^2\right) = 1 - P\left((z - \mathrm{E}Z)^2 \geq a^2\right)$$
$$\geq 1 - \sigma_Z^2/a^2,$$

where the inequality is established using Markov's inequality. If $\sigma_Z^2 = 0$, then

$$P(\mathrm{E}Z - a < z < \mathrm{E}Z + a) = 1 \quad \forall\, a > 0,$$

and since only $z = \mathrm{E}Z$ satisfies the inequality $\forall\, a > 0$, $P(z = \mathrm{E}Z) = 1$ when $\sigma_Z^2 = 0$. ■

The result on the linear relationship between $X$ and $Y$ when $\rho_{XY} = +1$ or $-1$, or equivalently, when $\sigma_{XY}$ achieves its upper or lower bound, is as follows.

**Theorem 3.35** (Correlation Bounds and Linearity) *If $\rho_{XY} = +1$ or $-1$, then $P(y = a_1 + bx) = 1$ or $P(y = a_2 - bx) = 1$, respectively, where $a_1 = \mathrm{E}Y - (\sigma_Y/\sigma_X)\mathrm{E}X$, $a_2 = \mathrm{E}Y + (\sigma_Y/\sigma_X)\mathrm{E}X$, and $b = (\sigma_Y/\sigma_X)$.*

**Proof** Define $Z = \lambda_1(X - \mathrm{E}X) + \lambda_2(Y - \mathrm{E}Y)$, and note that $\mathrm{E}Z = 0$. It follows immediately that $\sigma_Z^2 = \mathrm{E}Z^2 = \mathrm{E}[\lambda_1(X - \mathrm{E}X) + \lambda_2(Y - \mathrm{E}Y)]^2 = \lambda_1^2\mathrm{E}(X - \mathrm{E}X)^2 + \lambda_2^2\mathrm{E}(Y - \mathrm{E}Y)^2 + 2\lambda_1\lambda_2\sigma_{XY} \geq 0\ \forall\ \lambda_1, \lambda_2$, which can be represented in matrix terms as

$$\sigma_Z^2 = [\lambda_1\ \lambda_2]\begin{bmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{XY} & \sigma_Y^2\end{bmatrix}\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix} \geq 0 \quad \forall\ (\lambda_1, \lambda_2).$$

If $\rho_{XY} = +1$ or $-1$, then $\sigma_{XY}$ achieves either its (nominal) upper or lower bound, respectively, or equivalently, $\sigma_{XY}^2 = \sigma_X^2\sigma_Y^2$. It follows that the above $2 \times 2$ matrix is singular, since its determinant would be zero. Then the columns of the matrix are linearly dependent, so that there exist nonzero values of $\lambda_1$ and $\lambda_2$ such that

$$\begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

and for these $\lambda$ values, the quadratic form above, and thus $\sigma_Z^2$, achieves the value 0. A solution for $\lambda_1$ and $\lambda_2$ is given by $\lambda_1 = \sigma_{XY}/\sigma_X^2$ and $\lambda_2 = -1$ (which can be validated by substituting these values for $\lambda_1$ and $\lambda_2$ in the linear equations, and noting that $\sigma_Y^2 = \sigma_{XY}^2/\sigma_X^2$ under the prevailing assumptions). Since $\sigma_Z^2 = 0$ at these values of $\lambda_1$ and $\lambda_2$, it follows from Theorem 3.34 that $P(z = 0) = 1$ (recall that $EZ = 0$).

Given the definition of $Z$, substituting the above solution values for $\lambda_1$ and $\lambda_2$ obtains an equivalent probability statement, $P(y = (EY - (\sigma_{XY}/\sigma_X^2)EX) + (\sigma_{XY}/\sigma_X^2)x) = 1$. If $\rho_{XY} = +1$, then $\sigma_{XY} = \sigma_X\sigma_Y$, yielding $P(y = a_1 + bx) = 1$ in the statement of the theorem, while if $\rho_{XY} = -1$, then $\sigma_{XY} = -\sigma_X\sigma_Y$, yielding $P(y = a_2 - bx) = 1$ in the statement of the theorem. ■

The theorem implies that when $\rho_{XY} = +1$ (or $-1$), the event that the outcome of $(X, Y)$ is on a straight line with positive (or negative) slope occurs with probability 1. As a diagrammatic illustration, if $(X, Y)$ is a discrete bivariate random variable, then the situation where $\rho_{XY} = +1$ would be exemplified by Figure 3.12.

Note in Figure 3.12 that $f(x, y)$ assumes positive values only for points along the line $y = a + bx$, reflecting the fact that $P(y = a + bx) = 1$. This situation illustrates what is known as a **degenerate random variable** and a **degenerate density function**. The defining characteristic of a degenerate random variable is that it is an $n$-variate random variable $(X_1, \ldots, X_n)$ whose components satisfy one or more linear functional relationships with probability one, i.e., if $P(a_i + \sum_{j=1}^n b_{ij}x_j = 0) = 1$, for $i = 1, \ldots, m$, then $(X_1, \ldots, X_n)$ is a degenerate random

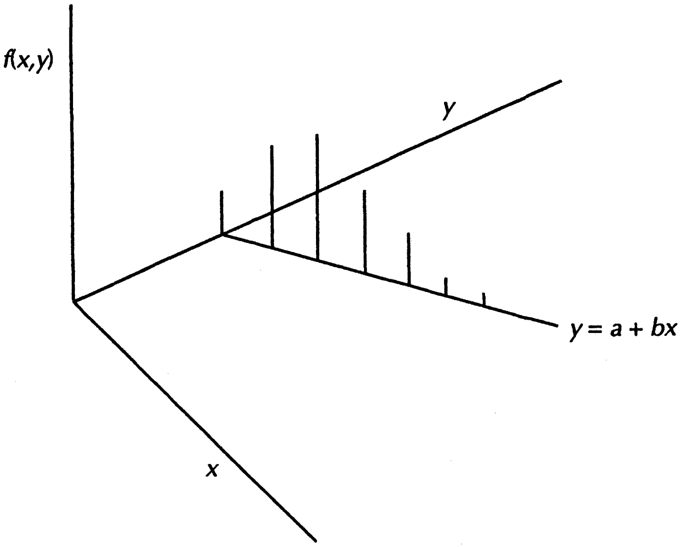


**Figure 3-12**
$\rho_{XY} = +1$, discrete case.

variable.[20] A characteristic of the accompanying degenerate density function for $(X_1, \ldots, X_n)$ is that the entire mass of probability (a mass of 1) is concentrated on a collection of points that lie on a hyperplane of dimension less than $n$, the hyperplane being defined by the collection of linear functional relationships.

Degeneracy causes no particular difficulty in the discrete case—probabilities of events for the degenerate random variable $(X_1, \ldots, X_n)$ can be calculated in the usual way by summing the degenerate density function over the outcomes in the event of interest. However, degeneracy in the continuous case results in $f(x_1, \ldots, x_n)$ *not* being a density function according to our original definition of the concept. For a heuristic description of the problem, examine the diagrammatic illustration in Figure 3.13 for a degenerate bivariate random variable in the continuous case. Intuitively, since there is no volume under the graph of $f(x,y)$, $\int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x,y)dy\,dx = 0 \;\forall\, x_1 \leq x_2$ and $\forall\, y_1 \leq y_2$, and $f(x,y)$ cannot be used in the usual way to assign probabilities to events for $(X, Y)$. However, there *is area* below the graph of $f(x,y)$ and above the line $y = a + bx$ representing the probability mass of 1 distributed over a segment (or perhaps, all) of the line. Since only subsets of the set $\{(x,y) : y = a + bx, x \in R(x)\}$[21] are assigned nonzero probability, the degenerate density function can be used to assign probabilities to events by use of line integrals,[22] which essentially integrates $f(x,y)$ over subsets of points along the line $y = a + bx$. The general concept of line integrals is beyond the scope of our study, but in essence, the relevant integral in the

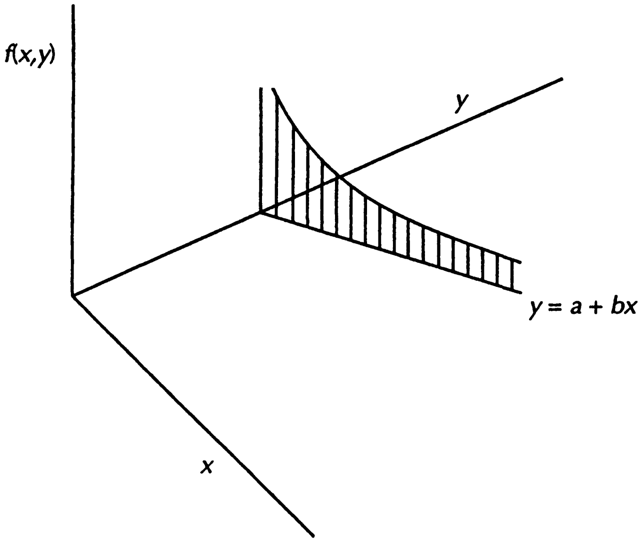


**Figure 3-13**
$\rho_{XY} = +1$, continuous case.

[20] The concept of degeneracy can be extended by calling $(X_1, \ldots, X_n)$ degenerate if the components satisfy one or more functional relationships (not necessarily linear) with probability 1. We will not examine this generalization here.

[21] Equivalently, $\{(x,y) : x = b^{-1}(y - a), y \in R(Y)\}$.

[22] For an introduction to the concept of line integrals, see E. Kreyzig, (1979), *Advanced Engineering Mathematics*, 4th ed. New York: Wiley, Chapter 9.

current context is of the form $\int_{x \in A} f(x, a + bx)dx$. Note the linear relationship linking $y$ and $x$ is explicitly accounted for by substituting $a + bx$ for $y$ in $f(x, y)$, which converts $f(x, y)$ into a function of the single variable $x$. Then the function of $x$ is integrated over the points in the event $A$ for $x$, which determines the probability of the event $B = \{(x, y) : y = a + bx, x \in A\}$ for the bivariate random variable $(X, Y)$.

Having introduced the concept of degeneracy, we can alternatively characterize $\rho_{XY} = +1$ or $-1$ as a case where the bivariate random variable $(X, Y)$, and its accompanying joint density function are *degenerate*, with $X$ and $Y$ satisfying, respectively, a positively sloped or negatively sloped linear functional relationship, with probability one. What can be said about the relationship between $X$ and $Y$ when $| \rho_{XY} |< 1$? The closer $| \rho_{XY} |$ is to one, the closer the relationship between $X$ and $Y$ is to being linear, where "closeness" can be interpreted as follows. Define the random variable $\hat{Y} = a + bX$ to represent predictions of $Y$ outcomes based on a linear function of $X$. We will choose the coefficients $a$ and $b$ so that $\hat{Y}$ is the *best linear* prediction of $Y$, where *best* is taken to mean "minimum expected squared distance between outcomes of $Y$ and outcomes of $\hat{Y}$."

**Theorem 3.36** (Best Linear Prediction of Y Outcomes) *Let $(X,Y)$ have moments of at least the second order, and let $\hat{Y} = a+bX$. Then the choices of a and b that minimize* $\mathrm{E}d^2(Y,\hat{Y}) = \mathrm{E}(Y-(a+bX))^2$ *are given by* $a = \mathrm{E}Y - (\sigma_{XY}/\sigma_X^2)\mathrm{E}X$ *and* $b = (\sigma_{XY}/\sigma_X^2)$.

**Proof** Left to the reader. ■

Now define $V = Y - \hat{Y}$ to represent the deviations between outcomes of $Y$ and outcomes of the *best* linear prediction of $Y$ outcomes, $\hat{Y}$ as defined in Theorem 3.36. Since $\mathrm{E}Y = \mathrm{E}\hat{Y}$, $\mathrm{E}V = 0$. It follows that

$$\sigma_Y^2 = \mathrm{E}(Y - \mathrm{E}Y)^2 = \mathrm{E}(\hat{Y} - \mathrm{E}\hat{Y} + V)^2 = \sigma_{\hat{Y}}^2 + \sigma_V^2 + \sigma_{\hat{Y}V},$$

where

$$\sigma_V^2 = \mathrm{E}V^2 = \mathrm{E}d^2(Y, \hat{Y}) = \mathrm{E}d^2(Y, a + bX) = \sigma_Y^2 - \sigma_{XY}^2/\sigma_X^2 = \sigma_Y^2\left[1 - \rho_{XY}^2\right],$$

$$\sigma_{\hat{Y}}^2 = \mathrm{E}(\hat{Y} - \mathrm{E}\hat{Y})^2 = \sigma_Y^2\rho_{XY}^2,$$

$$\sigma_{\hat{Y}V} = \mathrm{E}((\hat{Y} - \mathrm{E}\hat{Y})V) = (\sigma_{XY}/\sigma_X^2)\,\mathrm{E}((X - \mathrm{E}X)V) = 0.$$

Thus, the variance of $Y$ is decomposed into a proportion $\rho_{XY}^2$ due to $\hat{Y}$ and a proportion $(1 - \rho_{XY}^2)$ due to $V$, i.e.,

$$\sigma_Y^2 = \sigma_{\hat{Y}}^2 + \sigma_V^2 = \sigma_Y^2\rho_{XY}^2 + \sigma_Y^2\left(1 - \rho_{XY}^2\right).$$

We can now interpret values of $\rho_{XY} \in (-1, 1)$. Specifically, $\rho_{XY}^2$ is the proportion of the variance in $Y$ that is explained by the best linear prediction of the form $\hat{Y} = a + bX$, and the proportion of the variance unexplained is $(1 - \rho_{XY}^2)$. Relatedly, $\sigma_Y^2(1 - \rho_{XY}^2)$ is precisely the expected squared distance between outcomes of $Y$ and outcomes of the best linear prediction $\hat{Y} = a + bX$. Thus, the closer | $\rho_{XY}$ | is to 1, the more the variance in $Y$ is explained by the linear function $a+bX$, and the closer is the expected squared distance between $Y$ and $\hat{Y} = a + bX$. It is in this sense that the higher the value of | $\rho_{XY}$ |, the closer is the linear association between $Y$ and $X$. If $\rho_{XY} = 0$, the random variables are said to be **uncorrelated.** In this case, Theorem 3.36 indicates that the best linear predictor is E$Y$—there is effectively no linear association with $X$ whatsoever. The reader should note that $Y$ and $X$ can be interchanged in the preceding argument, leading to an analogous interpretation of the degree of linear association between $X$ and $\hat{X} = a+bY$ (for appropriate changes in the definitions of $a$ and $b$).

## 3.7 Means and Variances of Linear Combinations of Random Variables

It is useful in practice to be able to determine the mean and variance of a random variable that is defined as a linear combination of other random variables. While this determination can be accomplished from first principles by applying the basic definitions of the mean and variance to the linear function of the random variables, there are certain general results that facilitate and expedite the process. In particular, we will see that the mean and variance of a linear combination of random variables can be expressed as simple functions of the means, variances, and covariances of the random variables involved in the linear combination. Our first result concerns the determination of the mean.

**Theorem 3.37** *Let* $Y = \sum_{i=1}^{n} a_i X_i$, *where the* $a_i$*'s are real constants. Then* $\mathrm{E}Y = \sum_{i=1}^{n} a_i \mathrm{E}X_i$.

**Proof**

$$\mathrm{E}Y = \mathrm{E}\sum_{i=1}^{n} a_i X_i = \sum_{i=1}^{n} \mathrm{E}a_i X_i \qquad \text{(Theorem 3.9)},$$

$$= \sum_{i=1}^{n} a_i \mathrm{E}X_i. \qquad \text{(Theorem 3.6)} \qquad \blacksquare$$

Letting

$$\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix},$$

the matrix representation of Theorem 3.37 is

*If* $Y = \mathbf{a}'\mathbf{X}$, *then* $\mathrm{E}Y = \mathrm{E}\mathbf{a}'\mathbf{X} = \mathbf{a}'\mathrm{E}\mathbf{X}$.

Regarding the variance of the linear combination of random variables, we have the following result.

**Theorem 3.38** *Let* $Y = \sum_{i=1}^{n} a_i X_i$ *where the* $a_i$*'s are real constants. Then*

$$\sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_{X_i}^2 + 2\sum_{i<j}\sum a_i a_j \sigma_{X_i X_j}.$$

**Proof**

$$\sigma_Y^2 = \mathrm{E}(Y - \mathrm{E}Y)^2 = \mathrm{E}\left(\sum_{i=1}^{n} a_i(X_i - \mathrm{E}X_i)\right)^2$$

$$= \mathrm{E}\left[\sum_{i=1}^{n} a_i^2(X_i - \mathrm{E}X_i)^2 + 2\sum_{i<j}\sum a_i a_j (X_i - \mathrm{E}X_i)(X_j - \mathrm{E}X_j)\right]$$

$$= \sum_{i=1}^{n} a_i^2\sigma_{X_i}^2 + 2\sum_{i<j}\sum a_i a_j \sigma_{X_iX_j},$$

where the last equality follows from Theorem 3.9 and Theorem 3.6. ■

We can represent the result of Theorem 3.38 very compactly in matrix notation once we define the notion of a **covariance matrix**.

**Definition 3.21**
***Covariance matrix***

The covariance matrix of an $n$-variate random variable $\mathbf{X} = [X_1, \ldots, X_n]'$ is the $n \times n$ symmetric matrix $\mathbf{Cov}(\mathbf{X}) = \mathrm{E}(\mathbf{X} - \mathrm{E}\mathbf{X})(\mathbf{X} - \mathrm{E}\mathbf{X})'$.

In order to appreciate the full informational content of the covariance matrix, note that, by definition,

$$\underset{(n\times n)}{\mathbf{Cov}(\mathbf{X})} = \mathrm{E}\begin{bmatrix} X_1 - \mathrm{E}X_1 \\ \vdots \\ X_n - \mathrm{E}X_n \end{bmatrix}[(X_1 - \mathrm{E}X_1)\cdots(X_n - \mathrm{E}X_n)]$$

$$= \mathrm{E}\begin{bmatrix} (X_1 - \mathrm{E}X_1)^2 & (X_1 - \mathrm{E}X_1)(X_2 - \mathrm{E}X_2) & \cdots & (X_1 - \mathrm{E}X_1)(X_n - \mathrm{E}X_n) \\ (X_2 - \mathrm{E}X_2)(X_1 - \mathrm{E}X_1) & (X_2 - \mathrm{E}X_2)^2 & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ (X_n - \mathrm{E}X_n)(X_1 - \mathrm{E}X_1) & \cdots & & (X_n - \mathrm{E}X_n)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_n} \\ \sigma_{X_2X_1} & \sigma_{X_2}^2 & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ \sigma_{X_nX_1} & \cdots & \cdots & \sigma_{X_n}^2 \end{bmatrix}.$$

Thus, the covariance matrix associated with an $n$-variate random variable has the variance of the $i$th random variable displayed in the $(i, i)$th (diagonal entry) position in the matrix, while the covariance between the $i$th and $j$th random variables is displayed in the $(i, j)$th position (off-diagonal entry) in the matrix. Since $\sigma_{X_iX_j} = \mathrm{E}(X_i - \mathrm{E}X_i)(X_j - \mathrm{E}X_j) = \mathrm{E}(X_j - \mathrm{E}X_j)(X_i - \mathrm{E}X_i) = \sigma_{X_jX_i}$, the covariance matrix is *symmetric*, i.e., the $(i, j)$th entry is exactly equal to the $(j, i)$th entry ∀ $i \neq j$.

Let the $n \times 1$ vectors **a** and **X** be as defined preceding Theorem 3.38. The matrix representation of Theorem 3.38 is then given as follows:

*If* $\mathbf{Y} = \mathbf{a}'\mathbf{X}$, *then* $\sigma^2_{\mathbf{Y}} = \mathbf{a}'\,\mathbf{Cov}(\mathbf{X})\mathbf{a}$.

Note that it is necessarily the case that the *covariance matrix is a positive semidefinite matrix* because $\sigma^2 = \mathbf{a}'\,\mathbf{Cov}(\mathbf{X})\mathbf{a} \geq 0$ for any $a$, which necessarily follows from the fact that variances cannot be negative. (Recall that a matrix **Z** is positive semidefinite iff $\mathbf{a}'\mathbf{Z}\mathbf{a} \geq 0$ for all choices of the vector **a**).

The preceding results can be extended to the case where **Y** is a vector defined by linear combinations of the $n$-variate random variable **X**. We first extend the results corresponding to the mean of **Y**.

**Theorem 3.39** *Let* $\mathbf{Y} = \mathbf{AX}$, *where* **A** *is a* $k \times n$ *matrix of real constants, and* **X** *is an* $n \times 1$ *vector of random variables. Then* $\mathrm{E}\mathbf{Y} = \mathrm{E}\mathbf{AX} = \mathbf{A}\mathrm{E}\mathbf{X}$.

**Proof** The matrix **A** can be represented as a collection of $k$ $(1 \times n)$ row vectors

$$\underset{(k\times n)}{\mathbf{A}} = \begin{bmatrix} a_{1\cdot} \\ \vdots \\ a_{k\cdot} \end{bmatrix} \quad \text{where} \quad \underset{(1\times n)}{a_{i\cdot}} = (a_{i1}, a_{i2}, \ldots, a_{in}).$$

Then

$$\mathrm{E}\mathbf{AX} = \mathrm{E}\begin{bmatrix} a_{1\cdot}\mathbf{X} \\ \vdots \\ a_{k\cdot}\mathbf{X} \end{bmatrix} = \begin{bmatrix} a_{1\cdot}\mathrm{E}\mathbf{X} \\ \vdots \\ a_{k\cdot}\mathrm{E}\mathbf{X} \end{bmatrix} = \mathbf{A}\mathrm{E}\mathbf{X}$$

by Theorem 3.37. ■

A useful corollary to Theorem 3.39 concerns the generalization where **X** is an $n \times \ell$ *matrix* of random variables.

**Corollary 3.6** *Let* $\mathbf{Y} = \mathbf{AX}$, *where* **A** *is a* $k \times n$ *matrix of real constants and* **X** *is an* $n \times \ell$ *matrix of random variables. Then* $\mathrm{E}\mathbf{Y} = \mathbf{A}\mathrm{E}\mathbf{X}$.

**Proof** This follows directly from Theorem 3.39 applied columnwise to the matrix **AX**. ■

If we *postmultiply* rather than *premultiply* a random matrix **X** by a conformable matrix of constants, we obtain a result on expectation qualitatively similar to the preceding result.

**Corollary 3.7** *Let* $\mathbf{Y} = \mathbf{XB}$, *where* **X** *is an* $n \times \ell$ *matrix of random variables and* **B** *is an* $\ell \times m$ *matrix of real constants. Then* $\mathrm{E}\mathbf{Y} = (\mathrm{E}\mathbf{X})\mathbf{B}$.

**Proof**

$$
\begin{aligned}
\mathbf{EY} &= \mathbf{EXB} \\
&= \mathbf{E(B'X')'} && \text{(property of matrix transpose)} \\
&= \mathbf{(EB'X')'} && \text{(E is an elementwise operation)} \\
&= \mathbf{(B'EX')'} && \text{(Corollary 3.6)} \\
&= \mathbf{(EX)B} && \text{(property of matrix transpose)}
\end{aligned}
$$

If a random matrix **X** is *both* premultiplied and postmultiplied by conformable matrices of real constants, then the previous two corollaries can be combined into the following result. ■

**Corollary 3.8** *Let* **A** *be a* $k \times n$ *matrix of real constants, let* **X** *be an* $n \times \ell$ *matrix of random variables, and let* **B** *be an* $\ell \times m$ *matrix of real constants. Then* **EAXB** = **A(EX)B**.

**Proof** Let **Z** = **XB**. Then by Corollary 3.6, **EAXB** = **EAZ** = **AEZ** = **AEXB**, which equals **A(EX)B** by Corollary 3.7. ■

When **Y** = **AX** is a vector of two or more random variables, we can define a variance for each $Y_i$, as well as a covariance for each pair $(Y_i, Y_j)$, $i \neq j$. We are led to a generalization of Theorem 3.38 that involves the definition of the *covariance matrix* of the $k \times 1$ random vector **Y** = **AX**.

**Theorem 3.40** *Let* **Y** = **AX**, *where* **A** *is a* $k \times n$ *matrix of real constants and X is an* $n \times 1$ *vector of random variables. Then* **Cov (Y)** = **Cov (AX)** = **A Cov (X)A′**.

**Proof** By definition,

$$
\begin{aligned}
\mathbf{Cov(Y)} &= \mathbf{E(Y-EY)(Y-EY)'} \\
&= \mathbf{EA(X-EX)(X-EX)'A'} && \text{(substitution and Theorem 3.39)} \\
&= \mathbf{A[E(X-EX)(X-EX)']A'} && \text{(Corollary 3.8)} \\
&= \mathbf{A\,Cov(X)A'} && \text{(by definition).}
\end{aligned}
$$

■

We illustrate the use of some of the above theorems in the following example, where we also introduce the notion of a **correlation matrix** (see part (g)).

**Example 3.28** Your company sells two brands of video cassettes: high quality (HQ) and standard (S). The price of the HQ cassette is \$4, while the standard cassette sells for \$3. The quantities of the cassettes sold on any given day are represented by the bivariate random variable $\mathbf{Q} = (Q_{HQ}, Q_S)'$, where

$$
\mathbf{EQ} = \begin{bmatrix} 10 \\ 30 \end{bmatrix} \quad \text{and} \quad \mathbf{Cov(Q)} = \begin{bmatrix} 2 & -3 \\ -3 & 5 \end{bmatrix}.
$$

**a.** What is the expected value of the revenue obtained from the sale of cassettes on any given day?

**Answer**: Revenue (in dollars) is defined as

$$R = 4Q_{HQ} + 3Q_S = [4 \quad 3]\begin{bmatrix} Q_{HQ} \\ Q_S \end{bmatrix},$$

and Theorem 3.37 applies. Therefore,

$$ER = [4 \quad 3]\begin{bmatrix} 10 \\ 30 \end{bmatrix} = 130.$$

**b.** What is the variance associated with daily revenue?
**Answer**: Theorem 3.38 applies here. We have

$$\sigma_R^2 = [4 \quad 3]\begin{bmatrix} 2 & -3 \\ -3 & 5 \end{bmatrix}\begin{bmatrix} 4 \\ 3 \end{bmatrix} = 5.$$

**c.** Production costs per cassette are \$2.50 and \$2 for the high-quality and standard cassette, respectively. Define the expected value of the vector $\begin{bmatrix} R \\ C \end{bmatrix}$, where $C = 2.50Q_{HQ} + 2Q_S$ represents the total cost of cassettes sold on any given day.
**Answer**: Theorem 3.39 can be used here (we could also apply 3.37 to obtain E$C$, since we already know E$R$ from above).

$$E\begin{bmatrix} R \\ C \end{bmatrix} = \begin{bmatrix} 4 & 3 \\ 2.5 & 2 \end{bmatrix}\begin{bmatrix} 10 \\ 30 \end{bmatrix} = \begin{bmatrix} 130 \\ 85 \end{bmatrix}$$

**d.** What is the covariance matrix of $\begin{bmatrix} R \\ C \end{bmatrix}$?
**Answer**: Using Theorem 3.40,

$$\mathbf{Cov}\begin{bmatrix} R \\ C \end{bmatrix} = \begin{bmatrix} 4 & 3 \\ 2.5 & 2 \end{bmatrix}\begin{bmatrix} 2 & -3 \\ -3 & 5 \end{bmatrix}\begin{bmatrix} 4 & 2.5 \\ 3 & 2 \end{bmatrix} = \begin{bmatrix} 5 & 3.5 \\ 3.5 & 2.5 \end{bmatrix}.$$

**e.** What is the expected level of profit on any given day?
**Answer**: Profit is defined as $\Pi = R - C$, and Theorem 3.37 implies that

$$E\Pi = [1 \quad -1]\begin{bmatrix} 130 \\ 85 \end{bmatrix} = 45.$$

**f.** What is the variance of daily profit?
**Answer**: Applying Theorem 3.38 results in

$$\sigma_\Pi^2 = [1 \quad -1]\begin{bmatrix} 5 & 3.5 \\ 3.5 & 2.5 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = .5.$$

**g.** A matrix of correlations (or, **correlation matrix**) for $\mathbf{X} = (X_1, \ldots, X_n)'$ can be defined by pre- and post-multiplication of the covariance matrix by the inverse of the diagonal matrix of standard deviations, i.e., (reader please verify):

$$\mathbf{Corr}(\mathbf{X}) = \begin{bmatrix} \sigma_{X_1} & & \\ & \ddots & \\ & & \sigma_{X_n} \end{bmatrix}^{-1} \mathbf{Cov}(\mathbf{X}) \begin{bmatrix} \sigma_{X_1} & & \\ & \ddots & \\ & & \sigma_{X_n} \end{bmatrix}^{-1}.$$

The $(i, j)$th entry of the correlation matrix is the correlation between $X_i$ and $X_j$. Define the correlation matrix for **Q**.

**Answer:**

$$\mathbf{Corr(Q)} = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{5} \end{bmatrix}^{-1} \begin{bmatrix} 2 & -3 \\ -3 & 5 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{5} \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 1 & -.949 \\ -.949 & 1 \end{bmatrix}$$

Note that $\rho_{Q_{HQ}, Q_S} = -.949$, which is given by the off-diagonal elements in this $(2 \times 2)$ case, while the diagonal elements are ones because these values represent the correlation of a random variable with itself. □

## 3.8 Necessary and Sufficient Conditions for Positive Semidefiniteness

*Proof that the symmetric matrix* $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ *(recall that symmetry implies* $a_{12} = a_{21}$*) is positive semidefinite iff* $a_{11} \geq 0$, $a_{22} \geq 0$, *and* $a_{11}a_{22} - a_{12}a_{21} \geq 0$.

**Proof** The matrix **A** will be positive semidefinite iff the characteristic roots of **A** are nonnegative (e.g., F. A. Graybill (1983), *Matrices with Applications in Statistics.* Belmont, CA: Wadsworth, p. 397). The characteristic roots of **A** are found by solving the determinantal equation

$$\begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = 0$$

for $\lambda$, which can be represented as $(a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0$ or $\lambda^2 - (a_{11} + a_{22})\lambda + (a_{11}a_{22} - a_{12}a_{21}) = 0$.

Solutions to this equation can be found by employing the quadratic formula[23] to obtain

$$\lambda = \frac{(a_{11} + a_{22}) \pm \sqrt{(a_{11} + a_{22})^2 - 4(a_{11}a_{22} - a_{12}a_{21})}}{2}.$$

For $\lambda$ to be nonnegative, it must be the case that the numerator term $\geq 0$. Note the term under the square root sign *must* be nonnegative, since it can be rewritten as $(a_{11} - a_{22})^2 + 4a_{12}a_{21}$, and because $a_{12} = a_{21}$ (by symmetry of **A**), two nonnegative numbers are being added together. If $(a_{11} + a_{22}) > 0$, then $\lambda \geq 0$ only if $a_{11}a_{22} - a_{12}a_{21} \geq 0$, since otherwise the square root term would be larger than $(a_{11} + a_{22})$, and when subtracted from $(a_{11} + a_{22})$, would result

[23] Recall that the solutions to the quadratic equation $ax^2 + bx + c = 0$ are given by

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

in a negative $\lambda$. Also, the term $(a_{11} + a_{22})$ cannot be negative or else at least one of the solutions for $\lambda$ would necessarily be negative. Furthermore, both $a_{11}$ and $a_{22}$ must be nonnegative, for if one were negative then there is no value for the other that would result in both solutions for $\lambda$ being positive. Thus, necessity is proved. Sufficiency follows immediately, since $a_{11} \geq 0$, $a_{22} \geq 0$, and $a_{11}a_{22} - a_{12}a_{21} \geq 0$ imply both solutions of $\lambda \geq 0$. ■

## Key Words, Phrases, and Symbols

expectation of a random variable, E$X$

expectation of a function of a random variable, E$g(X)$

Jensen's inequality

expectation of a function of a multivariate random variable, E$g(X_1, \ldots, X_n)$

expectation of a matrix of random variables

multivariate Jensen's inequality

conditional expectation $\mathrm{E}(Y \mid x \in B)$, $\mathrm{E}(g(Y) \mid x \in B)$, $\mathrm{E}(Y \mid x = b)$, $\mathrm{E}(g(Y) \mid x = b)$

double expectation theorem

regression function of $Y$ on $X$

regression curve of $Y$ on $X$

moments of a random variable

$\mu'_r$, the $r$th moment about the origin

$\mu$, the mean of a random variable

$\mu_r$, the $r$th moment about the mean or $r$th central moment

variance of a random variable $\sigma^2$, or var$(X)$

standard deviation of a random variable $\sigma$, or std$(X)$

Markov's inequality

Chebyshev's inequality

symmetric density function

skewed density function

skewed to the left

skewed to the right

median, med$(X)$

quantile of $X$

mode, mode$(X)$

unimodal

moment-generating function, MGF, $M_X(t)$

MGF uniqueness theorem

characteristic function

cumulant-generating function $\psi_X(t)$

marginal MGF, marginal cumulant-generating function

cumulants, $\kappa_r$

$\mu'_{r,s}$, the $(r, s)$th joint moment about the origin

$\mu_{r,s}$, the $(r, s)$th joint moment about the mean

covariance between two random variables $\sigma_{XY}$ or cov$(X, Y)$

covariance bound

correlation between two random variables $\rho_{XY}$

correlation bound

degenerate random variable

degenerate density function

uncorrelated

means and variances of linear combinations of random variables $\mathrm{E}Y = \mathrm{E}\mathbf{a}'\mathbf{X} = \mathbf{a}'\mathrm{E}\mathbf{X}$

covariance matrix

$\mathbf{Cov(Y)}$

$\mathbf{Cov(AY) = A\,Cov(Y)A'}$

$\mathbf{EAX = AEX}$

$\mathbf{EXB = (EX)B}$

$\mathbf{EAXB = A(EX)B}$

correlation matrix

$\mathbf{Corr(X)}$

## Problems

**1.** A small domestic manufacturer of television sets places a three-year warranty on its picture tubes. During the warranty period, the manufacturer will replace the television set with a new one if the picture tube fails. The time in years until picture tube failure can be represented as the outcome of a random variable $X$ with probability density function

$$f(x) = .005e^{-.005x} I_{(0,\infty)}(x).$$

The times that picture tubes operate until failure can be viewed as independent random variables. The company sells 100 television sets in a given period.

a. What is the expected number of television sets that will be replaced due to picture tube failure?

b. What is the expected operating life of a picture tube?

**2.** A small rural bank has two branches located in neighboring towns in eastern Washington. The numbers of certificates of deposit that are sold at the branch in Tekoa and the branch in Oakesdale in any given week can be viewed as the outcome of the bivariate random variable $(X, Y)$ having joint probability density function

$$f(x, y) = \left[\frac{x^3y^3}{(3025)^2}\right] I_{\{0,1,2,\ldots,10\}}(x) I_{\{0,1,2,\ldots,10\}}(y).$$

a. Are the random variables independent?

b. What is the expected number of certificate sales by the Oakesdale branch?

c. What is the expected number of combined certificate sales for both branches?

d. What is the answer to b. *given that* Tekoa branch sells 4 certificates? Potentially helpful result:

$$\sum_{x=1}^{n} x^3 = \frac{n^2(n+1)^2}{4}.$$

**3.** The weekly number of luxury and compact cars sold by "Honest" Abe Smith at the Auto Mart, a local car dealership, can be represented as the outcome of a bivariate random variable $(X, Y)$ with the nonzero values of its joint probability density function given by

| | | | | $Y$ | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| $X$ | 0 | .20 | .15 | .075 | .05 | .03 |
| | 1 | .10 | .075 | .04 | .03 | .02 |
| | 2 | .05 | .03 | .02 | .01 | .01 |
| | 3 | .04 | .03 | .02 | .01 | .01 |

Al receives a base salary of \$100/week from the dealership, and also receives a commission of \$100 for every compact car sold and \$200 for every luxury car sold.

a. What is the expected value of the weekly commission that Al obtains from selling cars? What is the expected value of his total pay received for selling cars?

b. What is the expected value of his commission from selling compact cars? What is the expected value of his commission from selling luxury cars?

c. Given that Al sells four compact cars, what is the expected value of his commission from selling luxury cars?

d. If 38 percent of Al's total pay goes to federal and state taxes, what is the expected value of his pay after taxes?

**4.** The yield, in bushels per acre, of a certain type of feed grain in the midwest can be represented as the outcome of the random variable $Y$ defined by

$$Y = 3x_l^{.30}x_k^{.45}e^U,$$

where $x_l$ and $x_k$ are the per acre units of labor and capital utilized in production, and $U$ is a random variable with probability density function given by

$$f(u) = 2e^{-2u}I_{(0,\infty)}(u).$$

The price received for the feed grain is \$4/bushel, labor price per unit is \$10, and capital price per unit is \$15.

a. What is the expected yield per acre?

b. What is the expected level of profit per acre if labor and capital are each applied at the rate of 10 units per acre?

c. Define the levels of input usage that maximize expected profit. What is the maximum expected level of profit?

d. The acreage can be irrigated at a cost of \$125 per acre, in which case the yield per acre is defined by

$$Y = 5x_l^{.30}x_k^{.45}e^U.$$

If the producer wishes to maximize expected profit, should she irrigate?

**5.** The daily price per gallon and quantity sold (measured in *millions* of gallons) of unleaded gasoline on the wholesale spot market of a major commodity exchange is the outcome of a bivariate random variable $(P, Q)$ having the joint probability density function

$$f(p, q) = 2pe^{-pq}I_{[.5,1]}(p)I_{(0,\infty)}(q).$$

a. Define the regression curve of $q$ on $p$.

b. Graph the regression curve that you have defined in (a).

c. What is the expected value of the quantity of gasoline sold, given that price is equal to \$.75 per gallon?

d. What is the expected value of total dollar sales of gasoline on a given day?

**6.** The short-run production function for a particular agricultural crop is critically dependent on the level of rainfall during the growing season, the relationship being $Y = 30 + 3X - .075X^2$, where $y$ is yield per acre in bushels, and $x$ is inches of rainfall during the growing season.

a. If the expected value of rainfall is 20 inches, can the expected value of yield per acre be as high as 70 bushels per acre? Why or why not?

b. Suppose the variance of rainfall is 40 square inches. What is the expected value of yield per acre? How does this compare to the bound placed on $EY$ by Jensen's inequality?

**7.** For each of the densities below, indicate whether the mean and variance of the associated random variable exist. In addition, find the median and mode, and indicate whether or not each density is symmetric.

a. $f(x) = 3x^2 I_{[0,1]}(x)$

b. $f(x) = 2x^{-3} I_{[1,\infty)}(x)$

c. $f(x) = [\pi(1+x^2)]^{-1} I_{(-\infty,\infty)}(x)$

d. $f(x) = \binom{4}{x}(.2)^x(.8)^{4-x} I_{\{0,1,2,3,4\}}(x)$

**8.** The daily price of a certain penny stock is a random variable with an expected value of \$2. Then the probability $\leq .20$ that the stock price will be greater than or equal to \$10. True or false?

**9.** The miles per gallon attained by purchasers of a line of pickup trucks manufactured in Detroit are outcomes of a random variable with a mean of 17 miles per gallon and a standard deviation of .25 miles per gallon. How probable is the event that a purchaser attains between 16 and 18 miles per gallon with this line of truck?

**10.** The daily quantity of water demanded by the population of a large northeastern city in the summer months is the outcome of a random variable, $X$, measured in millions of gallons and having an MGF of $M_X(t) = (1 - .5t)^{-10}$ for $t < 2$.

a. Find the mean and variance of the daily quantity of water demanded.

b. Is the density function of water quantity demanded symmetric?

**11.** The annual return per dollar for two different investment instruments is the outcome of a bivariate random variable $(X_1, X_2)$ with joint moment-generating function $M_X(t) = \exp(u't + .5t'\Phi t)$, where

$$t = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}, \quad u = \begin{bmatrix} .07 \\ .11 \end{bmatrix}, \quad \text{and}$$

$$\Phi = \begin{bmatrix} .225 \times 10^{-3} & -.3 \times 10^{-3} \\ -.3 \times 10^{-3} & .625 \times 10^{-3} \end{bmatrix}.$$

a. Find the mean annual return per dollar for each of the projects.

b. Find the covariance matrix of $(X_1, X_2)$.

c. Find the correlation matrix of $(X_1, X_2)$. Do the outcomes of $X_1$ and $X_2$ satisfy a linear relationship $X_1 = \alpha_1 + \alpha_2 X_2$?

d. If an investor wishes to invest \$1,000 in a way that maximizes her expected dollar return on the investment, how should she distribute her investment dollars between the two projects? What is the variance of dollar return on this investment portfolio?

e. Suppose the investor wants to minimize the variance of her dollar return. How should she distribute the \$1,000? What is the expected dollar return on this investment portfolio?

f. Suppose the investor's utility function with respect to her investment portfolio is $U(M) = 5M^b$, where $M$ is the dollar return on her investment of \$1,000. The investor's objective is to maximize the expected value of her utility. If $b = 1$, define the optimal investment portfolio.

g. Repeat (f), but let $b = 2$.

h. Interpret the investment behavior differences in (f) and (g) in terms of investor attitude toward risk.

**12.** Stanley Statistics, an infamous statistician, wants you to enter a friendly wager with him. For \$1,000, he will let you play the following game. He will continue to toss a fair coin until the first heads appears. Letting $x$ represent the number of times the coin was tossed to get the first heads, Stanley will then pay you $\$2^x$.

a. Define a probability space for the experiment of observing how many times a coin must be tossed in order to observe the first heads.

b. What is the expected payment that you will receive if you play the game?

c. Do you want to play the game? Why or why not?

**13.** The city of Megalopolis operates three sewage treatment plants in three different locations throughout the city. The daily proportion of operating capacity exhibited by the three plants can be represented as the outcome of a trivariate random variable with the following probability density function:

$$f(x_1, x_2, x_3) = \frac{1}{3}(x_1 + 2x_2 + 3x_3) \prod_{i=1}^{3} I_{[0,1]}(x_i),$$

where $x_i$ is the proportion of operating capacity exhibited by plant $i$, $i = 1, 2, 3$.

a. What are the expected values of the capacity proportions for the three plants, i.e., what is $\mathrm{E}\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$?

b. What is the expected value of the average proportion of operating capacity across all three plants, i.e., what is $\mathrm{E}\left(\frac{1}{3}\sum_{i=1}^{3} X_i\right)$?

c. *Given* that plant 3 operates at 90 percent of capacity, what are the expected values of the proportions of capacity for plants 1 and 2?

d. If the daily capacities of plants 1 and 2 are 100,000 gallons of sewage each, and if the capacity of plant 3 is 250,000 gallons, then what is the expected daily number of gallons of sewage treated by the city of Megalopolis?

**14.** The average price and total quantity sold of an economy brand of ballpoint pen in a large western retail market during a given sales period is represented by the outcome of a bivariate random variable having a probability density function

$$f(p, s) = 10pe^{-ps}I_{[.10,.20]}(p)I_{[0,\infty)}(s),$$

where $p$ is the average price, in dollars, of a single pen and $s$ is the total quantity sold, measured in 10,000-pen units.

a. Define the regression curve of $S$ on $P$.

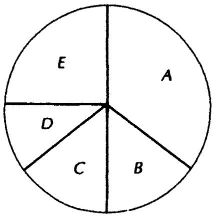
b. What is the expected quantity of pens sold, given that price is equal to \$0.12? (You may use the regression curve if you wish.)

c. What is the expected value of total revenue from the sale of ball point pens during the given sales period, i.e., what is $E(PS)$?

**15.** A game of chance is considered to be "equitable" or "fair" if a player's expected payoff is equal to zero. Examine the following games:

a. The player rolls a pair of fair dice. Let $Z$ represent the amount of money that the player bets on the game outcome. If the player rolls a 7 or 11, the player payoff is $2Z$ (i.e., he gets to keep his bet of \$$Z$, plus he receives an *additional* \$$2Z$). If the player does not roll a 7 or 11, he loses the \$$Z$ that he bet on the game. Is the game fair?

b. The player spins a spinner contained within a disk, , that is segmented into five pieces as




where $P(A) = 1/3$, $P(B) = 1/6$, $P(C) = 1/6$, $P(D) = 1/12$, $P(E) = 1/4$.

Each spin costs \$1. The payoffs corresponding to when the spinner lands in one of the five segments are given by:

| Segment | Payoff |
|---|---|
| A | \$0.60 |
| B | \$1.20 |
| C | \$1.20 |
| D | \$2.40 |
| E | \$0.80 |

Is the game fair?

c. A fair coin will be tossed repeatedly *until* heads occurs. If the heads occurs on the $j$th toss of the coin, the player will receive \$$2^j$. How much should the player be charged to play the game if the game is to be fair? (Note: This is a trick question and represents the famous "St. Petersburg paradox" in the statistical literature.)

**16.** The manager of a bakery is considering how many chocolate cakes to bake on any given day. The manager knows that the number of chocolate cakes that will be demanded by customers on any given day is a random variable whose probability density is given by

$$f(x) = \frac{x+1}{15}I_{\{0,1,2,3\}}(x) + \frac{7-x}{15}I_{\{4,5\}}(x).$$

The bakery makes a profit of \$1.50 on each cake that is sold. If a cake is not sold on a given day, the cake is thrown away (because of lack of freshness), and the bakery loses \$1. If the manager wants to maximize expected daily profit from the sale of chocolate cakes, how many cakes should be baked?

**17.** The daily price and quantity sold of wheat in a Northwestern market during the first month of the marketing year is the outcome of a bivariate random variable $(P, Q)$ having the probability density function

$$f(p, q) = .5pe^{-pq}I_{[3,5]}(p)I_{[0,\infty)}(q),$$

where $p$ is measured in \$/bushel, and $q$ is measured in units of 100,000 bushels.

a. Define the conditional expectation of quantity sold as a function of price, i.e., define $E(Q \mid p)$ (the regression curve of $Q$ on $P$).

b. Graph the regression curve you derived in (a). Calculate the values of $E(Q \mid p = 3.50)$ and $E(Q \mid p = 4.50)$.

**18.** In each case, calculate the expected value of the random variable $Y$:

a. $\mathrm{E}(Y \mid x) = 2x^2 + 3$, $f_X(x) = e^{-x} I_{[0,\infty)}(x)$.

b. $\mathrm{E}(Y \mid x) = 3x_1x_2$, $\mathrm{E}X_1 = 5$, $\mathrm{E}X_2 = 7$, $X_1$ and $X_2$ are independent.

**19.** The total daily dollar sales in the ACME supermarket is represented by the outcome of the random variable $S$ having a mean of 20, where $s$ is measured in thousands of dollars.

a. The store manager tells you the probability that sales will exceed $30,000 on any given day is .75. Do you believe her?

b. You are now given the information that the variance of the random variable $S$ is equal to 1.96. How probable is the sales event that $s \in (10, 30)$?

**20.** The first three moments about the *origin* for the random variable $Y$ are given as follows: $\mu'_1 = .5$, $\mu'_2 = .5$, $\mu'_3 = .75$.

a. Define the first three moments about the *mean* for $Y$.

b. Is the density of $Y$ skewed? Why or why not?

**21.** The random variable $Y$ has the PDF $f(y) = y^{-2} I_{[1,\infty)}(y)$.

a. Find the mean of $Y$.

b. Find the first *100* moments about the origin (i.e., $\mu'_1, \mu'_2, \ldots, \mu'_{100}$) for the random variable $Y$.

**22.** The moment-generating function of the random variable $Y$ is given by $M_Y(t) = (1 - .25t)^{-3}$ for $t < 4$.

a. Find the mean and variance of the random variable $Y$.

b. Is the PDF of $Y$ skewed? Why or why not?

c. It is known that the moment generating function of the PDF

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} I_{(0,\infty)}(x)$$

is given by $M_X(t) = (1 - \beta t)^{-\alpha}$ for $t < \beta^{-1}$. The $\Gamma(\alpha)$ in the preceding expression for the pdf is known as the *gamma function*, which for integer values of $\alpha$ is such that $\Gamma(\alpha) = (\alpha - 1)!$. Define the exact functional form of the probability density function for $Y$, if you can.

**23.** A gas station sells regular and premium fuel. The two storage tanks holding the two types of gasoline are refilled every week. The proportions of the available supplies of regular and premium gasoline that are sold during a given week in the summer is an outcome of a bivariate random variable having the joint density function

$$f(x, y) = \tfrac{2}{5}(3x + 2y) I_{[0,1]}(x) I_{[0,1]}(y),$$

where $x$ = proportion of regular fuel sold and $y$ = proportion of premium fuel sold.

a. Find the marginal density function of $X$. What is the probability that greater than 75 percent of the available supply of regular fuel is sold in a given week?

b. Define the regression curve of $Y$ on $X$, i.e., define $\mathrm{E}(Y \mid x)$. What is the expected value of $Y$, given that $x = .75$? Are $Y$ and $X$ independent random variables?

c. Regular gasoline sells for $1.25/gallon and premium gasoline sells for $1.40/gallon. Each storage tank holds 1,000 gallons of gasoline. What is the expected revenue generated by the sale of gasoline during a week in the summer, *given* that $x = .75$?

**24.** Scott Willard, a famous weatherman on national TV, states that the temperature on a typical late fall day in the upper midwest, measured in terms of both the Celsius and Fahrenheit scales, can be represented as the outcome of the bivariate random variable $(C, F)$ such that

$$\mathrm{E}\begin{bmatrix} C \\ F \end{bmatrix} = \begin{bmatrix} 5 \\ 41 \end{bmatrix} \quad \text{and} \quad \mathrm{cov}(C, F) = \begin{bmatrix} 25 & 45 \\ 45 & 81 \end{bmatrix}.$$

a. What is the correlation between $C$ and $F$?

b. To what extent is there a linear relationship between $C$ and $F$? Define the appropriate linear relationship if it exists.

c. Is $(C, F)$ a degenerate bivariate random variable? Is this a realistic result? Why or why not?

**25.** A fruit processing firm is introducing a new fruit drink, "Peach Passion," into the domestic market. The firm faces uncertain output prices in the initial marketing period and intends to make a short-run decision by choosing the level of production that maximize the expected value of utility:

$$\mathrm{E}(U(\pi)) = \mathrm{E}\pi - \alpha\,\mathrm{var}(\pi).$$

Profit is defined by $\pi = Pq - C(q)$, $p$ is the price received for a unit of Peach Passion, $U$ is utility, the cost function is defined by $c(q) = .5q^2$, $\alpha \geq 0$ is a risk aversion parameter, and the PDF of the uncertain output price is given by

$$f(p) = .048(5p - p^2)I_{[0,5]}(p).$$

a. If the firm were risk neutral, i.e., $\alpha = 0$, find the level of production that maximizes expected utility.

b. Now consider the case where the firm is risk averse, i.e., $\alpha > 0$. Graph the relationship between the optimal level of output and the level of risk aversion (i.e., the level of $\alpha$). How large does $\alpha$ have to be for optimal $q = 1$?

c. Assume that $\alpha = 1$. Suppose that the Dept. of Agriculture were to *guarantee* a price to the firm. What *guaranteed* price would induce the firm to produce the same level of output as in the case where price was uncertain?

**26.** A Seattle newspaper intends to administer two different surveys relating to two different anti-tax initiatives on the ballot in November. The proportion of surveys mailed that will actually be completed and returned to the newspaper can be represented as the outcome of a bivariate random variable $(X, Y)$ having the density function

$$f(x, y) = \frac{2}{3}(x + 2y)I_{[0,1]}(x)I_{[0,1]}(y),$$

where $x$ is the proportion of surveys relating to initiative I that are returned, and $y$ refers to the proportion of surveys relating to initiative II that are returned.

a. Are $X$ and $Y$ independent random variables?

b. What is the conditional distribution of $x$, given $y = .50$? What is the probability that less than 50 percent of the initiative I surveys are returned, given that 50 percent of the initiative II surveys are returned?

c. Define the regression curve of $X$ on $Y$. Graph the regression curve. What is the expected proportion of initiative I surveys returned, given that 50 percent of the initiative II surveys are returned?

**27.** An automobile dealership sells two types of four-door sedans, the "Land Yacht" and the "Mini-Rover." The number of Land Yachts and Mini-Rovers sold on any given day varies, with the probabilities of the various possible sales outcomes given by the following table:

| | | Number of Land Yachts Sold | | | |
|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 |
| Number of Mini-Rovers Sold | 0 | .05 | .05 | .02 | .02 |
| | 1 | .03 | .10 | .08 | .03 |
| | 2 | .02 | .15 | .15 | .04 |
| | 3 | .01 | .10 | .10 | .05 |

Land Yachts sell for \$22,000 each, and Mini-Rovers for \$7,500 each. These cars cost the dealership \$20,000 and \$6,500, respectively, which must be paid to the car manufacturer.

a. Define a random variable that represents daily profit above dealer car cost, i.e., total dollar sales − total car cost. (Let $x$ = number of Land Yachts sold and $y$ = number of Mini-Rovers sold). What is the expected value of daily profit above dealer car cost?

b. The daily cost (other than the cost of cars) of running the dealership is equal to \$4,000. What is the probability that total profit on a given day will be positive?

c. What is the expected number of Mini-Rovers sold on a day when no Land Yachts will be sold? What is this expected number on a day when two Land Yachts are sold? Are $X$ and $Y$ independent random variables? Why or why not?

**28.** The season average price per pound, $p$, and total season quantity sold, $q$, of sweet cherries in a regional market can be represented as the outcome of a bivariate random variable $(P, Q)$ with the joint probability density function

$$f(p, q) = .5qe^{-q(.5+p)}I_{(0,\infty)}(q)I_{(0,\infty)}(p),$$

where $p$ is measured in dollars, and $q$ is measured in millions of pounds.

a. Find the marginal density of $Q$. What is the expected value of quantity sold?

b. Define the regression curve of $P$ on $Q$. What is the expected value of $P$, given that $q = 1/2$?

c. If the government receives 10 percent of the gross sales of sweet cherries every season, what is the expected value of the revenue collected by the government from the sale of sweet cherries given that $q = 1/2$?

Hints $\int xe^{ax}dx = (e^{ax}/a^2)(ax - 1).$

**29.** The yield/acre of wheat on a given parcel of land can be represented as the outcome of a random variable

$Y$ defined by

$$Y = 10x^{1/3}e^{\varepsilon} \quad \text{for } x \in [8, 100],$$

where

$y$ = wheat output in bushes/acre

$x$ = pounds/acre of fertilizer applied

$\varepsilon$ = is a random variable having the probability density function,

$$f(\varepsilon) = 3e^{-3\varepsilon}I_{(0,\infty)}(\varepsilon).$$

a. If fertilizer is applied at the rate of 27 pounds per acre, what is the probability that greater than 50 bushels/acre of wheat will be produced?

b. You sign a forward contract to sell your wheat for \$3.00/bushel at harvest time. Fertilizer costs \$0.20/pound. If you apply fertilizer at a rate of 27 pounds/acre, what is your expected return over fertilizer cost, per acre?

c. What is the variance of $Y$ if fertilizer is applied at the rate of 27 pounds per acre? Does the variance of $Y$ change if a different level of fertilizer is applied? Why or why not?

**30.** Let $X$ have the moment generating function $M_X(t)$. Show that

a. $M_{(X+a)}(t) = e^{at}M_X(t)$.

b. $M_{bX}(t) = M_X(bt)$.

c. $M_{(X+a)/b}(t) = e^{(a/b)t}M_X(t/b)$.

**31.** The AJAX Disk Co. manufactures compact discs (CDs) for the music industry. As part of its quality-control program, the diameter of each disc is measured using an electronic measuring device. Letting $X_1$ represent the actual diameter of the disc and $X_2$ represent the *measured* diameter of the disc,

$$\mathrm{E}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 4.6775 \\ 4.6775 \end{bmatrix},$$

$$\mathbf{Cov(X)} = \begin{bmatrix} .00011 & .00010 \\ .00010 & .00010 \end{bmatrix},$$

where $x_1$ and $x_2$ are measured in inches.

a. What is the correlation between the actual diameter and the measured diameter of the CDs?

b. Assume that $P(x_1 \in [4.655, 4.700]) = 1$. Use a graph to elucidate the degree of linear association between $X_1$ and $X_2$ that is represented by the correlation value you calculated in (a).

c. Given the characteristics of $(X_1, X_2)$ indicated above, the manager of the quality control-department states that the difference between measured and actual disk diameters is no more than .01 inches with probability $\geq .90$. Do you agree? Why or why not?

**32.** An investor wishes to invest \$1,000 and is examining two investment prospects. The net dollar return per dollar invested in the two projects can be represented as the outcome of a bivariate random variable $(X_1, X_2)$ where

$$\mathrm{E}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} .15 \\ .07 \end{bmatrix} \quad \text{and} \quad \mathbf{Cov(X)} = \begin{bmatrix} .04 & -.001 \\ -.001 & .0001 \end{bmatrix}.$$

a. If the investor invests \$500 in each project, what is his/her expected net dollar return? What is the variance associated with the net dollar return?

b. Suppose the investor wishes to invest the \$1,000 so that his/her expected utility is maximized, where $\mathrm{E}(U(R)) = \mathrm{E}R - .01\,\mathrm{var}(R)$, $R = \alpha_1 X_1 + \alpha_2 X_2$ represents the total return on the investment, $\alpha_1 + \alpha_2 = 1{,}000$, and $\alpha_i \geq 0$ for $i = 1, 2$. How much money should he/she invest in each of the projects?

**33.** The length of time in minutes for an individual to be served at a local restaurant is the outcome of a random variable, $T$, having a mean of 6 and a variance of 1.5. How probable is the event that an individual will be served within 3 to 9 minutes?

**34.**

a. Find the moment-generating function of a random variable $X$ having the density function

$$f(x) = \frac{1}{8}\binom{3}{x}I_{\{0,1,2,3\}}(x).$$

(Hint: Use of the binomial theorem may be helpful in finding a compact representation of this function.)
Use the MGF to calculate the first two moments of $X$ about the origin. Calculate the variance of $X$.

b. Repeat (a), using the density function

$$f(x) = \frac{1}{10}e^{-x/10}I_{(0,\infty)}(x).$$

# 4 Parametric Families of Density Functions

Some specific functional forms of probability density functions that have been found to be useful in statistical applications are examined in this chapter. The selection includes a number of the more commonly used densities.[1] Our density function definitions will actually identify **parametric families** of density functions. That is, the algebraic expressions for the density functions will contain one or more unknowns, called **parameters**, which can be assigned values chosen from a set of admissible values called the **parameter space**. A *specific member* of a family of densities will be associated with each specific value of the parameters contained in the parameter space. The general notation $f(x; \theta)$ will be used to distinguish elements in the domain of the density function from elements in the parameter space of the parametric family of functions. In particular, the argument preceding the semicolon ($x$ in the present case) represents domain elements, while the argument following the semicolon ($\theta$ in this case) represents elements of the parameter space.

For each parametric family of densities examined, we will present a particular **parameterization** of the family that identifies the parameters used in the algebraic representation of the density functions as well as the collection of admissible values for the parameters (the latter collection being the afore-

[1] Johnson and Kotz provide an extensive survey of density functions that have been used in statistical applications. Norman L. Johnson and Samuel Kotz (1969, 1970, 1972), *Distribution in Statistics*, Vol. I–IV. New York: John Wiley and Sons.

mentioned parameter space). We will use various English and Greek letters to represent parameters. The parameter space will be generically represented by the Greek capital *omega*, $\Omega$. Be aware that, in general, parameterizations are *not unique*. Generally, the collection of densities in a parametric family can be **reparameterized**, meaning that an alternative set of parameters and an associated parameter space can be defined that equivalently identifies each and every density in a family of density functions. Possibilities for reparameterizations will not concern us currently, although we will revisit this issue later when we examine methods of statistical inference.

Each parametric family has its own distinguishing characteristics that make the PDFs appropriate candidates for specifying the probability space of some experiments and inappropriate for others. The characteristics include whether the PDFs are discrete or continuous, whether the use of the PDFs are restricted to nonnegative-valued and/or integer-valued random variables, whether the densities in the family are symmetric or skewed, and the degree of flexibility with which the density can distribute probability over events in the range of a random variable. Furthermore, the functional forms of some parametric families of densities follow deductively from the characteristics of certain types of experiments. We will point out major characteristics and application contexts for each of the parametric families presented. We will also introduce procedures for assessing the adequacy of the choice of a particular family for a given application later in Chapter 10.

## 4.1 Parametric Families of Discrete Density Functions

**Family Name: Uniform**

*Parameterization* $N \in \Omega = \{N : N \text{ is a positive integer}\}$

*Density Definition* $f(x; N) = \frac{1}{N} I_{\{1,2,\ldots,N\}}(x)$

*Moments* $\mu = (N+1)/2, \sigma^2 = (N^2-1)/12, \mu_3 = 0$

*MGF* $M_X(t) = \sum_{j=1}^{N} e^{jt}/N$

*Background and Application* The discrete uniform density function assigns equal probability to each of $N$ possible outcomes of an experiment. The density is used to construct a probability space for any experiment having $N$ possible outcomes that are all equally likely. The outcomes are coded $1, 2, \ldots, N$, and each outcome is assigned probability $f(x; N) = 1/N$.

**Example 4.1** The office manager has a box of six printer cables with different pin configurations to accommodate computer hook-up of the various types of printers used by the company. A new employee needs a printer cable to hook up her printer. She randomly chooses printer cables from the box, one at a time without replacement, in an attempt to hook up her printer. Letting the outcome of

$X$ denote the number of cables tried before the correct one is found, what is the probability density of $X$?

**Answer**: Let $A_i$ be the event that the correct cable is chosen on the $i$th try. Then

$$f(1) = P(A_1) = \frac{1}{6} \qquad \text{(classical probability)}$$

$$f(2) = P(A_2 \mid \bar{A}_1)P(\bar{A}_1) = \left(\frac{1}{5}\right)\left(\frac{5}{6}\right) = \frac{1}{6} \qquad \text{(multiplication rule)},$$

$$f(3) = P(A_3 \mid \bar{A}_1 \cap \bar{A}_2)P(\bar{A}_2 \mid \bar{A}_1)P(\bar{A}_1) \qquad \text{(extended multiplication rule)},$$

$$= \left(\frac{1}{4}\right)\left(\frac{4}{5}\right)\left(\frac{5}{6}\right) = \frac{1}{6},$$

$$\vdots$$

$$f(6) = (1)\left(\frac{1}{2}\right)\left(\frac{2}{3}\right)\left(\frac{3}{4}\right)\left(\frac{4}{5}\right)\left(\frac{5}{6}\right) = \frac{1}{6}.$$

Thus, $X$ has the uniform distribution with $N = 6$ (see Figure 4.1). □

**Family Name: Bernoulli**

*Parameterization* $p \in \Omega = \{p : 0 \le p \le 1\}$

*Density Definition* $f(x; p) = p^x(1-p)^{1-x} I_{\{0,1\}}(x)$

*Moments* $\mu = p, \sigma^2 = p(1-p), \mu_3 = 2p^3 - 3p^2 + p$

*MGF* $M_X(t) = pe^t + (1-p)$

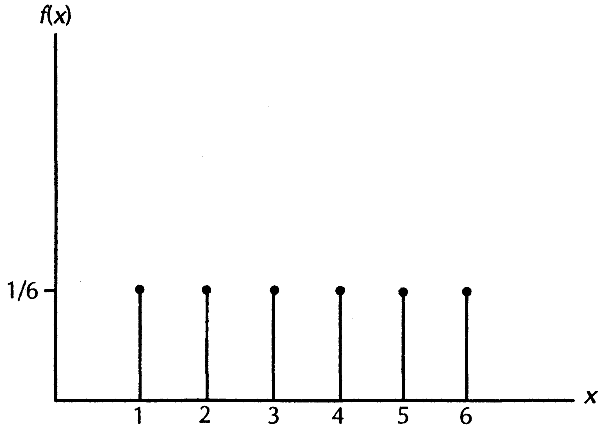


**Figure 4-1**
Uniform density, $N = 6$.

*Background and Application* The Bernoulli density, named after the Swiss mathematician Jacques Bernoulli (1654–1705), can be used to construct a probability space for an experiment that has two possible outcomes (e.g., cure versus no cure, defective versus nondefective, success versus failure) that *may* or *may not* be equally likely to occur. The two outcomes of the experiment are coded 0 and 1, where the event $x = 0$ is assigned probability $f(0; p) = 1 - p$ and the event $x = 1$ is assigned probability $f(1; p) = p$.

**Example 4.2** A shipment of video recorders to a local electronics retailer contains three defective and seven nondefective recorders. The recorders are placed on the store's shelves, and a customer randomly chooses a recorder to purchase. Coding the choice of a defective recorder as $x = 1$ and the choice of a nondefective recorder as $x = 0$, the Bernoulli density can be used to construct the probability space of the experiment by letting $p = .3$ (see Figure 4.2). □

**Family Name: Binomial**

*Parameterization* $(n, p) \in \Omega = \{(n, p) : n \text{ is a positive integer}, 0 \leq p \leq 1\}$

*Density Definition*

$$f(x; n, p) = \begin{cases} \dfrac{n!}{x!(n-x)!} p^x (1-p)^{n-x} & \text{for } x = 0, 1, 2, \ldots, n, \\ 0 \quad \text{otherwise} \end{cases}$$

*Moments* $\mu = np, \sigma^2 = np(1-p), \mu_3 = np(1-p)(1-2p)$

*MGF* $M_X(t) = (1 - p + pe^t)^n$

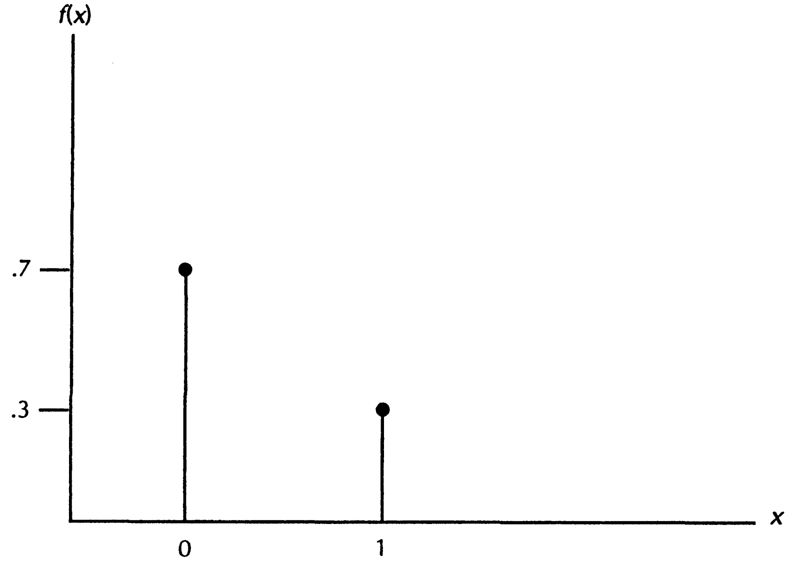


**Figure 4-2**
Bernoulli density, $p = .3$

*Background and Application* The binomial density function is used to construct a probability space for an experiment that consists of $n$ independent repetitions (also called *trials*) of a given experiment of the Bernoulli type (i.e., the experiment has two possible outcomes), with the observation of interest being *how many of the n Bernoulli trials result in one of the two types of outcomes*, say, type $A$ (e.g., how many successes, defectives, or cures occur in $n$ repetitions of the Bernoulli-type experiment?). The value of $x$ represents the total number of outcomes of type $A$ that occur in the n Bernoulli trials. The parameters $n$ and $p$ refer respectively to the number of trials and the probability of observing the type A outcome in the underlying Bernoulli-type experiment. It is assumed that the repetitions are executed in such a way that the outcome observed on any trial does not affect the probability of occurrence of outcomes on any other trial (which is the meaning of the phrase *independent repetitions* or *independent trials*).

The functional form of the density can be deduced directly from the characteristics of the experiment described above. Let $(X_1, \ldots, X_n)$ be a collection of $n$ independent random variables, where each $X_i$ has a Bernoulli density with the same value of $p$, as $f(x_i; p) = p^{x_i}(1 - p)^{1-x_i}I_{\{0,1\}}(x_i)$. Let $x_i = 1$ indicate that outcome $A$ occurs on the $i$th Bernoulli trial. Then the random variable $X = \sum_{i=1}^{n} X_i$ represents the number of Bernoulli trials that result in outcome $A$. Since the Bernoulli random variables are independent, the probability of obtaining a *particular* sequence of $x$ outcomes of type $A$ in a sequence of $n$ trials is $p^x(1-p)^{n-x}$. The number of *different* sequences of $n$ trials that result in $x$ outcomes of type $A$ is given by $\binom{n}{x}$ (which is the number of different ways of placing $x$ outcomes of type $A$ into the $n$ positions of the sequence). Since the different sequences are mutually exclusive, it follows that the probability of observing $x$ outcomes of type $A$ is given by the sum of the probabilities of the $\binom{n}{x}$ different sequences that result in the outcome $x$, which is represented by the binomial density function defined above. The binomial density assigns the probability $f(x; n, p)$ to the outcome $x \in R(X) = \{0, 1, 2, \ldots, n\}$

**Example 4.3** Upon analyzing the cash register receipts of a large department store over an extended period of time, it is found that 30 percent of the customers pay for their purchases by credit card, 50 percent pay by cash, and 20 percent pay by check. Of the next five customers that make purchases at the store, what is the probability that three of them will pay by credit card?

**Answer**: Assume that how a customer pays for her purchases is the outcome of a Bernoulli trial with $x_i = 1 \Rightarrow$ credit card, and $x_i = 0 \Rightarrow$ cash or check. Given that the respective probabilities of these outcomes are .30 and .70, and assuming that customers' payment methods are independent of one another, it follows that $X = \sum_{i=1}^{5} X_i$ represents the number of customers that pay by credit card, and $X$ has a binomial density function with $n = 5$ and $p = .3$. The graph of the density is given in Figure 4.3, where the following values hold:

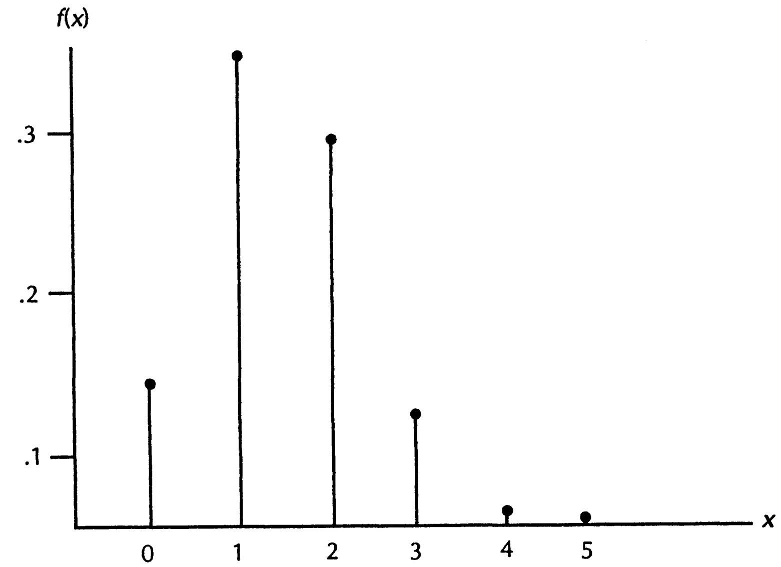


**Figure 4-3**
Binomial density,
$n = 5, p = .3$

| $x$ | $f(x)$ | $x$ | $f(x)$ |
|---|---|---|---|
| 0 | .1681 | 3 | .1323 |
| 1 | .3602 | 4 | .0284 |
| 2 | .3087 | 5 | .0024 |

Thus, $P(x = 3) = .1323$. □

**Family Name: Multinomial**

*Parameterization* $(n, p_1, \ldots, p_m) \in \Omega = \{(n, p_1, \ldots, p_m)$: $n$ is a positive integer, $0 \le p_i \le 1, \forall\, i, \sum_{i=1}^m p_i = 1\}$

*Density Definition*

$$f(x_1, \ldots, x_m; n, p_1, \ldots, p_m) = \begin{cases} \dfrac{n!}{\prod_{i=1}^m x_i!} \displaystyle\prod_{i=1}^m p_i^{x_i} & \text{for } x_i = 0, 1, 2, \ldots, n \ \forall\, i, \displaystyle\sum_{i=1}^m x_i = n, \\ 0 \quad \text{otherwise} \end{cases}$$

*Moments* $\mu_i = np_i, \sigma_i^2 = np_i(1 - p_i), \mu_{3,i} = np_i(1 - p_i)(1 - 2p_i)$,
$\text{Cov}(X_i, X_j) = -np_ip_j$

*MGF* $M_X(t) = (\sum_{i=1}^m p_i e^{t_i})^n$

*Background and Application* The multinomial density function is an extension of the binomial density function to the case where there is interest in more than two different types of outcomes for each trial of the underlying repeated experiment. In particular, the multinomial density function is used to construct a probability space for an experiment that consists of $n$ independent repetitions of a given experiment characterized by $m > 2$ different types of outcomes. The observation of interest is *how many of each of the m different types of outcomes of the experiment occur in n repetitions of the experiment*. The value of $x_i$ represents the total number of outcomes of type $A_i$ that occur in the $n$ repetitions. The parameters $n$ and $p_i$ refer, respectively, to the number of trials conducted and the probability of observing the type $A_i$ outcome in one trial. It is assumed that the repetitions are conducted in such a way that the outcome observed on any trial does not affect the probability of occurrence of outcomes on any other trial. The motivation for the density function definition is a direct extension of the arguments used in the binomial case upon recognizing that the number of different sequences of $n$ repetitions of the experiment that result in $x_i$ type $A_i$ outcomes, $i = 1, \ldots, m$, equals $(\frac{n!}{x_1! \ldots x_m!})$. The details are left to the reader. The multinomial density assigns the probability $f(x_1, \ldots, x_m; n, p_1, \ldots, p_m)$ to the outcome $(x_1, \ldots, x_m) \in R(X) = \{(x_1, \ldots, x_m) : x_i \in (0, 1, \ldots, n) \; \forall \, i, \; \sum_{i=1}^m x_i = n\}$.

It is useful to note that the marginal density of any of the $X_i$ variables is binomial with parameters $n$ and $p_i$. Furthermore, any subset of the random variables $(X_1, \ldots, X_m)$ has a multinomial density.

**Example 4.4** Recall Ex. 4.3. Of the next five customers entering the store, what is the probability that two will pay by credit card, two will pay by cash, and one will pay by check?

**Answer**: For *each* of the five experiments of observing customers' payment methods, we have *three* different types of outcomes that are of interest. In the specification of the multinomial density, we let $x_1$, $x_2$, and $x_3$ refer, respectively, to the number of payments made by credit card, cash, and check. The probabilities of observing a payment by credit card, cash, or check in any given trial is $p_1 = .3$, $p_2 = .5$, and $p_3 = .2$, respectively. Then the probability we seek is given by

$$f(2, 2, 1; 5, .3, .5, .2) = \left[\frac{5!}{2!2!1!}\right] (.3)^2 (.5)^2 (.2)^1 = .135. \qquad \square$$

**Family Name: Negative Binomial and Geometric**

*Parameterization* (*for the geometric density family,* $r = 1$)

$(r, p) \in \Omega = \{(r, p) : r \text{ is a positive integer}, 0 < p < 1\}$

*Density Definition*

$$f(x; r, p) = \begin{cases} \dfrac{(x-1)!}{(r-1)!(x-r)!} p^r (1-p)^{x-r} & \text{for } x = r, r+1, r+2, \ldots, \\ 0 \quad \text{otherwise} \end{cases}$$

*Moments* $\mu = r/p, \sigma^2 = r(1-p)/p^2, \mu_3 = r((1-p) + (1-p)^2)/p^3$

*MGF* $M_X(t) = e^{rt}p^r(1-(1-p)e^t)^{-r}$ for $t < -\ln(1-p)$

*Background and Application* The negative binomial density function (also sometimes referred to as the **Pascal distribution**) can be used to construct a probability space for an experiment that consists of independent repetitions of a given experiment of the Bernoulli type, just like the case of the binomial density, *except* the observation of interest is now *how many Bernoulli trials are necessary to obtain r outcomes of a particular type,* say type $A$ (e.g., how many Bernoulli trials are necessary to obtain r successes, defectives, or tails?). In comparing the binomial and negative binomial densities, notice that the roles of the *number of Bernoulli trials* and the *number of successes* are reversed with respect to what is the random variable and what is the parameter. For the negative binomial density, the value of $x$ represents the number of Bernoulli trials necessary to obtain $r$ outcomes of type $A$.

In order to motivate the density function definition, let $(X_1,\ldots,X_n)$ be a collection of $n$ independent random variables, where each $X_i$ has a Bernoulli density, precisely the same as in our discussion of the binomial density. Let the probability of obtaining an outcome of type $A$ be $p$ for each trial. Since the Bernoulli random variables are independent, the probability of obtaining a sequence of $x$ trials that result in $r$ outcomes of type $A$, *with the last trial being the rth such outcome,* is $p^r(1-p)^{x-r}$. The number of *different* sequences of $x$ trials that result in $r$ outcomes of type $A$ *and end with the rth outcome being of type A*, is given by $(x-1)!/((r-1)!(x-r)!)$ (which is the number of different ways of placing $r-1$ outcomes of type $A$ in the first $x-1$ positions of the sequence). Since the different sequences are mutually exclusive, it follows that the probability of needing $x$ Bernoulli trials to obtain $r$ outcomes of type $A$ is given by the sum of the probabilities of the $(x-1)!/((r-1)!(x-r)!)$ different sequences that result in the outcome $x$, this sum being represented by the negative binomial density function defined above. The negative binomial density assigns the probability $f(x;r,p)$ to the outcome $x \in R(X) = \{r, r+1, r+2, \ldots\}$.

The geometric family of densities is a subset of the family of negative binomial densities defined by setting $r = 1$. Thus, the geometric density is appropriate for assigning probability to events relating to *how many Bernoulli trials are necessary to get the first outcome of type A.* The geometric density function has a unique property in that it is the only discrete density for a nonnegative, integer-valued random variable for which $P[x > i + j \mid x > i] = P[x > j] \; \forall \; i$ and $j \in \{0, 1, 2, \ldots\}$. This conditional probability property is referred to as the **memoryless property**, meaning that in any experiment characterized by the geometric density, if the experiment has already resulted in $i$ trials without a type $A$ outcome, the experiment has "no memory" of this fact, since the probability that more than $j$ additional trials will be needed to obtain the first type $A$ outcome is precisely the same as if the first $i$ trials had never occurred. The proof that the geometric density has this property is left to the reader. The reader may wish to consult V. K. Rohatgi, (1976), *An Introduction to Probability Theory and Mathematical Statistics,* New York: John Wiley, p. 191, for a proof that the geometric density is the *only* density for nonnegative, integer-valued, random variables that has this property.

**Example 4.5** A salesperson has a quota of 10 sales per day that she is expected to meet for her performance to be considered satisfactory. If the probability is .25 that any given customer she contacts will make a purchase, and if purchase decisions are independent across consumers, what is the probability that the salesperson will meet her quota within 30 customer contacts?

**Answer**: The negative binomial density function can be applied with $p = .25$ and $r = 10$. The event of interest is $x \leq 30$, which has the probability

$$P(x \leq 30) = \sum_{x=10}^{30} \left( \frac{(x-1)!}{9!(x-10)!} \right) (.25)^{10}(.75)^{x-10} = .1966. \qquad \square$$

**Example 4.6** A machine produces envelopes, and the probability that any given envelope will be defective is $p = .001$. The production of envelopes from the machine can be viewed as a collection of independent Bernoulli trials (i.e., each envelope is either defective or nondefective, the probability of a defective envelope is .001, and the occurrence of a defective envelope does not affect the probability that other envelopes will be defective or nondefective). What is the probability that the first defective envelope will occur *after* 500 envelopes have been produced? *Given that* the machine has already produced 500 envelopes without a defective one, what is the probability that the first defective envelope will occur *after another* 500 envelopes have been produced?

**Answer**: The geometric density can be applied with $p = .001$. The first event of interest is $\{x > 500\}$. The probability of the event can be calculated by first noting that the cumulative distribution function for the geometric density is given by

$$F(b) = [1 - (1-p)^{\text{trunc}(b)}]I_{[1,\infty)}(b).$$

Then $P(x > 500) = 1 - F(500) = 1 - [1 - (.999)^{500}] = .6064$. The graph of $f(x)$ in this case has a set of spikes at $x = 1, 2, 3, \ldots,$ which decline *very* slowly, beginning with $f(1) = .001$, and with the image of $x = 250$ still being equal to $f(250) = .00078$. As $x \to \infty$, $f(x) \to 0$.

The second probability we seek is of the form $P(x > 1000 \mid x > 500)$. By the memoryless property of the geometric density, we know that this probability is equal to $P(x > 500) = .6064$, as calculated above. □

**Family Name: Poisson**

*Parameterization* $\lambda \in \Omega = \{\lambda : \lambda > 0\}$

*Density Definition*

$$f(x;\lambda)=\begin{cases}\dfrac{e^{-\lambda}\lambda^x}{x!} & \text{for } x=0,1,2,\ldots,\\ 0 & \text{otherwise}\end{cases}$$

*Moments* $\mu=\lambda, \sigma^2=\lambda, \mu_3=\lambda$

*MGF* $M_X(t)=e^{\lambda(e^t-1)}$

*Background and Application* When the number of independent and identical Bernoulli experiments is very large and $p$ is small, the Poisson density, named after the French mathematician Simeon Poisson (1781–1840), provides an approximation to the probability that $x=\sum_{i=1}^n x_i=c$ and thus provides an approximation to probabilities generated by the binomial density. In fact, the limit of the binomial density as $n\to\infty$ and $\lambda=np$, for $\lambda$ a fixed constant, is the Poisson density. We examine this situation in more detail to provide an example of how *limiting densities* arise.[2] In the discussion, we will need the following result.

**Lemma 4.1**

$$\boxed{e^v=\lim_{n\to\infty}(1+(v/n))^n.}$$

The nonzero values of the binomial density can be expressed alternatively as

$$f(x;n,p)=\frac{\prod_{i=1}^x(n-i+1)}{x!}p^x(1-p)^{n-x}.$$

Let $np=\lambda$ for some $\lambda>0$, so that $p=\lambda/n$ can be substituted in the binomial density expression to yield

$$\frac{\prod_{i=1}^x(n-i+1)}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x},$$

where we have suppressed the indicator function by assuming that $x$ is a nonnegative integer $\leq n$. Algebraically rearranging the above expression, and letting $n\to\infty$, yields

$$\begin{aligned}&\lim_{n\to\infty}\frac{\prod_{i=1}^x(n-i+1)}{n^x}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}\\&=\lim_{n\to\infty}\left[\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\ldots\left(\frac{n-x+1}{n}\right)\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}\right]\\&=\frac{e^{-\lambda}\lambda^x}{x!}\end{aligned}$$

[2] Limiting densities will be discussed further in Chapter 5.

since "the limit of a product equals the product of the limits" when all of the limits exist, and since

$$\lim_{n\to\infty}\frac{n-i}{n}=1\ \forall\, i,\ \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x}=1,\ \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n}=e^{-\lambda}\quad\text{(by Lemma 4.1)},$$

$$\lim_{n\to\infty}\frac{\lambda^x}{x!}=\frac{\lambda^x}{x!}.$$

Therefore, the binomial density converges to the Poisson density for $np=\lambda$, $\lambda>0$, $n\to\infty$. The usefulness of this result is that for *large* $n$, and thus for *small* $p=\lambda/n$, we can adopt the approximation

$$\binom{n}{x}p^x(1-p)^{n-x}\approx\left(\frac{e^{-np}(np)^x}{x!}\right),$$

that is, we replace the parameter $\lambda$ in the Poisson density by the product of the binomial density parameters $n$ and $p$. The approximation can be quite useful since the Poisson density is relatively easy to evaluate, whereas for large $n$, dealing with the factorial expressions in the binomial density can be cumbersome.

As a rule of thumb for indicating when the approximation is good, probabilities assigned by the Poisson density will be close to the true binomial probabilities when $n\geq 20$ and $p\leq .05$. The approximation is excellent when $n\geq 100$ and $\lambda=np\leq 10$.

**Example 4.7** A publishing company is typesetting a novel that is 300 pages long and averages 1,500 typed letters per page. If typing errors are as likely to occur for one letter as another, if a typing error occurring in one place does not affect the probability of a typing error occurring in any other place, and if the probability of mistyping a letter is small, then the total number of typing errors in the book can be viewed as the outcome of a random variable having, *approximately*, a Poisson density. For example, if the probability of a typing error for any given letter is $10^{-5}$, then $\lambda=np=4.5$ in the Poisson density. The probability of observing 10 or fewer errors in the book would be approximated as

$$P(x\leq 10)\approx\sum_{x=0}^{10}e^{-4.5}(4.5)^x/x!=.9933.$$

A partial graph (truncated after $x=10$) of the Poisson density in this case is given in Figure 4.4, where

| $x$ | $f(x)$ | $x$ | $f(x)$ |
|---|---|---|---|
| 0 | .0111 | 6 | .1281 |
| 1 | .0500 | 7 | .0824 |
| 2 | .1125 | 8 | .0463 |
| 3 | .1687 | 9 | .0232 |
| 4 | .1898 | 10 | .0104 |
| 5 | .1708 | | |

□

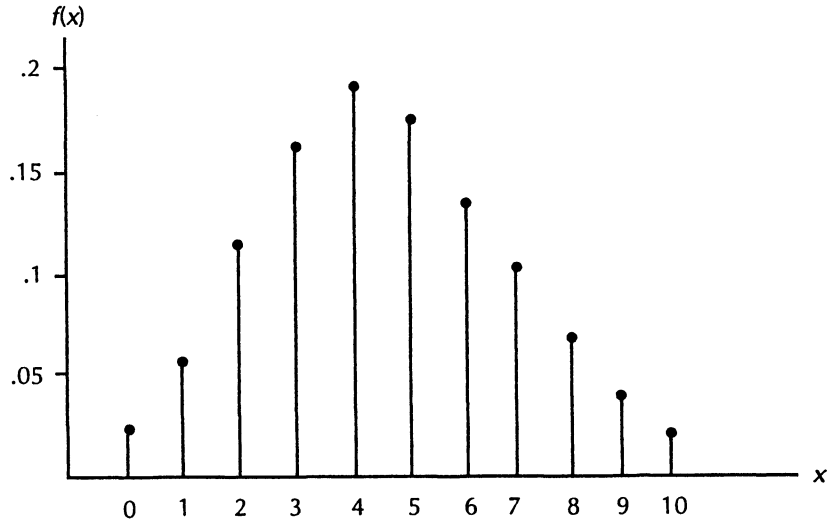


**Figure 4-4** Partial Poisson density, $\lambda = 4.5$

*Poisson Process* Besides serving as an approximation to the binomial density for large $n$ and small $p$, the Poisson density is important in its own right for constructing probability spaces for experiments whose outcomes are governed by the so-called **Poisson process**. The Poisson process refers to a particular type of experimental situation in which the number of occurrences of some specific type, say type $A$ (e.g., a traffic accident, a telephone call, the arrival of a customer at a checkout stand, a flaw in a length of wire) in a time, space, volume, or length dimension, possesses the following general probabilistic characteristics (we state the conditions for a time dimension; the conditions can be adapted to the other dimensions of measurement as well):

**Definition 4.1** ***Poisson process***

Let an experiment consist of observing the number of type $A$ outcomes that occur over a fixed interval of time, say $[0, t]$. The experiment is said to follow the Poisson process if:

1. the probability that precisely one type $A$ outcome will occur in a small time interval of length $\Delta t$ is *approximately* proportional to the length of the interval, as $\gamma[\Delta t] + o(\Delta t)$, where $\gamma > 0$ is the proportionality factor,[3]
2. the probability of two or more type $A$ outcomes occurring in a small time interval of length $\Delta t$ is negligible relative to the probability that one type $A$ outcome occurs, the negligible probability being of order of magnitude $o(\Delta t)$,

[3] $o(t)$ is a generic notation applied to any function of $\Delta t$ whose values approach zero at a rate faster than $\Delta t$, so that $\lim_{\Delta t \to 0}(o(\Delta t))/(\Delta t) = 0$. The "$o(\Delta t)$" stands for "of smaller order of magnitude than $\Delta t$." For example, $h(\Delta t) = (\Delta t)^2$ is a function to which we could affix the label $o(\Delta t)$, while $h(\Delta t) = (\Delta t)^{1/2}$ is not. More will be said about orders of magnitude in Chapter 5.

3. the numbers of type $A$ outcomes that occur in nonoverlapping time intervals are independent events.

To appreciate what is meant by the probability of two or more type $A$ outcomes occurring in a small interval of length $\Delta t$ being *negligible* relative to the probability that just one type $A$ outcome occurs, note that

$$\lim_{\Delta t \to 0} \frac{P(\geq 2 \text{ type } A)}{P(1 \text{ type } A)} = \lim_{\Delta t \to 0} \frac{o(\Delta t)}{\gamma[\Delta t] + o(\Delta t)} = \lim_{\Delta t \to 0} \frac{\frac{o(\Delta t)}{\Delta t}}{\frac{\gamma[\Delta t]}{\Delta t} + \frac{o(\Delta t)}{\Delta t}} = 0.$$

Thus, for small enough intervals $\Delta t$, $P(\geq 2 \text{ type } A)$ is negligible *relative to* $P(1 \text{ type } A)$.

We now indicate why the Poisson process leads to the Poisson density.

**Theorem 4.1** (Poisson Process $\Rightarrow$ Poisson Density) *Let $X$ represent the number of times event $A$ occurs in an interval of time $[0,t]$. If the experiment underlying $X$ follows the Poisson process, then the density of $X$ is the Poisson density.*

**Proof** Partition the interval $[0, t]$ into $n$ successive disjoint subintervals, each of length $\Delta t = t/n$, and denote these intervals by $I_j$, $j = 1, \ldots, n$. Let the random variable $X(I_j)$ denote the number of outcomes of type $A$ that occur within subinterval $I_j$, so that

$$X = \sum_{j=1}^{n} X(I_j).$$

Examine the event $x = k$, and note that $P(x = k) = P(A_n) + P(B_n)$, where $A_n$ and $B_n$ are the *disjoint* sets

$$A_n = \left\{ \sum_{j=1}^{n} x(I_j) = k, x(I_j) = 0 \quad \text{or} \quad 1, \ \forall\, j \right\},$$

$$B_n = \left\{ \sum_{j=1}^{n} x(I_j) = k, x(I_j) \geq 2, \quad \text{for 1 or more } j\text{'s} \right\}.$$

Since $B_n \subset \{x(I_j) \geq 2 \text{ for 1 or more } j\text{'s}\} \subset \cup_{j=1}^{n} \{x(I_j) \geq 2\}$, Boole's inequality implies

$$P(B_n) \leq \sum_{i=1}^{n} P\{x(I_j) \geq 2\} = \sum_{i=1}^{n} o\left(\frac{t}{n}\right) = t\left[\frac{o\left(\frac{t}{n}\right)}{\frac{t}{n}}\right],$$

so that $\lim_{n \to \infty} P(B_n) = 0$.

Now examine $P(A_n)$. For each subinterval, define a "success" as observing exactly one type $A$ outcome, and a "failure" otherwise. Then by property (1) of the Poisson process, $P(\text{success}) = \gamma[t/n] + o(t/n)$ and $P(\text{failure}) = 1 - \gamma[t/n] - o(t/n)$. Since events in one subinterval are independent of events in other subintervals by property (3) of the Poisson process, we can view the observations on the $n$ subintervals as a collection of independent Bernoulli trials, each trial yielding a success or a failure. It follows that probability can be assigned to event $A_n$ by the binomial density as

$$P(A_n) = \binom{n}{k}\left[\gamma\frac{t}{n} + o\left(\frac{t}{n}\right)\right]^k\left[1 - \gamma\frac{t}{n} - o\left(\frac{t}{n}\right)\right]^{n-k}.$$

We need the following extension of Lemma 4.1:

**Lemma 4.2**

$$\boxed{e^v = \lim_{n\to\infty}(1 + (v/n) \pm o(v/n))^n.}$$

Then, following a similar approach to the one used to demonstrate the convergence of the binomial to the Poisson density, it can be shown using Lemma 4.2 that $\lim_{n\to\infty} P(A_n) = (e^{-\gamma t}(\gamma t)^k)/k!$, which is the value of the Poisson density, where $\lambda = \gamma t$.

Finally, since $P(x = k) = P(A_n) + P(B_n) \ \forall\, n$, we have

$$P(x = k) = \lim_{n\to\infty}[P(A_n) + P(B_n)] = \frac{e^{-\gamma t}(\gamma t)^k}{k!},$$

and so the conditions of the Poisson process lead to assignments of probability via the Poisson density function. ■

Note that when use of the Poisson density is motivated from within the context of the Poisson process, $\gamma$ can be interpreted as the *mean rate of occurrence of the type A outcome per unit of time.* This follows from the fact that $EX = \lambda = \gamma t$ if $X$ has a Poisson density, and then $\gamma = EX/t$. In applications, the Poisson density would be chosen for constructing the probability space of an experiment when either the conditions of the Poisson process hold, or when conditions exist for the Poisson density to be a reasonable approximation to the binomial density. The Poisson PDF has been used to represent the probability of such random variables as the number of machine breakdowns in a work shift, the number of customer arrivals at a checkout stand during a period of time, the number of telephone calls arriving at a switchboard during a time interval, and the number of flaws in panes of glass.

**Example 4.8** The milk bottling machine in a milk processing plant has a history of breaking down, on average, once every two weeks. The chief of the repair and maintenance crew is scheduling vacations for the summer months and wants to know what the probability is that the bottling machine will break down more than three times during the next four weeks. What is the probability?

**Answer**: When viewed in the context of ever shorter time intervals (hours, minutes, seconds), breakdowns appear to be increasingly less likely, and it seems reasonable to assume that the probability of two breakdowns within a short-enough time interval would be negligible. Assuming the repair crew returns the machine to full operating performance after each breakdown, it is reasonable to assume that the event of a breakdown in any short interval of time is independent of a breakdown occurring in other intervals of time.[4] All told, it would appear that the conditions of the Poisson process are a reasonable approximation to this situation, and we endeavor to assign probability to the event $x > 3$ using the Poisson density. Since the average number of breakdowns is 1 every 2 weeks and since the chief is interested in a 4-week time interval, the Poisson density of relevance here is $(e^{-2}2^x)/x!I_{\{0,1,2,\ldots\}}(x)$, where $\lambda = \gamma t = 2$ was chosen to represent a rate of 2 breakdowns every 4 weeks ($\gamma = .5$ breakdowns per week times $t = 4$ weeks). The probability of more than 3 breakdowns in the 4-week period is then given by

$$P(x > 3) = 1 - \sum_{i=0}^{3} P(x = i) = 1 - \sum_{i=0}^{3} e^{-2}\frac{2^i}{i!} = .143. \qquad \square$$

**Family Name: Hypergeometric**

*Parameterization* $(M, K, n) \in \Omega = \{(M, K, n) : M = 1, 2, 3, \ldots; K = 0, 1, \ldots, M; n = 1, 2, \ldots, M\}$

*Density Definition*

$$f(x; M, K, n) = \begin{cases} \dfrac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}} & \text{for integer values } \max[0, n - (M - K)] \le x \le \min(n, K), \\ 0 & \text{otherwise} \end{cases}$$

*Moments*

$$\mu = \frac{nK}{M}, \quad \sigma^2 = n\left(\frac{K}{M}\right)\left(\frac{M-K}{M}\right)\left(\frac{M-n}{M-1}\right),$$

$$\mu_3 = n\left(\frac{K}{M}\right)\left(\frac{MK}{M}\right)\left(\frac{M-2K}{M}\right)\left(\frac{M-n}{M-1}\right)\left(\frac{M-2n}{M-2}\right)$$

*MGF* $M_X(t) = [((M - n)!(M - K)!)/M!]H(-n, -K, M - K - n + 1, e^t)$, where $H(\cdot)$ is the hypergeometric function

$$H(\alpha, \beta, r, Z) = 1 + \frac{\alpha\beta}{r}\frac{Z}{1!} + \frac{\alpha\beta(a + 1)(\beta + 1)}{r(r + 1)}\frac{Z^2}{2!} + \cdots$$

(Note: This MGF is not too useful in practice since moments are defined in terms of an infinite sum; to illustrate this fact, the reader should attempt to define $\mu$ by differentiating $M_X(t)$ once with respect to $t$ and evaluating the derivative at zero.)

*Background and Application* The hypergeometric density is used to construct a probability space for an experiment in which there are

[4] Ultimately, this assumption could be tested using a nonparametric test of hypothesis. We will examine tests of independence in our discussion of hypothesis testing procedures later in the text.

1. $M$ objects, of which $K$ of them are of one type, say type $A$;
2. the remaining $M - K$ objects are of a different type, say $B$; and
3. $n$ objects are randomly drawn *without replacement* from the original collection of $M$ objects and the number, $x$, of type $A$ outcomes in the collection of $n$ objects drawn is observed.

By **drawing randomly without replacement**, we mean that, at each draw, all of the objects that are still left in the collection have an equal probability of being chosen. The hypergeometric density assigns probability to the number of type $A$ objects drawn out of a total sample of $n$ objects.

To motivate the density function definition, note that the number of different ways of choosing the sample of $n$ objects is given by $\binom{M}{n}$, the number of different ways of choosing $x$ type $A$ items is $\binom{K}{x}$, and the number of different ways to choose $(n - x)$ type $B$ items is $\binom{M-K}{n-x}$. Then, since all possible samples having $x$ type $A$ and $n-x$ type $B$ outcomes are equally likely, the classical probability definition states that the probability of obtaining $x$ type $A$ outcomes from a random sample (without replacement) of $n$ objects from the aforementioned collection is $\binom{K}{x}\binom{M-K}{n-x} \div \binom{M}{n}$.

Note that the binomial and hypergeometric densities both assign probabilities to outcomes of the type *observe x type A outcomes from a total number of n observations*. The important difference between experiments for which the binomial density or the hypergeometric density applies is that in the former case, the $n$ trials are *independent* and *identical* and would correspond to randomly **drawing objects with replacement** (meaning once an object is drawn, and the observation is made, the object is placed back in the total collection of objects so that, at each draw, *all* of the original objects in the collection are equally probable to be chosen), whereas randomly drawing objects *without replacement* characterizes the experiment to which the hypergeometric density is applied.

**Example 4.9** Suppose a shipment of 1,000 computer memory modules contains 50 defectives. What is the probability of obtaining no defective modules in a random drawing, without replacement, of 5 modules from the shipment?

**Answer**: The appropriate hypergeometric density for this case has $M = 1{,}000$, $K = 50$, and $n = 5$. The probability assigned to the (elementary) event $x = 0$ is

$$f(0; 1{,}000, 50, 5) = \frac{\left(\frac{50!}{0!50!}\right)\left(\frac{950!}{5!945!}\right)}{\left(\frac{1000!}{5!995!}\right)} = .7734. \qquad \square$$

Regarding the graph of the hypergeometric density, $f(x; M, K, n)$ increases as $x$ increases, until a maximum value is reached at the greatest integer value of $x$ for which $x \le ((n+1)(K+1))/(M+2)$ (this can be shown by examining the values of $x$ for which the density is increasing, i.e., $f(x; M, K, n)/f(x - 1; M, K, n) \ge 1$). The value of the PDF declines thereafter. Some random variables to which the hypergeometric density has been applied include observations on the number of responses of a certain type in the context of auditing, quality control, and consumer or employee attitude surveys.

### Family Name: Multivariate Hypergeometric

*Parameterization*

$$(M, K_1, \ldots, K_m, n) \in \Omega = \{(M, K_1, \ldots, K_m, n) : M = 1, 2, \ldots;$$
$$K_i = 0, 1, \ldots, M \quad \text{for } i = 1, \ldots, m; \sum_{i=1}^{m} K_i = M; n = 1, 2, \ldots, M\}$$

*Density Definition*

$$f(x_1, \ldots, x_m; M, K_1, \ldots, K_m, n) = \begin{cases} \dfrac{\prod_{i=1}^{m} \binom{K_i}{x_i}}{\binom{M}{n}} & \text{for } x_i \in \{0, 1, 2, \ldots, n\} \ \forall\, i, \sum_{i=1}^{m} x_i = n, \\ 0 & \text{otherwise} \end{cases}$$

*Moments*

$$\mu_i = \frac{nK_i}{M}, \qquad \sigma_i^2 = n\left(\frac{K_i}{M}\right)\left(\frac{M - K_i}{M}\right)\left(\frac{M - n}{M - 1}\right),$$

$$\mu_{3,i} = n\left(\frac{K_i}{M}\right)\left(\frac{MK_i}{M}\right)\left(\frac{M - 2K_i}{M}\right)\left(\frac{M - n}{M - 1}\right)\left(\frac{M - 2n}{M - 2}\right)$$

*MGF* not useful

*Background and Application* The multivariate hypergeometric density is a generalization of the hypergeometric density in the same sense as the multinomial density is a generalization of the binomial density. In particular, we now consider the case where we are interested in *more than two* different types of outcomes for each object chosen from the original collection of objects. Letting $K_i, i = 1, \ldots, m$, refer to the number of objects of type $i$ that are in the collection, $M = \sum_{i=1}^{m} K_i$ represent the total number of objects in the collection, $n$ represent the number of objects randomly drawn *without replacement* from the collection, and $x_i$ be the number of outcome of type $i$, $i = 1, \ldots, m$, then an extension of the argument used to motivate the density definition in the hypergeometric case leads to the definition of the multivariate hypergeometric density function presented earlier.

Note that the marginal density of each $X_i$ is hypergeometric with parameters $(M, K_i, n)$. Furthermore, any subset of the random variables $(X_1, \ldots, X_m)$ also has a multivariate hypergeometric density, as the reader can verify.

**Example 4.10** As part of its quality assurance program, a large northwestern bank regularly interviews a randomly selected subset of the customers who transact business at one of its branches each week. Among other questions, the customers are asked to rank the overall service they received as being "excellent," "good," "average," "below average," or "poor." In one of the smaller rural branches, there were 100 customers during a particular week. If the bank randomly chooses 5 of the 100 customers to interview, and if the 100 customers were distributed across the rating categories as 50, 30, 10, 7, 3, respectively, what is the probability that the interviews will result in 2 "excellent," 2 "good," and 1 "average" ratings?

**Answer**: Use the multivariate hypergeometric density with $M = 100$, $K_1 = 50$, $K_2 = 30$, $K_3 = 10$, $K_4 = 7$, $K_5 = 3$, and $n = 5$. Then

$$P(x_1 = 2, x_2 = 2, x_3 = 1, x_4 = x_5 = 0) = \frac{\binom{50}{2}\binom{30}{2}\binom{10}{1}\binom{7}{0}\binom{3}{0}}{\binom{100}{5}} = .0708. \qquad \square$$

## 4.2 Parametric Families of Continuous Density Functions

**Family Name: Uniform**

*Parameterization* $(a, b) \in \Omega = \{(a, b) : -\infty < a < b < \infty\}$

*Density Definition* $f(x; a, b) = (1/(b - a))I_{[a,b]}(x)$

*Moments* $\mu = (a + b)/2, \sigma^2 = (b - a)^2/12, \mu_3 = 0$

*MGF*

$$M_X(t) = \begin{cases} \dfrac{e^{bt} - e^{at}}{(b - a)t} & \text{for } t \neq 0, \\ 1 \quad \text{for } t = 0 \end{cases}$$

*Background and Application* The continuous uniform density is used to construct probability spaces for experiments having an uncountable, infinite number of possible outcomes that are all equally likely in the interval $[a, b]$, for finite $b - a$. All interval subsets of $[a, b]$ of length $k$ are assigned equal probability, $k/(b - a)$. The continuous uniform density has important applications in the computer generation of random-variable outcomes for a wide array of probability distributions. We will examine these types of applications in Chapter 6, a preview of which is provided in the next example.

**Example 4.11** An efficiency analyst wishes to simulate the daily demand for teller services in a moderately sized branch of a regional bank. The probability that a customer will require the services of a teller is known to be .30, whereas the customer will utilize the services of a cash machine, loan officer, or investment banker with probability .70. The type of service demanded is independent across customers. The efficiency analyst concludes that of $n$ customers entering the branch on a given day, the number that utilize a teller's services is the outcome of a random variable $X$ having the binomial density $f(x) = \binom{n}{x}(.30)^x(.70)^{n-x}I_{\{0,1,\ldots,n\}}(x)$. The analyst has a computer random-number generator that produces outcomes of a random variable $Y$ having the uniform density $h(y) = I_{[0,1]}(y)$. How can the analyst simulate daily outcomes of $X$?

**Answer**: Let $F(b) = \sum_{x\leq b} f(x)$ be the CDF of $X$, and define $n+1$ intervals as $I_0 = [0, F(0))$ and $I_j = [F(j-1), F(j))$, for $j = 1, \ldots, n$. Note that the lengths of the intervals represent the respective probabilities assigned to the outcomes $x \in \{0, 1, \ldots, n\}$ by the aforementioned binomial density; for example, $P(x = j) = F(j) - F(j-1) = f(j)$, for $j \in \{1, 2, \ldots, n\}$. Then an outcome of $X$ can be simulated by first generating an outcome of $Y$ and then calculating $x = \sum_{j=0}^{n} jI_{I_j}(y)$. That these outcomes follow the appropriate binomial density can be motivated by the fact that $x = j$ iff $y \in I_j$ and $P_X(j) = P_Y(I_j) = \int_{y\in I_j} I_{[0,1]}(y)dy = f(j)$.

For a specific illustration, let $n = 5$. Then $I_0 = [0, .1681), I_1 = [.1681, .5283), I_2 = [.5283, .8370), I_3 = [.8370, .9693), I_4 = [.9693, .9977)$, and $I_5 = [.9977, 1)$. If the computer random-number generator were to generate an outcome $y = .6311$, then the analyst would simulate that 2 customers required teller's services on a day when 5 customers entered the branch (since $.6311 \in I_2$).

The reader might contemplate an extension of this simulation exercise to the case where the number of customers entering the branch on a given day is the outcome of a random variable $N$. □

### Family Name: Gamma

*Parameterization* $(\alpha, \beta) \in \Omega = \{(\alpha, \beta) : \alpha > 0, \beta > 0\}$

*Density Definition* $f(x; \alpha, \beta) = (1/(\beta^\alpha \Gamma(\alpha)))x^{\alpha-1}e^{-x/\beta}I_{(0,\infty)}(x)$,
where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$ is called the **gamma function**, having the property that if $\alpha$ is a positive integer, $\Gamma(\alpha)$ has values $\Gamma(\alpha) = (\alpha-1)!$, and if $\alpha = 1/2$, then $\Gamma(1/2) = \pi^{1/2}$. Also, for any real $\alpha > 0$, $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$.

*Moments* $\mu = \alpha\beta$, $\sigma^2 = \alpha\beta^2$, $\mu_3 = 2\alpha\beta^3$

*MGF* $M_X(t) = (1 - \beta t)^{-\alpha}$ for $t < \beta^{-1}$

*Background and Applications* The gamma family of density functions is a versatile collection of density functions that can be used to model a wide range of experiments whose outcomes are coded as *nonnegative* real numbers and whose respective probabilities are to be assigned via a density function that is *skewed to the right*. An extensive variety of density shapes are possible by altering the parameters $\alpha$ and $\beta$, a few of which are illustrated in Figures 4.5 and 4.6. It can be shown that the gamma density is strictly decreasing when $\alpha \leq 1$. The density increases to a maximum of $((\alpha-1)e^{-1})^{\alpha-1}/(\beta\Gamma(\alpha))$, at $x = (\alpha-1)\beta$, for $\alpha > 1$.

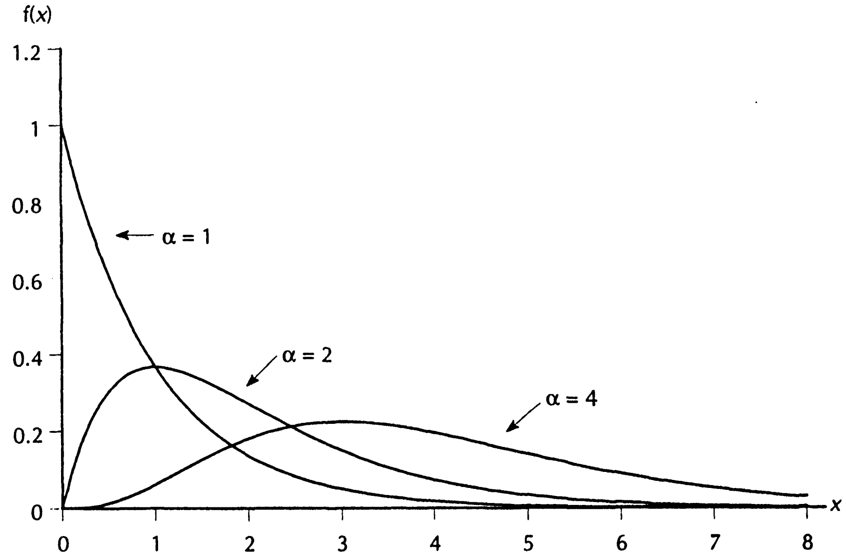


**Figure 4-5** Gamma densities, case I, $\beta = 1$

While its wide variety of shapes makes the gamma family a candidate for constructing the probability space of many experiments with nonnegative outcomes, the gamma family has *specific* uses with regard to waiting times between occurrences of events based on the Poisson process. In particular, let $Y$ have a Poisson density with parameter $\lambda$, and let $y$ refer generically to the number of successes that occur in a period of time $t$, so that $\gamma = \lambda/t$ is the rate of success of the Poisson process. If $X$ measures the time that passes until the Poisson process produces the *first* success, then $X$ has a gamma density with $\alpha = 1$ and $\beta = \gamma^{-1}$. If instead $X$ measures the time that passes until the Poisson process produces the $r$th success, then $X$ has a gamma density with

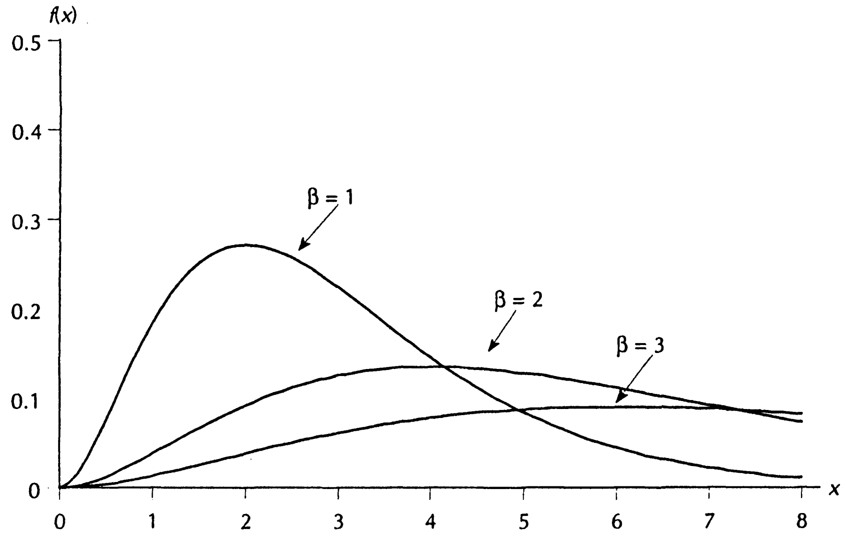


**Figure 4-6** Gamma densities, case II, $\alpha = 3$

$\alpha = r$ and $\beta = \gamma^{-1}$. Proofs of these propositions can be found in S. F. Arnold, *Mathematical Statistics*, p. 166.

**Example 4.12** Recall Ex. 4.8 regarding breakdowns of a milk bottling machine, which was assumed to be a Poisson process with rate $\gamma = \lambda/t = 2/4 = .5$. What is the probability that the machine will not breakdown in four weeks of use?
**Answer**: The event of interest is $\{x > 4\}$, that is, the first breakdown occurs after 4 weeks of operation. From the preceding discussion, we know that $x$ has a gamma density with $\alpha = 1$ and $\beta = \gamma^{-1} = 2$. Therefore, $P(x > 4) = \int_4^\infty (1/2)e^{-x/2}dx = -e^{-x/2}\,|_4^\infty = .1353$. Note this is precisely the same probability as the outcome 0 receives in the Poisson density of Ex. 4.8, which is as it should be, given that *no* breakdowns occurring in 4 weeks (0 outcome for the Poisson random variable) coincides with the *first* breakdown occurring after 4 weeks (an outcome greater than 4 for the gamma random variable). □

The gamma family of densities has an important **additivity property**, which we state in the following theorem.

**Theorem 4.2** (Gamma Additivity) *Let $X_1,\ldots,X_n$ be independent random variables with respective gamma densities* Gamma $(\alpha_i,\beta)$, $i = 1,\ldots,n$. *Then $Y = \sum_{i=1}^n X_i$ has the gamma density* Gamma $(\sum_{i=1}^n \alpha_i,\beta)$.

**Proof** Since $M_{X_i}(t) = (1-\beta t)^{-\alpha_i}$ for $t < \beta^{-1}$, $i = 1,\ldots,n$, and since the $X_i$'s are independent, Theorem 3.26 implies

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n}(1-\beta t)^{-\alpha_i} = (1-\beta t)^{-\sum_{i=1}^{n}\alpha_i} \quad \text{for } t < \beta^{-1}.$$

Thus, by the MGF uniqueness theorem, $Y$ has the density Gamma$(\sum_{i=1}^n \alpha_i, \beta)$. ■

Therefore, the sum of independent gamma random variables has a gamma distribution as long as the underlying gamma densities share the same $\beta$ parameter value.

Scaling a gamma random variable by a positive constant results in a random variable that also has a gamma distribution, as demonstrated in the next theorem.

**Theorem 4.3** (Scaling of Gamma Random Variables) *Let $X$ have a gamma density* Gamma $(\alpha,\beta)$, *and let $c > 0$. Then $Y = cX$ has a gamma density* Gamma $(\alpha,\beta c)$.

**Proof** Since $M_X(t) = (1-\beta t)^{-\alpha}$ for $t < \beta^{-1}$, Theorem 3.26 implies that

$$M_Y(t) = M_{cX}(t) = M_X(ct) = (1-\beta ct)^{-\alpha} \quad \text{for } t < (\beta c)^{-1},$$

which by the MGF uniqueness theorem indicates that $Y$ has the gamma density Gamma$(\alpha, \beta c)$. ■

We will also have use for the following property of gamma PDFs.

**Theorem 4.4** (Gamma Inverse Additivity) *Let $Y = X_1 + X_2$, where $Y$ has the gamma density* Gamma $(\alpha,\beta)$, *$X_1$ has the gamma density* Gamma $(\alpha_1,\beta)$, *$\alpha > \alpha_1$, and $X_1$ and $X_2$ are independent. Then $X_2$ has the gamma density* Gamma $(\alpha - \alpha_1,\beta)$.

**Proof** Since $M_Y(t) = (1 - \beta t)^{-\alpha}$ and $M_{X_1}(t) = (1 - \beta t)^{-\alpha_1}$ for $t < \beta^{-1}$, Theorem 3.26 implies that

$$M_Y(t) = (1 - \beta t)^{-\alpha} = (1 - \beta t)^{-\alpha_1} M_{X_2}(t) = M_{X_1}(t) M_{X_2}(t),$$

which in turn implies

$$M_{X_2}(t) = (1 - \beta t)^{-\alpha}/(1 - \beta t)^{-\alpha_1} = (1 - \beta t)^{-(\alpha-\alpha_1)} \quad \text{for } t < \beta^{-1}.$$

By the MGF uniqueness theorem, $X_2$ has the gamma density Gamma$(\alpha - \alpha_1, \beta)$. ■

Its wide variety of density shapes has resulted in the gamma family's being applied to a myriad of nonnegative-valued random variables suspected of having a right-skewed PDF. Some specific applications include the waiting times between customer arrivals or machine breakdowns or telephone calls, the breaking strength of manufactured construction materials, the operating lives of electronic equipment and other objects, and the length of time required to service a customer at a store. We now examine two important subfamilies of the gamma family of densities that are defined by special choices of the parameters $\alpha$ and $\beta$.

### Gamma Subfamily Name: Exponential

*Parameterization* $\theta \in \Omega = \{\theta : \theta > 0\}$

*Density Definition* The gamma density, with $\alpha = 1$ and $\beta = \theta$.

$$f(x; \theta) = \frac{1}{\theta} e^{-x/\theta} I_{[0,\infty)}(x)$$

*Moments* $\mu = \theta$, $\sigma^2 = \theta^2$, $\mu_3 = 2\theta^3$

*MGF* $M_X(t) = (1 - \theta t)^{-1}$ for $t < \theta^{-1}$

*Background and Application* The exponential density is used to construct a probability space for experiments that have a real-valued sample space given by the *nonnegative* subset of the real line, $[0, \infty)$, and in which interval events of fixed length $d > 0$ of the form $[t, t + d]$ are to be assigned probability that monotonically decreases as $t$ increases. A specific application concerns the experiment of observing the time that passes until a Poisson process with rate $\gamma = \lambda/t$ produces the *first* success, in which case the exponential density with $\theta = \gamma^{-1}$ is appropriate (recall our previous discussion regarding the relationship between the gamma density and the Poisson process).

A prominent application of the exponential density is in representing the operating lives until failure of various objects. In this regard, the following property of the exponential density is of notable importance.

**Theorem 4.5** (Memoryless Property of Exponential Density) *If X has an exponential density, then* $P(x > s + t \mid x > s) = P(x > t)$ $\forall$ *t and s* > 0.

**Proof**

$$P(x > s + t \mid x > s) = \frac{P(x > s + t)}{P(x > s)} = \frac{\int_{s+t}^{\infty} \frac{1}{\theta} e^{-x/\theta} dx}{\int_{s}^{\infty} \frac{1}{\theta} e^{-x/\theta} dx}$$

$$= \frac{e^{-(s+t)/\theta}}{e^{-s/\theta}} = e^{-t/\theta} = P(x > t) \qquad \blacksquare$$

Interpreted in the context of operating life, the **memoryless property** implies that *given* that the object has already functioned for s units of time without failing (the meaning of the conditioning event $x > s$), the probability that it will function for at least an *additional* $t$ units of time (the meaning of the event $x > s + t$) is the same as the unconditional probability that it would function for at least $t$ units of time (the meaning of the event $x > t$). In effect, the object is "**as good as new**" after functioning $s$ units of time, since the probability of functioning for at least another $t$ units of time is the same as if it had not previously functioned at all (i.e., as if it were new).[5]

While the memoryless property certainly is not applicable to the lifetimes of all objects, the assumption is appropriate in the modeling of the lifetimes of certain electronic components, fuses, jeweled watch bearings, and other objects that are not subject to significant wear and are essentially "as good as new" if they are still functioning. Furthermore, the assumption is appropriate for objects, such as machinery, that receive periodic maintenance that reconditions the object to essentially new status.

If the memoryless assumption is not appropriate, the more versatile gamma family of densities can be considered for constructing the probability space. It can be shown that members of the gamma family, other than the exponential family, exhibit "**wear-out**" effects in the sense that $P(x > s+t \mid x > s)$ *declines* as $s$ increases for $t > 0$ and $\alpha > 1$, that is, the probability that the object functions for at least $t$ units of time beyond the $s$ units of time for which it has already functioned declines as the value of $s$ increases. For $\alpha < 1$, the conditional probability actually increases as s increases—this is referred to in the literature as the "**work-hardening**" effect. The graphs of some exponential densities are displayed in Figure 4.7.

**Example 4.13** The lifetime of a fuse your company manufactures is the outcome of a random variable with mean = 500 hours. The fuse is "as good as new" while functioning. What is the probability that the fuse functions for at least 1,000 hours?

[5] The exponential density is the *only* density for continuous nonnegative-valued random variables that has the memoryless property. See V. K. Rohatgi (1976) *An Introduction to Probability Theory and Mathematical Statistics.* New York: John Wiley, p. 209.

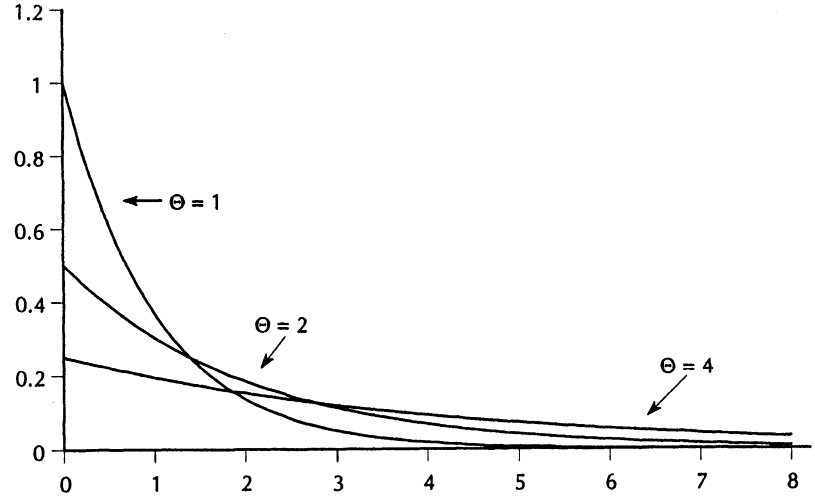


**Figure 4-7**
Exponential densities

**Answer:** The exponential density is appropriate with $\theta = 500$. Then,

$$P(x \geq 1000) = \int_{1000}^{\infty} \frac{1}{500} e^{-x/500} I_{(0,\infty)}(x) dx = -e^{-x/500} \Big|_{1000}^{\infty} = .1353. \qquad \square$$

**Gamma Subfamily Name: Chi-Square**

*Parameterization* $v \in \Omega = \{v : v \text{ is a positive integer}\}$

*Density Definition* The gamma density, with $\alpha = v/2$ and $\beta = 2$.

$$f(x; v) = \frac{1}{2^{v/2}\Gamma(v/2)} x^{(v/2)-1} e^{-x/2} I_{(0,\infty)}(x)$$

*Moments* $\mu = v$, $\sigma^2 = 2v$, $\mu_3 = 8v$

*MGF* $M_X(t) = (1 - 2t)^{-v/2}$ for $t < \frac{1}{2}$

*Background and Application* The parameter $v$ of the chi-square density is called the **degrees of freedom**. The reason for this label will be clarified later in the chapter when we show that the sum of the squares of $v$ independent random variables, each having a density called the standard normal (to be discussed in Section 4.3), will have a chi-square density with $v$ degrees of freedom. The chi-square density with $v$ degrees of freedom is often indicated by the notation $\chi^2_v$ or $\chi^2(v)$. We will utilize the former. The relationship between the chi-square density and the normal density makes the chi-square density especially important in applications concerning hypothesis testing and confidence interval estimation, which is its primary application context, as will be seen in later chapters. Note that for $v = 2$, the $\chi^2$ density is equivalent to the exponential density with $\theta = 2$. Also, $\chi^2_v$ is a valid PDF even for noninteger values of $v > 0$, in which case the PDF $\chi^2_v$ is referred to as the *nonintegral chi-square density*. Our use of $\chi^2_v$ will be restricted to integer-valued $v$. Some chi-square densities are graphed in Figure 4.8.

There are two important properties of the $\chi^2$ density that we note in the following two corollaries.

**Corollary 4.1** (Chi-Square Additivity) *Let $X_1, \ldots, X_k$ be independent random variables having chi-square densities with $v_1, \ldots, v_k$ degrees of freedom, respectively. Then $Y = \sum_{i=1}^k X_i$ has a chi-square density with degrees of freedom $v = \sum_{i=1}^k v_i$.*

**Proof** This follows from Theorem 4.2, with $\beta = 2$. ■

Thus the sum of independent chi-square random variables also has a chi-square distribution.

**Corollary 4.2** *Let $X_1$ and $X_2$ be independent random variables, where $X_1$ has a chi-square density with $v_1$ degrees of freedom, and $Y = X_1 + X_2$ has a chi-square density with $v > v_1$ degrees of freedom. Then $X_2$ has a chi-square density with $v - v_1$ degrees of freedom.*

**Proof** This follows from Theorem 4.4 with $\beta = 2$, $\alpha = v/2$, and $\alpha_1 = v_1/2$. ■

Later in our study of hypothesis testing it will be useful to know upper- and lower-bound values $h$ and $\ell$ for which $P(x \geq h) = \alpha$ and $P(x \leq \ell) = \alpha$ are true, where $X$ has a $\chi^2$ density with $v$ degrees of freedom and $\alpha \in (0, 1)$. We are diagrammatically examining the upper and lower tails of the $\chi^2_v$ density in Figure 4.9. Such events and their probabilities are identified in tables of the $\chi^2$ density available from many published sources, including this text's Appendix.

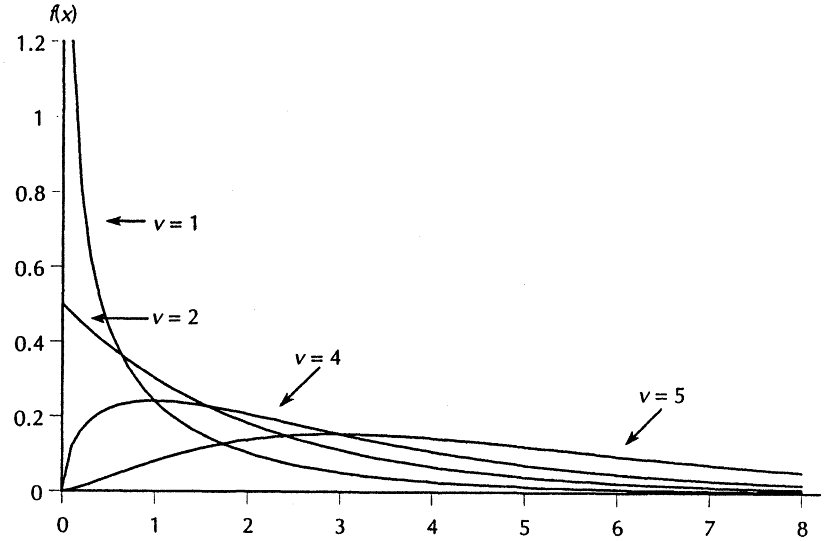


**Figure 4-8**
$\chi^2$ densities.

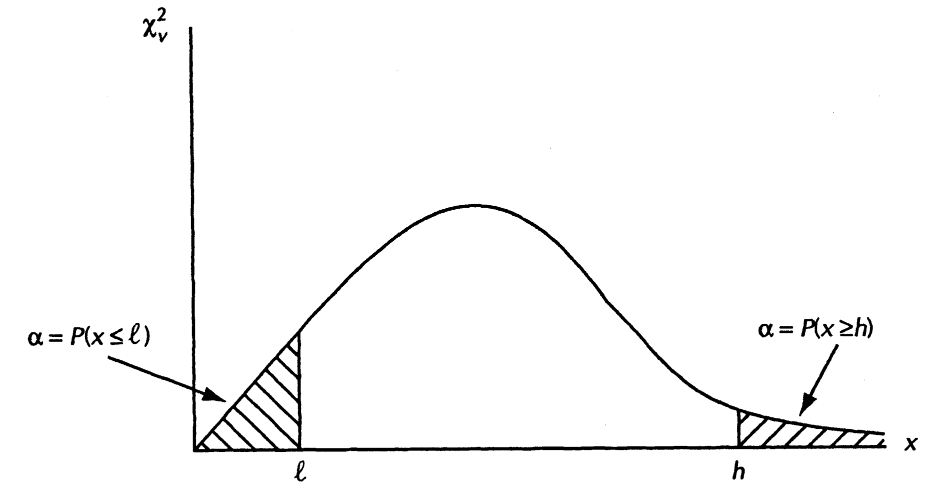


**Figure 4-9**
Upper and lower $\alpha$-level tails of a $\chi^2$ density

Typically, the events are identified for $\alpha$ = .01, .025, and .05. Of course, the appropriate values of $h$ and $\ell$ can be found by solving the following equations (by numerical methods, generally):

$$\int_h^\infty \frac{1}{2^{v/2}\Gamma\left(\frac{v}{2}\right)} x^{(v/2)-1} e^{-x/2} dx = \alpha,$$

$$\int_0^\ell \frac{1}{2^{v/2}\Gamma\left(\frac{v}{2}\right)} x^{(v/2)-1} e^{-x/2} dx = \alpha,$$

given any $\alpha$-level of interest $\in (0, 1)$. A number of computer programs exist that can be used to identify the appropriate $h$ or $\ell$, such as the IMSLIB package of mathematical subroutines, the SAS package, Shazam, or the GAUSS programming language.

**Example 4.14** Let $X$ have a $\chi^2$ density with 10 degrees of freedom. Find the values of $h$ and $\ell$ for which $P(x \geq h) = P(x \leq \ell) = .05$.
**Answer**: In the chi-square table for the row corresponding to 10 degrees of freedom, the values of $h$ and $\ell$ associated with the upper and lower .05 tails of the $\chi^2$ density are, respectively, 18.307 and 3.940. □

**Family Name: Beta**

*Parameterization* $(\alpha, \beta) \in \Omega = \{(\alpha, \beta) : \alpha > 0, \beta > 0\}$

*Density Definition*

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{(0,1)}(x),$$

where

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$$

is called the **beta function.** Some useful properties of the beta function include the fact that

$$B(\alpha,\beta) = B(\beta,\alpha) \quad \text{and} \quad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

so that the beta function can be evaluated in terms of the gamma function.

*Moments*

$$\mu = \alpha/(\alpha+\beta), \qquad \sigma^2 = \alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2],$$

$$\mu_3 = 2(\beta-\alpha)(\alpha\beta)/[(\alpha+\beta+2)(\alpha+\beta+1)(\alpha+\beta)^3]$$

*MGF* $M_X(t) = \sum_{r=1}^{\infty}(B(r+\alpha,\beta)/B(\alpha,\beta))(t^r/r!)$

*Background and Application* The beta density is a very versatile density (i.e., it can assume a large variety of shapes) for constructing probability spaces for experiments having a continuous real-valued sample space equal to the set [0, 1]. Table 4.1 provides a quick reference regarding the numerous shape characteristics of the beta density. The versatility of this density family makes it useful for representing PDFs for random variables associated with virtually any experiment whose outcomes constitute a continuum between 0 and 1. The density has obvious applications in modeling experiments whose outcomes are in the form of proportions, such as the proportion of time a certain machine is in a state of being repaired, the proportion of chemical impurities in a liquid product, or the proportion of respondents to a survey (a continuous approximation in the latter case). In Chapter 6 we will also see that the beta density has an

**Table 4.1** Summary of Beta Density Shapes

| Conditions on $\alpha$ and $\beta$ | Behavior of $f(x)$ |
|---|---|
| $\alpha < \beta$ | skewed to the right, $\mu_3 > 0$ |
| $\alpha > \beta$ | skewed to the left, $\mu_3 < 0$ |
| $\alpha = \beta$ | symmetric about $\mu = 1/2$ |
| $\alpha > 1$ and $\beta > 1$ | maximum value when $x = (\alpha-1)/(\alpha+\beta-2)$ and $f(x) \to 0$ if $x \to 1$ or $x \to 0$ |
| $\alpha < 1$ | $f(x) \to \infty$ as $x \to 0$ |
| $\beta < 1$ | $f(x) \to \infty$ as $x \to 1$ |
| $\alpha < 1$ and $\beta < 1$ | $f(x)$ is U-shaped, having minimum value when $x = (\alpha-1)/(\alpha+\beta-2)$, and $f(x) \to \infty$ when $x \to -\infty$ or $\infty$ |
| $(\alpha-1)(\beta-1) < 0$ | J-shaped |
| $\alpha = \beta = 1$ | uniform on (0,1) |

important application in assigning probabilities to events involving so-called *order statistics*. Figure 4.10 illustrates some of the beta density shapes that are possible.

Integration of the density cannot be accomplished in closed form for nonintegral values of the parameters $\alpha$ and $\beta$. Computer programs exist for integrating the beta density, for example, the program MDBETA in the IMSLIB set of computer routines or the procedure CDFBETA in GAUSS. Integrals of the beta density have also been extensively tabled by Pearson.[6] When $\alpha$ and $\beta$ are integers, there is a relationship between integrals of the beta density and the binomial density that is sometimes useful:

$$F(c) = \int_0^c \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}dx = \sum_{i=\alpha}^{n}\binom{n}{i}c^i(1-c)^{n-i},$$

where $n = \alpha + \beta - 1$ and $c \in [0, 1]$. In any case, when $\alpha$ and $\beta$ are integers, integration of the beta density can be accomplished in closed form. The following example illustrates the point.

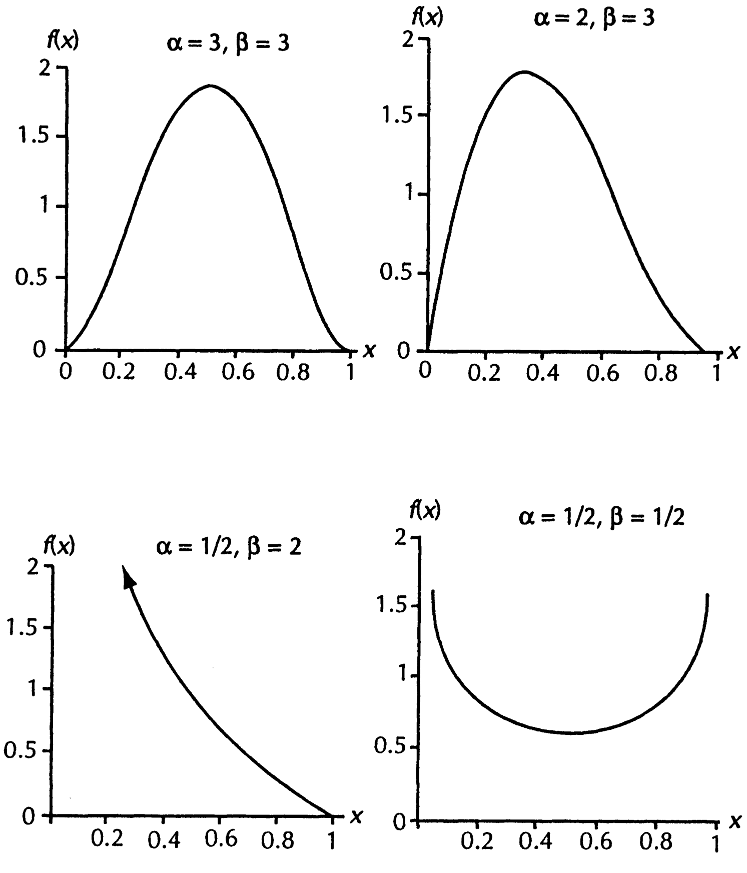


**Figure 4-10**
Beta densities

[6] K. Pearson, (1956), *Tables of the Incomplete Beta Function*. New York: Cambridge Univ. Press.

**Example 4.15** A wholesale distributor of heating oil has a storage tank that holds the distributor's inventory of heating oil. The tank is filled every Monday morning. The wholesaler is interested in the proportion of the tank's capacity that remains in inventory after the weekly sales of heating oil. Having studied this proportion over many weeks, management decided to view the proportion as the outcome of a random variable having the beta density with $\alpha = 4$ and $\beta = 3$. What is the probability that less than 20 percent of the storage capacity of the tank remains in inventory at the end of any work week?

**Answer**:

$$P(x < .20) = \int_0^{.2} \frac{1}{B(4,3)} x^3(1-x)^2 dx = \frac{\Gamma(7)}{\Gamma(4)\Gamma(3)} \int_0^{.2} (x^3 - 2x^4 + x^5) dx$$

$$= \frac{6!}{3!2!}\left[\frac{x^4}{4} - \frac{2x^5}{5} + \frac{x^6}{6}\right]\Bigg|_0^{.2} = .01696.$$ □

## 4.3 The Normal Family of Densities

The **normal family** of densities is the most extensively used density in applied statistics. We will begin our discussion of the normal family by examining the univariate case, and then we will proceed to the multivariate normal density.

**Family Name: Univariate Normal**

*Parameterization* $(a, b) \in \Omega = \{(a, b) : a \in (-\infty, \infty), b > 0\}$

*Density Definition* $f(x; a, b) = (1/\sqrt{2\pi}b) \exp[(-1/2)((x-a)/b)^2]$.

*Moments* $\mu = a, \sigma^2 = b^2, \mu_3 = 0$

*MGF* $M_X(t) = \exp[at + (1/2)b^2t^2]$

*Background and Application* The univariate normal family of densities is indexed by the two parameters $a$ and $b$. Given the special relationship between these parameters and the mean and variance of the density, the normal density function is usually represented alternatively as

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],$$

and the moment-generating function of the normal density is given by

$$M_X(t) = \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right].$$

The abbreviation $N(z; \mu, \sigma^2)$ is often used to signify that the random variable $Z$ has a normal distribution with mean $\mu$ and variance $\sigma^2$. When the random

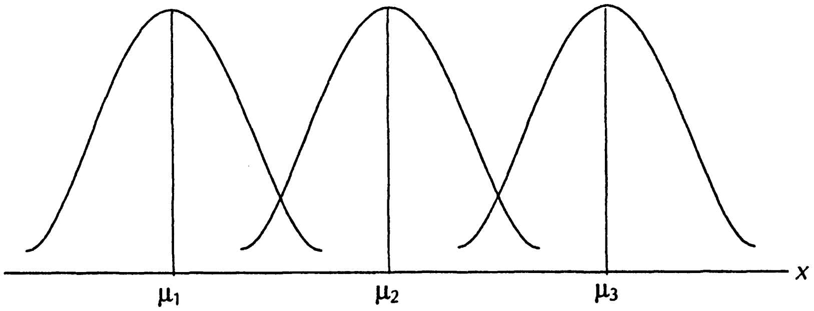


**Figure 4-11** Normal densities for fixed $\sigma$, $\mu_3 > \mu_2 > \mu_1$.

variable being referred to is not ambiguous, the abbreviation is often shortened to $N(\mu, \sigma^2)$. Once the mean and the variance of a normal distribution are numerically specified, then a unique member of the family of density functions is identified. The specific member of the family for which $\mu = 0$ and $\sigma^2 = 1$ is very important in applied statistics and is given the special name of the **standard normal density** or the **standard normal distribution** .

The normal density is symmetric about its mean, $\mu$, has points of inflection at $\mu - \sigma$ and $\mu + \sigma$, and has a characteristic bell shape. The bell becomes more spread out as the variance increases. We illustrate the general characteristics of a normal density in Figures 4.11 and 4.12.

A very useful property of any normally distributed random variable is that it can be easily transformed into a random variable having the *standard normal density.*

**Theorem 4.6** *Let X have the density $N(x;\mu,\sigma^2)$. Then $Z = (X - \mu)/\sigma$ has the density $N(z;0,1)$.*

**Proof** The MGF of $Z$ is defined by

$$M_Z(t) = \mathrm{E}\exp[tZ] = \mathrm{E}\exp\left[t\frac{(X-\mu)}{\sigma}\right] = e^{-t\mu/\sigma}M_X\left(\frac{t}{\sigma}\right) \qquad \text{(Theorem 3.26)}$$

$$= e^{-t_*\mu}M_X(t_*)\left(\text{substitute } t_* = \frac{t}{\sigma}\right)$$

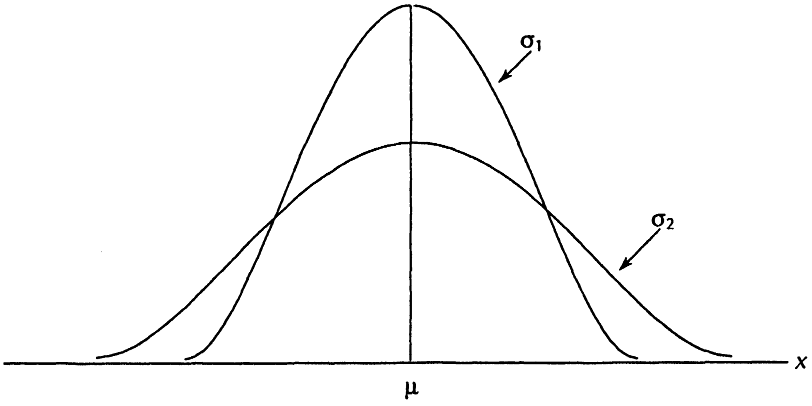


**Figure 4-12** Normal densities for fixed $\mu$, $\sigma_2 > \sigma_1$.

and since $M_X(t_*) = \exp(\mu t_* + (1/2)\sigma^2 t_*)$,

$$M_Z(t) = \exp[-t_*\mu]\exp\left[\mu t_* + (1/2)\sigma^2 t_*^2\right] = \exp[t^2/2] \qquad \left(\text{substituting } t_* = \frac{t}{\sigma}\right).$$

By the MGF uniqueness theorem, the MGF of $Z$ is identified with a standard normal density, so that $Z$ has an $N(z; 0, 1)$ density. (Note that $\exp[t^2/2] = \exp[\mu t + (1/2)\sigma^2 t^2]$, with $\mu = 0$ and $\sigma^2 = 1$). ■

In applications, Theorem 4.6 implies that the probability of an event $A$, $P_X(A)$, for a random variable $X$ having a normal density $N(x; \mu, \sigma)$ is equal to the probability $P_Z(B)$ of the equivalent event $B = \{z : z = (x - \mu)/\sigma, x \in A\}$, for a *standard normal random variable Z*. This accounts for the prevalence of published tables of the standard normal CDF since, in principle, the standard normal distribution is sufficient to assign probabilities to *all* events involving normally distributed random variables. The operation of subtracting the mean from a random variable, and then dividing by its standard deviation, is referred to as **standardizing a random variable**. The outcome value of a *standardized* random variable can be interpreted as a measure of the distance of the outcome from its mean measured in standard deviation units, (e.g., $z = 3$ would mean the outcome of $Z$ was 3 standard deviations from its mean). Thus, if a random variable having a normal density is *standardized*, the standardized random variable has a standard normal density. The random variable having the density $N(0, 1)$ is often referred to as a **standard normal random variable**.

**Example 4.16** The miles per gallon (mpg) achieved by a new pickup truck produced by a Detroit manufacturer can be viewed as a random variable having a normal density with a mean of 17 mpg and a standard deviation of .5 mpg. What is the probability that a new pickup will achieve between 16 and 18 mpg?
**Answer**: Let $X$ have the density $N(x; 17, .25)$. Then,

$$P(16 \le x \le 18) = P\left(\frac{16-17}{.5} \le \frac{x-17}{.5} \le \frac{18-17}{.5}\right) = P(-2 \le z \le 2)$$

$$= F(2) - F(-2) = .9772 - .0228 = .9544,$$

where $F(\cdot)$ is the CDF of the standard normal random variable. ■

An important observation should be made concerning the application of the normal density function in Ex. 4.16. Note that miles per gallon *cannot* be negative, and yet the normal density assigns nonzero probability to the event that $x < 0$, that is, $P(x < 0) = \int_{-\infty}^{0} N(x; \mu, \sigma^2)dx > 0 \ \forall \ \mu$ and $\sigma^2 > 0$. This is, in fact, illustrative of a situation that arises frequently in practice where the normal distribution is used in the construction of a probability space for an experiment whose outcomes assume only nonnegative values. The empirical justification for this seeming misuse of the normal density is that $P(x < 0)$ should be negligible in these cases, given the relevant values of $\mu$ and $\sigma^2$, in which case the anomaly of $P(x < 0) > 0$ can be ignored for all practical purposes. In Ex. 4.16, $P(x < 0) < 1 \times 10^{-10}$.

There is a relationship between standard normal random variables and the $\chi^2$ density that will prove to be very important for developing hypothesis-testing procedures. We develop this relationship in the following two theorems.

**Theorem 4.7** *If X has the density N(0,1), then $Y = X^2$ has a $\chi^2$ density with 1 degree of freedom.*

**Proof** The MGF of $Y$ is defined as

$$\begin{aligned} M_Y(t) &= \mathrm{E}\exp(Yt) = \mathrm{E}\exp(X^2t) = \int_{-\infty}^{\infty} \exp(x^2t)\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}x^2\right]dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}x^2(1-2t)\right]dx \\ &= (1-2t)^{-1/2}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}(1-2t)^{-1/2}}\exp\left[-\frac{1}{2}\left(\frac{x}{(1-2t)^{-1/2}}\right)^2\right]dx \\ &= (1-2t)^{-1/2}\underbrace{\int_{-\infty}^{\infty} N(x;0,(1-2t)^{-1})dx}_{1} = (1-2t)^{-1/2} \quad \text{for } t < \frac{1}{2}. \end{aligned}$$

Therefore, by the MGF uniqueness theorem, $Y = X^2$ has a $\chi^2$ density with 1 degree of freedom. ■

**Theorem 4.8** (Sums of Squares of Independent Standard Normal Random Variables) *Let $(X_1,\ldots,X_n)$ be independent random variables, each having the density N(0,1). Then $Y = \sum_{i=1}^n X_i^2$ has a $\chi^2$ density with n degrees of freedom.*

**Proof** The random variables $X_i^2$, $i = 1,\ldots,n$, are independent by Theorem 2.9. Then, from Theorems 3.26 and 4.7,

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i^2}(t) = \prod_{i=1}^{n}(1-2t)^{-1/2} = (1-2t)^{-n/2} \quad \text{for } t < \frac{1}{2},$$

which by the MGF uniqueness theorem implies that $Y$ has a $\chi^2$ density with $n$ degrees of freedom. ■

In words, Theorem 4.8 is often stated as *the sum of squares of n independent standard normal random variables has a $\chi^2$ density with n degrees of freedom.* We can now motivate why the parameter $v$ in a $\chi^2$ density is labeled **degrees of freedom**. In particular, $v$ represents the number, or *degree*, of *freely* varying (i.e., independent) standard normal random variables whose sum of squares represents a $\chi_v^2$-distributed random variable.

It will be useful in our later study of hypothesis testing to know the values of $h$ and $\ell$ for which the following statements are true:

$P(x \geq h) = \alpha = 1 - F(h)$ and $P(x \leq \ell) = \alpha = F(\ell)$, where $F(\cdot)$ refers to the CDF of a random variable having an $N(\mu, \sigma^2)$ density and $\alpha \in (0, 1)$. We diagrammatically examine the upper and lower tails of the normal family of densities in Figure 4.13. The probabilities of such events are extensively tabled for the case of the *standard* normal density, where typically the values of the cumulative standard normal density, $F(c)$, are given for numerous choices of $c$ (see Appendix). The probabilities of these events for arbitrary normal distributions can be obtained from knowledge of the standard normal distribution via standardization. For the event $x \leq \ell$,

$$P(x \leq \ell) = P\left(\frac{x-\mu}{\sigma} \leq \frac{\ell-\mu}{\sigma}\right) = P(z \leq \ell_*),$$

where $\ell_* = (\ell - \mu)/\sigma$ and $z = (x - \mu)/\sigma$ can be interpreted as the outcome of a *standard* normal random variable using Theorem 4.6. Thus, the value of $P(x \leq \ell)$ can be obtained from the standard normal table as equal to the probability that $z \leq \ell_* = (\ell - \mu)/\sigma$. Similarly,

$$P(x \geq h) = P\left(\frac{x-\mu}{\sigma} \geq \frac{h-\mu}{\sigma}\right) = P(z \geq h_*), \quad \text{where } h_* = \frac{h-\mu}{\sigma}.$$

**Example 4.17** Find the probability that $x \geq 5.29$ where $X$ is a random variable having a normal distribution with mean 2 and variance 4.

**Answer:** Note that

$$P(x \geq 5.29) = P\left(\frac{x-2}{2} \geq \frac{5.29-2}{2}\right) = P(z \geq 1.645) = .05,$$

which was found from the table of the standard normal CDF. □

A number of computer programs exist that will numerically integrate the standard normal density over the interval $(-\infty, c]$ chosen by the user. The reader can utilize SAS, Shazam, or the GAUSS programming language, for example.

The normal density is often used in modeling experiments for which a symmetric, bell-shaped probability density is suspected. The normal density has been found to be a useful representation of event probabilities for literally

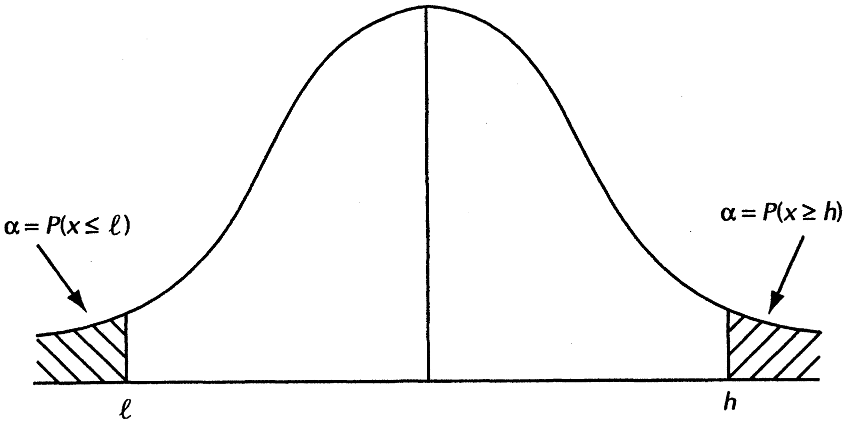


**Figure 4-13** Upper and lower $\alpha$-level tails of $N(\mu, \sigma^2)$

thousands of real-world experiments, which is attributable in large part to the fact that, under general conditions, certain useful functions of a collection of independent random variables, such as sums and averages, have *approximately* normal densities when the collection is large enough, even if the original random variables in the collection do *not* have normal densities. In fact, the normal density was originally discovered by A. de Moivre (1667–1745) as an *approximation* to the binomial density (recall that the binomial density applies to the *sum* of independent Bernoulli random variables). These results are based on *central limit theorems*, which will be examined in Chapter 5.

Examples of applications include fill weights of food and beverage containers, employee aptitude test scores, labor hours required to construct prefabricated homes, numbers of insurance claims filed during a time period, and weight gains of meat animals in a feedlot. In Chapter 10 we will examine statistical tests that can be used to assess the validity of the normal PDF for characterizing event probabilities in real-world experiments.

**Family Name: Multivariate Normal Density**

*Parameterization*

$$\mathbf{a} = (a_1, \ldots, a_n)' \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nn} \end{bmatrix}$$

$(\mathbf{a}, \mathbf{B}) \in \Omega = \{(\mathbf{a}, \mathbf{B}) : \mathbf{a} \in R^n, \mathbf{B}$ is a symmetric, $(n \times n)$, positive definite matrix$\}$.

*Density Definition*

$$f(\mathbf{x}; \mathbf{a}, \mathbf{B}) = \frac{1}{(2\pi)^{n/2}|\mathbf{B}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{a})'\mathbf{B}^{-1}(\mathbf{x}-\mathbf{a})\right]$$

*Moments*

$$\underset{(n\times 1)}{\boldsymbol{\mu}} = \mathbf{a}, \qquad \underset{n\times n}{\mathbf{Cov}\,(\mathbf{X}) = \mathbf{B}} \qquad \underset{(n\times 1)}{\boldsymbol{\mu}_3} = [\mathbf{0}]$$

*MGF* $M_{\mathbf{X}}(t) = \exp[\mathbf{a}'\mathbf{t} + (1/2)\mathbf{t}'\mathbf{B}\mathbf{t}]$, where $\mathbf{t} = (t_1, \ldots, t_n)'$.

*Background and Application* The $n$-variate normal family of densities is indexed by $n + n(n+1)/2$ parameters consisting of $n$ elements in the $(n \times 1)$ vector $\mathbf{a}$ and $n(n+1)/2$ elements representing the distinct elements in the $(n \times n)$ symmetric matrix $\mathbf{B}$ (while there are $n^2$ number of elements in the matrix $\mathbf{B}$, only $n(n+1)/2$ of these elements are distinct (or different) given the symmetry of $\mathbf{B}$, i.e., $b_{ij} = b_{ji}$ $\forall\, i \neq j$). Given the special relationship between the mean vector, the covariance matrix, and the parameters $\mathbf{a}$ and $\mathbf{B}$, the $n$-variate normal density is most often represented by

$$N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right],$$

where $\mathbf{\Sigma}$ is a popular notation for the covariance matrix of $\mathbf{X}$ (The lack of summation range indicators and the problem context will have to be relied upon to distinguish between when $\Sigma$ designates a covariance matrix and when $\Sigma$ signifies summation.) When it is clear which random vector is being referred to, the notation $N(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma})$ is often shortened to $N(\boldsymbol{\mu}, \mathbf{\Sigma})$. The MGF of $X$ is represented by

$$M_X(t) = \exp(\boldsymbol{\mu}'\mathbf{t} + .5\mathbf{t}'\mathbf{\Sigma}\mathbf{t}).$$

In order to illustrate graphically some of the characteristics of the multivariate normal density, we will temporarily concentrate on the bivariate case. Thus, in the above formula, $n = 2$, $\boldsymbol{\mu}$ is a $(2 \times 1)$ column vector, and $\mathbf{\Sigma}$ is a $(2 \times 2)$ positive, definite covariance matrix. The graph of the bivariate normal density is a three-dimensional bell of sorts, such as the illustration in Figure 4.14.a. The mode of the normal density occurs at $\mathbf{x} = \boldsymbol{\mu}$, which in the bivariate case occurs at $x_1 = \mathrm{E}X_1$ and $x_2 = \mathrm{E}X_2$. **Iso-density contours** (i.e., the collection of $(x_1, x_2)$ points resulting in a fixed value of the joint density function, $f(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) = c$) are in the form of ellipses (ellipsoids in higher dimensions, e.g., a "football" in three dimensions) with center at $\boldsymbol{\mu}$, so that in the bivariate case the center of the ellipse is at $(\mathrm{E}X_1, \mathrm{E}X_2)$. One can think of these ellipses as being formed by "slicing" through the density at a certain height, removing the top portion of the density, and then projecting the exposed elliptical top onto the $(x_1, x_2)$-plane (see Figure 4.14.b).

The shape and orientation of the ellipse (or ellipsoid) are determined by the elements of the covariance matrix, $\mathbf{\Sigma}$. The major (larger) axis of the ellipse as measured from the origin, $\boldsymbol{\mu}$, is in the direction of the characteristic vector of $\mathbf{\Sigma}^{-1}$ associated with the smallest characteristic root of $\mathbf{\Sigma}^{-1}$.[7] The length of an axis is given by $2k/\sqrt{\lambda_i}$, where $k = \sqrt{2\pi}|\mathbf{\Sigma}|^{.5}c$, $c$ is the chosen value of the density, and $\lambda_i$ is either the smallest or largest characteristic root of $\mathbf{\Sigma}^{-1}$.[8]

In the bivariate case, the slope $dx_1/dx_2$, of the major axis of the ellipse is positive if $\sigma_{12} > 0$ and negative if $\sigma_{12} < 0$. The slope increases in absolute value as the ratio $\sigma_1/\sigma_2$ increases, holding $\rho$ (the correlation coefficient) constant. As $|\sigma_{12}| \to \sigma_1\sigma_2$, so that $\rho \to \pm 1$, the length of the minor axis of the ellipse $\to 0$, and the ellipse concentrates on the major axis, which approaches the line $x_1 = \mathrm{E}X_1 \pm (\sigma_1/\sigma_2)(x_2 - \mathrm{E}X_2)$ (compare to Theorem 3.35). If $\sigma_1 = \sigma_2$, the principal axis is given by the line $x_1 = \mathrm{E}X_1 \pm (\sigma_1/\sigma_2)(x_2 - \mathrm{E}X_2)$, for any magnitude of the correlation.

An illustration of some of the myriad of possibilities for the graphs of the iso-density contours is given in Figure 4.15.

[7] Recall that the characteristic roots and vectors of a square matrix **A** are the scalars, $\lambda$, and associated vectors, **p**, that satisfy the equation $[\mathbf{A} - \lambda\mathbf{I}]\mathbf{p} = [\mathbf{0}]$. There will be as many roots and associated vectors as there are rows (or columns) in the square matrix **A**.

[8] These arguments extend in a natural way to higher dimensions, in which case we are examining $n$ axes of the ellipsoid. See B. Bolch, and C. Huang (1974), *Multivariate Statistical Methods for Business and Economics*. Englewood Cliffs, NJ: Prentice-Hall, p. 19–23, for the matrix theory underlying the derivation of the results on axis length and orientation discussed here.

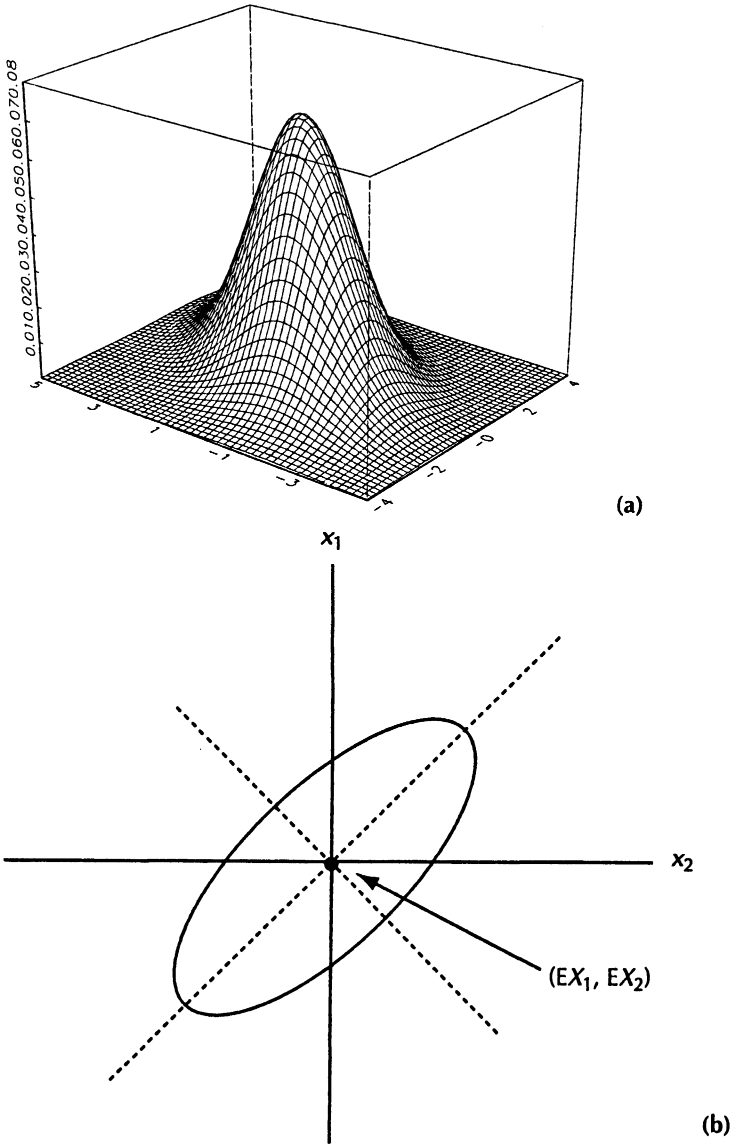


**Figure 4-14**
Bivariate normal density; (a) shows a view of normal density in three dimensions, and (b) illustrates an iso-density ellipse.

The multivariate normal density is often used in modeling experiments characterized by a PDF that is symmetric about its mean vector, $\mu$, is bell-shaped (when viewed in three or fewer dimensions), and is such that the highest density weighting (mode) occurs at the mean vector with density values declining as $\mathbf{x}$ becomes more distant from $\mathbf{EX}$ in any direction. Also, under rather general conditions, certain important vector functions of a collection of independent multivariate random variables, such as sums and averages, have approximately a normal distribution when the collection is large enough, even

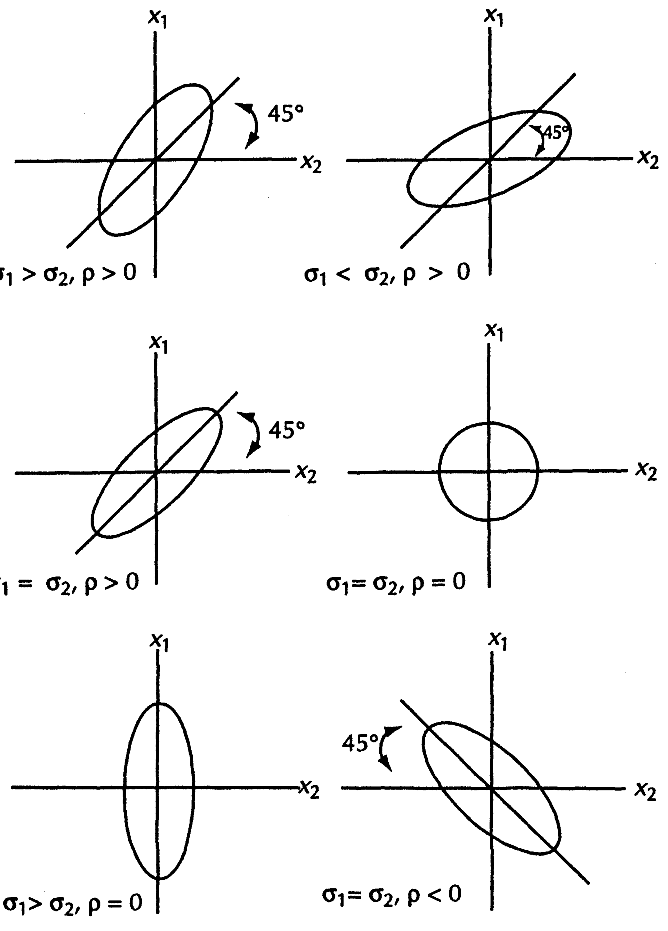

$x_1$
45°
$x_2$
$\sigma_1 > \sigma_2, \rho > 0$
$x_1$
45°
$x_2$
$\sigma_1 < \sigma_2, \rho > 0$
$x_1$
45°
$x_2$
$\sigma_1 = \sigma_2, \rho > 0$
$x_1$
$x_2$
$\sigma_1 = \sigma_2, \rho = 0$
$x_1$
$x_2$
$\sigma_1 > \sigma_2, \rho = 0$
$x_1$
45°
$x_2$
$\sigma_1 = \sigma_2, \rho < 0$

**Figure 4-15**
Iso-density ellipses of bivariate normal density (all origins at $(EX_1, EX_2)$

if the original multivariate random variables in the collection do *not* have normal probability distributions. These results are known as *multivariate central limit theorems*, which we will study in Chapter 5 and which account for much of the motivation for the assumption of multivariate normality in empirical work.

*Marginal Densities* A useful property of the multivariate normal family is that the marginal density of any subset of the random variables $(X_1, \ldots, X_n)$ and the conditional density of any subset of the random variables $(X_1, \ldots, X_n)$, *given* an elementary event for the remaining random variables, are in the normal family and are easy to identify. We first present an important theorem concerning the PDF of a linear combination of normally distributed variables. The theorem will facilitate our examination of marginal and conditional densities and is also useful in its own right.

**Theorem 4.9** (PDF of Linear Combinations of Normal Random Variables) *Let* $\mathbf{X}$ *be an* $n$*-variate random variable having the density function* $N(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$*. Let* $\mathbf{A}$ *be any* $(k \times n)$ *matrix of real constants with rank* $k$*, and let* $\mathbf{b}$ *be any* $(k \times 1)$ *vector of real constants. Then the* $(k \times 1)$ *random vector* $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$ *has the density* $N(\mathbf{y};\mathbf{A}\boldsymbol{\mu} + \mathbf{b},\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.

**Proof** The MGF of $\mathbf{Y}$ is defined as

$$
\begin{aligned}
M_{\mathbf{Y}}(\mathbf{t}) &= \mathrm{E}\exp(\mathbf{t}'\mathbf{Y}) = \mathrm{E}\exp(\mathbf{t}'(\mathbf{AX}+\mathbf{b}))\\
&= \exp(\mathbf{t}'\mathbf{b})\mathrm{E}\exp(\mathbf{t}'\mathbf{AX}) = \exp(\mathbf{t}'\mathbf{b})\exp\left(\mathbf{t}'\mathbf{A}\boldsymbol{\mu} + \tfrac{1}{2}\mathbf{t}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'\mathbf{t}\right)\\
&= \exp\left(\mathbf{t}'(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}) + \tfrac{1}{2}\mathbf{t}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'\mathbf{t}\right),
\end{aligned}
$$

where the next-to-last equality follows from the fact that $\mathbf{X}$ is normally distributed, and $\mathrm{E}\exp(\mathbf{t}'\mathbf{AX}) = \mathrm{E}\exp(\mathbf{t}_*'\mathbf{X}) = M_{\mathbf{X}}(\mathbf{t}_*) = \exp[\mathbf{t}_*'\boldsymbol{\mu} + (1/2)\mathbf{t}_*'\boldsymbol{\Sigma}\mathbf{t}_*]$, with $\mathbf{t}_*' = \mathbf{t}'\mathbf{A}$. Thus, $M_{\mathbf{Y}}(\mathbf{t})$ identifies the multivariate normal density with mean $\mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ and covariance matrix $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ as the density of $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$. ■

**Example 4.18** A firm uses two variable inputs in the manufacture of an output and also uses "just-in-time" production methods so that inputs arrive precisely when they are needed to produce the output. The weekly average input and selling prices, $\mathbf{r}$ and $p$, during the spring quarter can be viewed as the outcomes of a trivariate normal density function $f(r_1, r_2, p;\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$
\boldsymbol{\mu} = \begin{bmatrix} .50 \\ 1.25 \\ 5 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} .05 & .02 & .01 \\ .02 & .10 & .01 \\ .01 & .01 & .40 \end{bmatrix}.
$$

What is the density function of the bivariate random variable $(\Pi, C)$, where $\Pi$ represents profit above variable cost and $C$ represents variable cost in a week where 100 units of input 1 and 150 units of input 2 are utilized, and 100 units of output are produced and sold?

**Answer**: The random variables $\Pi$ and $C$ can be defined as

$$
\mathbf{Y} = \begin{bmatrix} \Pi \\ C \end{bmatrix} = \begin{bmatrix} -100 & -150 & 100 \\ 100 & 150 & 0 \end{bmatrix} \begin{bmatrix} R_1 \\ R_2 \\ P \end{bmatrix} = \mathbf{A}X + \mathbf{b},
$$

where $\mathbf{b} = [\mathbf{0}]$ and $\mathbf{A}$ is the bracketed $(2 \times 3)$ matrix following the second equality sign. Then, Theorem 4.9 implies that $[\Pi, C]'$ is bivariate normally distributed as $N(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$, with $\boldsymbol{\mu}_* = \mathbf{A}\boldsymbol{\mu} = [\,262.5 \quad 237.5\,]'$ and

$$
\boldsymbol{\Sigma}_* = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \begin{bmatrix} 6850 & -3100 \\ -3100 & 3350 \end{bmatrix}. \qquad \square
$$

A useful implication of Theorem 4.9 for generating or *simulating* outcomes of multivariate normally distributed random variables on the computer is that *any* such random variable can be represented in terms of linear combinations of independent random variables having the standard normal density. Specifically, if the $n \times 1$ random variable $\mathbf{Z}$ has the PDF $N([\mathbf{0}], \mathbf{I})$ so that $Z_1, \ldots, Z_n$ are independent $N(0, 1)$ random variables, then the $n \times 1$ random variable $\mathbf{Y}$ with PDF $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be represented in terms of $\mathbf{Z}$ as $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}$, where $\mathbf{A}$ is chosen so that $\mathbf{A}\mathbf{A}' = \boldsymbol{\Sigma}$.[9] The utility of this representation stems from the fact that many modern statistical software packages are capable of generating independent outcomes of a random variable having the $N(0, 1)$ density; such programs are referred to as "standard normal random-number generators." We will further pursue the notion of simulating random-variable outcomes in Chapter 6.

We now state an important result concerning the marginal density of subsets of the random variables $(X_1, \ldots, X_n)$ when the $n$-variate random variable has an $n$-variate normal density.

**Theorem 4.10** (Marginal Densities for $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$) *Let* $\mathbf{Z}$ *have the density* $N(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma})$, *where*

$$\mathbf{Z} = \left[\begin{array}{c} \underset{(m\times 1)}{\mathbf{Z}_{(1)}} \\ \hline \underset{(n-m)\times 1}{\mathbf{Z}_{(2)}} \end{array}\right], \quad \boldsymbol{\mu} = \left[\begin{array}{c} \underset{(m\times 1)}{\boldsymbol{\mu}_{(1)}} \\ \hline \underset{(n-m)\times 1}{\boldsymbol{\mu}_{(2)}} \end{array}\right], \quad \textit{and} \quad \boldsymbol{\Sigma} = \left[\begin{array}{c|c} \underset{(m\times m)}{\boldsymbol{\Sigma}_{11}} & \underset{(m\times(n-m))}{\boldsymbol{\Sigma}_{12}} \\ \hline \underset{((n-m)\times m)}{\boldsymbol{\Sigma}_{21}} & \underset{((n-m)\times(n-m))}{\boldsymbol{\Sigma}_{22}} \end{array}\right]$$

*Then the marginal PDF of* $\mathbf{Z}_{(1)}$ *is* $N(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_{11})$, *and the marginal PDF of* $\mathbf{Z}_{(2)}$ *is* $N(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_{22})$.

**Proof** Let

$$\mathbf{A} = \left[\underset{(m\times m)}{\mathbf{I}} \;\Big|\; \underset{m\times(n-m)}{[\;\mathbf{0}\;]}\right]$$

and $\mathbf{b} = [0]$ in Theorem 4.9, where $\mathbf{I}_{(m\times m)}$ is the $(m \times m)$ identity matrix. It follows that $\mathbf{Z}_{(1)} = \mathbf{A}\mathbf{Z}$ has the normal density $N(\boldsymbol{\mu}_{(1)}, \boldsymbol{\Sigma}_{11})$. The result for $\mathbf{Z}_{(2)}$ is proven similarly by letting

$$\mathbf{A} = \left[\underset{(n-m)\times m}{[\;\mathbf{0}\;]} \;\Big|\; \underset{(n-m)\times(n-m)}{\mathbf{I}}\right]. \qquad \blacksquare$$

**Example 4.19** Referring to Ex. 4.18, partition $\mathbf{X}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as

$$\mathbf{X} = \left[\begin{array}{c} \mathbf{X}_{(1)} \\ \hline \mathbf{X}_{(2)} \end{array}\right] = \left[\begin{array}{c} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \hline \mathbf{X}_3 \end{array}\right], \quad \boldsymbol{\mu} = \left[\begin{array}{c} \boldsymbol{\mu}_{(1)} \\ \hline \boldsymbol{\mu}_{(2)} \end{array}\right] = \left[\begin{array}{c} \mu_1 \\ \mu_2 \\ \hline \mu_3 \end{array}\right] = \left[\begin{array}{c} .50 \\ 1.25 \\ \hline 5 \end{array}\right]$$

and

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right] = \left[\begin{array}{cc|c} .05 & .02 & .01 \\ .02 & .10 & .01 \\ \hline .01 & .01 & .40 \end{array}\right].$$

[9] The symmetric matrix square root $\boldsymbol{\Sigma}^{1/2}$ of $\boldsymbol{\Sigma}$ could be chosen for $\mathbf{A}$. Alternatively, there exists a lower triangular matrix, called the Cholesky decomposition of $\boldsymbol{\Sigma}$, which satisfies $\mathbf{A}\mathbf{A}' = \boldsymbol{\Sigma}$. Either choice of $\mathbf{A}$ can be calculated straightforwardly on the computer, such as through the use of the GAUSS programming language.

It follows from Theorem 4.10 that the marginal densities of $(X_1, X_2)$ and $X_3$ are given by

$$f_{X_1X_2}(x_1, x_2) = N\left(\begin{bmatrix} .50 \\ 1.25 \end{bmatrix}, \begin{bmatrix} .05 & .02 \\ .02 & .10 \end{bmatrix}\right) \quad \text{and} \quad f_{X_3}(x_3) = N(5, .4). \qquad \square$$

The reader may have noticed that there is an intuitive way of interpreting the implementation of Theorem 4.10. Specifically, by simply "crossing out" any entries in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that refer in any way to random variables that are not of interest, one is left with the mean vector and covariance matrix of the normal density for the random variables that are of interest. For example, in Ex. 4.19, if $(X_1, X_2)$ is the random vector of interest, then by "crossing out" $\mu_{(2)} = \mu_3$, $\Sigma_{12}$, $\Sigma_{21}$, and $\Sigma_{22}$, one is left with the mean vector and covariance matrix of the normally distributed random vector $(X_1, X_2)$.

The reader should note that the order of the random variables in the $\mathbf{X}$ vector is arbitrary. For example, we might have $\mathbf{X} = [X_1, X_2, X_3, X_4]'$ or, alternatively, the same random variables might be listed as $\mathbf{X} = [X_3, X_1, X_4, X_2]'$. The point is that Theorem 4.10 can be applied to obtain the marginal density function of *any* subset of the random variable $(X_1, \ldots, X_n)$ by simply ordering them appropriately in the definition of $\mathbf{Z}$ in the theorem. Of course, the entries in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ must be correspondingly ordered so that random variables are associated with their appropriate means, variances, and covariances.

**Example 4.20** The annual percentage return on three investment instruments is the outcome of a trivariate random variable $[X_1, X_2, X_3]'$ having the density $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{bmatrix} 2 \\ 7 \\ 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 3 \end{bmatrix}.$$

To identify the marginal density of the returns on investments 1 and 3 using Theorem 4.10, first reorder the random variables as $\mathbf{X} = [X_1, X_3, X_2]'$ so that the corresponding mean vector and covariance matrix of the trivariate normal density of $\mathbf{X}$ are now

$$\boldsymbol{\mu}_* = \begin{bmatrix} 2 \\ 1 \\ 7 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\Sigma}_* = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Then, a straightforward application of Theorem 4.10 (with the appropriate interpretation of the symbols $\mathbf{Z}_{(1)}$ and $\mathbf{Z}_{(2)}$ used in that theorem) implies that $[X_1, X_3]'$ has the density

$$N\left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 4 & 0 \\ 0 & 3 \end{bmatrix}\right).$$

The reader should note that our intuitive "crossing out" procedure discussed above can be applied here, where one "crosses out" any entries in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ referring to $X_2$. □

*Conditional Densities* The definition of conditional densities when dealing with $n$-variate normally distributed random variables is somewhat more involved than the case of defining marginal densities. We present the case of conditioning on an elementary event for a subset of the random variables.

**Theorem 4.11** (Conditional Densities for $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$) *Let* $\mathbf{Z}$ *be as defined in Theorem 4.10, and let*

$$\underset{(n\times 1)}{\mathbf{z}^0} = \begin{bmatrix} \underset{(m\times 1)}{\mathbf{z}^0_{(1)}} \\ \hline \underset{(n-m)\times 1}{\mathbf{z}^0_{(2)}} \end{bmatrix}$$

*be a vector of constants. Then*

$$f(\mathbf{z}_{(1)} \mid \mathbf{z}_{(2)} = \mathbf{z}^0_{(2)}) = N(\boldsymbol{\mu}_{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{z}^0_{(2)} - \boldsymbol{\mu}_{(2)}), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}),$$

$$f(\mathbf{z}_{(2)} \mid \mathbf{z}_{(1)} = \mathbf{z}^0_{(1)}) = N(\boldsymbol{\mu}_{(2)} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{z}^0_{(1)} - \boldsymbol{\mu}_{(1)}), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}).$$

**Proof** We prove the result for the case of $\mathbf{Z}_{(1)}$; the case for $\mathbf{Z}_{(2)}$ can be proved analogously. By definition,

$$f(\mathbf{z}_{(1)} \mid \mathbf{z}_{(2)} = \mathbf{z}^0_{(2)}) = \frac{f(\mathbf{z}_{(1)}, \mathbf{z}^0_{(2)})}{f_{\mathbf{Z}_{(2)}}(\mathbf{z}^0_{(2)})}$$

$$= \frac{\frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}\begin{bmatrix}\mathbf{z}_{(1)} - \boldsymbol{\mu}_{(1)} \\ \mathbf{z}^0_{(2)} - \boldsymbol{\mu}_{(2)}\end{bmatrix}'\boldsymbol{\Sigma}^{-1}\begin{bmatrix}\mathbf{z}_{(1)} - \boldsymbol{\mu}_{(1)} \\ \mathbf{z}^0_{(2)} - \boldsymbol{\mu}_{(2)}\end{bmatrix}\right)}{\frac{1}{(2\pi)^{(n-m)/2}|\boldsymbol{\Sigma}_{22}|^{1/2}}\exp(-\frac{1}{2}(\mathbf{z}^0_{(2)} - \boldsymbol{\mu}_{(2)})'\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{z}^0_{(2)} - \boldsymbol{\mu}_{(2)}))}.$$

The following lemma on partitioned determinants and partitioned inversion will be useful.

**Lemma 4.3**
***Partitioned Inversion and Partitioned Determinants***

Partition the $(n \times n)$ matrix $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \underset{(m\times m)}{\boldsymbol{\Sigma}_{11}} & \underset{m\times(n-m)}{\boldsymbol{\Sigma}_{12}} \\ \hline \underset{(n-m)\times m}{\boldsymbol{\Sigma}_{21}} & \underset{(n-m)\times(n-m)}{\boldsymbol{\Sigma}_{22}} \end{array}\right].$$

**a.** If $\boldsymbol{\Sigma}_{11}$ is nonsingular, then $|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{11}| \cdot |\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}|$.

**b.** If $\boldsymbol{\Sigma}_{22}$ is nonsingular, then $|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{22}| \cdot |\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}|$.

**c.** If $|\boldsymbol{\Sigma}| \neq 0$, $|\boldsymbol{\Sigma}_{11}| \neq 0$, and $|\boldsymbol{\Sigma}_{22}| \neq 0$, then

$$\boldsymbol{\Sigma}^{-1} = \left[\begin{array}{c|c} (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{-1} & -(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \hline -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{-1} & (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \end{array}\right].$$

**d.** The diagonal blocks in the partitioned matrix of part (c) can also be expressed as

$$(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}=\Sigma_{11}^{-1}+\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1}$$

and

$$(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}=\Sigma_{22}^{-1}+\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}$$

(see F. A. Graybill (1983), *Matrices with Applications in Statistics*, 2nd ed., Belmont, CA: Wadsworth, pp. 183–186, for further discussion and proofs).

Utilizing Lemma 4.3, note that

$$\frac{|\Sigma_{22}|^{1/2}}{|\Sigma|^{1/2}}=|\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|^{-1/2}$$

and

$$\exp\left[-\tfrac{1}{2}\begin{bmatrix}\mathbf{z}_{(1)}-\mu_{(1)}\\ \mathbf{z}^0_{(2)}-\mu_{(2)}\end{bmatrix}'\Sigma^{-1}\begin{bmatrix}\mathbf{z}_{(1)}-\mu_{(1)}\\ \mathbf{z}^0_{(2)}-\mu_{(2)}\end{bmatrix}+\tfrac{1}{2}(\mathbf{z}^0_{(2)}-\mu_{(2)})'\Sigma_{22}^{-1}(\mathbf{z}^0_{(2)}-\mu_{(2)})\right]$$

$$=\exp[-\tfrac{1}{2}(\mathbf{z}_{(1)}-(\mu_{(1)}+\Sigma_{(12)}\Sigma_{22}^{-1}(\mathbf{z}^0_{(2)}-\mu_{(2)})))'\Phi^{-1}(\mathbf{z}_{(1)}-(\mu_{(1)}+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{z}^0_{(2)}-\mu_{(2)})))],$$

where $\Phi=[\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]$. Then the expression for the conditional density function reduces to

$$f(\mathbf{z}_{(1)}\,|\,\mathbf{z}_{(2)}=\mathbf{z}^0_{(2)})=\frac{1}{(2\pi)^{m/2}|\Phi|^{1/2}}\exp[-\tfrac{1}{2}(\mathbf{z}_{(1)}-\gamma)'\Phi^{-1}(\mathbf{z}_{(1)}-\gamma)],$$

where $\gamma=\mu_{(1)}+\Sigma_{12}\Sigma_{22}^{-1}\left(\mathbf{z}^0_{(2)}-\mu_{(2)}\right)$, which is the normal density $N(\gamma,\Phi)$. ■

**Example 4.21** Let $X=(X_1,X_2,X_3)$ be the trivariate random variable representing percentage returns on investments in Ex. 4.20. Suppose we wanted to define the conditional density of returns on investment instrument 1, *given* $x_2=1$ and $x_3=2$. By an appropriate interpretation of Theorem 4.11 (i.e., let $\mathbf{z}_{(1)}=x_1$ and $\mathbf{z}_{(2)}=(x_2,x_3)'$), we know that $X_1$ will have a normal density with mean

$$\mathrm{E}(x_1\,|\,x_2=1,x_3=2)=2+[1\quad 0]\begin{bmatrix}1&1\\1&3\end{bmatrix}^{-1}\begin{bmatrix}-6\\1\end{bmatrix}=-7.5$$

and variance

$$\sigma^2_{(X_1|x_2=1,x_3=2)}=4-[1\quad 0]\begin{bmatrix}1&1\\1&3\end{bmatrix}^{-1}\begin{bmatrix}1\\0\end{bmatrix}=2.5,$$

that is, the *conditional* density of $X_1$ is $f(x_1\,|\,x_2=1,x_3=2)=N(-7.5,2.5)$.

To find the conditional density of returns on investments 1 and 2 *given* $x_3=0$, Theorem 4.11 (with $\mathbf{z}_{(1)}=(x_1,x_2)'$ and $\mathbf{z}_{(2)}=x_3$) implies that the mean vector is equal to

$$\mathrm{E}\left[\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \,\middle|\, x_3 = 0\right] = \begin{bmatrix} 2 \\ 7 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}[3]^{-1}[-1] = \begin{bmatrix} 2 \\ \frac{20}{3} \end{bmatrix},$$

and the covariance matrix is

$$\mathbf{Cov}((X_1, X_2) \mid x_3 = 0) = \begin{bmatrix} 4 & 1 \\ 1 & 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix}[3]^{-1}[0 \quad 1] = \begin{bmatrix} 4 & 1 \\ 1 & \frac{2}{3} \end{bmatrix}. \qquad \square$$

In the special case where $X_1$ is a scalar and $\mathbf{X}_2$ is a $(k \times 1)$ vector, the conditional expectation of $X_1$ expressed as a *function* of $\mathbf{x}_2$ defines the *regression function* of $X_1$ on $\mathbf{X}_2$, which in this case can also be referred to as the **regression hyperplane** of $X_1$ on $\mathbf{X}_2$. If $X_2$ is also a scalar, the regression hyperplane is a **regression line**. The regression hyperplane is thus given by

$$\mathrm{E}(X_1 \mid \underset{(k \times 1)}{\mathbf{x}_2}) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2) = a + \mathbf{B}\mathbf{x}_2,$$

where $\mathbf{B} = \Sigma_{12}\Sigma_{22}^{-1}$ and $a = \mu_1 - \mathbf{B}\mu_2$. The special case of the regression line can be written as

$$\mathrm{E}(X_1 \mid x_2) = \mu_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2 - \mu_2) = a + bx_2,$$

where $\rho$ is the correlation coefficient, $b = \rho\sigma_1/\sigma_2$, and $a = \mu_1 - b\mu_2$. Figure 4.16 illustrates the regression line of $X_1$ on $X_2$.

The particular functional form of the normal density is the reason why the regression function becomes a regression *hyperplane* and the regression curve becomes a regression *line*. In fact, the normal density belongs to a collection

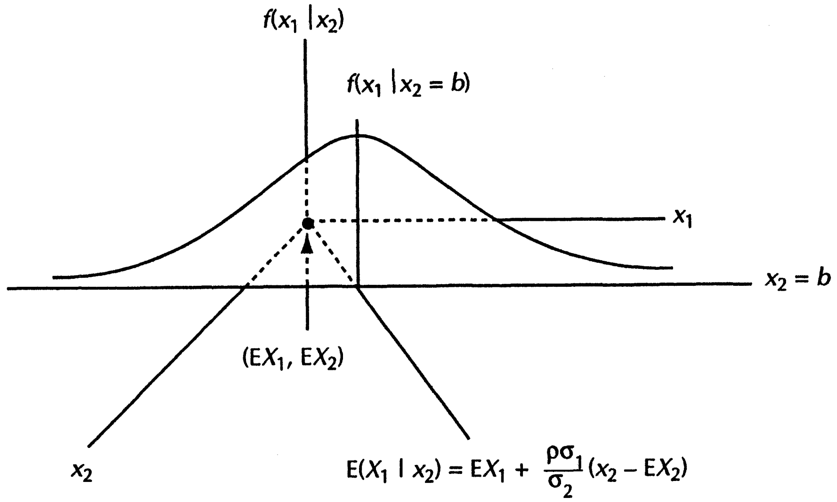

$f(x_1 \mid x_2)$
$f(x_1 \mid x_2 = b)$
$x_1$
$x_2 = b$
$(\mathrm{E}X_1, \mathrm{E}X_2)$
$x_2$
$\mathrm{E}(X_1 \mid x_2) = \mathrm{E}X_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2 - \mathrm{E}X_2)$

**Figure 4-16**
Regression line of $X_1$ on $X_2$.

of families of distributions referred to as *elliptically contoured distributions*, all of which are associated with linear regression functions (S. Cambanis, et al. (1981), *On the Theory of Elliptically Contoured Distributions J. Multivariate Analysis*, 11, p. 368).

While not true in general, in the case of a normal distribution, zero covariance implies independence of random variables, as stated in the following theorem.

**Theorem 4.12** (Cov (**X**) = [**0**] $\Rightarrow$ Independence When **X** has PDF $N(\mu,\Sigma)$) *Let* $\mathbf{X} = [X_1, \ldots, X_n]'$ *have the density* $N(\mu,\Sigma)$. *Then* $(X_1, \ldots, X_n)$ *are independent iff* $\Sigma$ *is a diagonal matrix.*

**Proof** The *only if* part of the theorem follows immediately from the fact that independence of random variables implies zero covariance. To see the *if* part, suppose $\Sigma$ is a diagonal matrix. Then the density of **X** can be written as

$$N(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)\right]$$

$$= \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}\sigma_i} \exp\left[-\frac{1}{2}\frac{(x_i - \mu_i)^2}{\sigma_i^2}\right] = \prod_{i=1}^{n} N(x_i; \mu_i, \sigma_i^2),$$

since

$$|\Sigma| = \prod_{i=1}^{n} \sigma_i^2, \quad \text{and} \quad \Sigma^{-1} = \begin{bmatrix} \sigma_1^{-2} & & \\ & \ddots & \\ & & \sigma_n^{-2} \end{bmatrix}.$$

Because the joint density factors into the product of the $n$ marginal densities, $(X_1, \ldots, X_n)$ are independent. ■

A variation on the preceding theme is the case where two *vectors* $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent by virtue of the covariances between the elements in $\mathbf{X}_1$ and the elements in $\mathbf{X}_2$ being zero.

**Theorem 4.13** ($\mathbf{Z}_{(1)}$ and $\mathbf{Z}_{(2)}$ Independent $\Leftrightarrow$ $\Sigma_{12} = [0]$ for $N(\mu,\Sigma)$) *Let*

$$\underset{(n\times 1)}{\mathbf{Z}} = \begin{bmatrix} \underset{(m\times 1)}{\mathbf{Z}_{(1)}} \\ \hline \underset{(n-m)\times 1}{\mathbf{Z}_{(2)}} \end{bmatrix}$$

*have the multivariate normal density identified in Theorem 4.10. Then the vectors* $\mathbf{Z}_{(1)}$ *and* $\mathbf{Z}_{(2)}$ *are independent iff* $\Sigma_{12} = \Sigma_{21}' = [\mathbf{0}]$.

**Proof** The *only if* part of the theorem follows immediately from the fact that the independence of random variables implies zero covariance. To see the *if* part, suppose $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21} = [0]$. Then the PDF of $\mathbf{Z}$ can be written as

$$\begin{aligned}
N(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\right] \\
&= \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}_{11}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{z}_{(1)}-\boldsymbol{\mu}_{(1)})'\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{z}_{(1)}-\boldsymbol{\mu}_{(1)})\right] \\
&\quad \times \frac{1}{(2\pi)^{(n-m)/2}|\boldsymbol{\Sigma}_{22}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{z}_{(2)}-\boldsymbol{\mu}_{(2)})'\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{z}_{(2)}-\boldsymbol{\mu}_{(2)})\right] \\
&= \prod_{i=1}^{2} N\left(\mathbf{z}_{(i)}; \boldsymbol{\mu}_{(i)}, \boldsymbol{\Sigma}_{ii}\right)
\end{aligned}$$

since

$$|\boldsymbol{\Sigma}| = \left|\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & [0] \\ \hline [0] & \boldsymbol{\Sigma}_{22} \end{array}\right| = |\boldsymbol{\Sigma}_{11}| \cdot |\boldsymbol{\Sigma}_{22}| \quad \text{and} \quad \boldsymbol{\Sigma}^{-1} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11}^{-1} & [0] \\ \hline [0] & \boldsymbol{\Sigma}_{22}^{-1} \end{array}\right].$$

Given that the joint density of $\mathbf{Z}$ factors into the product of the marginal densities, $\mathbf{Z}_{(1)}$ and $\mathbf{Z}_{(2)}$ are independent random vectors. ■

## 4.4 The Exponential Class of Densities

The majority of the discrete and continuous density function families that we have examined so far are special cases of the **exponential class** of density functions.[10] We will see later that problems of statistical inference involving experiments having probability spaces that involve density families from the *exponential* class often lead to somewhat simplified procedures for statistical inference.

**Definition 4.2**
***Exponential class of densities***

The density function $f(\mathbf{x}; \boldsymbol{\Theta})$ is a member of the exponential class of density functions iff

$$f(\mathbf{x}; \boldsymbol{\Theta}) = \begin{cases} \exp\left(\Sigma_{i=1}^{k} c_i(\boldsymbol{\Theta}) g_i(\mathbf{x}) + d(\boldsymbol{\Theta}) + z(\mathbf{x})\right) & \text{for } \mathbf{x} \in A, \\ 0 \quad \text{otherwise,} \end{cases}$$

where $\mathbf{x} = (x_1, \ldots, x_n)'$; $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_k)'$; $c_i(\boldsymbol{\Theta})$, $i = 1, \ldots, k$, and $d(\boldsymbol{\Theta})$ are real-valued functions of $\boldsymbol{\Theta}$ that do not depend on $\mathbf{x}$; $g_i(\mathbf{x})$, $i = 1, \ldots, k$, and $z(\mathbf{x})$ are real-valued functions of $\mathbf{x}$ that do not depend on $\boldsymbol{\Theta}$; and $A \subset R^n$ is a

[10] We warn the reader that some authors refer to this collection of densities as the exponential *family* rather than the exponential *class*. The collection of densities referred to in this section is a broader concept than that of a parametric *family*, since densities are distinguished not only by parameter values but also by different choices of functional forms. We (and others) use the term *class* to distinguish this broader density collection from that of a parametric family and to also avoid confusion with the exponential density family discussed in Section 4.2.

set of $n$-tuples contained in $n$-dimensional real space whose definition does not depend on the parameter vector $\Theta$.

In order to check whether a given density family belongs to the exponential class, one must determine whether there exists definitions of $c_i(\Theta)$, $d(\Theta)$, $g_i(\mathbf{x})$, $z(\mathbf{x})$, and $A$ such that the density can be equivalently represented in the exponential form presented in Definition 4.2. Of the densities we have studied, the Bernoulli, binomial and multinomial (for known values of $n$), negative binomial (for known values of $r$), Poisson, geometric, gamma (including exponential and chi-square), Beta, univariate normal, and multivariate normal density families all belong to the exponential class of densities. The discrete and continuous uniform, and the hypergeometric density families are *not* members of the exponential class (why?). We now present a number of examples that illustrate how membership in the exponential class of densities can be verified. In general, there is no standard method of verifying whether a density family belongs to the exponential class, and so the verification process must rely on the ingenuity of the analyst.

**Example 4.22** Normal Family $\subset$ Exponential Class

*Univariate case*: Let $k = 2$, let $n = 1$, and define

$$c_1(\Theta) = \frac{\mu}{\sigma^2}, \qquad c_2(\Theta) = -\frac{1}{2\sigma^2}, \qquad g_1(x) = x, \qquad g_2(x) = x^2,$$

$$d(\Theta) = -\frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \ln(2\pi\sigma^2)\right), \qquad z(x) = 0, \qquad A = R.$$

Substitution into Definition 4.2 yields

$$f(x;\Theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],$$

which is the univariate normal density.

*Multivariate case* (Suppose an $n$-variate case.) Let $k = n + (n(n+1)/2)$, and note that the multivariate normal density can be written as

$$N(\mathbf{x};\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}\left(\Sigma_{i=1}^n x_i^2\Sigma^{-1}_{(ii)} + 2\sum_{i<j}\sum x_i x_j \Sigma^{-1}_{(ij)} - 2\boldsymbol{\mu}'\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}\right)\right],$$

where $\Sigma^{-1}_{(ij)}$ refers to the $(i, j)$th entry in the $\Sigma^{-1}$ matrix. Define

$$\underset{(1\times k)}{\mathbf{c}(\Theta)'} = -\left(-\boldsymbol{\mu}'\Sigma^{-1} \mid \tfrac{1}{2}\Sigma^{-1}_{(11)}, \Sigma^{-1}_{(12)}, \ldots, \Sigma^{-1}_{(1n)}, \tfrac{1}{2}\Sigma^{-1}_{(22)}, \Sigma^{-1}_{(23)}, \ldots, \Sigma^{-1}_{(2n)}, \ldots, \tfrac{1}{2}\Sigma^{-1}_{(nn)}\right),$$

$$\underset{(1\times k)}{\mathbf{g}(\mathbf{x})'} = \left(\mathbf{x}' \mid x_1^2, x_1x_2, \ldots, x_1x_n, x_2^2, x_2x_3, \ldots, x_2x_n, \ldots, x_n^2\right),$$

$$d(\Theta) = -\tfrac{1}{2}\left(\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} + \ln(2\pi)^n|\Sigma|\right), \qquad z(\mathbf{x}) = 0, \qquad A = R^n.$$

It then follows that $\exp[\mathbf{c}(\Theta)'\mathbf{g}(\mathbf{x}) + d(\Theta) + z(\mathbf{x})] = N(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$. □

**Example 4.23** Bernoulli Family ⊂ Exponential Class

Let $k = 1$, let $n = 1$, and define

$$c_1(\Theta) = \ln\left(\frac{p}{1-p}\right), \qquad g_1(x) = x, \qquad d(\Theta) = \ln(1-p), \qquad z(x) = 0, \qquad A = \{0, 1\}.$$

Substitution into Definition 4.2 yields $f(x; p) = p^x(1-p)^{1-x}I_{\{0,1\}}(x)$. □

**Example 4.24** Gamma Family ⊂ Exponential Class

Let $k = 2$, let $n = 1$, and define

$$c_1(\Theta) = \alpha - 1, \qquad c_2(\Theta) = -\frac{1}{\beta}, \qquad g_1(x) = \ln x, \qquad g_2(x) = x,$$

$$d(\Theta) = -\ln(\beta^\alpha\Gamma(\alpha)), \qquad z(x) = 0, \qquad A = (0, \infty).$$

Substitution into Def. 4.2 yields

$$f(x; \alpha, \beta) = \frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta}I_{(0,\infty)}(x).$$ □

For now, we will be content simply to acknowledge the existence of the exponential class of densities, with the recognition that it encompasses many families of densities that are commonly used in applications. Later we will examine certain general properties of the class of densities that facilitate the construction and evaluation of parameter estimation procedures and statistical hypothesis tests.

## Key Words, Phrases, and Symbols

parameters
parameter space
parametric families of densities
reparameterized
(discrete) uniform family
Bernoulli family
binomial family
multinomial family
negative binomial family
Pascal distribution
geometric family
Poisson family
Poisson process
Poisson density approximation to binomial

mean rate of occurrence
memoryless property
hypergeometric family
drawing randomly without replacement
drawing objects with replacement
(continuous) uniform family
gamma family
gamma function, $\Gamma(\alpha)$
additivity property
exponential subfamily
memoryless property (or "as good as new")
wear-out effect
work-hardening effect
chi-square subfamily, $\chi^2_v$
degrees of freedom
chi-square additivity
beta family
beta function, $B(\alpha, \beta)$
univariate normal family
standard normal density
standardizing a random variable
standard normal random variable
multivariate normal family
iso-density contours
regression function (hyperplane) or curve (line) of $X_1$ on $X_2$
exponential class of densities

## Problems

**1.** A shipment of 100 videocassettes contains $k$ defective cassettes. You randomly sample 20 videocassettes, without replacement, from the shipment of 100 videocassettes. Letting $p = k/100$, the probability that you will obtain less than 3 defective cassettes in your sample of 20 cassettes is then given by

$$P(x \le 3) = \sum_{x=0}^{3} \binom{20}{x} p^x (1-p)^{20-x} I_{\{0,1,2,\ldots,20\}}(x).$$

True or false?

**2.** The daily price, $p$, and quantity demanded, $q$, of gasoline on a European spot market can be viewed (approximately) as the outcome of a bivariate normal random variable, where the bivariate normal density has mean vector and covariance matrix as follows:

$$\mu = \begin{bmatrix} 2.50 \\ 100 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} .09 & -1 \\ -1 & 100 \end{bmatrix}.$$

The price is measured in U.S. dollars per gallon of gasoline, and the quantity demanded is measured in thousands of gallons.

a. What is the probability that greater than 110,000 gallons of gasoline will be demanded on any given day?

b. What is the probability that the price of gasoline will be between \$2.00 and \$3.00 on any given day?

c. Define the regression function of $Q$ on $p$. Graph the regression function. What is the expected daily quantity of gasoline demanded given that price is equal to \$3.00?

d. Given that the price equals \$3.00, what is the probability that quantity demanded will exceed 110,000 gallons? What is this probability on a day when price equals \$2.00?

**3.** Show that the following probability density functions are members of the exponential class of densities:

a. binomial family, for a fixed value of $n$

b. Poisson family

c. negative binomial family for a fixed value of $r$

d. multinomial family, for a fixed value of $n$

e. beta family

**4.** For each PDF family below, show whether or not the family belongs to the exponential class of densities.

a. $f(x; \beta) = \beta x^{-(\beta+1)} I_{(1,\infty)}(x)$, $\beta \in \Omega = (0, \infty)$. (This is a subfamily of the *Pareto family* of PDFs.)

b. $f(x; \Theta) = (1/2\Theta) \exp(-|x|/\Theta)$, $\Theta \in \Omega = (0, \infty)$. (This is a subfamily of the *double exponential family* of PDFs.)

c. $f(x; \mu, \sigma) = (1/x\sqrt{2\pi}\sigma) \exp\left(-(\ln(x) - \mu)^2/(2\sigma^2)\right) I_{(0,\infty)}(x)$, $(\mu, \sigma) \in \Omega = \{(\mu, \sigma) : \mu \in (-\infty, \infty), \sigma > 0\}$. (This is the *log-normal family* of PDFs.)
Hint: Expanding the square in the exponent of $e$ may be helpful as an alternative representation of the exponent.

d. $f(x; r) = ((1 - r)/r) r^x I_{\{1,2,3,\ldots\}}(x)$, $r \in \Omega = (0, 1)$.

**5.** Prove that if $X$ has the geometric density, then the "memoryless property" $P(x > s + t \mid x > s) = P(x > t)$ holds for every choice of positive integers $s$ and $t$.

**6.** Prove that the CDF of the geometric family of densities can be defined by

$$F(b) = [1 - (1 - p)^{\text{trunc}(b)}] I_{[1,\infty)}(b).$$

**7.** The quantity of wheat demanded per day in a midwestern market during a certain marketing period is represented by

$Q = 100{,}000 - 12{,}500p + V \quad \text{for } p \in [2, 6],$

where

$Q$ is quantity demanded in bushels,

$p$ is price/bushel,

$V$ is *approximately normally* distributed.

You know that the expected quantity demanded is given by

$\mathrm{E}Q = 100{,}000 - 12{,}500p \quad \text{for } p \in [2, 6],$

and thus is a function of $p$, and the variance of quantity demanded is $\mathrm{var}(Q) = 16 \times 10^6$.

a. What is the mean and variance of $V$?

b. If $p = 4$, what is the probability that more than 50,000 bushels of wheat will be demanded?

c. If $p = 4.50$, what is the probability that more than 50,000 bushels of wheat will be demanded?

d. For quantity demanded to be greater than 50,000 bushels with probability .95, what does $p$ have to be?

e. Is it possible that $V$ could actually be normally distributed instead of only *approximately* normally distributed? Explain.

**8.** An investor has $10,000, which she intends to invest in a portfolio of three stocks that she feels are good investment prospects. During the investor's planning horizon, the weekly closing prices of the stocks can be viewed as the outcome of a trivariate normal random variable with

$$\mathbf{EX} = \begin{bmatrix} 27 \\ 10 \\ 18 \end{bmatrix} \quad \text{and} \quad \mathbf{Cov(X)} = \begin{bmatrix} 9 & 2 & 1 \\ 2 & 1 & -1 \\ 1 & -1 & 4 \end{bmatrix}.$$

The current price of the stocks are $23, $11, and $19, respectively.

a. If she invests her $10,000 equally among the three stocks, what is the expected value of her portfolio? What is the variance of the portfolio value?

b. What is the probability density function of the portfolio value? What is the probability that the closing value of her portfolio for a given week will exceed $11,000?

c. What is the probability density function of the value of stock 1? If she invests the $10,000 entirely in stock 1, what is the probability that the closing value of her portfolio will exceed $11,000?

d. What is the conditional density function of the value of stock 1 *given* that stock 3 has a value of $17? If she invests the $10,000 entirely in stock 1, what is the conditional probability that the closing value of her portfolio will exceed $11,000, *given* that stock 3 has a value of $17?

e. If she divides her $10,000 equally among stocks 1 and 2, what is the conditional probability that this portfolio will have a closing value exceeding $11,000 *given* that stock 3 has a value of $17?

**9.** Let $Y$ have a chi-square distribution with 15 degrees of freedom, let $X$ have a chi-square distribution with 5 degrees of freedom, and let $Y = X + Z$, where $X$ and $Z$ are independent random variables.

a. Calculate $P(y > 27.488)$.

b. Calculate $P(6.262 < y < 27.488)$.

c. Find $c$ such that $P(y > c) = .05$.

d. Find $c$ such that $P(z > c) = .05$.

**10.** Let $Y$ have the density $N(5, 36)$, $X$ have the density $N(4, 25)$, let $Y$ and $X$ be independent random variables, and define $W = X - Y$.

a. Calculate $P(y > 10)$.

b. Calculate $P(-10 < y < 10)$.

c. Calculate $P(w > 0)$.

d. Find $c$ such that $P(w > c) = .95$.

**11.** Let $X$ be a bivariate random variable having the probability density $N(\mu, \Sigma)$, with

$$\mu = \begin{bmatrix} 5 \\ 8 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}.$$

a. Define the regression curve of $X_1$ on $X_2$. What is $E(X_1 \mid x_2 = 9)$?

b. What is the conditional variance of $X_1$ *given* that $x_2 = 9$?

c. What is the probability that $x_1 > 5$? What is the probability that $x_1 > 5$, *given* that $x_2 = 9$?

*In problems 12–20 below, identify the most appropriate parametric family of density functions from those presented in this chapter on which to base the probability space for the experiment described, and answer the questions using the probability space you define:*

**12.** WAYSAFE, a large retail supermarket, has a standard inspection policy that determines whether a shipment of produce will be accepted or rejected. Specifically, the employees examine 5 percent of the objects in any shipment received, and if no defective produce is found in any of the items examined, the shipment is accepted. Otherwise, it is rejected. The items chosen for inspection are drawn randomly, one at a time, without replacement, from the objects in the shipment.

A shipment of 1,000 5-lb. bags of potatoes are received at the loading dock. Suppose that in reality, 2 percent of the 5-lb. bags have defective potatoes in them. The "objects" that are being inspected are bags of potatoes, with a bag of potatoes being defective if any of the potatoes in the bag are defective.

a. Define an appropriate probability space for the inspection experiment.

b. What is the probability that WAYSAFE will accept the shipment of potatoes?

c. What is the expected number of defective bags of potatoes when choosing 5 percent of the bags for inspection in the manner described above?

d. If the inspection policy is changed so that 10 percent of the objects will be inspected, and the shipment will be accepted only if no defectives are found, what is the probability that the shipment will be accepted?

**13.** The FLAMES-ARE-US Co. manufactures butane cigarette lighters. Your top-of-the-line lighter, which has the brand name "SURE-FLAME," costs \$29.95. As a promotional strategy, the SURE-FLAME lighter carries a guarantee that if it takes more than five attempts before the lighter actually lights, then the customer will be given \$1,000,000. The terms of the guarantee require that the demonstration of failure of a SURE-FLAME lighter to light within five attempts must be witnessed by an official of the company, and each original buyer of a new SURE-FLAME lighter is allowed only one attempt at being awarded the \$1,000,000. The lighter is such that each attempt at lighting the lighter has a probability of success (it lights) equal to .95, and the outcomes of attempts to light the lighter are independent of one another.

a. Define the appropriate probability space for the experiment of observing the number of attempts necessary to obtain the first light with the SURE-FLAME lighter.

b. What is the probability that a buyer who attempts to demonstrate the failure of the SURE-FLAME to light in five attempts will actually be awarded \$1,000,000?

c. What is the expected number of attempts required to obtain a light with the SURE- FLAME lighter?

d. What is the expected value of the award paid to any consumer who attempts to claim the \$1,000,000?

**14.** An instructor in an introductory economics class has constructed a multiple-choice test for the midterm examination. The test consists of 20 questions worth 5 points each. For each question, the instructor lists four possible answers, of which only one is correct. John Partytime, a student in the class, has not attended class regularly and admits (to himself) that he is really not prepared to take the exam. Nonetheless, he has decided to take the exam, and his strategy is to randomly choose one of the four answers for each question. He feels very confident that this course of action will result in an exam score considerably higher than zero.

a. Define a probability space that can be used to assign probability to events involving the number of questions that John answers correctly.

b. What is the probability that John receives a zero on the exam?

c. What is the probability that John receives at least 25 points on the exam?

d. What is John's expected score on the exam?

**15.** The liquid crystal display in the new Extime brand of digital watches is such that the probability it continues to function for at least $x$ hours before failure is constant (for any given choice of $x$), regardless of how long the display has already been functioning. The expected value of the number of hours the display functions before failure is known to be 30,000.

a. Define the appropriate probability space for the experiment of observing the number of hours a display of the type described above functions before failure.

b. What is the probability that the display functions for at least 20,000 hours?

c. If the display has already functioned for 10,000 hours, what is the probability that it will continue to function for at least another 20,000 hours?

d. The display has a rather unique guarantee in the sense that any purchaser of an Extime watch, whether the watch is new or used, has a warranty on the display for two years from the date of purchase , during which time if the display fails, it will be replaced free of charge. Assuming that the number of hours the watch operates in a given period of time is essentially the same for all buyers of the Extime watch, is it more likely that a buyer of a used watch will be obtaining a free display replacement than a buyer of a new watch, given an equal period of watch ownership? Explain.

**16.** The Department of Transportation in a foreign county establishes gas-mileage standards that automobiles sold must meet or else a "gas guzzler" tax is imposed on the sale of the offending types of automobile. For the "compact, 4-door" class of automobiles, the target average gas mileage is 25 miles per gallon.

Achievement of the standard is tested by randomly choosing 20 cars from a manufacturer's assembly line and then examining the distance between the vector of 20 observed measurements of gas mileage/gallon and a $(20 \times 1)$ vector of targeted gas mileages for these cars. Letting $\mathbf{X}$ represent the $20 \times 1$ vector of observed gas mileages, and letting $\mathbf{t}$ represent the $20 \times 1$ vector of targeted gas mileages (i.e., $t = [25, 25, \ldots, 25]'$), the distance measure is

$$D(\mathbf{x}, \mathbf{t}) = [(\mathbf{x} - \mathbf{t})'(\mathbf{x} - \mathbf{t})]^{1/2}.$$

If $D(\mathbf{x}, \mathbf{t}) \leq 6$, then the type of automobile being tested is judged to be consistent with the standard; otherwise, the type of automobile will be taxed.

Specific Motors Company is introducing a new four-door compact into the market and has requested that this type of automobile be tested for adherence to the gas-mileage standard. The engineers at Specific Motors know that the miles per gallon achieved by their compact four-door automobile can be represented by a normal distribution with mean 25 and variance 1.267, so that the target gas mileage is achieved *on average.*

a. What is the probability that a car randomly chosen from Specific Motor's assembly line will be within one mile per gallon of the gas-mileage standard?

b. What is the probability that Specific Motor's compact four-door will be judged as being consistent with the gas-mileage standard?

c. A neighboring country uses a simpler test for determining whether the gas-mileage standard is met. It also has a target of 25 miles per gallon, but its test involves forming the simple average of the 20 randomly observed miles per gallon and then simply testing whether the calculated average is within 1 mile per gallon of 25 miles per gallon. That is, the gas-mileage standard will be judged to have been met if

$$\frac{1}{20} \sum_{i=1}^{20} x_i \in [24, 26].$$

What is the probability that Specific Motors will pass this alternative test?

**17.** An appliance manufacturer is conducting a survey of consumer satisfaction with appliance purchases. All customers who have purchased one of the company's appliances within the last year will be mailed a customer satisfaction survey. The company is contemplating the proportion of surveys that customers will actually return. It is known from considerable past experience with these types of surveys that the expected proportion of returned surveys is equal to .40, with a variance of .04.

a. What is the probability that more than 50 percent of the surveys wiil be returned?

b. What is the probability that less than 25 percent of the surveys will be returned?

c. What is the median level of response to this type of survey? What is the mode?

**18.** Customers arrive at the rate of four per minute at a large bank branch in downtown Seattle. In its advertising, the bank stresses that customers will receive service promptly with little or no waiting.

a. What is the probability that there will be more than 25 customers entering the bank in a 5-minute period?

b. What is the expected number of customers who will enter the bank during a five-minute period?

c. If the bank staff can service 20 customers in a 5-minute interval, what is the probability that the customer load will exceed capacity in a 5-minute interval, so that some customers will experience delays in obtaining service?

**19.** The accounts of the Excelsior company are being audited by an independent accounting firm. The company has 200 active accounts, of which 140 are current accounts, 45 are past due 60 or more days, and 15 accounts are delinquent. The accounting firm will randomly choose five different accounts in its auditing procedure.

a. What is the probability that none of the accounts chosen will be delinquent?

b. What is the probability that at most one of the accounts chosen will be delinquent?

c. What is the probability that there will be three current, one past due, and one delinquent accounts chosen?

d. What are the expected numbers of the various types of accounts that will be chosen?

**20.** The Stonebridge Tire Co. manufactures passenger-car tires. The manufacturing process results in either first-quality tires, blemished tires, or defective tires. The proportions of the tires manufactured that fall in the three categories are .88, .09, and .03, respectively. The manufacturing process is such that the classification of a given tire is unaffected by the classifications of any other tires produced.

a. A lot of 12 tires is taken from the assembly line and inspected for shipment to a tire retailer that sells only first-quality tires. What is the probability that all of the tires will be first-quality tires?

b. For the lot of tires in part (a), what is the probability that there will be no defective tires among the 12 tires?

c. What are the expected number of first-quality tires, blemished tires, and defective tires in the lot of 12 tires from part (a)?

d. What is the probability that the 12 tires will contain 8 first-quality tires, 3 blemished tires, and 1 defective tire?

# 5 Basic Asymptotics

## 5.1 Introduction

In this chapter we establish results relating to the probability characteristics of functions of $n$-variate random variables when $n$ is large. In particular, certain types of functions of an $n$-variate random variable $X_{(n)} = (X_1, \ldots, X_n)$, say $Y_n = g(X_1, \ldots, X_n)$, may converge in various ways to a constant, or the distribution of $g(X_{(n)})$ may approach a "limiting" distribution as $n \to \infty$.

There are at least three interrelated reasons why the study of the "asymptotic behavior" of $g(X_{(n)})$ is an important endeavor. In practice, functions such as $g(X_{(n)})$ will be used to represent point estimation, hypothesis testing, or confidence-set estimation procedures (Chapters 7–10), and $n$ will refer to the number of data observations relating to the experiment being analyzed. In order to be able to evaluate and/or compare the merits of these statistical procedures, and indeed to be able to define the latter two types of procedures at all, it is necessary to establish the probability characteristics of $g(X_{(n)})$. Unfortunately, it is often the case in statistical and econometric practice that the actual probability density or distribution of $g(X_{(n)})$ is too difficult to derive, or intractable to work with analytically, when $n$ is finite. Asymptotic theory identifies methods that *provide tractable approximations to the probability distribution of* $g(X_{(n)})$ when $n$ is sufficiently large, and thereby it also *provides a means of evaluating, comparing, and/or defining various statistical inference procedures.*

Asymptotic theory also *provides the principal rationale for the prevalent use of the normal probability distribution in statistical analyses.*

The material in this chapter is generally of an advanced and technical nature, and the proofs of some of the theorems will be deferred to a more advanced course of study. Some readers, upon first reading, may wish to skim the chapter for main results, while others may wish to move on to Chapter 6 and refer back to the results in this chapter as needed. Readers intending to emphasize econometrics in their graduate work are well advised to study this chapter in detail.

## 5.2 Elements of Real Analysis

In this section, we present a number of results from real analysis that facilitate the development and understanding of various types of asymptotic probability behavior. In particular, the concepts of sequences, limits, continuity of a function, and orders of magnitude of a sequence will be examined.

We begin with the notion of a **sequence**. In the definition, we refer to the set of **natural numbers**, which is simply the set of positive integers in their natural order, 1, 2, 3, ....

**Definition 5.1** ***Sequence***

Let $A$ be any set. A sequence in $A$ is a function having the natural numbers, $N$, for its domain, and its range contained in $A$, i.e., $f: N \rightarrow A$, is a sequence in $A$.

When utilizing the concept of a sequence, we (and others) will often suppress the function aspect of its definition and concentrate on the collection of image elements of the function. Thus, given a sequence defined by $\{(n, y): y = f(n), n \in N\}$, we will equivalently refer to the collection of image elements $\{y_1, y_2, y_3, \ldots\}$ as the sequence, where $y_n = f(n)$. The subscripts on the elements of the set $\{y_1, y_2, y_3, \ldots\}$ serve to define the order of the elements in the sequence. Furthermore, we will utilize the notation $\{y_n\}$ as an abbreviation for the sequence $\{y_1, y_2, y_3, \ldots\}$.[1] In the following numerous examples of sequences, we continue to use $N$ to denote the set of natural numbers.

**Example 5.1** The following are sequences in $R$:

**a.** $\{2, 4, 8, \ldots\}$, which is defined by the function $y = 2^n$, $n \in N$.
**b.** $\{1, 1/3, 1/9, 1/27, \ldots\}$, which is defined by the function $y = (1/3)^{n-1}$, $n \in N$.
**c.** $\{-3, -1, 1, 3, \ldots\}$, which is defined by the function $y = 2n - 5, n \in N$. □

[1] Another common abbreviated notation that is sometimes used to denote a sequence is given by $(y_n)$. The notation we have adopted is more prevalent in the statistics literature. While there will be no confusion in this text, in general, the reader will have to rely on the context of a discussion to determine whether $\{y_n\}$ refers to a sequence or to a set containing the single element $y_n$.

**Example 5.2** Sequence of Matrices

Let $\mathbf{X}_n$ be an $(n \times 2)$ matrix whose $i$th row is defined by the $(1 \times 2)$ vector $[1, i]$, so that

$$\mathbf{X}_n = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & n \end{bmatrix}.$$

Then

$$\left\{ \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & \frac{3}{2} \\ \frac{3}{2} & \frac{5}{2} \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 2 & \frac{14}{3} \end{bmatrix}, \ldots \right\}$$

is a sequence of *matrices* $\{\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \ldots\}$ defined by the function

$$\mathbf{Y}_n = \frac{1}{n}\mathbf{X}_n'\mathbf{X}_n, \qquad n \in N,$$

where the $n$th element of the sequence is defined as

$$\mathbf{Y}_n = \begin{bmatrix} 1 & \frac{(\sum_{i=1}^n i)}{n} \\ \frac{(\sum_{i=1}^n i)}{n} & \frac{(\sum_{i=1}^n i^2)}{n} \end{bmatrix} = \begin{bmatrix} 1 & \frac{(n+1)}{2} \\ \frac{(n+1)}{2} & \frac{(n+1)(2n+1)}{6} \end{bmatrix}. \qquad \square$$

Later in this chapter we will frequently encounter **sequences of random variables**. In this case, the set $A$ in Def. 5.1 is a collection of random variables, and the function $f: N \to A$ defining the sequence places the random variables in $A$ in a specific order. That is, the sequence of random variables $\{Y_1, Y_2, Y_3, \ldots\}$ is simply an ordered collection of random variables. In our study of asymptotics, the elements in the sequence of random variables will often be defined as functions of other random variables, such as $Y_n = g_n(X_1, \ldots, X_n)$, and we will be interested in studying the characteristics of the sequence of probability distributions associated with the $Y_n$'s as $n \to \infty$.

The following examples are of sequences of random variables. We introduce the notation $Y \sim f(y)$ to indicate that $Y$ **has probability density** $f(y)$, or that $Y$ **is distributed as** $f(y)$. The acronym **iid** stands for **independent and identically distributed**, meaning that the random variables in a collection are independent and each of the random variables has the same PDF or probability distribution.

**Example 5.3** Let $X_1, \ldots, X_n$ be iid random variables, each with PDF $N(\mu, \sigma^2)$, where $X_i$ represents the miles per gallon obtained from the $i$th automobile of a certain type tested for fuel efficiency. Examine the sequence of random variables

$$\{Y_1, Y_2, Y_3, \ldots\}, \qquad \text{where } Y_n = n^{-1}\sum_{i=1}^n X_i, \quad n \in N.$$

Note that the $n$th element of the sequence represents the *average* miles per gallon obtained from $n$ of the automobiles tested, and

$$Y_n \sim N\left(\mu, \tfrac{\sigma^2}{n}\right),$$

so that we can define a **sequence of probability density functions** associated with the sequence of random variables as

$$\left\{N(\mu,\sigma^2), N\left(\mu, \tfrac{\sigma^2}{2}\right), N\left(\mu, \tfrac{\sigma^2}{3}\right), \ldots\right\}. \qquad \square$$

**Example 5.4** Let $X_1, \ldots, X_n$ be iid Bernoulli-type random variables each with density function $p^z(1-p)^{1-z}I_{\{0,1\}}(z)$, where $X_i$ indicates whether the $i$th customer entering a store makes a purchase $(x_i = 1)$ or not $(x_i = 0)$. Examine the sequence of random variables

$$\{Y_1, Y_2, Y_3, \ldots\}, \qquad \text{where } Y_n = \sum_{i=1}^{n} X_i, \quad n \in N.$$

Note that the $n$th element of the sequence represents how many of the first $n$ customers make a purchase, and

$$Y_n \sim \binom{n}{y_n} p^{y_n}(1-p)^{n-y_n} I_{\{0,1,2,\ldots,n\}}(y_n),$$

that is, $Y_n$ has a binomial distribution with parameters $n$ and $p$, say Bin$(n,p)$. We can define a sequence of density functions associated with the sequence of random variables as

$$\{\text{Bin}(1,p), \text{Bin}(2,p), \text{Bin}(3,p), \ldots\}. \qquad \square$$

### Limit of a Sequence

We now examine the concept of the **limit of a real number sequence**. We begin with a sequence whose elements are scalars and then extend the result to a sequence whose elements are vectors of real numbers.

**Definition 5.2** ***Limit of a real number sequence***

Let $\{y_n\}$ be a sequence whose elements are (scalar) real numbers. Suppose there exists a real number, $y$, such that for every real $\varepsilon > 0$ there exists an integer $N(\varepsilon)$ for which $n \geq N(\varepsilon) \Rightarrow |y_n - y| < \varepsilon$. Then $y$ is the **limit of the sequence** $\{y_n\}$, and the sequence $\{y_n\}$ is said to **converge to** $y$ as $n \to \infty$. The existence of the limit is denoted by $y_n \to y$ or $\lim_{n\to\infty} y_n = y$. If the limit does not exist, the sequence is said to be **divergent**.

The definition of the limit implies that for a sufficiently large choice of $n$, $y_n$ (and $y_{n+1}, y_{n+2}, y_{n+3}, \ldots$) becomes arbitrarily close to the number $y$. This is so since, by the definition, we can choose $\varepsilon > 0$ to be arbitrarily small and yet

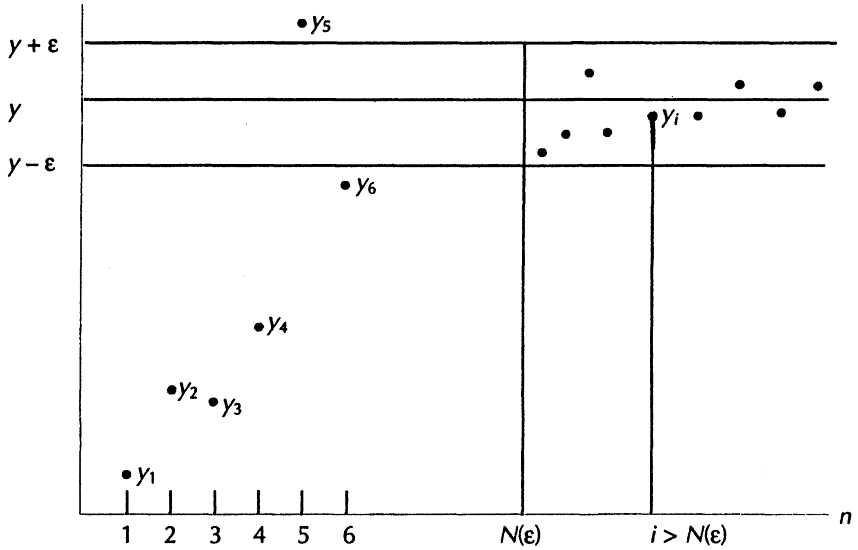


**Figure 5-1** Illustration of the sequence $\{y_n\}$ for which $\lim_{n\to\infty} y_n = y$.

there exists an $n$ large enough (namely $n \geq N(\varepsilon)$) such that $y - \varepsilon < y_n < y + \varepsilon$. Figure 5.1 provides a graphical illustration of the limit concept.

In the figure, it is seen that for all elements $y_i \in \{y_n\}$ for which $i$ is large enough, i.e., for $i > N(\varepsilon)$, the value of $y_i$ is contained in the interval $(y - \varepsilon, y + \varepsilon)$. In other words, $y_i$ is within $\varepsilon$-distance of $y$ for $i > N(\varepsilon)$. Furthermore, for every choice of $\varepsilon > 0$, no matter how small, there exists an $N(\varepsilon)$ for which a figure such as Figure 5.1 could be drawn.

It can be shown that for the limit of a sequence of real numbers to exist, it is necessary (but not sufficient) that the sequence is *bounded* (Bartle, Real Analysis, p. 93), as defined next.

**Definition 5.3** ***Bounded sequence of real numbers***

The sequence of real numbers $\{y_n\}$ is **bounded** iff there exists a finite number $m > 0$ such that $|y_n| \leq m \; \forall \; n \in N$; otherwise the sequence is said to be **unbounded**.

Thus, for a sequence of real numbers to be bounded, there must exist a positive number that is larger than the absolute value of each and every number in the sequence. For a sequence that has no limit and is *also* unbounded, if $\forall \; m > 0$ there exists a positive integer $N(m)$ such that $y_n > m$ (or $y_n < -m$) $\forall \; n > N(m)$, we write $y_n \to \infty$ (or $y_n \to -\infty$) to denote that the sequence *diverges to infinity* (or to *negative infinity*).

**Example 5.5** Boundedness and Existence of a Limit for Sequences $\{y_n\}$ of Real Numbers

**a.** $y_n = 3 + n^{-2}$, $n \in N$. This sequence is bounded, since $|y_n| \leq 4 \; \forall \; n \in N$. Also, the sequence has a limit, where $y_n \to 3$. This follows since, $\forall \; \varepsilon > 0$,

$|y_n - 3| < \varepsilon \ \forall \ n > \varepsilon^{-1/2}$, and there always exists an integer $N(\varepsilon) \geq \varepsilon^{1/2}$ (e.g., $\text{trunc}(\varepsilon^{-1/2}) + 1$).

**b.** $y_n = \sin(n)$, $n \in N$ (let $n$ represent degrees). The sequence is bounded, since $|\sin(x)| \leq 1 \ \forall \ x$. The sequence does *not* have a limit, since $\sin(x)$ cycles between the values of $-1$ and $1$.

**c.** $y_n = n^2 - 3n + 1$, $n \in N$. The sequence is not bounded, since there does not exist a finite number $m > 0$ for which $n^2 - 3n + 1 \leq m \ \forall \ n \in N$. Since the sequence is unbounded, the sequence does not have a limit. Since, $\forall \ m > 0$, $n^2 - 3n + 1 > m$ when $n > N(m) = \text{trunc}([3 + \sqrt{5 + 4m}]/2) + 1$ (use the quadratic formula), $y_n \to \infty$, that is, the sequence diverges to infinity. □

The preceding example illustrates that boundedness of a sequence is not sufficient for the existence of a limit for the sequence. We add that it is proper to speak of *the* limit of a sequence since the limit will be *unique* if it exists at all (Bartle, Real Analysis, p. 93).

The limit concept can be extended to the sequence whose elements are real-valued vectors or matrices. We introduce the notation $Y_n[i, j]$ to indicate the $(i, j)$th element of the $(m \times k)$ matrix $\mathbf{Y}_n$ in a sequence of matrices, and similarly $y_n[i]$ denotes the $i$th element of the $(m \times 1)$ vector $\mathbf{y}_n$ in a sequence of vectors. The extension of the limit concept amounts to viewing the matrix sequence as encompassing $mk$ sequences, $\{Y_n[i, j]\}$, one for each matrix element, with each to be examined for convergence (by Def. 5.2). Limits of vector sequences follow by letting $k = 1$.

**Definition 5.4** ***Limit of a real-valued matrix sequence***

Let $\{\mathbf{Y}_n\}$ be a sequence whose elements are $(q \times k)$ real-valued matrices. Suppose there exists an $(q \times k)$ matrix of real numbers $\mathbf{Y}$ such that $Y_n[i, j] \to Y[i, j]$ for $i = 1, \ldots, q$ and $j = 1, \ldots, k$. Then the matrix $\mathbf{Y}$ is the limit of the matrix sequence $\{\mathbf{Y}_n\}$, and the sequence $\{\mathbf{Y}_n\}$ is said to converge to $\mathbf{Y}$ as $n \to \infty$. The existence of the limit is denoted by $\mathbf{Y}_n \to \mathbf{Y}$, or by $\lim_{n\to\infty} \mathbf{Y}_n = \mathbf{Y}$. If the limit does not exist, the sequence is said to be divergent.

The definition of the limit implies that for a sufficiently large choice of $n$, the *matrix* $\mathbf{Y}_n$ (and $\mathbf{Y}_{n+1}, \mathbf{Y}_{n+2}, \mathbf{Y}_{n+3}, \ldots$) becomes arbitrarily close to the *matrix* $\mathbf{Y}$ element by element. The definition also implies that for the real-valued matrix to have a limit, the sequence of matrices must be bounded elementwise, i.e., $|Y_n[i, j]| \leq m \ \forall \ n \in N$ and $\forall \ i, j$ (or else $Y_n[i, j] \not\to Y[i, j]$ for some $i$ and $j$, and then $\mathbf{Y}_n \not\to \mathbf{Y}$). Regarding divergence to infinity (or negative infinity), since there are essentially $qk$ convergence conditions involved when examining sequences of $(q \times k)$ matrices, patterns of divergence and convergence of the various elements of the matrices can be quite diverse.

**Example 5.6** Boundedness and Limits of Matrices

**a.** Recall the sequence of matrices in Ex. 5.2. In this case, only the sequence $\{Y_n[1, 1]\}$ is bounded. All other sequences of matrix elements are unbounded and, in fact, diverge to infinity, i.e., $Y_n[i,j] \to \infty$ for $(i,j) \neq (1,1)$. Since all of the sequences of matrix elements must be bounded for the matrix sequence to converge, the matrix sequence does not have a limit.

**b.** Let $\{\mathbf{Y}_n\}$ be a sequence of matrices such that

$$\mathbf{Y}_n = \begin{bmatrix} 3n^{-1} & n^{-1} \\ 3 & 1+n^{-1} \end{bmatrix}, \quad n \in N.$$

All four sequences of matrix elements are bounded, since $|3n^{-1}| \leq 3$, $|n^{-1}| \leq 1$, $|3| \leq 3$, and $|1+n^{-1}| \leq 2 \ \forall \ n \in N$. Furthermore, limits exist for all four sequences of matrix elements, since $3n^{-1} \to 0$, $n^{-1} \to 0$, $3 \to 3$, and $1+n^{-1} \to 1$. Thus, $\mathbf{Y}_n \to \mathbf{Y} = \left[\begin{smallmatrix} 0 & 0 \\ 3 & 1 \end{smallmatrix}\right]$. □

One might be interested in a sequence $\{y_n\}$ that is defined via a function of the elements of other sequences. For example, we may be interested in the sequence $\{y_n\}$ that is defined by adding corresponding elements in the sequences $\{x_n\}$ and $\{z_n\}$, as $y_n = x_n + z_n$. Of course, we can analyze the properties of the sequence $\{y_n\}$ directly to establish whether the sequence converges, but if the convergence properties of $\{x_n\}$ and $\{z_n\}$ are known, the following lemma can expedite the analysis if the functions are defined via addition, subtraction, or multiplication. We precede the statement of the lemma with definitions for adding, subtracting, and multiplying sequences. Note that $\mathbf{X}[.,i]$ refers to the $i$th column of the matrix $\mathbf{X}$.

**Definition 5.5**
***Adding, subtracting, and multiplying sequences***

Let $\{\mathbf{X}_n\}$ and $\{\mathbf{Z}_n\}$ be sequences of conformable, real-valued matrices.

**a.** Summation: The summation of $\mathbf{X}_n$ and $\mathbf{Z}_n$, $\mathbf{X}_n + \mathbf{Z}_n$, is a sequence $\mathbf{Y}_n$ defined by $\mathbf{Y}_n = \mathbf{X}_n + \mathbf{Z}_n$, $\forall \ n$.

**b.** Difference: The difference between $\{\mathbf{X}_n\}$ and $\{\mathbf{Z}_n\}$, $\{\mathbf{X}_n\} - \{\mathbf{Z}_n\}$, is a sequence $\{\mathbf{Y}_n\}$ defined by $\mathbf{Y}_n = \mathbf{X}_n - \mathbf{Z}_n \ \forall \ n$.

**c.** Product: The product of $\{\mathbf{X}_n\}$ and $\{\mathbf{Z}_n\}$, $\{\mathbf{X}_n\}\{\mathbf{Z}_n\}$, is a sequence $\{\mathbf{Y}_n\}$ defined by $\mathbf{Y}_n = \mathbf{X}_n\mathbf{Z}_n$, $\forall \ n$.

**Lemma 5.1**
***Combinations of sequences***

Let $\{\mathbf{X}_n\}$ and $\{\mathbf{Z}_n\}$ be convergent sequences of conformable, real-valued matrices such that $\mathbf{X}_n \to \mathbf{X}$ and $\mathbf{Z}_n \to \mathbf{Z}$. Then

**a.** $\mathbf{X}_n + \mathbf{Z}_n \to \mathbf{X} + \mathbf{Z}$,

**b.** $\mathbf{X}_n - \mathbf{Z}_n \to \mathbf{X} - \mathbf{Z}$,

**c.** $\mathbf{X}_n\mathbf{Z}_n \to \mathbf{X}\mathbf{Z}$,

**d.** if $\{a_n\}$ is a sequence in $R$ that converges to $a$, then $a_n\mathbf{X}_n \to a\mathbf{X}$,

**e.** if $\{b_n\}$ is a sequence of nonzero numbers in $R$ that converges to $b \neq 0$, then $b_n^{-1}\mathbf{X}_n \to b^{-1}\mathbf{X}$,

**f.** $\sum_{i=1}^{k} \mathbf{X}_n[.,i] \to \sum_{i=1}^{k} \mathbf{X}[.,i]$,

g. if $\{\mathbf{Z}_n\}$ is a sequence of nonsingular matrices that converges to the nonsingular matrix $\mathbf{Z}$, then $\mathbf{Z}_n^{-1} \to \mathbf{Z}^{-1}$ and $\mathbf{Z}_n^{-1}\mathbf{X}_n \to \mathbf{Z}^{-1}\mathbf{X}$.

**Proof:** Bartle, Real Analysis, pp. 100–101, and Defs. 5.4 and 5.5.

The reader should note that since the sequences $\{\mathbf{X}_n\}$ and $\{\mathbf{Z}_n\}$ can themselves be defined in terms of combinations of other sequences, the lemma actually implies convergence results involving more than just two sequences. For example, letting $\mathbf{X}_n = \mathbf{A}_n + \mathbf{B}_n \to \mathbf{A} + \mathbf{B}$, then $\mathbf{A}_n + \mathbf{B}_n + \mathbf{Z}_n \to \mathbf{A} + \mathbf{B} + \mathbf{Z}$ and $(\mathbf{A}_n + \mathbf{B}_n)\mathbf{Z}_n \to (\mathbf{A} + \mathbf{B})\mathbf{Z}$.

**Example 5.7** Convergence Properties of Sequences

a. Let $\{x_n\}$ and $\{z_n\}$ be defined as $x_n = 3 + n^{-1/2}$ and $z_n = 2\exp(-2/n)$ for $n \in N$, respectively. Note that $x_n \to 3$ and $z_n \to 2$. Then, using Lemma 5.1, $x_n + z_n \to 5$, $x_n - z_n \to 1$, and $x_n z_n \to 6$. Let $\{a_n\}$ be defined by $a_n = 5(n+1)/n$ for $n \in N$, and note that $a_n \to 5$. Also, define the vector sequence $\{y_n\}$ by

$$\underset{(2\times 1)}{\mathbf{y}_n} = \begin{bmatrix} x_n \\ z_n \end{bmatrix}$$

so that $\mathbf{y}_n \to \left[\begin{smallmatrix} 3 \\ 2 \end{smallmatrix}\right]$. Then, from Lemma 5.1, $\mathbf{a}_n \mathbf{y}_n \to \left[\begin{smallmatrix} 15 \\ 10 \end{smallmatrix}\right]$ and $\mathbf{a}_n^{-1}\mathbf{y}_n \to \left[\begin{smallmatrix} 3/5 \\ 2/5 \end{smallmatrix}\right]$

b. Let $\{\mathbf{W}_n\}$ be a matrix sequence defined by

$$\mathbf{W}_n = \begin{bmatrix} 2 + n^{-1} & \dfrac{2}{n} \\ 3 & \dfrac{(n+1)}{n} \end{bmatrix} \quad \text{for } n \in N,$$

and let $\{\mathbf{x}_n\}$ be a vector sequence defined by

$$\mathbf{x}_n = \begin{pmatrix} 1 + n^{-1} \\ 2\exp(n^{-1}) \end{pmatrix} \quad \text{for } n \in N.$$

Note that $\mathbf{W}_n \to \left[\begin{smallmatrix} 2 & 0 \\ 3 & 1 \end{smallmatrix}\right]$ and $\mathbf{x}_n \to \left[\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right]$. Using Lemma 5.1, it follows that

$$\mathbf{W}_n^{-1} \to \begin{bmatrix} 2 & 0 \\ 3 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} .5 & 0 \\ -1.5 & 1 \end{bmatrix}, \quad \mathbf{W}_n\mathbf{x}_n \to \begin{bmatrix} 2 \\ 5 \end{bmatrix}, \quad \text{and} \quad \mathbf{W}_n^{-1}\mathbf{x}_n \to \begin{bmatrix} .5 \\ .5 \end{bmatrix}.$$

Note further that $\mathbf{W}_n[.,1] + \mathbf{W}_n[.,2] + \mathbf{x}_n \to \left[\begin{smallmatrix} 3 \\ 6 \end{smallmatrix}\right]$. □

### Continuous Functions

Continuous functions will play a prominent role in a number of important asymptotic results. We define the concept of a continuous function below using two alternative but completely equivalent characterizations. We remind the reader that $d(x, w) = [(x - w)'(x - w)]^{1/2}$ is the distance between the points $x$ and $w$.

**Definition 5.6** *Continuous function*[2]

The function $g\colon A \to R$, for $A \subset R^m$, is continuous at the point $x \in A$ iff either

**a.** $\forall \varepsilon > 0, \exists \delta(\varepsilon) > 0$ such that $w \in A$ and $d(x, w) < \delta(\varepsilon)$ implies $|g(w) - g(x)| < \varepsilon$, or

**b.** $\forall$ sequence $\{x_n\}$ in $A$ for which $x_n \to x$, it is also true that $g(x_n) \to g(x)$.

The *vector* function $g\colon A \to R^k$ is continuous at the point $x \in A$ iff each coordinate function $g_j(x)$ is continuous at the point $x$, $j = 1, \ldots, k$.

The function $g$ is said to be continuous on (the set) $B \subset A$ if the function is continuous at every point in $B$.

Intuitively, a function is continuous at a point $x$ if, when the function is evaluated at domain elements that are closer and closer to $x$, the value of the function is closer and closer to the value of the function at $x$ (in an elementwise vector comparison sense if $k \geq 2$). In the simple case where $m = k = 1$, the graph of a function that is continuous on an interval set $B = (a, b)$ can be drawn $\forall x \in B$ "without lifting the pencil from the paper," i.e., the graph is an unbroken curve.

**Example 5.8** Continuity Properties of Functions

**a.** Let $f\colon (0, \infty) \to R$ be defined by $y = x^{-1}$. Intuitively we expect $f$ to be continuous on $(0, \infty)$, since its graph is an unbroken curve. To demonstrate formally that $f$ is continuous on $(0, \infty)$, let $x_0 \in (0, \infty)$, and note that

$$|f(x) - f(x_0)| = |x^{-1} - x_0^{-1}| = \frac{|x_0 - x|}{|xx_0|} = \frac{|x_0 - x|}{xx_0},$$

where we have eliminated the absolute value operator in the denominator because in the domain of $f$, $x > 0$. It follows that if $|x - x_0| < \delta$, then $|f(x) - f(x_0)| < \delta / xx_0$.

Now choose any $\varepsilon > 0$. For $f$ to be continuous at $x_0$, it must be the case that $\delta$ can be chosen such that $|x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon$. Such choices of $\delta$ exist, one being $\delta = x_0^2 \varepsilon / (1 + x_0 \varepsilon)$. Since the argument can be applied $\forall\, x_0 \in (0, \infty)$ and $\forall\, \varepsilon > 0$, $f$ is continuous on $(0, \infty)$.

**b.** Let $f\colon R^2 \to R^2$ be defined by the coordinate functions

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} x[1]^2 + 2x[2] \\ x[1] \end{bmatrix}.$$

Let $\mathbf{x}_* \in R^2$, and let $\{\mathbf{x}_n\}$ be any sequence in $R^2$ for which $\mathbf{x}_n \to \mathbf{x}_*$. To demonstrate continuity of $f$ at $\mathbf{x}_*$, we will show that $\mathbf{x}_n \to \mathbf{x}_* \Rightarrow \mathbf{f}(\mathbf{x}_n) \to \mathbf{f}(\mathbf{x}_*)$.

[2] This definition can be altered to provide definitions for *continuity from the right* and *continuity from the left*. For continuity from the right, the condition $w \geq x$ is added in part (a). The condition $x_n \geq x\ \forall\, n$ is added to part (b). For continuity from the left, the conditions become $w \leq x$ and $x_n \leq x$, respectively, $\forall\, n$.

Examine $f_1(\mathbf{x})$ first. Note that $f_1(\mathbf{x}_n) = x_n[1]^2 + 2x_n[2] = (x_n[1] \cdot x_n[1]) + 2x_n[2]$ can be interpreted as the summation of the sequence $\{x_n[1]\}\{x_n[1]\}$ and the sequence $2\{x_n[2]\}$. Since $x_n[1] \to x_*[1]$ and $x_n[2] \to x_*[2]$ by assumption, it follows from Lemma 5.1 that $\{x_n[1]\}\{x_n[1]\} \to x_*[1]^2, 2\{x_n[2]\} \to 2x_*[2]$, and thus $f_1(\mathbf{x}_n) \to x_*[1]^2 + 2x_*[2] = f_1(\mathbf{x}_*)$. Verifying convergence of the second coordinate function is straightforward, since $f_2(\mathbf{x}_n) = x_n[1] \to x_*[1] = f_2(\mathbf{x}_*)$. Thus, $f$ is continuous at $\mathbf{x}_*$, and since the above argument holds for any $\mathbf{x}_* \in R^2$, $f$ is continuous on $R^2$.

**c.** Let $f\colon\ R \to R$ be defined by $f(x) = I_{[0,\infty)}(x)$. Intuitively, since the graph of the function has a break at $x = 0$, we expect that the function is not continuous on $R$. To demonstrate formally that $f$ is discontinuous at $x = 0$, it will be shown that there does *not* exist a $\delta(\varepsilon) > 0$ for every $\varepsilon > 0$ such that $|x - 0| < \delta(\varepsilon) \Rightarrow |f(x) - f(0)| < \varepsilon$. Choose any $\varepsilon \in (0, 1)$, and note that $f(0) = 0$. Given *any* choice of $\delta(\varepsilon) > 0$, let $x = \delta(\varepsilon)/2$, so that $|x - 0| < \delta(\varepsilon)$. Then $|f(x) - f(0)| = 1 \not< \varepsilon$. Therefore, $f$ is not continuous on $R$. (Note: $f$ *is* continuous on $(0, \infty)$ and on $(-\infty, 0)$, as the reader might wish to verify as an exercise.) □

### Convergence of Function Sequence

When we speak of **convergence of a sequence of functions**, we are referring to a case where the function definitions themselves can change as $n$ changes, so that we can conceptualize a sequence of image values $f_1(x), f_2(x), f_3(x), \ldots$ for *each* value of $x$ in the *common* domain of the sequence of functions. Interest centers on whether there exists a "limit" function definition, $f(x)$, such that $f_n(x) \to f(x)$ for all $x$ in some subset of the common domain of the sequence of functions. The subset of $x$ values on which $\{f_n(x)\}$ converges to $f(x)$ could be the entire domain, or some smaller subset of points, or the null set (i.e., $\{f_n(x)\}$ does not converge for any $x$).

We formalize the concept of convergence of a sequence of functions in the next definition. In the definition, the notation $\{f_n\}$ refers to the sequence of function definitions, while $\{f_n(x)\}$ refers to the sequence of image values generated by the sequence of function definitions when evaluated at the point $x$.

**Definition 5.7**
***Convergence of a function sequence***

Let $\{f_n\}$ be a sequence of functions, $f_n\colon\ D \to R^\ell$, having common domain $D \subset R^m$. Let $f\colon\ D_0 \to R^\ell$ be a function with domain $D_0 \subset D$. The function sequence $\{f_n\}$ is said to converge on $D_0$ to $f$ if $f_n(x) \to f(x)\ \forall\ x \in D_0$. If $\{f_n\}$ converges to $f$ on $D_0$, $f$ is called the **limiting function of** $\{f_n\}$ **on** $D_0$, and $\{f_n\}$ is said to be *convergent on* $D_0$.

Intuitively, if $f$ is the limit function of $\{f_n\}$ on $D_0$, then for large enough $n$, $f_n(x) \approx f(x)$ for $x \in D_0$ since $f_n(x)$ converges to $f(x)$ on $D_0$. Then $f(x)$ can be viewed as an approximation to $f_n(x)$ on $D_0$ when $n$ is large.

**Example 5.9** Convergence of Function Sequences

**a.** Let the function sequence $\{f_n\}$ be defined by $f_n(x) = n^{-1} + 2x^2$ for $x \in R$. Define the function $f\colon R \to R$ by $f(x) = 2x^2$. Then $f$ is the limiting function of $\{f_n\}$ on $R$. To see this, note that $f_n(x) \to 2x^2 = f(x)\ \forall\ x \in R$. For large $n$, $f_n(x) \approx f(x)\ \forall\ x \in R$.

**b.** Let the vector function sequence $\{f_n\}$ be defined by

$$\mathbf{f}_n(\mathbf{x}) = \begin{bmatrix} f_{1n}(\mathbf{x}) \\ f_{2n}(\mathbf{x}) \end{bmatrix} \begin{bmatrix} \dfrac{(2x_1^2 + 3nx_1x_2)}{n} \\ x_1^2 + x_2^2 \end{bmatrix} \quad \text{for } (x_1, x_2) \in R^2.$$

Define the function $f\colon R^2 \to R^2$ by

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} 3x_1x_2 \\ x_1^2 + x_2^2 \end{bmatrix} \quad \text{for } (x_1, x_2) \in R^2.$$

Then $f$ is the limiting function of $\{f_n\}$ on $R^2$. To see this, note that $f_{1n}(\mathbf{x}) = 2x_1^2/n + 3x_1x_2 \to 0 + 3x_1x_2 = f_1(\mathbf{x})$, and $f_{2n}(\mathbf{x}) = x_1^2 + x_2^2 = f_2(\mathbf{x})\ \forall\ \mathbf{x} \in R^2$, so that $f_n(\mathbf{x}) \to f(\mathbf{x})\ \forall\ \mathbf{x} \in R^2$.

**c.** Let the function sequence $\{f_n\}$ be defined by $f_n(x) = x - 2\exp(-nx)$ for $x \in R$. Let $f\colon R \to R$ be defined by $f(x) = x - 2I_{\{0\}}(x)$ for $x \in D_0 = [0, \infty]$. Then $f$ is the limiting function of $\{f_n\}$ on the set $[0, \infty)$. The sequence $\{f_n\}$ does not converge for $x < 0$. To justify these conclusions, note that, by Lemma 5.1, $\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} x - \lim_{n\to\infty} 2\exp(-nx) = x - 2I_{\{0\}}(x)\ \forall\ x \geq 0$. When $x < 0$, $2\exp(-nx) \to \infty$, and thus $f_n(x)$ does not converge. □

### Order of Magnitude of a Sequence

In analyses involving sequences, it is sometimes useful to be able to characterize or compare sequences and/or terms in a sequence relative to their order of magnitude in addition to, or in lieu of, any examination of convergence properties of the sequences involved. In particular, when the definition of a sequence contains a number of terms, the orders of magnitude of the terms will distinguish which terms make dominant contributions to the magnitude of the sequence as $n$ increases. This concept is defined below.

**Definition 5.8** ***Order of magnitude of a sequence***

Let $\{x_n\}$ be a real number sequence, and let $\{\mathbf{W}_n\}$ be a real-valued matrix sequence.

**a.** $\mathrm{O}(n^k)$: The sequence $\{x_n\}$ is said to be **at most of order** $n^k$, denoted by $\mathrm{O}(n^k)$, if there exists a finite real number $c$ such that $|n^{-k}x_n| \leq c\ \forall\ n \in N$.

**b.** $\mathrm{o}(n^k)$: The sequence $\{x_n\}$ is said to be **of order smaller than** $n^k$, denoted by $\mathrm{o}(n^k)$, if $n^{-k}x_n \to 0$.

**c.** If $\{\mathbf{W}_n[i, j]\}$ is $\mathrm{O}(n^k)$ (or $\mathrm{o}(n^k)$) $\forall\ i$ and $j$, then the matrix sequence $\{\mathbf{W}_n\}$ is said to be $\mathrm{O}(n^k)$ (or $\mathrm{o}(n^k)$).

Intuitively, a sequence $\{x_n\}$ is $O(n^k)$ if the corresponding sequence $\{n^{-k}x_n\}$ is such that all elements of $n^{-k}x_n$ are bounded in absolute value by the positive number $c$. A sequence $\{x_n\}$ is $o(n^k)$ if the product sequence $\{n^{-k}x_n\}$ converges to zero. Note that if $\{x_n\}$ is $O(n^k)$, then $\{x_n\}$ is $o(n^{k+\varepsilon})$ $\forall$ $\varepsilon > 0$, and if $\{x_n\}$ is $o(n^k)$, then it is also $O(n^k)$. Notationally, the case $O(n^0)$ or $o(n^0)$ is most often represented as $O(1)$ or $o(1)$.

**Example 5.10** Order of Magnitude of a Sequence

**a.** Let $\{x_n\}$ be defined by $x_n = 3n^3 - n^2 + 2$, for $n \in N$. Then $\{x_n\}$ is $O(n^3)$, since $n^{-3}x_n = 3 - n^{-1} + 2n^{-3}$ is bounded. Also, $\{x_n\}$ is $o(n^{3+\varepsilon})$ for any $\varepsilon > 0$ since $n^{-3-\varepsilon}x_n = 3n^{-\varepsilon} - n^{-1-\varepsilon} + 2n^{-3-\varepsilon} \to 0$.

**b.** Let $\{x_n\}$ be defined by $x_n = 3 + n^{-1}$, for $n \in N$. Then $\{x_n\}$ is $O(1)$, since $x_n = 3 + n^{-1}$ is bounded, and $\{x_n\}$ is $o(n^{\varepsilon})$ $\forall$ $\varepsilon > 0$, since $n^{-\varepsilon}x_n = 3n^{-\varepsilon} + n^{-1-\varepsilon} \to 0$.

**c.** Let the vector sequence $\{\mathbf{x}_n\}$ be defined by

$$\begin{bmatrix} x_n[1] \\ x_n[2] \end{bmatrix} = \begin{bmatrix} 3n^{-1} \\ n^{-1} \end{bmatrix}.$$

Then the vector (or $(2 \times 1)$ matrix) sequence $\{\mathbf{x}_n\}$ is $o(1)$ and $O(1)$, since $x_n \to \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. □

Some useful results regarding the order of magnitude of sums and products of sequences are given in the following lemma.

**Lemma 5.2**

Let $\{x_n\}$ and $\{z_n\}$ be real number sequences. The following relationships between orders of magnitude hold:

| IF | | THEN | |
|---|---|---|---|
| $\{x_n\}$ | $\{z_n\}$ | $\{x_n + z_n\}$ | $\{x_n z_n\}$ |
| $O(n^k)$ | $O(n^m)$ | $O(n^{\max(k,m)})$ | $O(n^{k+m})$ |
| $o(n^k)$ | $o(n^m)$ | $o(n^{\max(k,m)})$ | $o(n^{k+m})$ |
| $O(n^k)$ | $o(n^m)$ | $O(n^{\max(k,m)})$ | $o(n^{k+m})$ |

**Proof**: H. White (1984), *Asymptotic Theory for Econometricians.* Orlando, Academic Press, p. 15.

Given that the sequences referred to in Lemma 5.2 can be functions of other sequences, the results can be extended in a myriad of ways to an arbitrary finite number of sequences. In the following lemma we state some useful extensions.

**Lemma 5.3**

Let $\{\mathbf{X}_n\}$ and $\{\mathbf{W}_n\}$ be $(m \times \ell)$ and $(r \times m)$ matrix sequences, respectively.

**a.** If $\{\mathbf{X}_n\}$ is such that $\{\mathbf{X}_n[.,i]\}$ is $O(n^{k_i})$ (or $o(n^{k_i})$) for $i = 1,\ldots,\ell$, then $\{\sum_{i=1}^{\ell} \mathbf{X}_n[.,i]\}$ is $O(n^{k\max})$ (or $o(n^{k\max})$), where $k\max = \max(k_1,\ldots,k_\ell)$.

b. (Special case of (a)): If $\{\mathbf{X}_n\}$ is $O(n^k)$ (or $o(n^k)$), then $\{\sum_{i=1}^{\ell} \mathbf{X}_n[.,i]\}$ is $O(n^k)$ (or $o(n^k)$).

c. If $\{\mathbf{X}_n\}$ is $O(n^k)$ and $\{\mathbf{W}_n\}$ is $\left\{\begin{smallmatrix} O(n^d) \\ o(n^d) \end{smallmatrix}\right\}$, then $\{\mathbf{W}_n\mathbf{X}_n\}$ is $\left\{\begin{smallmatrix} O(n^{k+d}) \\ o(n^{k+d}) \end{smallmatrix}\right\}$ and $\{n^{-v}\mathbf{W}_n\mathbf{X}_n\}$ is $\left\{\begin{smallmatrix} O(n^{k+d-v}) \\ o(n^{k+d-v}) \end{smallmatrix}\right\}$.

d. (Special case of (c)): If $\{\mathbf{X}_n\}$ is $O(n^k)$ (or $o(n^k)$), then $\{\mathbf{X}'_n\mathbf{X}_n\}$ is $O(n^{2k})$ (or $o(n^{2k})$) and $\{n^{-v}\mathbf{X}'_n\mathbf{X}_n\}$ is $O(n^{2k-v})$ (or $o(n^{2k-v})$).

**Proof**: This follows from Lemma 5.2 and mathematical induction.

We illustrate the application of some of the preceding results in the following example.

**Example 5.11** Orders of Magnitude for Combinations of Sequences

a. Let $\{x_n\}$ be defined by $x_n = 2n^{-1} + 5n$, and let $\{z_n\}$ be defined by $z_n = n^2 + 2n$. Note that $\{x_n\}$ is $O(n^1)$ and $\{z_n\}$ is $O(n^2)$. It follows immediately, from Lemma 5.2, that $\{x_n + z_n\} = \{n^2 + 7n + 2n^{-1}\}$ is $O(n^2)$ (and $o(n^{2+\varepsilon})$ for $\varepsilon > 0$) and $\{x_n z_n\} = \{5n^3 + 10n^2 + 2n + 4\}$ is $O(n^3)$ (and $o(n^{3+\varepsilon})$ for $\varepsilon > 0$).
b. Let $\{\mathbf{x}_n\}$ and $\{\mathbf{W}_n\}$ be respectively defined by

$$\mathbf{x}_n = \begin{bmatrix} 7+n^{-1} \\ n^{-1} \end{bmatrix} \quad \text{and} \quad \mathbf{W}_n = \begin{bmatrix} n^2+2n+1 & 3n^2+7 \\ n^2 & n^2+n \end{bmatrix}$$

so that $\mathbf{x}_n$ is $O(1)$ and $\mathbf{W}_n$ is $O(n^2)$. From Lemma 5.3, it follows immediately that

$$\{\mathbf{W}_n\mathbf{x}_n\} = \left\{\begin{bmatrix} 7n^2 + 18n + 9 + 8n^{-1} \\ 7n^2 + 2n + 1 \end{bmatrix}\right\}$$

is $O(n^2)$ and $o(n^{2+\varepsilon})$ for $\varepsilon > 0$, and $\{\mathbf{x}'_n\mathbf{x}_n\} = \{49 + 14n^{-1} + 2n^{-2}\}$ is $O(1)$ and $o(n^{\varepsilon})$ for $\varepsilon > 0$. It also follows, for example, that $\{n\mathbf{x}'_n\mathbf{x}_n\}$ and $\{n^{-1}\mathbf{W}_n\mathbf{x}_n\}$ are both $O(n^1)$ and $o(n^{1+\varepsilon})$ for $\varepsilon > 0$. □

## 5.3 Types of Random-Variable Convergence

The convergence concepts examined in the preceding section relate to *nonrandom* sequences of real numbers, vectors, or matrices. We now extend these convergence concepts to sequences of random variables, such as $\{X_n\} = \{X_1, X_2, X_3, \ldots\}$.

A fundamental difference between a non-random sequence $\{x_n\}$ and a random sequence $\{X_n\}$ is that the elements of a random sequence are random variables as opposed to fixed numbers and thus are capable of assuming any real numbers within the ranges of the respective random variables in the sequence. Thus, while $\{x_n\}$ refers to only one sequence of real numbers, $\{X_n\}$ refers to a collection of random variables capable of defining many different real number sequences, depending on the outcomes of the random variables $X_1, X_2, X_3,$ .... Since any *particular* real number sequence associated with $\{X_n\}$ can be thought of as an outcome of the random variables involved in the sequence, it is then meaningful to define probabilities of various types of events involving outcomes of the random variables in the sequence. For example, one might be interested in the probability that the outcomes of a sequence of scalar random variables $\{X_n\}$ are bounded by a particular value $m > 0$, i.e., $P(|x_n| \leq m, \forall\, n \in N)$, or in the probability that the outcomes of the sequence converge to a limit $c$, i.e., $P(\lim_{n\to\infty} x_n = c)$. Since the sequence is random, convergence and boundedness questions cannot be verified as being unequivocally true or false on the basis of a given sequence of real numbers, but they can be assigned a probability of occurrence in the context of the probability space for the outcomes of the random variables involved.

We will examine four basic types of random-variable convergence:

1. convergence in distribution;
2. convergence in probability;
3. convergence in mean square (or convergence in quadratic mean); and
4. almost-sure convergence (convergence with probability 1).

### Convergence in Distribution

The concept of convergence in distribution involves the question of whether the sequence of cumulative distribution functions associated with the random variables in the sequence $\{Y_n\}$ converges to a limiting cumulative distribution function (recall Def. 5.7). In many cases of practical interest, convergence in distribution can also be characterized in terms of the convergence of a sequence of density functions to a limiting density function. The usefulness of the concept lies in establishing an *approximation* to the true CDF or PDF for $Y_n$ when $n$ is large enough (where "large enough" means that the CDF or PDF of $Y_n$ is close to its limiting CDF or PDF). Such approximating CDFs or PDFs can be extremely useful when the true CDF or PDF for $Y_n$ is very difficult (or impossible) to define or is intractable to work with, but the limiting CDF or PDF is easier to define and analyze.

We first characterize convergence in distribution in terms of a sequence of CDFs. We alert the reader to the fact that *all results presented henceforth can be interpreted in terms of multivariate random variables,* unless the context is explicitly defined in terms of scalar random variables.

**Definition 5.9**
***Convergence in distribution (CDFs)***

Let $\{Y_n\}$ be a sequence of random variables, and let $\{F_n\}$ be the associated sequence of cumulative distribution functions corresponding to the random variables. If there exists a cumulative distribution function $F$ such that $F_n(y) \to F(y)\ \forall\, y$ at which $F$ is continuous, then $F$ is called the **limiting CDF of** $\{Y_n\}$. Letting $Y$ have the distribution $F$, i.e., $Y \sim F$, we then say that $Y_n$

converges in distribution (or converges in law) to the random variable $Y$, and we denote this convergence by $Y_n \xrightarrow{d} Y$ (or $Y_n \xrightarrow{L} Y$). We also write $Y_n \xrightarrow{d} F$ as a short hand notation for $Y_n \xrightarrow{d} Y \sim F$, which is read "$Y_n$ converges in distribution to $F$."

If $Y_n \xrightarrow{d} Y$, then as $n$ becomes large, the CDF of $Y_n$ is approximated ever more closely by its limiting CDF, the CDF of $Y$ (see Figure 5.2). Note that it is admissible in the definition that $Y = c$, i.e., the random variable can be degenerate so that $P(y = c) = 1$. When the limiting CDF is associated with a degenerate random variable, we say that the sequence of random variables **converges in distribution to a constant**, and we denote this by $Y_n \xrightarrow{d} c$.

In the case of either nonnegative, integer-valued, discrete random variables or continuous random variables, convergence of a sequence of PDFs is sufficient for establishing convergence in distribution.

**Theorem 5.1** (Convergence in Distribution (Densities)) *Let $\{Y_n\}$ be a sequence of either continuous or nonnegative, integer-valued, discrete random variables, and let $\{f_n\}$ be the associated sequence of probability density functions corresponding to the random variables. Let there exist a density function $f$ such that $f_n(y) \to f(y)$ $\forall\, y$, except perhaps on a set of points $A$ such that $P_Y(A) = 0$ in the continuous case, where $Y \sim f$. It follows that $Y_n \xrightarrow{d} Y$ (or $Y_n \xrightarrow{L} Y$).*

**Proof** The discrete case is left to the reader. For the continuous case, see H. Scheffé (1947), "A useful convergence theorem for probability distributions." *Ann. Math. Stat.*, 18, pp. 434–438. ■

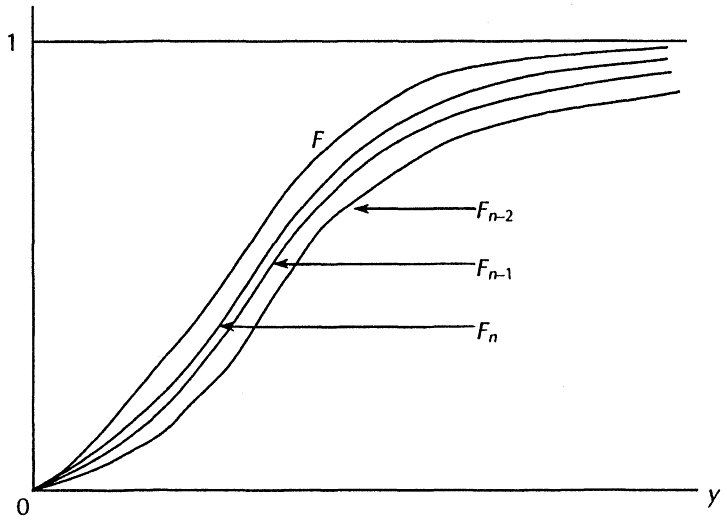


**Figure 5-2** Convergence in distribution (CDF).

If the PDFs of the elements of $\{Y_n\}$ converge to the PDF $f$ as indicated in Theorem 5.1, then $f$ is referred to as the **limiting density of** $\{Y_n\}$. The notation $Y_n \xrightarrow{d} f$ is sometimes used as short hand for $Y_n \to Y \sim f$, where it is understood that the limiting CDF of $\{Y_n\}$ in this case coincides with the CDF associated with the PDF $f$. The term **limiting distribution** is often used generically to refer to either the limiting CDF or the limiting density. If $Y_n \xrightarrow{d} Y$ and if the PDFs $\{f_n\}$ converge to a limiting density, then as $n$ becomes large, the PDF of $Y_n$ is approximated ever more closely by its limiting density, the density of $Y$ (for example, see Figure 5.3).

The following example illustrates the concept of convergence in distribution characterized through convergence of CDFs. It is also designed to provide some insight into the qualifier "such that $F$ is continuous at $y$" used in Def. 5.9.

**Example 5.12** Convergence of CDFs

Let $\{Y_n\}$ be such that the cumulative distribution function for $Y_n$ is defined by

$$F_n(y) = \tfrac{n}{2}(y - \tau + n^{-1})I_{[\tau-n^{-1},\tau+n^{-1})}(y) + I_{[\tau+n^{-1},\infty)}(y),$$

and let $Y$ have the cumulative distribution function

$$F(y) = I_{[\tau,\infty)}(y).$$

Then $Y_n \xrightarrow{d} Y$. To see this, note that

$$\lim_{n\to\infty} F_n(y) = \tfrac{1}{2}I_{\{\tau\}}(y) + I_{(\tau,\infty)}(y)$$

(see Figure 5.4) which agrees with $F(y)$ for all points at which $F$ is *continuous*, which is all points except $\tau$. Thus, by definition, the limiting distribution of $Y_n$ is $F(y)$. □

Note that $F(y)$ in Ex. 5.12 makes intuitive sense as a limiting distribution for $Y_n$ when interpreted in the context of its original purpose—that of providing an

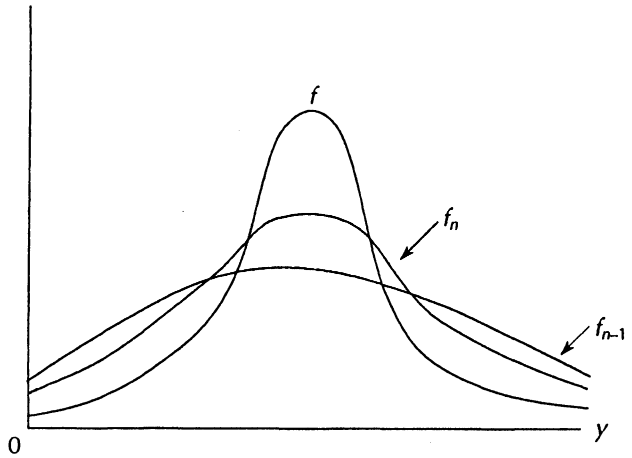


**Figure 5-3**
Convergence in distribution (densities).

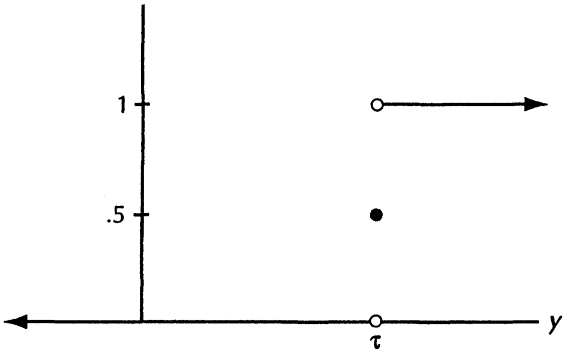


**Figure 5-4**
Graph of $\lim_{n\to\infty} F_n(y)$.

approximation to $F_n(y)$ for large $n$. Viewed in this way, the fact that $\lim_{n\to\infty} F_n(y)$ does not agree with $F(y)$ at $\tau$ is immaterial, since it is not $\lim_{n\to\infty} F_n(y)$ that we wish to approximate in any case.[3] The reader will note that as $n$ increases, $F_n(y)$ is ever more closely approximated by $F(y)$. In particular, from the definition of $F_n(y)$, it follows that $P(y_n > \tau + n^{-1}) = P(y_n < \tau - n^{-1}) = 0 \;\forall\; n$, implying $P(\tau - n^{-1} \leq y_n \leq \tau + n^{-1}) = 1 \;\forall\; n$. We would like to add that when $n \to \infty, P(y_n = \tau) \to 1$, i.e., the sequence of random variables with distribution functions $\{F_n\}$ converges in distribution to a degenerate random variable at $\tau$. Note that $F(y)$ implies this degeneracy.

The fact that the limiting distribution is degenerate in Ex. 5.12 when all of the distributions $F_n(y)$ are nondegenerate results in the limiting distribution's being inadequate for approximating certain characteristics of the distribution of $Y_n \sim F_n(y)$. For example, $F(y)$ clearly provides no information regarding the degree of spread in the outcomes of $Y_n$. We will return to this question shortly, where the notion of an *asymptotic distribution* will be introduced as an alternative to using a *degenerate* limiting distribution for approximating the distribution of random variables $Y_n$ in a sequence $\{Y_n\}$.

In the next example, we illustrate the concept of convergence in distribution characterized through convergence of density functions.

**Example 5.13** Convergence of Densities

Let $\{X_n\}$ be a sequence of normal random variables such that $X_n \sim N(0, 1) \;\forall\; n$. Let $\{Z_n\}$ be a sequence defined by $Z_n = (3 + n^{-1})X_n + (2n/(n-1))$, which then has an associated sequence of density functions $\{f_n\}$ defined by $f_n = N(2n/(n-1), (3 + n^{-1})^2)$. Since $f_n \to N(2, 9) = f$ (note that $f_n(z) \to f(z) \;\forall\; z \in R$), then $N(2, 9)$ is the limiting density of $\{Z_n\}$, i.e., $Z_n \xrightarrow{d} N(2, 9)$. □

We now formally extend the uniqueness theorem of moment-generating functions (recall Theorem 3.27) in a way that allows one to identify limiting

[3] Note that $\lim_{n\to\infty} F_n(y)$ is not a cumulative distribution function.

distributions of sequences of random variables through an examination of limiting moment-generating functions. This result is quite useful when, as is often the case, the limit of a sequence of moment-generating functions is more easily defined than the limit of a sequence of CDFs or density functions.

**Theorem 5.2** (Convergence in Distribution (MGFs)) *Let $\{Y_n\}$ be a sequence of random variables having an associated sequence of moment generating functions $\{M_{Y_n}(t)\}$. Let Y have the moment-generating function $M_Y(t)$. Then $Y_n \xrightarrow{d} Y$ iff $M_{Y_n}(t) \to M_Y(t)\ \forall\ t \in (-h,h)$, for some $h > 0$.*

**Proof** See E. Lukacs (1970), *Characteristic Functions*. London: Griffin, pp. 49–50, for a proof of this theorem for the more general case characterized by convergence of characteristic functions (which subsumes Theorem 5.2 as a special case). In the multivariate case, $t$ will be a vector, and convergence of the MGF must hold $\forall\ t_i \in (-h, h)$, and $\forall\ i$. ■

The theorem implies that if we can discover that $\lim_{n\to\infty} M_{Y_n}(t)$ exists and is equal to the moment-generating function $M_Y(t)\ \forall\ t$ in an open interval containing zero, then the distribution associated with $M_Y(t)$ is the limiting distribution of the sequence $\{Y_n\}$. We illustrate the use of Theorem 5.2 in the following example.

**Example 5.14** $Z_n = (X_n - n)/\sqrt{2n} \xrightarrow{d} N(0,1)$ if $X_n \sim \chi^2_n$

Let $\{X_n\}$ be a sequence of random variables, where $X_n \sim \chi^2_n\ \forall\ n$. The sequence $\{X_n\}$ then has an associated sequence of MGFs given by $\{M_{X_n}(t)\}$, where $M_{X_n}(t) = (1-2t)^{-n/2}\ \forall\ n$. Let the random sequence $\{Z_n\}$ be defined by $Z_n = (X_n - n)/\sqrt{2n}$, which has an associated sequence of MGFs given by $\{M_{Z_n}(t)\}$, where $M_{Z_n}(t) = (1 - \sqrt{2/n}\,t)^{-n/2}\exp(-\sqrt{n/2}\,t)$. Note that the elements in the sequence $\{Z_n\}$ are *standardized* $\chi^2$ random variables (recall $\mathrm{E}X_n = n$ and $\mathrm{var}(X_n) = 2n$). We now show that $\ln M_{Z_n}(t) \to t^2/2$. To see this, first note that

$$\ln M_{Z_n}(t) = -\frac{n}{2}\ln\left(1 - \sqrt{\frac{2}{n}}\,t\right) - \sqrt{\frac{n}{2}}\,t.$$

Expressing the first term on the right-hand side of the equality in terms of a Taylor series expansion around $t = 0$ obtains

$$\ln M_{Z_n}(t) = \left[\sqrt{\frac{n}{2}}\,t + \frac{t^2}{2} + o(1)\right] - \sqrt{\frac{n}{2}}\,t = \frac{t^2}{2} + o(1) \to \frac{t^2}{2}.$$

It follows that

$$\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty} \exp\left(\ln M_{Z_n}(t)\right) = \exp\left(\lim_{n\to\infty} \ln M_{Z_n}(t)\right) = \exp\left(\frac{t^2}{2}\right)$$

since $\exp(\cdot)$ is a continuous function (recall Def. 5.6, especially part (b). Since $\exp(t^2/2)$ is the MGF of a standard normal distribution, we know by Theorem 5.2 that the sequence $\{Z_n\}$ of standardized $\chi^2_n$ random variables has an $N(0,1)$ limiting distribution, i.e., $Z_n \xrightarrow{d} N(0,1)$. □

*Asymptotic Distributions* We now introduce the concept of an **asymptotic distribution** for the elements in a random sequence. In the most general sense of the term, an *asymptotic distribution* for $Z_n$ in the sequence $\{Z_n\}$ is any distribution that provides an approximation to the true distribution of $Z_n$ for large $n$. In this general sense, if $\{Z_n\}$ has a limiting distribution, then since the limiting distribution can be interpreted as an approximation to the distribution of $Z_n$ for large $n$, the limiting distribution might be considered an *asymptotic distribution* for $Z_n$. The purpose of introducing the additional concept of asymptotic distributions is to generalize the concept of approximating distributions for large $n$ to include cases where either the sequence $\{Z_n\}$ has *no* limiting distribution, or the limiting distribution is *degenerate* (and hence not particularly useful as an approximation to the distribution of a nondegenerate $Z_n$).

We will focus on asymptotic distributions that apply to random variables defined by $g(X_n, \Theta_n)$, where $X_n \xrightarrow{d} X$ for nondegenerate $X$. This context covers most cases of practical interest.

**Definition 5.10**
***Asymptotic distribution for $g(X_n, \Theta_n)$ when $X_n \xrightarrow{d} X$***

Let the sequence of random variables $\{Z_n\}$ be defined by $Z_n = g(X_n, \Theta_n)$, where $X_n \xrightarrow{d} X$ for *nondegenerate* $X$, and $\{\Theta_n\}$ is a sequence of real numbers, matrices, and/or parameters. Then an *asymptotic distribution for* $Z_n$ is given by the distribution of $g(X, \Theta_n)$, denoted by $Z_n \overset{a}{\sim} g(X, \Theta_n)$ and meaning "$Z_n$ is **asymptotically distributed as** $g(X, \Theta_n)$.

Before proceeding to examples, we provide some general intuition underlying Def. 5.10. Operationally, Def. 5.10 implies that if $Z_n$ can be defined as a function, $g(X_n, \Theta_n)$, of some random variable $X_n$ for which $X_n \xrightarrow{d} X$ and $X$ is nondegenerate, then an asymptotic distribution for $Z_n$ is given by the probability distribution associated with the same function *applied to* $X$, $g(X, \Theta_n)$. The problem of identifying asymptotic distributions is then equivalent to the problem of identifying distributions of functions of $X$ since $Z_n = g(X_n, \Theta_n)$ and $X_n \xrightarrow{d} X \Rightarrow Z_n \overset{a}{\sim} g(X, \Theta_n)$.

In order to motivate why this approach to approximating the distribution of $Z_n$ for large $n$ makes sense, first note that a correspondence between events for $Z_n$ and events for $X_n$ can be established as

$$z_n \in A \Longleftrightarrow x_n \in B_n = \{x_n : z_n = g(x_n, \Theta_n), z_n \in A\}.$$

Note further that since $X_n \xrightarrow{d} X$, we can approximate the probability of the event $B_n$ for $X_n$ using the nondegenerate limiting distribution of $\{X_n\}$ as $P_{X_n}(x_n \in B_n) \approx P_X(x \in B_n)$, for large $n$. It follows from the equivalence of the events $A$ and $B_n$ that, for large $n$,

$$P_{Z_n}(z_n \in A) = P_{X_n}(x_n \in B_n) \approx P_X(x \in B_n) = P_X(g(x, \Theta_n) \in A).$$

Thus, an approximation to $P_{Z_n}(z_n \in A)$ can be calculated using the probability distribution of $g(X, \Theta_n)$ in the same sense, and with the same degree of accuracy, that $P_{X_n}(x_n \in B_n)$ can be approximated using the probability distribution of $X$.

We emphasize that a limiting distribution for $\{Z_n\} = \{g(X_n, \Theta_n)\}$ need *not* exist for the preceding approximation arguments to hold.

**Example 5.15** Asymptotic Distribution for $n^{-1}X_n$, where $X_n \sim \text{Bin}(n, \lambda/n)$

Let $\{X_n\}$ be such that $X_n$ has a binomial distribution with parameters $n$ and $p = \lambda/n$ for some fixed value of $\lambda > 0$, i.e., $X_n \sim \text{Bin}(n, \lambda/n)$. Note that $M_{X_n}(t) = (1+p(e^t-1))^n = (1+(\lambda(e^t-1)/n))^n \rightarrow e^{\lambda(e^t-1)}$ (recall that $\lim_{n\rightarrow\infty}(1+(a/n))^n = e^a$), which by Theorem 5.2 implies that $X_n \xrightarrow{d} X \sim (e^{-np}(np)^x/x!)I_{\{0,1,2,\ldots\}}(x)$ (Poisson distribution with parameter $\lambda = np$). Note that this result is in agreement with the relationship established between the Binomial and Poisson distributions in Chapter 4. Define the sequence $\{Z_n\}$ by $Z_n = g(X_n, n) = n^{-1}X_n$, and note that since $M_{Z_n}(t) = (1+(\lambda/n)(e^{t/n}-1))^n \rightarrow 1$, then by Theorem 5.2 the *limiting distribution* of $\{Z_n\}$ exists and is *degenerate* at 0 (i.e., an MGF equal to the constant function 1 is associated with the density for which $P(z = 0) = 1$). On the other hand, the *asymptotic distribution* of $Z_n$ as defined in Def. 5.10 is given by the distribution of $g(X, n) = n^{-1}X$, so that $Z_n \overset{a}{\sim} (e^{-np}(np)^{nx}/(nx)!)I_{\{0,1/n,2/n,\ldots\}}(x)$. □

**Example 5.16** $Y_n \overset{a}{\sim} N(n, 2n)$ if $Y_n \sim \chi^2_n$

In Ex. 5.14, it was demonstrated that $W_n \xrightarrow{d} W \sim N(0,1)$, where $W_n = (Y_n - n)/\sqrt{2n}$ and $Y_n \sim \chi^2_n$. It follows from Def. 5.10, using results on linear combinations of normally distributed random variables, that

$$Y_n = g(W_n, n) = \sqrt{2n}\, W_n + n \overset{a}{\sim} \sqrt{2n}\, W + n \sim N(n, 2n).$$

Thus, for large $n$, a $\chi^2$-distributed random variable is approximately normally distributed with mean $n$ and variance $2n$. Note that $\{Y_n\}$ does *not* have a limiting distribution. As an illustration of the use of the approximation, note that if $n = 100$, then referring to the $\chi^2$-distribution with 100 degrees of freedom, $P(\chi_{100} \leq 124.34) = .95$. Using the asymptotic distribution $N(100, 200)$, the approximation to this probability is given by $\hat{P}(\chi_{100} \leq 124.34) \approx .9573$. □

*Convergence in Distribution for Continuous Functions* The final result involving the concept of convergence in distribution that we will present here facilitates identification of the limiting distribution of a continuous function, $g(X_n)$, when $\{X_n\}$ is such that $X_n \xrightarrow{d} X$.

**Theorem 5.3** *Let $X_n \xrightarrow{d} X$, and let the random variable $g(X)$ be defined by a function $g(x)$ that is continuous, except perhaps on a set of points assigned probability zero by the probability distribution of $X$. Then $g(X_n) \xrightarrow{d} g(X)$.*

**Proof** See the proof of Theorem 5.17 and R. Serfling (1980), *Approximation Theorems of Mathematical Statistics*. New York: Wiley, pp. 24–25. ■

Thus, if $X_n \xrightarrow{d} X$, the *limiting distribution* of $g(X_n)$ is given by the distribution of $g(X)$ if $g$ is continuous with probability 1 (with respect to the distribution of $X$). Note that this result applies to a function of $X_n$ that does not depend on any arguments other than $X_n$ itself. In particular, the function $g$ cannot depend on $n$ or on any other sequence of real numbers or matrices whose elements change with $n$, so, for example, $g(x_n) = 3 + x_n$ would be admissible, while $g(x_n, n) = n^{1/2}x_n + 2n$ would not.

**Example 5.17** Let $\{Z_n\}$ be such that $Z_n \xrightarrow{d} Z \sim N(0,1)$. It follows immediately, both from Theorem 5.3 and from results on linear combinations of normally distributed random variables, that $g(Z_n) = 2Z_n + 5 \xrightarrow{d} 2Z + 5 \sim N(5,4)$. One could also demonstrate via Theorem 5.3 that $g(Z_n) = Z_n^2 \xrightarrow{d} Z^2 \sim \chi_1^2$ (since the square of a standard normal random is a $\chi^2$ random variable). □

**Example 5.18** Let $\{\mathbf{Z}_n\}$ be a sequence of bivariate random variables such that $\mathbf{Z}_n \xrightarrow{d} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = [2\ 3]'$ and $\boldsymbol{\Sigma} = \left[\begin{smallmatrix} 4 & 0 \\ 0 & 9 \end{smallmatrix}\right]$. Then it follows, both from Theorem 5.3 and from and results on linear combinations of normally distributed random variables, that $\mathbf{g}(\mathbf{Z}_n) = \mathbf{A}\mathbf{Z}_n \xrightarrow{d} \mathbf{A}\mathbf{Z} \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$. For example, if

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \qquad \text{then } \mathbf{A}\mathbf{Z}_n \xrightarrow{d} \left(\begin{bmatrix} 5 \\ -1 \end{bmatrix}, \begin{bmatrix} 13 & -5 \\ -5 & 13 \end{bmatrix}\right).$$

One could also demonstrate via Theorem 5.3 that $g(\mathbf{Z}_n) = (\mathbf{Z}_n - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Z}_n - \boldsymbol{\mu}) \xrightarrow{d} (\mathbf{Z} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \boldsymbol{\mu}) \sim \chi_2^2$, based on the fact that the two independent, standard normal random variables, represented by the $(2\times 1)$ vector $\boldsymbol{\Sigma}^{-1/2}(\mathbf{Z}-\boldsymbol{\mu})$, are squared and summed. □

### Convergence in Probability

Referring to a random sequence of *scalars*, $\{Y_n\}$, the concept of convergence in probability involves the question of whether outcomes of the random variable $Y_n$ are close to the outcomes of some random variable $Y$ with high probability when $n$ is large enough. If so, the outcomes of $Y$ can serve as an approximation to the outcomes of $Y_n$ for large enough $n$. Stated more rigorously, the issue is whether the sequence of probabilities associated with the sequence of events $\{(y_n, y): |y_n - y| < \varepsilon\}$ converges to 1 for every choice of $\varepsilon > 0$. When $\mathbf{Y}_n$ and $\mathbf{Y}$ are $(m \times k)$ *matrices*, convergence in probability of $\mathbf{Y}_n$ to $\mathbf{Y}$ requires that each element of $\mathbf{Y}_n$ converge in probability to the corresponding element of $\mathbf{Y}$. The vector case is subsumed by letting $k = 1$.

**Definition 5.11** ***Convergence in probability***

The sequence of random variables $\{Y_n\}$ *converges in probability* to the random variable $Y$ iff

**a.** Scalar case: $\lim_{n\to\infty} P(|y_n - y| < \varepsilon) = 1\ \forall\, \varepsilon > 0$,

**b.** Matrix case: $\lim_{n\to\infty} P(|y_n[i,j] - y[i,j]| < \varepsilon) = 1\ \forall\, \varepsilon > 0, \forall\, i$ and $j$.

Convergence in probability will be denoted by $Y_n \xrightarrow{p} Y$, or $\operatorname{plim} Y_n = Y$, the latter notation meaning the **probability limit of** $Y_n$ *is* $Y$.

Convergence in probability imposes a restriction on the joint probability distribution of $(Y_n, Y)$ as $n \to \infty$. In particular, as $n \to \infty$, the joint distribution of $(Y_n, Y)$ approaches a degenerate distribution defined by linear restrictions as $P(y_n = y) = 1$. To motivate this interpretation, note that Def. 5.11 implies that for any arbitrarily small $\varepsilon > 0$,

$$P\left(y_n[i,j] \in (y[i,j] - \varepsilon, y[i,j] + \varepsilon)\right) \to 1, \quad \forall\, i \text{ and } j.$$

It then follows from Bonferroni's probability inequality that

$$P(y_n[i,j] \in (y[i,j] - \varepsilon, y[i,j] + \varepsilon), \quad \forall\, i \text{ and } j) \to 1.$$

Thus, outcomes of $Y_n$ are arbitrarily close to outcomes of $Y$ with probability approaching 1 as $n \to \infty$. Therefore, for large enough $n$, observing outcomes of $Y$ is essentially equivalent to observing outcomes of $Y_n$,[4] which motivates the idea that the random variable $Y$ can serve as an approximation to the random variable $Y_n$.

Note that the random variable $Y$ in Def. 5.11 could be *degenerate*, i.e., $Y$ could be a number or matrix of numbers. Using the notation introduced in Def. 5.11, we denote this situation by $Y_n \xrightarrow{p} c$, where $c$ is a scalar, or matrix of constants. This situation is referred to as "$Y_n$ converges in probability to a constant."

**Example 5.19** Let $\{Y_n\}$ have an associated sequence of density functions, $\{f_n\}$, defined by $f_n(y) = n^{-1}I_{\{0\}}(y) + (1 - n^{-1})I_{\{1\}}(y)$. Note that

$$\lim_{n\to\infty} P(|y_n - 1| = 0) = \lim_{n\to\infty}(1 - n^{-1}) = 1,$$

so that $\lim_{n\to\infty} P(|y_n - 1| < \varepsilon) = 1\ \forall\, \varepsilon > 0$, and $\text{plim}\, Y_n = 1$. Thus, $Y_n$ converges in probability to 1. □

**Example 5.20** Let $Y \sim N(0,1)$, and let $\{Z_n\}$ have an associated sequence of density functions, $\{f_n\}$, with $\mathrm{E}Z_n = 0$ and $\text{var}(Z_n) = n^{-1}$. Define the random sequence $\{Y_n\}$ by $Y_n = Y + Z_n$, and assume $Y$ and $Z_n$ are independent, so that $\mathrm{E}Y_n = 0$ and $\text{var}(Y_n) = 1 + n^{-1}$. Then $Y_n \xrightarrow{p} Y$ (or $\text{plim}\, Y_n = Y$) since, $\forall\, \varepsilon > 0$,

$$\lim_{n\to\infty} P(|y_n - y| < \varepsilon) = \lim_{n\to\infty} P(|z_n| < \varepsilon) = 1,$$

which follows by an application of Markov's inequality. □

[4] The reader may be wondering why we don't simply replace the uncountably infinite collection of probability statements

$$P(|y_n - y| < \varepsilon), \quad \forall\, \varepsilon > 0,$$

in the definition of convergence in probability with the definition

$$\lim_{n\to\infty} P(y_n = y) = 1.$$

The problem is that such a convergence definition would not be useful in the examination of any sequence of nondegenerate, *continuous* random variables, since $P(y_n = y) = 0\ \forall\, n \Rightarrow \lim_{n\to\infty} P(y_n = y) = 0$, no matter how close the outcomes of $Y_n$ become to outcomes of $Y$ as $n \to \infty$.

**Example 5.21** Let $\{\mathbf{Y}_n\}$ be such that $\mathrm{E}\mathbf{Y}_n = \left[\begin{smallmatrix}2\\3\end{smallmatrix}\right]$ and $\mathbf{Cov}(\mathbf{Y}_n) = n^{-1}\left[\begin{smallmatrix}2&1\\1&1\end{smallmatrix}\right]$. Note by Markov's inequality that, $\forall\, \varepsilon > 0$,

$$\lim_{n\to\infty} P(|y_n[1] - 2| < \varepsilon) \geq \lim_{n\to\infty}\left(1 - \frac{2/n}{\varepsilon^2}\right) = 1,$$

$$\lim_{n\to\infty} P(|y_n[2] - 3| < \varepsilon) \geq \lim_{n\to\infty}\left(1 - \frac{n^{-1}}{\varepsilon^2}\right) = 1,$$

so that $Y_n[1] \xrightarrow{p} 2$ and $Y_n[2] \xrightarrow{p} 3$. From Def. 5.11, it follows that $\mathbf{Y}_n \xrightarrow{p} \mathbf{Y} = \left[\begin{smallmatrix}2\\3\end{smallmatrix}\right]$. □

Since Def. 5.11 states that $\operatorname{plim} \underset{(m\times k)}{\mathbf{Y}_n} = \underset{(m\times k)}{\mathbf{Y}}$ iff $\operatorname{plim} Y_n[i,j] = Y[i,j]\ \forall\, i,j$, it follows that "the probability limit of a matrix is the matrix of probability limits." Thus, just as for the expectation operator, the plim is an *element wise* operator.

**Definition 5.12** ***Probability limits of matrices (and vectors for k = 1)***

Let $\{\mathbf{Y}_n\}$ be a sequence of $(m \times k)$ random matrices. Then

$$\operatorname{plim}\begin{bmatrix} Y_n[1,1] & \ldots & Y_n[1,k] \\ \vdots & \ddots & \vdots \\ Y_n[m,1] & \ldots & Y_n[m,k] \end{bmatrix} = \begin{bmatrix} \operatorname{plim} Y_n[1,1] & \ldots & \operatorname{plim} Y_n[1,k] \\ \vdots & \ddots & \vdots \\ \operatorname{plim} Y_n[m,1] & \ldots & \operatorname{plim} Y_n[m,k] \end{bmatrix}.$$

At this point we emphasize a fundamental difference in concept between *convergence in distribution* and *convergence in probability* (which will also apply when comparing convergence in distribution to other types of convergence we will study). For the convergence in distribution concept $Y_n \xrightarrow{d} Y$, it is immaterial whether outcomes of the random variables in the sequence $\{Y_n\}$ are related in any way to outcomes of $Y$. In particular, it makes no difference whether $Y_n$ and $Y$ are even referring to the same type of experiment. The convergence $Y_n \xrightarrow{d} Y$ states only that the random variables in the sequence $\{Y_n\}$ have a limiting distribution that is equal to the probability distribution of $Y$. There is no implication that the outcomes of $Y_n$ necessarily emulate the outcomes of $Y$ in any way as $n \to \infty$ (although they might). This is simply a result of the fact that random variables with the same probability distributions are not necessarily the same random variable. For example, $X \sim f(x) = (1/2)I_{\{0,1\}}(x)$ could refer to tossing a fair coin where $x = 1$ is a head and $x = 0$ is a tail, while $Y \sim h(y) = (1/2)I_{\{0,1\}}(y)$ could be referring to the rolling of a fair die, where $y = 0$ stands for a roll of 3 or less and $y = 1$ stands for a roll of 4 or more. Clearly, the PDFs of $X$ and $Y$ are equal, i.e., $f(t) = h(t)\ \forall\, t$. Just as clearly, outcomes of $X$ and $Y$ are not related in any sense, since whether $x = 1$ (or 0) has no bearing on whether $y = 1$ (or 0). Indeed, $X$ and $Y$ refer to two different experiments.

The situation is quite different for convergence in probability, where the convergence referred to involves the outcomes of the respective random variables $Y_n$ and $Y$ themselves, and not merely convergence of their probability distributions. That is, convergence in probability implies that, for large $n$, outcomes of the random variable $Y_n$ in the sequence $\{Y_n\}$ are close to the outcomes of $Y$ with high probability, and for large enough $n$, observing an outcome of $Y$ is a very good approximation to observing the corresponding outcome of $Y_n$. Thus, for large enough $n$, outcomes of the experiment represented by $Y_n$ can be treated essentially as outcomes of the experiment represented by $Y$.

It is useful to note that if a real-valued sequence of numbers, or matrices, $\{a_n\}$ converges (in the sense of real analysis) to a given number, or matrix, i.e., $a_n \to a$, then it can also be stated that $\text{plim}\, a_n = a$, since $\{a_n\}$ can be interpreted as a sequence of *degenerate* random variables.

**Theorem 5.4** *Let* $\{a_n\}$ *be such that* $\underset{(m\times k)}{\mathbf{a}_n} \to \underset{(m\times k)}{\mathbf{a}}$. *Then* $\text{plim}\, \mathbf{a}_n = \mathbf{a}$.

**Proof** See the proof of Theorem 5.16. ■

As an illustration of Theorem 5.4, note that in Ex. 5.5.a, we could write $\text{plim}\, y_n = 3$, or in Ex. 5.6.b, we could write $\text{plim}\, \mathbf{Y}_n = \left[\begin{smallmatrix} 0 & 0 \\ 3 & 1 \end{smallmatrix}\right]$.

*Convergence in Probability for Continuous Functions* When a sequence of random variables is defined via a *continuous* function, $g(X_n)$, of the random variables in another sequence $\{X_n\}$, the plim operator acts analogously to the lim operator of real analysis (note the analogy to Theorem 5.3).

**Theorem 5.5** *Let* $X_n \xrightarrow{p} X$, *and let the random variable* $g(X)$ *be defined by a function* $g(x)$ *that is continuous, except perhaps on a set of points assigned probability zero by the probability distribution of* $X$. *Then* $g(X_n) \xrightarrow{p} g(X)$, *or equivalently,* $\text{plim}\, g(X_n) = g(\,\text{plim}\, X_n)$.

**Proof** See the proof of Theorem 5.17 and R. Serfling, Approximation Theorems, p. 24–25. ■

Theorem 5.5 says that for functions that are continuous with probability 1 (with respect to the distribution of $X$), *the probability limit of the function is the function of the probability limit.* The theorem can greatly simplify finding the probability limits of complicated functions of $X_n$, especially when convergence is to a constant.

**Example 5.22**

**a.** Let $\{X_n\}$ be a positive-valued random variable such that $X_n \xrightarrow{p} 3$. Then $Y_n = g(X_n) = \ln(X_n) + (X_n)^{1/2}$ is such that $\text{plim}\, Y_n = \text{plim}\, g(X_n) = g(\text{plim}\, X_n) = \ln(\text{plim}(X_n)) + (\text{plim}\, X_n)^{1/2} = \ln(3) + 3^{1/2} = 2.8307$.

**b.** Let $\{\mathbf{X}_n\}$ be such that $\underset{(k\times 1)}{\mathbf{X}_n} \xrightarrow{p} \underset{(k\times 1)}{\mathbf{X}} \sim N([\mathbf{0}], \mathbf{I})$, and let $\mathbf{Y}_n = g(\mathbf{X}_n) = \mathbf{X}_n'\mathbf{X}_n$. Then $\mathbf{Y}_n \xrightarrow{p} g(\mathbf{X}) = \mathbf{X}'\mathbf{X} \sim \chi^2_k$. □

Through examining a number of special cases of Theorem 5.5, we can establish a catalogue of some useful properties of the plim operator.

**Theorem 5.6** (plim properties) *For conformable* $\mathbf{X}_n$, $\mathbf{Y}_n$, *and constant matrix* $\mathbf{A}$.

**a.** $\operatorname{plim} \mathbf{A}\mathbf{X}_n = \mathbf{A}(\operatorname{plim} \mathbf{X}_n)$;
**b.** $\operatorname{plim} \sum_{i=1}^{m} X_n[i] = \sum_{i=1}^{m} \operatorname{plim} X_n[i]$ (*the plim of a sum = the sum of the plims*);
**c.** $\operatorname{plim} \prod_{i=1}^{m} X_n[i] = \prod_{i=1}^{m} \operatorname{plim} X_n[i]$ (*the plim of a product = the product of the plims*);
**d.** $\operatorname{plim} \mathbf{X}_n\mathbf{Y}_n = (\operatorname{plim} \mathbf{X}_n)(\operatorname{plim} \mathbf{Y}_n)$;
**e.** $\operatorname{plim} \mathbf{X}_n^{-1}\mathbf{Y}_n = (\operatorname{plim} \mathbf{X}_n)^{-1} \operatorname{plim} \mathbf{Y}_n$ ($\operatorname{plim} X_n$ *being nonsingular*).

**Proof** All of the results follow from Theorem 5.5 and from the fact that the functions being analyzed are continuous functions. Note in particular that the matrix inverse function is continuous at all points for which the matrix is nonsingular. ■

**Example 5.23** Let $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$, and let $\{\mathbf{X}_n\}$ be such that $\operatorname{plim} \mathbf{X}_n = \begin{bmatrix} 2 \\ 5 \end{bmatrix}$. Then, $\operatorname{plim} \mathbf{A}\mathbf{X}_n = \mathbf{A} \operatorname{plim} \mathbf{X}_n = \begin{bmatrix} 9 \\ 7 \end{bmatrix}$, $\operatorname{plim}(X_n[1] + X_n[2]) = \operatorname{plim} X_n[1] + \operatorname{plim} X_n[2] = 2 + 5 = 7$, and $\operatorname{plim}(X_n[1] \cdot X_n[2]) = \operatorname{plim} X_n[1] \cdot \operatorname{plim} X_n[2] = 2 \cdot 5 = 10$. □

**Example 5.24** Let $\{\mathbf{Y}_n\}$ be such that $\operatorname{plim} \mathbf{Y}_n = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ and $\{\mathbf{X}_n\}$ be such that $\operatorname{plim} \mathbf{X}_n = \begin{bmatrix} 3 & 1 \\ 2 & 1 \end{bmatrix}$. Then

$$\operatorname{plim} \mathbf{X}_n\mathbf{Y}_n = (\operatorname{plim} \mathbf{X}_n)(\operatorname{plim} \mathbf{Y}_n) = \begin{bmatrix} 3 & 1 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 7 \\ 4 & 5 \end{bmatrix},$$

$$\operatorname{plim} \mathbf{X}_n^{-1}\mathbf{Y}_n = (\operatorname{plim} \mathbf{X}_n)^{-1} \operatorname{plim} \mathbf{Y}_n = \begin{bmatrix} 1 & -1 \\ -2 & 3 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 4 & -1 \end{bmatrix}. \quad \square$$

**Example 5.25** Let $\{\mathbf{X}_n\}$ be such that $\operatorname{plim} \mathbf{X}_n = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$ and $\{\mathbf{Y}_n\}$ be such that $\underset{(2\times1)}{\mathbf{Y}_n} \xrightarrow{p} \underset{(2\times1)}{\mathbf{Y}} \sim \mathbf{N}([\mathbf{0}], \mathbf{I})$. Then

$$\mathbf{X}_n\mathbf{Y}_n \xrightarrow{p} (\operatorname{plim} \mathbf{X}_n)\mathbf{Y} \sim \mathbf{N}\left([\mathbf{0}], \begin{bmatrix} 13 & 14 \\ 14 & 20 \end{bmatrix}\right),$$

$$\mathbf{X}_n^{-1}\mathbf{Y}_n \xrightarrow{p} (\operatorname{plim} \mathbf{X}_n)^{-1}\mathbf{Y} \sim \mathbf{N}\left([\mathbf{0}], \begin{bmatrix} .3125 & -.2188 \\ -.2188 & .2031 \end{bmatrix}\right). \quad \square$$

*Relationships Between Convergence in Probability and in Distribution* There are a number of useful results on convergence in distribution that also incorporate the convergence in probability concept. One such result has use in situations where it is relatively easy to show (or it is known) that $X_n \xrightarrow{d} X$ and $(X_n - Y_n) \xrightarrow{p} [0]$ but where it is more difficult to demonstrate directly that $Y_n \xrightarrow{d} X$.

**Theorem 5.7** *Let* $\{(\mathbf{X}_n, \mathbf{Y}_n)\}$ *be a sequence of pairs of* $(m \times k)$ *random matrices for which* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *and* $(\mathbf{X}_n - \mathbf{Y}_n) \xrightarrow{p} [\mathbf{0}]$. *Then* $\mathbf{Y}_n \xrightarrow{d} \mathbf{X}$.

**Proof** Y. S. Chow and H. Teicher (1978), *Probability Theory*, New York: Springer-Verlag, New York, p. 249. ■

**Example 5.26** Let $\{X_n\}$ be such that $X_n \sim N(0, (n-1)/n)$, so that $X_n \xrightarrow{d} X \sim N(0,1)$. Let $\{Z_n\}$ be such that $Z_n \sim \chi^2_n$, with $X_n$ and $Z_n$ independent, and define $\{Y_n\}$ by $Y_n = (1 + n^{-1})X_n + n^{-1}Z_n - 1$. Then, by Theorem 5.6, $Y_n \xrightarrow{d} X \sim N(0,1)$. To see this, first note that $X_n - Y_n = 1 - n^{-1}(X_n + Z_n)$, so that $\text{plim}(X_n - Y_n) = 1 - \text{plim}(X_n/n) - \text{plim}(Z_n/n)$ (since the plim of a sum is the sum of the plims). Now, note that $\mathrm{E}(X_n/n) = 0 \;\forall\, n$ and $\text{var}(X_n/n) = n^{-3}(n-1) \to 0$ implies $\text{plim}\, X_n/n = 0$ by Chebyshev's inequality. Similarly, since $\mathrm{E}(Z_n/n) = 1 \;\forall\, n$ and $\text{var}(Z_n/n) = (2/n) \to 0$, $\text{plim}(Z_n/n) = 1$ by Chebyshev's inequality. Then $\text{plim}(X_n - Y_n) = 0$, and since $X_n \xrightarrow{d} X \sim N(0,1)$, $Y_n \xrightarrow{d} X \sim N(0,1)$, by Theorem 5.7, with $m = k = 1$. □

The following corollary to Theorem 5.7 indicates that convergence in probability of a sequence of random variables implies convergence in distribution of the sequence.

**Corollary 5.1** $Y_n \xrightarrow{p} Y \Rightarrow Y_n \xrightarrow{d} Y$.

**Proof** This follows immediately from Theorem 5.7 upon defining $X_n = X = Y \;\forall\, n$. ■

Therefore, one way of discovering the limiting distribution of a sequence of scalar or multivariate random variables is to discover the probability limit of the sequence, in which case the limiting distribution is identical to the distribution of the random variable representing the probability limit.

**Example 5.27** Let $\{Y_n\}$ be defined by $Y_n = (2+n^{-1})X+3$, where $X \sim N(1,2)$. Using properties of the plim operator, it follows that $\text{plim}\, Y_n = \text{plim}(2+n^{-1})X + \text{plim}(3) = 2X+3 \sim N(5,8)$. Then, Corollary 5.1 implies that $Y_n \xrightarrow{d} N(5,8)$. □

The converse of Corollary 5.1 is generally not true, as indicated in the discussion immediately following Def. 5.12. However, in the special case where a sequence of random variables converges in distribution to a *constant*, the converse of Corollary 5.1 does hold.

**Theorem 5.8** $Y_n \xrightarrow{d} c \Rightarrow Y_n \xrightarrow{p} c$.

**Proof** Let $\{Y_n\}$ be a sequence of scalar random variables and suppose $Y_n \xrightarrow{d} c$, so that $F_n(y) \to F(y) = I_A(y)$, where $A = \{y\colon y \geq c\}$. Then as $n \to \infty$, $P(|y_n - c| < \varepsilon) \geq F_n(c+\tau) - F_n(c-\tau) \to 1$, for $\tau \in (0,\varepsilon)$ and $\forall\, \varepsilon > 0$, which implies that $Y_n \xrightarrow{p} c$. The multivariate case can be proven similarly using marginal CDFs and the elementwise nature of the plim operator. ■

Thus, in the case of convergence to a constant, the notions of convergence in probability and convergence in distribution are equivalent. Otherwise, convergence in probability is the more stringent type of convergence.

The concepts of convergence in distribution and convergence in probability can be combined to produce a very versatile extension of Theorem 5.3 that is useful for deriving the limiting distribution of a much wider variety of functions of $X_n$. In particular, with reference to a random sequence $\{X_n\}$ for which $X_n \xrightarrow{d} X$, the extension facilitates the discovery of the limiting distribution for a function of $X_n$ that is continuous with probability 1 (with respect to the distribution of $X$) when the function *also* depends on a convergent sequence of numbers or matrices $\{a_n\}$, and/or on a sequence of other random variables $\{Y_n\}$ that converges in probability to a *constant* matrix $\mathbf{y}$.

**Theorem 5.9** *Let* $\{\mathbf{X}_n\}$, $\{\mathbf{Y}_n\}$, *and* $\{\mathbf{a}_n\}$ *be such that* $\underset{(k\times m)}{\mathbf{X}_n} \xrightarrow{d} \underset{(k\times m)}{\mathbf{X}}$, $\underset{(\ell\times q)}{\mathbf{Y}_n} \xrightarrow{p} \underset{(\ell\times q)}{\mathbf{y}}$, *and* $\underset{(j\times p)}{\mathbf{a}_n} \rightarrow \underset{(j\times p)}{\mathbf{a}}$. *Let the set B be such that the probability distribution of* $\mathbf{X}$ *assigns* $P(\mathbf{x} \in B) = 1$, *and let the random variable* $g(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{a}_n)$ *be defined by a (possibly vector) function g that is continuous at every point in* $B \times \mathbf{y} \times \mathbf{a}$. *Then* $g(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{a}_n) \xrightarrow{d} g(\mathbf{X}, \mathbf{y}, \mathbf{a})$.

**Proof** This follows from the proof in V. Fabian and J. Hannan (1985), Introduction to Probability and Mathematical Statistics. New York: John Wiley, p. 159, and from Theorem 5.4. ■

Theorem 5.9 reduces the problem of identifying the limiting distribution of $g(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{a}_n)$ to the problem of identifying the distribution of $g(\mathbf{X}, \mathbf{y}, \mathbf{a})$. Note that either of the arguments $\mathbf{Y}_n$ or $\mathbf{a}_n$ could be a *ghost* in the function $g(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{a}_n)$, meaning that the value of the function $g$ is not affected by $\mathbf{Y}_n$ or $\mathbf{a}_n$ (in which case $\mathbf{Y}_n$ or $\mathbf{a}_n$ could be completely ignored in Theorem 5.9).

**Example 5.28** Let $\{\mathbf{Z}_n\}$ be a sequence of bivariate random variables such that $\mathbf{Z}_n \xrightarrow{d} \mathbf{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = [2\ 3]'$ and $\boldsymbol{\Sigma} = \left[\begin{smallmatrix} 4 & 0 \\ 0 & 9 \end{smallmatrix}\right]$. Let $\{\boldsymbol{\mu}_n\}$ be any sequence of $(2 \times 1)$ vectors such that $\boldsymbol{\mu}_n \rightarrow \boldsymbol{\mu}$ (e.g., $\boldsymbol{\mu}_n = [2 + n^{-1} \mid 3/(1 + \exp(-n))]'$ and let $\boldsymbol{\Sigma}_n$ be any sequence of nonsingular $(2 \times 2)$ matrices such that $\boldsymbol{\Sigma}_n \xrightarrow{p} \boldsymbol{\Sigma}$, (for example,

$$\boldsymbol{\Sigma}_n = \begin{bmatrix} 4+n^{-1} & n^{-1} \\ n^{-1} & 9+n^{-1} \end{bmatrix}.$$

Then it follows from Theorem 5.9 that $g(\mathbf{Z}_n, \boldsymbol{\Sigma}_n, \boldsymbol{\mu}_n) = (\mathbf{Z}_n - \boldsymbol{\mu}_n)' \boldsymbol{\Sigma}_n^{-1} (\mathbf{Z}_n - \boldsymbol{\mu}_n) \xrightarrow{d} (\mathbf{Z} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{Z} - \boldsymbol{\mu}) = g(\mathbf{Z}, \boldsymbol{\Sigma}, \boldsymbol{\mu}) \sim \chi_2^2$ (based on the fact that two independent, standard, normal random variables, represented by the $(2 \times 1)$ vector $\boldsymbol{\Sigma}^{-1/2}(\mathbf{Z} - \boldsymbol{\mu})$, are squared and summed). Compare the generality of this result to the result found in Ex. 5.18. □

**Example 5.29** Let $X_n \xrightarrow{d} X \sim \theta^{-1}\exp(-x/\theta)I_{(0,\infty)}(x)$ (i.e., exponential with parameter $\theta$), $Y_n \xrightarrow{p} 4$, and $a_n \rightarrow 2$. Then $Z_n = g(X_n, Y_n, a_n) = (Y_n)^2 X_n/a_n \xrightarrow{d} 4^2 X/2 = 8X = g(X, 4, 2) \sim$ Gamma $(1, 8\theta)$. □

A number of special cases of Theorem 5.9 provide a number of useful results that are collectively referred to as **Slutsky's theorems**. The results refer to convergence in distribution of sums and products of random matrices.

**Theorem 5.10** (Slutsky's Theorems) *Let* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *and* $\mathbf{Y}_n \xrightarrow{p} \mathbf{c}$. *Then, for conformable* $\mathbf{X}_n$ *and* $\mathbf{Y}_n$,

**a.** $\mathbf{X}_n + \mathbf{Y}_n \xrightarrow{d} \mathbf{X} + \mathbf{c}$,

**b.** $\mathbf{Y}_n\mathbf{X}_n \xrightarrow{d} \mathbf{c}\mathbf{X}$,

**c.** $\mathbf{Y}_n^{-1}\mathbf{X}_n \xrightarrow{d} \mathbf{c}^{-1}\mathbf{X}$ (*if* $\mathbf{c}^{-1}$ *exists*).

**Proof** Each function on the left-hand side of (a), (b), and (c) is of the form $g(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{a}_n)$ and satisfies the conditions of Theorem 5.9, with $\mathbf{a}_n$ being a ghost in the definition of the function $g$. ■

*Order of Magnitude in Probability* It is sometimes useful to be able to characterize or compare random sequences and/or terms in a random sequence relative to their *order of magnitude* in addition to, or in lieu of, any examination of convergence in probability of the sequences involved. The order-of-magnitude concept is especially useful in sorting out which random terms in the definition of a random sequence make dominant contributions to the magnitude of the sequence outcome and which terms are irrelevant as $n$ increases. The order of magnitude of a random sequence, in terms of probability, is described in the following definition. The concept can be thought of as a random or probabilistic counterpart to the order-of-magnitude concept in real analysis (recall Def. 5.8).

**Definition 5.13**
***Order of magnitude in probability***

Let $\{X_n\}$ be a sequence of random scalars, and let $\{\mathbf{W}_n\}$ be a real-valued, random matrix sequence.

**a.** $O_p(n^k)$: The sequence $\{X_n\}$ is said to be **at most of order $n^k$ in probability**, denoted by $O_p(n^k)$, iff for every $\varepsilon > 0$ there exists a corresponding positive constant $c(\varepsilon) < \infty$ such that $P(n^{-k}|X_n| \leq c(\varepsilon)) \geq 1 - \varepsilon, \forall\, n$.

**b.** $o_p(n^k)$: The sequence $\{X_n\}$ is said to be **of order smaller than $n^k$ in probability**, denoted by $o_p(n^k)$, iff $n^{-k}X_n \xrightarrow{p} 0$.

**c.** If $\{W_n[i,j]\}$ is $O_p(n^k)$ (or $o_p(n^k)$) $\forall\, i$ and $j$, then the random matrix sequence $\{\mathbf{W}_n\}$ is said to be $O_p(n^k)$ (or $o_p(n^k)$).

Given Def. 5.13, a random sequence of scalars, $\{X_n\}$, is $O_p(n^k)$ iff one can always find a finite interval within which the outcomes of $n^{-k}X_n$ will occur with probability arbitrarily close to (but not necessarily equal to) 1 for each term in the sequence. The random sequence $\{X_n\}$ is $o_p(n^k)$ iff $n^{-k}X_n$ converges in probability to zero. If a random sequence is $O_p(1) \equiv O_p(n^o)$, the random sequence is said to be **bounded in probability**. Note that the ranges of the random variables in a sequence need *not* be finite for the sequence to be $O_p(n^k)$, or even $O_p(1)$, as the next example illustrates.

**Example 5.30** Let $\{X_n\}$ be such that $X_i \sim N(0, 1)$, $\forall\, i$, with all terms in the sequence being independent random variables. Define $\{Z_n\}$ as $Z_n = \sum_{i=1}^{n} X_i$. Then $\{X_n\}$ itself is $O_p(1)$ and $\{Z_n\}$ is $O_p(n^{1/2})$. To see this, first note that since $X_i \sim N(0, 1)$, there always exists a constant $c(\varepsilon) > 0$ large enough such that $\int_{-c(\varepsilon)}^{c(\varepsilon)} N(x; 0, 1)dx \geq 1-\varepsilon$ for any choice of $\varepsilon > 0$, so that $\{X_n\}$ is $O_p(1)$ (i.e., $\{X_n\}$ is bounded in probability). Now, note that $n^{-1/2}Z_n = n^{-1/2}\sum_{i=1}^{n} X_i \sim N(0, 1)$, whence it follows from the preceding argument that $\{Z_n\}$ is $O_p(n^{1/2})$. It also follows that $\{X_n\}$ is $o_p(n^{\delta})$ and $\{Z_n\}$ is $o_p(n^{1/2+\delta})\ \forall\, \delta > 0$. Finally, in the sequence defined by $Y_n = n^{-1/2}(X_n + Z_n)$, note that $n^{-1/2}X_n$ is $o_p(1)$, while $n^{-1/2}Z_n$ is $O_p(1)$, implying that as $n \to \infty$, $n^{-1/2}Z_n$ is the dominant random term in the definition of $Y_n$ while $n^{-1/2}X_n$ is stochastically irrelevant as $n \to \infty$. □

Rules for determining the order of magnitude of the sum or product of sequences of random variables are analogous to the rules introduced for the case of sequences of real numbers or matrices and will not be repeated here. In particular, *Lemmas 5.2 and 5.3 apply with* O *and* o *changed to* $O_p$ *and* $o_p$, *respectively.* The reader should also note the following relationship between O (or o) and $O_p$ (or $o_p$) for sequences of real numbers or matrices.

**Theorem 5.11** *Let $\{a_n\}$ be a sequence of real numbers or matrices. Then if $\{a_n\}$ is $O(n^k)$ (or $o(n^k)$) then $\{a_n\}$ is $O_p(n^k)$ (or $o_p(n^k)$).*

**Proof** This follows directly from definitions upon interpreting the sequence of real numbers or matrices as a sequence of degenerate random variables or random matrices. Details are left to the reader. ■

**Example 5.31** Let $\{x_n\}$ be $O(n^1)$, $\{Y_n\}$ be $O_p(n^2)$, and $\{Z_n\}$ be $O_p(n^1)$. Then $\{x_nY_n\}$ is $O_p(n^3)$, $\{x_n + Y_n + Z_n\}$ is $O_p(n^2)$, $\{n^{-1}x_n\}$ is $O_p(1)$, $\{n^{-2}x_n(Y_n + Z_n)\}$ is $O_p(n^1)$, $\{x_nY_nZ_n\}$ is $O_p(n^4)$, and $\{n^{-5}x_nY_nZ_n\}$ is $o_p(1)$. □

### Convergence in Mean Square (or Convergence in Quadratic Mean)

Referring to a random sequence of *scalars* $\{Y_n\}$ the concept of convergence in mean square involves the question of whether, for some random variable $Y$, the sequence of expectations $\{E(Y_n - Y)^2\}$ converges to zero as $n \to \infty$. Since $E(Y_n - Y)^2$ can be interpreted as the expected squared distance between outcomes of $Y_n$ and outcomes of $Y$, convergence in mean square implies that outcomes of $Y_n$ and $Y$ are "close" to one another for large enough $n$, and arbitrarily close when $n \to \infty$. When $\mathbf{Y}_n$ and $\mathbf{Y}$ are $m \times k$ matrices, the question of convergence in mean square concerns whether the preceding convergence of expectations to zero occurs elementwise between corresponding entries in $\mathbf{Y}_n$ and $\mathbf{Y}$.

**Definition 5.14** ***Convergence in mean square (or convergence in quadratic mean)***

The sequence of random variables $\{Y_n\}$ **converges in mean square** to the random variable, $Y$, iff

**a.** Scalar case: $\lim_{n\to\infty} \mathrm{E}(Y_n - Y)^2 = 0$,

**b.** Matrix case: $\lim_{n\to\infty} \mathrm{E}(Y_n[i,j] - Y[i,j])^2 = 0$, $\forall$ $i$ and $j$.

Convergence in mean square will be denoted by $Y_n \xrightarrow{m} Y$.

As in the case of convergence in probability, convergence in mean square imposes restrictions on the characteristics of the joint probability distribution of $(Y_n, Y)$ as $n \to \infty$. In particular, first- and second-order moments of corresponding entries in $Y_n$ and $Y$ converge to one another as indicated in the necessary and sufficient conditions for mean square convergence presented in the following theorem.

**Theorem 5.12** $Y_n \xrightarrow{m} Y$ *iff* $\forall$ *i and j*:

**a.** $\mathrm{E}Y_n[i,j] \to \mathrm{E}Y[i,j]$,
**b.** $\mathrm{var}(Y_n[i,j]) \to \mathrm{var}(Y[i,j])$,
**c.** $\mathrm{cov}(Y_n[i,j], Y[i,j]) \to \mathrm{var}(Y[i,j])$.

**Proof** We provide a proof for the scalar case, which suffices to prove the matrix case given the elementwise definition of mean-square convergence (Def. 5.14.b).
*Necessity*:

**a.** $\mathrm{E}Y_n \to \mathrm{E}Y$ follows from the fact that

$$|\mathrm{E}Y_n - \mathrm{E}Y| = |\mathrm{E}(Y_n - Y)| \le \mathrm{E}|Y_n - Y| \le [\mathrm{E}|Y_n - Y|^2]^{1/2} \to 0.$$

To see this, note that the first inequality follows from the fact that $(y_n - y) \le |y_n - y|$. Regarding the second inequality, note that $g(z) = z^2$ is a convex function on $R$, and letting $Z = |Y_n - Y|$, Jensen's inequality implies $(\mathrm{E}|Y_n - Y|)^2 \le \mathrm{E}|Y_n - Y|^2$ (recall $g(\mathrm{E}Z) \le \mathrm{E}g(Z)$). Convergence to zero occurs because $\mathrm{E}|Y_n - Y|^2 = \mathrm{E}(Y_n - Y)^2 \to 0$ by convergence in mean square.

**b.** $\mathrm{E}Y_n^2 = \mathrm{E}(Y_n - Y)^2 + \mathrm{E}Y^2 + 2\mathrm{E}[Y(Y_n - Y)]$, and since by the Cauchy-Schwartz inequality, $|\mathrm{E}[Y(Y_n - Y)]| \le [\mathrm{E}Y^2\mathrm{E}(Y_n - Y)^2]^{1/2}$, it follows that

$$\mathrm{E}(Y_n - Y)^2 + \mathrm{E}Y^2 - 2\left[\mathrm{E}Y^2\mathrm{E}(Y_n - Y)^2\right]^{1/2} \le \mathrm{E}Y_n^2 \le \mathrm{E}(Y_n - Y)^2 + \mathrm{E}Y^2 + 2\left[\mathrm{E}Y^2\mathrm{E}(Y_n - Y)^2\right]^{1/2}.$$

Then, since $\mathrm{E}(Y_n - Y)^2 \to 0$ by mean-square convergence, $\mathrm{E}Y_n^2 \to \mathrm{E}Y^2$. It follows that $\mathrm{var}(Y_n) \to \mathrm{var}(Y)$, since $\mathrm{var}(Y_n) = \mathrm{E}Y_n^2 - (\mathrm{E}Y_n)^2 \to \mathrm{E}Y^2 - (\mathrm{E}Y)^2 = \mathrm{var}(Y)$.

**c.** First note that

$$\begin{aligned}\mathrm{E}(Y_n - Y)^2 &= \mathrm{E}Y_n^2 - 2\mathrm{E}Y_nY + \mathrm{E}Y^2\\ &= \mathrm{var}(Y_n) + (\mathrm{E}Y_n)^2 - 2\,[\mathrm{cov}(Y_n, Y) + \mathrm{E}Y_n\mathrm{E}Y] + \mathrm{var}(Y) + (\mathrm{E}Y)^2.\end{aligned}$$

If $\mathrm{E}(Y_n - Y)^2 \to 0$, then by (a) and (b), the preceding equality implies $\mathrm{cov}(Y_n, Y) \to \mathrm{var}(Y)$.

*Sufficiency*:

From the expression defining $\mathrm{E}(Y_n - Y)^2$ in the preceding proof of the necessity of (c), it follows directly that (a), (b), and (c) $\Rightarrow \mathrm{E}(Y_n - Y)^2 \to 0$, which implies $Y_n \xrightarrow{m} Y$. ■

The conditions in Theorem 5.12 simplify when **Y** is a degenerate random variable equal to the constant matrix, **c**.

**Corollary 5.2** $\mathbf{Y}_n \xrightarrow{m} \mathbf{c}$ *iff* $\mathrm{E}Y_n[i,j] \to c[i,j]$ *and* $\mathrm{var}\,(Y_n[i,j]) \to 0\ \forall\ i$ *and* $j$.

**Proof** This follows directly from Theorem 5.12 upon letting $\mathbf{Y} = \mathbf{c}$, and noting that $\mathrm{var}(c[i,j]) = [0]$ and $\mathrm{cov}(Y_n[i,j], c[i,j]) = [0]$. ■

**Example 5.32** Let $\mathrm{E}\mathbf{Y}_n = \begin{bmatrix}2+3n^{-1}\\ 1+n^{-1}\end{bmatrix}$ and $\mathbf{Cov}(\mathbf{Y}_n) = n^{-2}\begin{bmatrix}2 & 1\\ 1 & 1\end{bmatrix}$. Since $\mathrm{E}\mathbf{Y}_n \to \begin{bmatrix}2\\1\end{bmatrix}$ and $\mathrm{Diag}(\mathbf{Cov}(\mathbf{Y}_n)) = \begin{bmatrix}2/n^2\\ 1/n^2\end{bmatrix} \to \begin{bmatrix}0\\0\end{bmatrix}$, it follows by Corollary 5.2 that $\mathbf{Y}_n \xrightarrow{m} \begin{bmatrix}2\\1\end{bmatrix}$, which in turn implies by Theorem 5.12 that $\mathbf{Y}_n \xrightarrow{p} \begin{bmatrix}2\\1\end{bmatrix}$ and $\mathbf{Y}_n \xrightarrow{d} \begin{bmatrix}2\\1\end{bmatrix}$. □

In addition to the convergence of second-order moments, convergence in mean square implies that the correlation between corresponding entries in $\mathbf{Y}_n$ and **Y** converge to 1 if the entries in **Y** have nonzero variances.

**Corollary 5.3** $Y_n \xrightarrow{m} Y \Rightarrow \mathrm{corr}\,(Y_n[i,j], Y[i,j]) \to 1$ *when* $\mathrm{var}\,(Y[i,j]) > 0,\ \forall\ i,j$.

**Proof** This follows directly from Theorem 5.12 since in the scalar case,

$$\mathrm{corr}(Y_n, Y) = \frac{\mathrm{cov}(Y_n, Y)}{[\mathrm{var}(Y_n)\,\mathrm{var}(Y)]^{1/2}} \to \frac{\mathrm{var}(Y)}{\mathrm{var}(Y)} = 1.$$

The matrix case follows by applying the preceding result elementwise. ■

The corollary indicates that if $\mathbf{Y}_n \xrightarrow{m} \mathbf{Y}$, and if none of the entries in **Y** is degenerate, then corresponding entries in $\mathbf{Y}_n$ and **Y** tend to be *perfectly positively correlated* as $n \to \infty$. It follows that as $n \to \infty$, the outcomes of $Y_n[i,j]$ and $Y[i,j]$ tend to exhibit properties that characterize a situation in which two random variables have perfect positive correlation and *equal* variances, including the fact that $P(|y_n[i,j] - y[i,j]| < \varepsilon) \to 1\ \forall\ \varepsilon > 0$ (recall the discussion of correlation in Chapter 3). If $Y[i,j]$ were degenerate, then although the correlation is undefined, the preceding probability convergence result still holds since, in this case, $Y_n[i,j]$ tends toward a degenerate random variable $(\mathrm{var}(Y_n[i,j]) \to 0)$ having an outcome equal to the appropriate scalar value of $y[i, j]$. Thus, if $Y_n \xrightarrow{m} Y$, then outcomes of $Y_n$ emulate outcomes of $Y$ ever more closely as $n \to \infty$, closeness being measured in terms of probability as well as expected squared distance.

From the preceding discussion, the reader will have anticipated the next result.

**Theorem 5.13** $Y_n \xrightarrow{m} Y \Rightarrow Y_n \xrightarrow{p} Y \Rightarrow Y_n \xrightarrow{d} Y$.

**Proof** (scalar case—matrix case by elementwise extension) Note that $(Y_n - Y)^2$ is a nonnegative-valued random variable, and letting $a = \varepsilon^2 > 0$, we have by Markov's inequality that

$$P((y_n - y)^2 \geq \varepsilon^2) \leq \mathrm{E}(Y_n - Y)^2/\varepsilon^2.$$

Thus, $P(|y_n - y| \geq \varepsilon) \leq \mathrm{E}(Y_n - Y)^2/\varepsilon^2$, or $P(|y_n - y| < \varepsilon) \geq 1 - \mathrm{E}(Y_n - Y)^2/\varepsilon^2$. By mean square convergence, $\mathrm{E}(Y_n - Y)^2 \to 0$, so that

$$\lim_{n\to\infty} P(|y_n - y| < \varepsilon) = 1 \quad \forall\, \varepsilon > 0.$$

Thus, $\operatorname{plim} Y_n = Y$. Convergence in distribution follows from Corollary 5.1. ■

Theorem 5.13 indicates that convergence in mean square is a *sufficient* condition for both convergence in probability and convergence in distribution. The result of the theorem can be quite useful as a tool for establishing the latter two types of convergence in cases where convergence in mean square is relatively easy to demonstrate. Furthermore, the theorem allows convergence in probability or in distribution to be demonstrated, even in cases where the distributions of $Y_n$ and/or $Y$ are not fully known, so long as the appropriate convergence properties of the relevant sequences of expectations can be established.

**Example 5.33** Let $Y \sim N(0, 1)$, $\mathrm{E}Y_n = 0\ \forall\ n$, $\mathrm{var}(Y_n) \to 1$, and $\mathrm{cov}(Y_n, Y) \to 1$. Then, since (recall the proof of Theorem 5.12.c) $\mathrm{E}(Y_n - Y)^2 = \mathrm{var}(Y_n) + \mathrm{var}(Y) - 2\,\mathrm{cov}(Y_n, Y) + (\mathrm{E}Y_n - \mathrm{E}Y)^2 \to 0$, it follows that $Y_n \xrightarrow{m} Y$, which implies that $Y_n \xrightarrow{p} Y$, and $Y_n \xrightarrow{d} N(0, 1)$. Note that while we did not know the forms of the probability distributions associated with the random variables in the sequence $\{Y_n\}$, we can nonetheless establish convergence in probability and in distribution via convergence in mean square. □

The following example demonstrates that convergence in mean square is *not* a necessary condition for convergence in probability (or in distribution), so that convergence in probability (or in distribution) does *not* imply convergence in mean square.

**Example 5.34** $Y_n \xrightarrow{p} Y$ and/or $Y_n \xrightarrow{d} Y \not\Rightarrow Y_n \xrightarrow{m} Y$

Let $\{Y_n\}$ be such that $P(y_n = 0) = 1 - n^{-2}$ and $P(y_n = n) = n^{-2}$. Then

$$\lim_{n\to\infty} P(y_n = 0) = 1,$$

so that plim $Y_n = 0$ and $Y_n \xrightarrow{d} 0$. However, $\mathrm{E}(Y_n - 0)^2 = 0(1 - n^{-2}) + n^2(n^{-2}) = 1$ $\forall\, n$, so that $Y_n \overset{m}{\not\to} 0$. □

As a further illustration of differences between convergence in mean square and convergence in probability or in distribution, we now provide an example showing that convergence in probability and convergence in distribution do *not* imply convergence of the first- and second-order moments of $Y_n$ and $Y$.

**Example 5.35** $Y_n \xrightarrow{p} Y$ and/or $Y_n \xrightarrow{d} Y \not\Rightarrow \mathrm{E}Y_n^r \to \mathrm{E}Y^r$

Let $\{Y_n\}$ be such that $Y_n \sim f_n(y) = (1-n^{-1})I_{\{0\}}(y) + n^{-1}I_{\{n\}}(y)$, and note that $f_n(y) \to f(y) = I_{\{0\}}(y)\, \forall\, y$, and thus $Y_n \xrightarrow{d} Y \sim f(y)$. Furthermore, since $\lim_{n\to\infty} P(|y_n - 0| < \varepsilon) = 1\ \forall\, \varepsilon > 0$, then $Y_n \xrightarrow{p} 0$. Now, note that $\mathrm{E}Y_n = 1\ \forall\, n$ and $\mathrm{E}Y_n^2 = n\ \forall\, n$, but $\mathrm{E}Y = 0$ and $\mathrm{E}Y^2 = 0$ when $Y \sim f(y)$. Thus, *neither* $Y_n \xrightarrow{p} Y$ *nor* $Y_n \xrightarrow{d} Y$ implies that $\mathrm{E}Y_n \to \mathrm{E}Y$ or $\mathrm{E}Y_n^2 \to \mathrm{E}Y^2$. □

In summary, convergence in mean square is a *more stringent* type of convergence than either convergence in probability or convergence in distribution. In addition to implying the latter two types of convergence, convergence in mean square *also* implies convergence of first- and second-order moments about the origin and mean and convergence to 1 of the correlation between corresponding entries in $Y_n$ and $Y$ (when var$(Y) > 0$).

### *Almost-Sure Convergence (or Convergence with Probability 1)

Referring to a sequence of random variables $\{Y_n\}$ which could be scalars or matrices, the concept of *almost-sure convergence* involves the question of whether, for some random variable Y, the outcome of the random *sequence* converges to the outcome of Y *with probability 1*. That is, does $P(y_n \to y) = P(\lim_{n\to\infty} y_n = y) = 1$?

**Definition 5.15** ***Almost sure convergence (or convergence with probability 1***

The sequence of random variables $\{Y_n\}$ **converges almost surely** to the random variable Y iff

**a.** Scalar case: $P(y_n \to y) = P(\lim_{n\to\infty} y_n = y) = 1$,

**b.** Matrix case: $P(y_n[i,j] \to y[i,j]) = P(\lim_{n\to\infty} y_n[i,j] = y[i,j]) = 1,\ \forall\, i$ and $j$.

Almost-sure convergence will be denoted by $Y_n \xrightarrow{as} Y$, or by **aslim** $Y_n = Y$, the latter notation meaning *the almost-sure limit of* $Y_n$ *is* $Y$.

Almost-sure convergence is the stochastic counterpart to the real-analysis concept of the limit of a sequence. Almost-sure convergence accommodates the fact that when dealing with sequences of nondegenerate random variables, more than one sequence outcome (perhaps an infinite number) is under consideration for convergence. If $Y_n \xrightarrow{as} Y$, then a real-analysis–type limit is (essentially) certain to be achieved by outcomes of the sequence $\{Y_n\}$, the limit being represented by the outcome of $Y$. If $Y$ is degenerate and equal to the constant $c$, then $Y_n \xrightarrow{as} c$ implies that the outcomes of $\{Y_n\}$ are (essentially) certain to converge to the value $c$.

Note that almost-sure convergence is defined in terms of an event involving an *infinite* collection of random variables contained in the sequence $\{Y_n\}$ and in $Y$, the event being that the sequence of outcomes $y_1, y_2, y_3, \ldots$ has a limit that equals the outcome $y$. This is notably different than in the case of either convergence in probability or convergence in mean square, which both relate to sequences of *marginal* probability distributions and outcomes of *bivariate* random variables $(Y_n, Y)$ for $n = 1, 2, 3, \ldots$. Through counterexamples, we will see that neither of the latter two types of convergence implies that $\lim_{n\to\infty} y_n = y$ occurs with probability 1.

In order to clarify the additional restrictions imposed on the sequence $\{Y_n\}$ by almost-sure convergence relative to convergence in probability, we provide an intuitive description of why $Y_n \xrightarrow{p} c$ does *not* imply that $\lim_{n\to\infty} Y_n$ exists with probability 1. Note that for a sequence of outcomes of $\{Y_n\}$ to have a limit, $c$, it must be the case that, $\forall\ \varepsilon > 0$, there exists an integer $N(\varepsilon)$ such that for $\forall\ n \geq N(\varepsilon)$, $|y_n - c| < \varepsilon$ (recall Def. 5.2). The definition of almost-sure convergence ensures that the outcomes $\{y_n\}$ are generated in such a way that the limit is achieved *with probability 1*, and so the preceding restriction on the outcomes of $\{Y_n\}$ will be met with probability 1 if $Y_n \xrightarrow{as} c$. However, if $Y_n \xrightarrow{p} c$, so that $P(|y_n - c| < \varepsilon) \to 1, \forall\ \varepsilon > 0$, then all that can be said is that, for any fixed large value of $n$, the probability that $y_n$ is close to $c$ is high, but *not* necessarily equal to 1. Since $Y_n \xrightarrow{p} c$ does not even imply that $P(|y_n - c| < \varepsilon) = 1$ for *any* fixed value of $n$, it certainly does not imply that there exists an $N(\varepsilon)$ for which $P(|y_n - c| < \varepsilon, \forall\ n \geq N(\varepsilon)) = 1$. Thus, the existence of a limit for $\{y_n\}$ with probability 1 is not implied by $Y_n \xrightarrow{p} c$.

An alternative and equivalent characterization of almost-sure convergence that follows directly from the fundamental definition of a limit (Def. 5.2) is presented in the next theorem. The alternative characterization facilitates both comparisons with convergence in probability and proofs of other results involving the concept of almost-sure convergence.

**Theorem 5.14** (Alternative Characterization of Almost-Sure Convergence (Scalar Case))
$P(\lim_{n\to\infty} y_n = y) = 1 \Leftrightarrow \lim_{n\to\infty} P(|y_i - y| < \varepsilon, i \geq n) = 1, \forall\ \varepsilon > 0$.

**Proof** *Necessity*: $\lim_{n\to\infty} y_n = y$ *implies* that, for every $\varepsilon > 0$, there exists an integer $N(\varepsilon)$ such that $|y_i - y| < \varepsilon\ \forall\ i \geq N(\varepsilon)$ (recall Def. 5.2). It follows that $P(\lim_{n\to\infty} y_n = y) \leq P(|y_i - y| < \varepsilon, i \geq N(\varepsilon))\ \forall\ \varepsilon > 0$. If $Y_n \xrightarrow{as} Y$, then the left-hand side of the preceding inequality is 1, which implies that the right-hand side is also 1. It follows that, for $\forall\ \varepsilon > 0$, $\lim_{n\to\infty} P(|y_i - y| < \varepsilon, i \geq n) = 1$ since the values of $P(|y_i - y)| < \varepsilon, i \geq n)$ must all be ones for $n$ large enough if $\lim_{n\to\infty} y_n = y$.

*Sufficiency*: R. Serfling, Approximation Theorems, pp. 6–7. ■

The alternative characterization is extended to cases where $\mathbf{Y}_n$ and $\mathbf{Y}$ are matrices, by applying the characterization elementwise to the respective entries in $\mathbf{Y}_n$ and $\mathbf{Y}$. Theorem 5.14 leads to a relatively straightforward demonstration that almost-sure convergence *implies* convergence in probability.

**Theorem 5.15** $Y_n \xrightarrow{as} Y \Rightarrow Y_n \xrightarrow{p} Y$.

**Proof** Suppose $Y_n \xrightarrow{as} Y$. Then $\forall\, \varepsilon > 0$, $\lim_{n\to\infty} P(|y_i - y| < \varepsilon, i \geq n) = 1$ by Theorem 5.14. Since $|y_i - y| < \varepsilon, i \geq n \Rightarrow |y_n - y| < \varepsilon$, it follows that

$$P(|y_i - y| < \varepsilon, i \geq n) \leq P(|y_n - y| < \varepsilon\}).$$

Then since the left-hand side has a limiting value of 1, $\forall\, \varepsilon > 0$, by almost-sure convergence, it follows that the right-hand side has a limiting value of 1, $\forall\, \varepsilon > 0$, implying convergence in probability. ■

The converse of Theorem 5.15 is not true, that is, convergence in probability does *not* imply almost-sure convergence. Furthermore, convergence in mean square and convergence in distribution also do *not* imply almost-sure convergence. These facts are demonstrated by the following counterexample.

**Example 5.36** $Y_n \xrightarrow{m} Y$ and/or $Y_n \xrightarrow{p} Y$ and/or $Y_n \xrightarrow{d} Y \not\Rightarrow Y_n \xrightarrow{as} Y$

Let $\{Y_n\}$ be a sequence of independent random variables such that $Y_n \sim f_n(y) = (1 - n^{-1})I_{\{0\}}(y) + n^{-1}I_{\{1\}}(y)$. Since $f_n(y) \to f(y) = I_{\{0\}}(y)$, $\{Y_n\}$ converges in distribution to the *constant* 0, and thus it is also true that $Y_n \xrightarrow{p} 0$. Furthermore, since $\mathrm{E}Y_n = n^{-1} \to 0$ and $\mathrm{var}(Y_n) = n^{-1} - n^{-2} \to 0$, it follows that $Y_n \xrightarrow{m} 0$. However, it does *not* follow that $Y_n \xrightarrow{as} 0$. To see this, note that $\forall\, \varepsilon \in (0, 1)$ and $\forall$ integer $s > n$,

$$\begin{aligned} P(|y_i| < \varepsilon, n \leq i \leq s) &= \prod_{i=n}^{s}(1 - i^{-1}) = \prod_{i=n}^{s}\frac{i-1}{i} \\ &= \frac{n-1}{n}\cdot\frac{n}{n+1}\cdot\frac{n+1}{n+2}\cdot\ldots\cdot\frac{s-1}{s} \\ &= \frac{n-1}{s} \to 0 \quad \text{as } s \to \infty. \end{aligned}$$

Therefore,

$$\lim_{n\to\infty} P(|y_i| < \varepsilon, i \geq n) = 0$$

since the probability value equals 0 $\forall\, n \geq 1$, and so the limit of the sequence of probability values equals 0 and *not 1*. Thus, $Y_n \overset{as}{\not\to} 0$. □

It is also true that almost-sure convergence does *not* imply convergence in mean square, as the following counterexample demonstrates.

**Example 5.37** $Y_n \xrightarrow{\text{as}} Y \not\Rightarrow Y_n \xrightarrow{\text{m}} Y$

Let $\{Y_n\}$ be a sequence of independent random variables such that $Y_n \sim f_n(y) = (1 - n^{-2})I_{\{0\}}(y) + n^{-2}I_{\{n\}}(y)$. Note that $\forall\, \varepsilon \in (0, n)$ and $\forall$ integer $s > n$,

$$P(|y_i| < \varepsilon, n \leq i \leq s) = \prod_{i=n}^{s} (1 - i^{-2}) = \prod_{i=n}^{s} \frac{i^2 - 1}{i^2}$$

$$= \frac{(n-1)(s+1)}{ns},$$

so that

$$P(|y_i| < \varepsilon, i \geq n) = \frac{n-1}{n}.$$

Then, $\forall\, \varepsilon > 0$,

$$\lim_{n \to \infty} P(|y_i| < \varepsilon, i \geq n) = 1,$$

so that $Y_n \xrightarrow{\text{as}} 0$. Now, note that $\mathrm{E}Y_n = n^{-1} \to 0$, but $\mathrm{var}(Y_n) = 1 - n^{-2} \not\to 0$. Therefore, $Y_n \overset{\text{m}}{\not\to} 0$ by Corollary 5.2. □

Similar to the case of convergence in probability (Theorem 5.4), it is useful to note that if a real-valued sequence of numbers, or matrices, converges (in the sense of real analysis) to a given number, or matrix, i.e., $a_n \to a$, then it can also be stated that $a_n \xrightarrow{\text{as}} a$.

**Theorem 5.16** *Let $\{\mathbf{a}_n\}$ be such that $\mathbf{a}_n \to \mathbf{a}$. Then $\mathbf{a}_n \xrightarrow{\text{as}} \mathbf{a}$.*

**Proof** The proof is immediate from the definition of almost sure convergence, since if $a_n[i, j] \to a[i, j]\ \forall\, i$ and $j$, then $P(\lim_{n\to\infty} a_n = a) = 1$, which implies that $a_n \xrightarrow{\text{as}} a$. ■

Similar to the cases of convergence in probability and convergence in distribution, a useful result for establishing almost-sure convergence of *continuous* functions of $X_n$, when $X_n \xrightarrow{\text{as}} X$, can be stated as follows (recall Theorems 5.3 and 5.5).

**Theorem 5.17** *Let $X_n \xrightarrow{\text{as}} X$, and let the random variable $g(X)$ be defined by a function $g(x)$ that is continuous, except perhaps on a set of points assigned probability zero by the probability distribution of $X$. Then $g(X_n) \xrightarrow{\text{as}} g(X)$ or, equivalently, $\operatorname{aslim} g(X_n) = g(\operatorname{aslim} X_n)$.*

**Proof** Assume that $X_n \xrightarrow{\text{as}} X$. For outcomes of $\{X_n\}$ for which $x_n \to x$ *and* $g$ is continuous at $x$, it *must* be the case that $g(x_n) \to g(x)$ (recall Def. 5.6). Then defining the sets

$$A = \{(\{x_n\}, x) \colon x_n \to x\} \quad \text{and} \quad B = \{(\{x_n\}, x) \colon g \text{ is continuous at } x\},$$

it follows that

$$P(g(x_n) \rightarrow g(x)) \geq P(A \cap B) \geq 1 - P(\bar{A}) - P(\bar{B}) = 1$$

since $P(\bar{A}) = 0$ by almost-sure convergence of $\{X_n\}$ to $X$, and $P(\bar{B}) = 0$ since $g$ is continuous with probability 1. Thus $g(X_n) \xrightarrow{\text{as}} g(X)$. ■

Note that all of the properties of the plim operator listed in Theorem 5.6 apply equally well to the aslim operator since they can all be justified as special cases of Theorem 5.17.

**Example 5.38** Let $\mathbf{X}_n \xrightarrow{\text{as}} \begin{bmatrix}2\\1\end{bmatrix}$. Then $g_1(\mathbf{X}_n) = X_n[2]/X_n[1] \xrightarrow{\text{as}} 1/2$, $g_2(\mathbf{X}_n) = X_n[2] - X_n[1] \xrightarrow{\text{as}} -1$, and $g_3(\mathbf{X}_n) = g_2(\mathbf{X}_n)g_1(\mathbf{X}_n) \xrightarrow{\text{as}} -1/2$. □

**Example 5.39** Let $\underset{(2\times 1)}{\mathbf{X}_n} \xrightarrow{\text{as}} \underset{(2\times 1)}{\mathbf{X}}$, where $X[1] = 3$ and $X[2] \sim N(1, 2)$. Then $g(\mathbf{X}_n) = X_n[1](1 + X_n[2]) \xrightarrow{\text{as}} 3(1 + X[2]) \sim N(6, 18)$. □

The final result we present in this subsection provides a necessary and sufficient condition for almost-sure convergence *to occur*. The criterion will be useful in proving strong laws of large numbers, which we will examine shortly.

**Theorem 5.18** (Cauchy's Criterion for Almost-Sure Convergence) *A sequence of random variables* $\{Y_n\}$ *converges almost surely to some (possibly degenerate) random variable iff*

$$\lim_{n\rightarrow\infty} P(\max_{m>n} |y_m - y_n| < \varepsilon) = 1, \quad \forall\, \varepsilon > 0.^{5}$$

**Proof** Y. S. Chow and H. Teicher (1978), *Probability Theory*. New York: Springer-Verlag, p. 68. ■

The Cauchy criterion states that for almost-sure convergence to occur, it is necessary and sufficient that the distance between the outcomes of the $n$th term in the random sequence and all subsequent terms beyond the $n$th be arbitrarily small with probability approaching 1 as $n \rightarrow \infty$. This makes intuitive sense, since for $\{y_n\}$ to converge to some value, *eventually* (i.e., for all values of $n$ large enough) all the values in the sequence must be arbitrarily close to the limit value and, thus, arbitrarily close to each other.

In summary, almost-sure convergence is a more stringent type of convergence than either convergence in probability or convergence in distribution. In addition to implying the latter two types of convergence, almost-sure convergence also implies that the sequence of outcomes of $\{Y_n\}$ *converges to a limit* represented by an outcome of $Y$ with probability 1.

[5] In the event that max does not exist, max is replaced by sup (supremum, i.e., the smallest upper bound) in the statement of the theorem.

$$\begin{tikzcd}
\boxed{Y_n \xrightarrow{as} Y} \arrow[d, Rightarrow] & \\
\boxed{Y_n \xrightarrow{p} Y} \arrow[r, Rightarrow, shift left] & \boxed{Y_n \xrightarrow{d} Y} \arrow[l, Rightarrow, shift left, "(Y = c)"] \\
\boxed{Y_n \xrightarrow{m} Y} \arrow[u, Rightarrow] &
\end{tikzcd}$$

**Figure 5-5** Convergence mode relationships.

### Relationships Between Convergence Modes

Gathering the results presented for the four types of random-variable convergence discussed in this section, we can summarize the relationships between the various types of convergence in Figure 5.5.

## 5.4 Laws of Large Numbers

In this section we examine results concerning the convergence behavior of the specific sequence of scalar random variables $\{\bar{X}_n\}$ whose $n$th term is defined by

$$\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i,$$

where $X_i$ is the $i$th term in another random sequence $\{X_n\}$. Thus, $\{\bar{X}_n\}$ is a sequence whose $n$th term is given by the simple average of the first $n$ terms in the sequence $\{X_n\}$. In the context of a sample of observations from an experiment, $\bar{X}_n$ will be referred to as the *sample mean*—a context that will be examined in Chapter 6. A convergence result for $\{\bar{X}_n\}$ that uses the concept of *convergence in probability* is referred to as a **weak law of large numbers** (WLLN), whereas a convergence result using *almost-sure convergence* is referred to as a **strong law of large numbers** (SLLN).

The types of convergence we will be examining can take either of two forms. When all of the means of the random variables in $\{X_n\}$ are equal to the same number, $\mu$, we examine conditions for which $\bar{X}_n \xrightarrow{as} \mu$ (a SLLN) or $\bar{X}_n \xrightarrow{p} \mu$ (a WLLN). When the means of the random variables in $\{X_n\}$ are not necessarily equal, we examine conditions for which $\bar{X}_n - \bar{\mu}_n \xrightarrow{as} 0$ (a SLLN) or $\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0$ (a WLLN), where $\bar{\mu}_n = n^{-1} \sum_{i=1}^{n} \mu_i$ and $\mu_i = \mathrm{E}X_i$ is the $i$th term in the real number sequence $\{\mu_n\}$. Thus, in the case of equal means, we are examining convergence of $\{\bar{X}_n\}$ to the common (constant) mean of the random variables in $\{X_n\}$, whereas in the case of unequal means, we are examining whether the difference between the outcomes of $\bar{X}_n$ and the average mean, $\bar{\mu}_n$, converges to zero in probability or almost surely as $n \to \infty$.[6]

The reader may wonder why the convergence behavior of such a specific sequence of random variables as $\{\bar{X}_n\}$ deserves explicit attention. The answer lies in the fact that a large number of important parameter-estimation and hypothesis-testing procedures in econometrics, and statistics in general, can be defined in terms of averages of random variables. The laws of large numbers are then useful for analyzing the asymptotic behavior of these procedures when the samples of data being analyzed are relatively large.

### Weak Laws of Large Numbers (WLLN)

There is a variety of conditions that can be placed on the random variables in the sequence $\{X_n\}$ that ensure either $\bar{X}_n \xrightarrow{p} \mu$ or $\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0$. These conditions relate to the independence, homogeneity of distribution, and variances and covariances of the random variables in the sequence $\{X_n\}$.

The basic idea underlying weak laws of large numbers is to have the distribution of the random variable $\bar{Y}_n = n^{-1} \sum_{i=1}^{n} (X_i - \mu_i)$ collapse and become degenerate on zero as $n \to \infty$. For this to happen, the random variable $\sum_{i=1}^{n} (X_i - \mu_i)$ must be of smaller order of magnitude (in probability) than $n$. The intuition underlying the degeneracy of $\bar{Y}_n$ in the limit is perhaps clearest in the case where the $X_i$'s are iid and $\text{var}(X_i) = \sigma^2$ exists. Then the distribution of $\sum_{i=1}^{n} (X_i - \mu_i)$ has a variance, or spread, given by $n\sigma^2$, which is expanding by a factor of $n$. The expanding spread of the distribution is counteracted via scaling the random variable by the factor $n^{-1}$ (actually, any $n^{-\delta}$ with $\delta > 1/2$ will do), leading to a distribution (of $\bar{Y}_n$) having variance $\sigma^2/n \to 0$ and implying the degeneracy of $\bar{Y}_n$ at zero as $n \to \infty$. As will be seen ahead, neither the existence of $\sigma^2$ nor the iid condition is necessary for $\bar{Y}_n$ to be degenerate at zero in the limit.

*IID Case* The only WLLN we will examine that does *not* require the existence of the variances of the random variables in the sequence $\{X_n\}$ is Khinchin's WLLN, as follows.

**Theorem 5.19** (Khinchin's WLLN) *Let $\{X_n\}$ be a sequence of iid random variables, and suppose $\mathrm{E}X_i = \mu < \infty$, $\forall$ i. Then $\bar{X}_n \xrightarrow{p} \mu$.*

**Proof** A general proof involving characteristic functions can be found in D. S. G. Pollock (1979), The Algebra of Econometrics. New York: John Wiley, p. 332. An alternative proof based entirely on probability inequalities is given by C. R. Rao, Statistical Inference, pp. 112–113. Our proof requires the additional assumption that $M_{X_i}(t)$ exists, but a general proof is analogous with the characteristic function replacing the MGF.

[6] The concepts of SLLNs and WLLNs can be generalized to the case where $\{\mu_n\}$ is a sequence of constants that are not necessarily the means of the $X_i$'s. See Y. S. Chow and H. Teicher, *Probability Theory*, p. 121.

The moment-generating function of $\bar{X}_n$ is given by

$$M_{\bar{X}_n}(t) = \mathrm{E}\exp\left(\frac{t}{n}\sum_{i=1}^{n} X_i\right) = \prod_{i=1}^{n}\mathrm{E}\exp\left(X_i\frac{t}{n}\right) = \prod_{i=1}^{n} M_{X_i}\left(\frac{t}{n}\right) = \left[M_{X_1}\left(\frac{t}{n}\right)\right]^n$$

since the $X_i$'s are *independent* and *identically* distributed. It follows that

$$\lim_{n\to\infty} M_{\bar{X}_n}(t) = \lim_{n\to\infty}\left[1 + \frac{nM_{X_1}(t/n) - n}{n}\right]^n = \exp\left(\lim_{n\to\infty} n\left[M_{X_1}\left(\frac{t}{n}\right) - 1\right]\right)$$

by Lemma 5.4.

**Lemma 5.4** $\boxed{\lim_{n\to\infty}[1 + a_n/n]^n = \exp(\lim_{n\to\infty} a_n).}$

By L'Hospital's rule,

$$\lim_{n\to\infty}\frac{M_{X_1}(t/n) - 1}{n^{-1}} = \lim_{n\to\infty}(-n^2)\frac{dM_{X_1}(t/n)}{d(t/n)}\left(\frac{-t}{n^2}\right) = t\mu,$$

since the first derivative of $M_{X_1}(t_*) \to \mu$ as $t_* = t/n \to 0$. Then $\lim_{n\to\infty} M_{\bar{X}_1}(t) = e^{t\mu}$, which is the MGF of a random variable that is degenerate at $\mu$. Therefore, by Theorems 5.2 and 5.8, $\bar{X}_n \xrightarrow{p} \mu$. ■

**Example 5.40** Let $\{X_n\}$ be a sequence of iid random variables, with $X_i \sim \mathrm{Gamma}(\alpha,\beta)$. It follows from Khinchin's WLLN that $\bar{X}_n \xrightarrow{p} \alpha\beta$. If $\alpha = 2$ and $\beta = 4$, it follows that $\bar{X}_n \xrightarrow{p} 8$. □

**Example 5.41** Let $\{X_n\}$ be a sequence of iid random variables, with $X_i \sim p^x(1-p)^{1-x}I_{\{0,1\}}(x)$. It follows from Khinchin's WLLN that $\bar{X}_n \xrightarrow{p} p$. If $p = .6$, it follows that $\bar{X}_n \xrightarrow{p} .6$. □

**Example 5.42** Let $f(x) = 2x^{-3}I_{[1,\infty)}(x)$, and suppose that the random variables in the sequence $\{X_n\}$ are iid, each with density function $f(x)$. Note that if $X \sim f(x)$, then

$$\mathrm{E}X = \int_1^\infty x(2x^{-3})dx = 2,$$

but

$$\mathrm{E}X^2 = \int_1^\infty x^2(2x^{-3})dx = \int_1^\infty 2x^{-1}dx = 2\ln x\,|_1^\infty \to \infty,$$

so that the $\mathrm{var}(X) = \mathrm{E}X^2 - (\mathrm{E}X)^2$ does *not* exist. Nonetheless, by Khinchin's theorem we know that $\bar{X}_n \xrightarrow{p} 2$. □

Khinchin's WLLN can be used to provide support for the relative-frequency definition of probability, as follows:

**Theorem 5.20** (Convergence in Probability of Relative Frequency) *Let $\{S,\Upsilon,P\}$ be the probability space of an experiment, and let $A$ be any event contained in $S$. Let* *an outcome of $N_A$ be the number of times that event $A$ occurs in $n$ independent and identical repetitions of the experiment. Then the relative frequency of event $A$ occurring is such that $(N_A/n) \xrightarrow{p} P(A)$.*

**Proof** Without loss of generality, we assume that $S$ is a real-valued sample space. Let the $n$ iid random variables $Z_1,\ldots,Z_n$ represent the $n$ independent and identical repetitions of the experiment, and define $X_i = I_A(Z_i)$, for $i = 1,\ldots,n$, so that $X_1,\ldots,X_n$ are $n$ iid (Bernoulli) random variables for which $x_i = 1$ indicates the occurrence and $x_i = 0$ indicates the nonoccurrence of event $A$ on the $i$th repetition. Since $\mathrm{E}X_i = P(A)$, $\forall\, i$, it follows from Khinchin's WLLN that $\bar{X}_n \xrightarrow{p} P(A)$. Then, since $N_A \equiv \sum_{i=1}^{n} X_i$ and $(N_A/n) \equiv \bar{X}_n$, we can also conclude that $(N_A/n) \xrightarrow{p} P(A)$. ■

Theorem 5.20 implies that as the number of independent, identical repetitions of an experiment → ∞, the probability that the relative frequency of event $A$ is arbitrarily close to the true probability of event $A$ approaches 1. Thus, the WLLN provides support for the notion that the relative frequency of the occurrence of an event can be used as the measure of the probability of the event as $n \to \infty$. Note, however, that we did *not* conclude that $P(\lim_{n\to\infty}(n_A/n) = P(A)) = 1$, that is, we cannot conclude that the *limit* of the relative frequency *exists* and *is equal to $P(A)$ with probability 1.* The latter result involves the notion of almost-sure convergence and will be dealt with in our subsequent discussion of SLLNs.

*Non-IID Case* WLLNs that relax the iid assumption of Khinchin's WLLN can be defined by imposing various other conditions on the variances and covariances of the random variables in the sequence $\{X_n\}$. The WLLN that we will present follows from the necessary and sufficient conditions governing the existence of a WLLN for a sequence of random variables stated in the next theorem.

**Theorem 5.21** (Necessary and Sufficient Conditions for WLLN) *Let $\{X_n\}$ be a sequence of random variables with finite variances (not necessarily independent), and let $\{\mu_n\}$ be the corresponding sequence of their expectations. Then*

$$\lim_{n\to\infty} P(|\bar{x}_n - \bar{\mu}_n| < \varepsilon) = 1, \quad \forall\, \varepsilon > 0 \quad \textit{iff}\ \mathrm{E}\left[\frac{(\bar{X}_n - \bar{\mu}_n)^2}{1 + (\bar{X}_n - \bar{\mu}_n)^2}\right] \to 0.$$

**Proof** *Sufficiency*; For any choice of $b > 0$ and $a \geq b$, $a/(a+1) \geq b/(b+1)$. It follows that, $\forall\, \varepsilon > 0$,

$$P\left((\bar{x}_n - \bar{\mu}_n)^2 \geq \varepsilon^2\right) \leq P\left(\frac{(\bar{x}_n - \bar{\mu}_n)^2}{1 + (\bar{x}_n - \bar{\mu}_n)^2} \geq \frac{\varepsilon^2}{1+\varepsilon^2}\right) \leq E\left[\frac{(\bar{X}_n - \bar{\mu}_n)^2}{1 + (\bar{X}_n - \bar{\mu}_n)^2}\right] \Big/ \left[\frac{\varepsilon^2}{1+\varepsilon^2}\right],$$

where the first inequality follows because the event on the left-hand side *implies* the event on the right, and the second inequality is an application of Markov's inequality. If the expectation of the bracketed term → 0 as $n \to \infty$, then, $\forall\, \varepsilon > 0$,

$$P\left((\bar{x}_n - \bar{\mu}_n)^2 < \varepsilon^2\right) = P(|\bar{x}_n - \bar{\mu}_n| < \varepsilon) \rightarrow 1$$

so that $(\bar{X}_n - \bar{\mu}_n) \xrightarrow{p} 0$.

*Necessity*: See B. V. Gnedenko (1968), The Theory of Probability. New York: Chelsea Publishing Co., p. 246–248. ■

Any condition placed on the random variables in the sequence $\{X_n\}$ that result in the convergence of the expectation of $(\bar{X}_n - \bar{\mu}_n)^2/[1 + (\bar{X}_n - \bar{\mu}_n)^2]$ to zero results in $(\bar{X}_n - \bar{\mu}_n) \xrightarrow{p} 0$ (or $\bar{X}_n \xrightarrow{p} \mu$ in the equal-means case), by Theorem 5.21. We present one such condition now.

**Theorem 5.22** (WLLN for Non-IID Case) *Let $\{X_n\}$ be a sequence of random variables with respective means given by $\{\mu_n\}$. If* $\text{var}(\bar{X}_n) \rightarrow 0$, *then* $(\bar{X}_n - \bar{\mu}_n) \xrightarrow{p} 0$.

**Proof** The result follows directly from Theorem 5.21 upon recognizing that

$$0 \leq \mathrm{E}\left[\frac{(\bar{X}_n - \bar{\mu}_n)^2}{1 + (\bar{X}_n - \bar{\mu}_n)^2}\right] \leq \mathrm{E}\left[(\bar{X}_n - \bar{\mu}_n)^2\right] = \text{var}(\bar{X}_n).$$ ■

**Example 5.43** Let $\{X_n\}$ be a sequence of Gamma-distributed random variables for which $\mathrm{E}X_i = 2^{-i}$, $\text{var}(X_i) = 4$, and $\sigma_{ij} = 0$, $\forall\, i \neq j$. Then, since $\text{var}(\bar{X}_n) = 4/n \rightarrow 0$, it follows by Theorem 5.22 that $\bar{X}_n - (1 - .5^n)/n \xrightarrow{p} 0$, where $\bar{\mu}_n = (1 - .5^n)/n$. □

**Example 5.44** Let $\{X_n\}$ be a sequence of independent, Beta-distributed, random variables for which $\mathrm{E}X_i = .4\ \forall\, i$. Note that the variance of a Beta distribution exhibits a finite upper bound, say $\sigma_i^2 \leq \tau$, since $P(x \in (0,1)) = 1$. Then, *for any* variances of the random variables, $\text{var}(\bar{X}_n) = n^{-2}\sum_{i=1}^n \sigma_i^2 \leq n^{-1}\tau \rightarrow 0$. By Theorem 5.22, $\bar{X}_n \xrightarrow{p} .4$. (Note: Khinchin's theorem cannot be used here since it is not known whether the $X_i$'s have identical distributions.) □

**Example 5.45** Let the sequence of random variables $\{X_n\}$ be such that $X_i \sim N(1, 1 + i^{-1})$ with $\sigma_{ij} = \rho^{|i-j|}$, $\rho \in (0,1)$ and $i \neq j$. Since $\sigma_i^2 \leq 2\ \forall\, i$, $\sum_{i=1}^n \sigma_i^2 \leq 2n$, and thus $\sum_{i=1}^n \sigma_i^2$ is $o(n^2)$. Also, given $i$, $\sum_{j>i}^n \sigma_{ij} = \sum_{j>i}^n \rho^{|i-j|} \rightarrow \rho/(1-\rho)$, so that $\sum_{j>i}^n \sigma_{ij}$ is $o(n^1)$. Then, since $\bar{\mu}_n = 1\ \forall\, n$, and $\text{var}(\bar{X}_n) = n^{-2}[\sum_{i=1}^n \sigma_i^2 + 2\sum_{i=1}^n \sum_{j>i}^n \sigma_{ij}] = o(1)$ so that $\text{var}(\bar{X}_n) \rightarrow 0$, it follows from Theorem 5.22 that $\bar{X}_n \xrightarrow{p} 1$. □

With reference to Ex. 5.43, it would seem reasonable to characterize the convergence in probability as $\bar{X}_n \xrightarrow{p} 0$, since $\bar{\mu}_n = (1 - .5^n)/n \rightarrow 0$. More generally, if $\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0$ and $\bar{\mu}_n \rightarrow c$, we can alternatively state that $\bar{X}_n \xrightarrow{p} c$, as indicated below.

**Theorem 5.23** $\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0$ *and* $\bar{\mu}_n \rightarrow c \Rightarrow \bar{X}_n \xrightarrow{p} c$.

**Proof** With reference to Theorems 5.4 and 5.5, we know that since $\bar{\mu}_n \xrightarrow{p} c$ or, equivalently, $(\bar{\mu}_n - c) \xrightarrow{p} 0$, and since

$$g(\bar{X}_n - \bar{\mu}_n, \bar{\mu}_n - c) = (\bar{X}_n - \bar{\mu}_n) + (\bar{\mu}_n - c) = \bar{X}_n - c$$

is a continuous function of $(\bar{X}_n - \bar{\mu}_n)$ and $(\bar{\mu}_n - c)$, then

$$\text{plim}(\bar{X}_n - c) = \text{plim}(\bar{X}_n - \bar{\mu}_n) + \text{plim}(\bar{\mu}_n - c) = 0,$$

so that by Def. 5.11, $\bar{X}_n \xrightarrow{p} c$. ■

### *Strong Laws of Large Numbers (SLLN)

As in the case of the WLLN, there is a variety of conditions that can be placed on the random variables in the sequence $\{X_n\}$ that ensure either $\bar{X}_n \xrightarrow{as} \mu(\mathrm{E}X_i = \mu, \forall\, i)$ or $\bar{X}_n - \bar{\mu}_n \xrightarrow{as} 0\ (\mathrm{E}X_i = \mu_i, \forall\, i)$. The conditions relate to the independence, homogeneity of distribution, and variances and covariances of the random variables in the sequence.

The basic idea underlying strong laws of large numbers is to have the joint distribution of $Y_n = n^{-1} \sum_{i=1}^{n} (X_i - \mu_i)$, $n = 1, 2, \ldots$, be such that the event $y_n \to 0$ is assigned probability 1. For this to happen, it is known from Theorem 5.14, as well as from the basic concept of a limit itself, that the event $\{|y_i| < \varepsilon, \forall\, i \geq n\}$ must have a probability approaching 1 as $n \to \infty\ \forall\, \varepsilon > 0$. Thus the *marginal* distribution of the (infinite) set of random variables $\{Y_n, Y_{n+1}, Y_{n+2}, \ldots\}$ must be approaching degeneracy on a zero vector as $n \to \infty$. Through various constraints on the spread (variance) and/or degree of dependence of the underlying random variables in the sequence $\{X_i\}$, this degenerate behavior can be attained.

******IID Case*** We begin by presenting a result known as **Kolmogorov's inequality**, which can be interpreted as a *generalization of Markov's inequality.*

**Theorem 5.24** (Kolmogorov's Inequality) *Let $X_1, \ldots, X_n$ be independent random variables for which $\mathrm{E}X_i = 0$ and $\sigma_i^2 < \infty\ \forall\, i$. Then, $\forall\, \varepsilon > 0$,*

$$P\left( \max_{1 \leq m \leq n} \left| \sum_{i=1}^{m} x_i \right| \geq \varepsilon \right) \leq \sum_{i=1}^{n} \frac{\sigma_i^2}{\varepsilon^2}.$$

***Proof** Let $W_j = \sum_{i=1}^{j} X_i$, and define the events

$$A_j = \{(x_1, \ldots, x_n) : |w_j| \geq \varepsilon, |w_i| < \varepsilon \text{ for } i < j\}, \quad j = 1, \ldots, n,$$

and

$$A = \{(x_1, \ldots, x_n) : \max_{1 \leq m \leq n} \left| \sum_{i=1}^{m} x_i \right| \geq \varepsilon\}.$$

The events $A_1,\ldots,A_n$ are disjoint, and $A = \cup_{i=1}^n A_i$, i.e., the $A_i$'s are a partition of $A$. Furthermore,

$$\text{(I)}\quad \mathrm{E}W_n^2 = \mathrm{E}\left[\sum_{i=1}^n X_i^2 + \sum\sum_{i\neq j} X_iX_j\right] = \sum_{i=1}^n \sigma_i^2 \geq \mathrm{E}W_n^2 I_A(X) = \sum_{i=1}^n \mathrm{E}W_n^2 I_{A_i}(X),$$

where $x = (x_1,\ldots,x_n)$, and the last equality holds because $I_A(x) = \sum_{i=1}^n I_{A_i}(x)$. Also note that $\forall\ j < n$, $W_n - W_j = \sum_{i=j+1}^n X_i = f(X_{j+1},\ldots,X_n)$ is independent of $W_j = g(X_1,\ldots,X_j)$ and of $W_j I_{A_j}(X) = h(X_1,\ldots,X_j)$ by the independence of $(X_1,\ldots,X_j)$ and $(X_{j+1},\ldots,X_n)$ (note that $X_{j+1},\ldots,X_n$ are *ghosts* in the function $I_{A_j}(X)$ given the definition of $A_j$). Therefore,

$$\mathrm{E}(W_n - W_j)W_j I_{A_j}(X) = [\mathrm{E}(W_n - W_j)][\mathrm{E}W_j I_{A_j}(X)] = 0$$

since $\mathrm{E}W_i = 0$, $\forall\ i$. It follows that

$$\mathrm{E}W_n^2 I_{A_j}(X) = E\left[W_n^2 - 2(W_n - W_j)W_j\right] I_{A_j}(X) = \mathrm{E}\left[W_j^2 + (W_n - W_j)^2\right] I_{A_j}(X)$$
$$\geq \mathrm{E}W_j^2 I_{A_j}(X) \geq \varepsilon^2 P(A_j),$$

where the last inequality follows from the fact that $w_j^2 \geq \varepsilon^2$ when $x \in A_j$. Using this result in (I), and recalling that the $A_i$'s are a partition of $A$,

$$\sum_{i=1}^n \sigma_i^2 \geq \varepsilon^2 \sum_{i=1}^n P(A_i) = \varepsilon^2 P(A) \quad \forall\ \varepsilon > 0. \qquad \blacksquare$$

Note that with $n = 1$ and $X_1$ defined to be a nonnegative-valued random variable with zero mean, Theorem 5.24 is a statement of Markov's inequality. Kolmogorov's inequality leads to **Kolmogorov's** SLLN for iid random variables.

**Theorem 5.25** (Kolmogorov's SLLN: Variances Exist) *Let $\{X_n\}$ be a sequence of iid random variables such that* $\mathrm{E}X_i = \mu$ *and* $\mathrm{var}(X_i) = \sigma^2 < \infty$, $\forall\ i$. *Then* $\bar{X}_n \xrightarrow{as} \mu$.

***Proof** Let $Y_i = (X_i - \mu)/i$, and note that $\mathrm{E}Y_i = 0$ and $\mathrm{var}(Y_i) = \sigma^2/i^2$. Define $W_k = \sum_{i=1}^k Y_i$, and examine the sums

$$W_{n+m} - W_n = \sum_{i=n+1}^{n+m} Y_i = \sum_{i=1}^m Y_{n+i}, \quad \text{for } m = 1, 2, \ldots, k.$$

By Kolmogorov's inequality (replace $x_i$ by $y_{n+i}$ in the statement of the inequality), it follows that, $\forall\ \varepsilon > 0$,

$$P\left(\max_{1\leq m\leq k} |w_{n+m} - w_n| \geq \varepsilon\right) = P\left(\max_{n+1\leq m\leq n+k} |w_m - w_n| \geq \varepsilon\right) \leq \sum_{i=n+1}^{n+k}\left[\frac{\sigma^2}{(i^2\varepsilon^2)}\right].$$

Letting $k \to \infty$, we have that[7]

$$P\left(\max_{m>n} |w_m - w_n| \geq \varepsilon\right) \leq \sum_{i=n+1}^{\infty} \left[\frac{\sigma^2}{(i^2\varepsilon^2)}\right] = \left(\frac{\sigma^2}{\varepsilon^2}\right) \sum_{i=n+1}^{\infty} i^{-2},$$

and since $\lim_{n\to\infty} \sum_{i=n+1}^{\infty} i^{-2} = 0$,[8] it follows from the Cauchy criterion for almost-sure convergence (Theorem 5.18) that $\{W_n\}$ converges almost surely to some random variable, say $W$. Now examine the Kronecker lemma from real analysis.

**Lemma 5.5**
***Kronecker's Lemma***

Let $\{a_n\}$ be a sequence of nondecreasing positive numbers, and let $\{z_n\}$ be a sequence of real numbers for which $\sum_{i=1}^{\infty} z_i/a_i$ converges. Then

$$\lim_{n\to\infty} a_n^{-1} \sum_{i=1}^{n} z_i = 0.$$

(See E. Lukacs (1968), Stochastic Convergence. Andover, MA: D. C. Heath and Co., p. 96).

Let $z_i = x_i - \mu$ and $a_i = i$, so that $y_i = z_i/a_i$ and $w_n = \sum_{i=1}^{n} z_i/a_i$. Since $w_n \to w$ with probability 1, and since by Kronecker's lemma $w_n \to w \Rightarrow n^{-1} \sum_{i=1}^{n} z_i \to 0$, then

$$n^{-1} \sum_{i=1}^{n} z_i = \bar{x}_n - \mu \to 0$$

with probability 1. ■

**Example 5.46** Let $\{X_n\}$ be a sequence of iid exponentially distributed random variables, i.e., $X_i \sim \theta^{-1} \exp(-x_i/\theta) I_{(0,\infty)}(x_i)\ \forall\, i$, where $\theta \leq c < \infty$. Note that $\mathrm{E}X_i = \mu = \theta$ and $\mathrm{var}(X_i) = \sigma^2 = \theta^2 \leq c^2 < \infty$, $\forall\, i$. Theorem 5.25 applies, so that $\bar{X}_n \xrightarrow{\text{as}} \theta$. □

The existence of variances is not necessary for a SLLN to hold in the iid case. The following theorem provides necessary and sufficient conditions for a SLLN to hold.

**Theorem 5.26** (Kolmogorov's SLLN (IID case)) *Let $\{X_n\}$ be a sequence of iid random variables. Then the condition $\mathrm{E}X_i = \mu < \infty$ is necessary and sufficient for $\bar{X}_n \xrightarrow{\text{as}} \mu$.*

**Proof** The proof of the theorem is somewhat difficult and can be found in C. Rao, Statistical Inference, pp. 115–116. ■

[7] If max does not exist, max is replaced by sup.

[8] $\sum_{i=1}^{n} i^{-p}$ is the so-called $p$ series that converges for $p > 1$ and diverges for $p \in (0, 1]$. Since the series converges for $p = 2$, it must be the case that $\sum_{i=n+1}^{\infty} i^{-2} \to 0$ as $n \to \infty$. See Bartle, Real Analysis, pp. 290–291.

**Example 5.47** Recall Ex. 5.42, where $\{X_n\}$ was a sequence of iid random variables for which $\mathrm{E}X_i = 2$ and for which the variance of the $X_i$'s did *not* exist. Nonetheless, by Kolmogorov's SLLN, we know that $\bar{X}_n \xrightarrow{\text{as}} 2$. □

Theorem 5.26 provides stronger support for the relative-frequency definition of probability than the WLLNs (recall Theorem 5.20), as the following theorem indicates.

**Theorem 5.27** (Almost-Sure Convergence Of Relative Frequency) *Let $\{S,\Upsilon,P\}$ be the probability space of an experiment, and let $A$ be any event contained in $S$. Let an outcome of $N_A$ be the number of times that event $A$ occurs in $n$ independent and identical repetitions of the experiment. Then the relative frequency of event $A$ occurring is such that $(N_A/n) \xrightarrow{\text{as}} P(A)$.*

**Proof** Without loss of generality, we assume that $S$ is a real-valued sample space. Let the $n$ iid random variables $Z_1,\ldots,Z_n$ represent the $n$ independent and identical repetitions of the experiment, and define $X_i = I_A(Z_i)$, for $i = 1,\ldots,n$, so that $X_1,\ldots,X_n$ are $n$ iid (Bernoulli) random variables for which $x_i = 1$ indicates the occurrence and $x_i = 0$ indicates the nonoccurrence of event $A$ on the $i$th repetition of the experiment. Note that $\mathrm{E}X_i = P(A) < \infty \ \forall\, i$, so that Theorem 5.26 is applicable. Then $\bar{X}_n \xrightarrow{\text{as}} P(A)$, and since $N_A \equiv \sum_{i=1}^{n} X_i$ and $N_A/n \equiv \bar{X}_n$, we can also conclude that $(N_A/n) \xrightarrow{\text{as}} P(A)$. ■

Theorem 5.27 implies that the relative frequency of the occurrence of event $A$ *achieves a limit* with probability 1 as $n \to \infty$. Furthermore, the value of the limit achieved equals the probability of the event $A$. Thus, the SLLN provides strong support for the notion that the relative frequency of the occurrence of an event can be used as the measure of the probability of the event as $n \to \infty$, since we are *essentially certain* that the relative frequency of an event will converge to the probability of the event.

**Non-IID Case* There are many other ways in which to place restrictions on $\{X_n\}$ so that $\bar{X}_n - \bar{\mu}_n \xrightarrow{\text{as}} 0$ or $X_n \xrightarrow{\text{as}} \mu$. We will present a SLLN for the nonidentically distributed case which can be applied whether or not the random variables in $\{X_n\}$ are independent. The theorem utilizes the concept of an asymptotic, nonpositively correlated sequence, defined below.

**Definition 5.16** *Asymptotic, nonpositively correlated sequence*[9]

The sequence of random scalars $\{X_n\}$, where $\mathrm{var}(X_i) = \sigma_i^2 < \infty \ \forall\, i$, is said to be **asymptotic nonpositively correlated** if there exists a sequence of constants $\{a_n\}$ such that $a_i \in [0, 1] \ \forall\, i$, $\sum_{i=0}^{\infty} a_i < \infty$, and $\mathrm{cov}(X_i, X_{i+t}) \le a_t\sigma_i\sigma_{i+t} \ \forall\, t > 0$.

[9] Some authors use the terminology "asymptotically uncorrelated" for this concept (e.g., H. White, Asymptotic Theory, p. 49). However, the concept does not rule out negative correlation.

Note that for $\sum_{i=0}^{\infty} a_i$ to be finite when $a_i \in [0,1]\ \forall\ i$, it must be the case that $a_n \to 0$ as $n \to \infty$. Since the $a_t$'s represent upper bounds to the correlations between $X_i$ and $X_{i+t}$, the definition implies that $X_i$ and $X_{i+t}$ cannot be positively correlated when $t \to \infty$.

**Example 5.48** An Asymptotic, Nonpositively Correlated Sequence

Let the sequence of random variables $\{X_n\}$ adhere to the (*first-order*) *autocorrelation process*:[10]

(I) $X_i = \rho X_{i-1} + \varepsilon_i$,

where the $\varepsilon_i$'s are iid with $\mathrm{E}\varepsilon_i = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma^2 \in (0,\infty)$, $\mathrm{cov}(X_{i-1}, \varepsilon_i) = 0\ \forall\ i$, $X_0 = 0$, and $|\rho| < 1$. It follows that $\mathrm{E}X_i = 0\ \forall\ i$, and $\mathrm{var}(X_i) = \sigma^2 \sum_{j=1}^{i} \rho^{2(j-1)}$ for $i \geq 1$. Note further that (I) implies

$$X_{i+t} = \rho^t X_i + \sum_{j=0}^{t-1} \rho^j \varepsilon_{t+i-j} \quad \text{for } t \geq 1.$$

To define the value of $\mathrm{corr}(X_{i+t}, X_i)$, let $\sigma_i^2 = \mathrm{var}(X_i)$, and note that

$$\begin{aligned}\mathrm{corr}(X_{i+t}, X_i) &= \mathrm{E}[(X_{i+t} - \mathrm{E}X_{i+t})(X_i - \mathrm{E}X_i)]/(\sigma_{i+t}\sigma_i) \\ &= \mathrm{E}\left[\left(\rho^t(X_i - \mathrm{E}X_i) + \sum_{j=0}^{t-1} \rho^j \varepsilon_{t+i-j}\right)(X_i - \mathrm{E}X_i)\right]/(\sigma_{i+t}\sigma_i) \\ &= \rho^t \sigma_i/\sigma_{i+t} \leq \begin{cases} 0 & \text{if } \rho^t \leq 0, \\ \rho^t & \text{if } \rho^t > 0, \end{cases}\end{aligned}$$

where the last inequality follows from the fact that

$$\sigma_i^2/\sigma_{i+t}^2 = \sum_{j=1}^{i} \rho^{2(j-1)} \Big/ \sum_{j=1}^{i+t} \rho^{2(j-1)} \leq 1 \quad \forall\ t \geq 1.$$

To demonstrate that $\{X_n\}$ is an asymptotic nonpositively correlated sequence, define $a_t = \rho^t$ if $\rho^t > 0$, and $a_t = 0$ if $\rho^t \leq 0$, so that $\mathrm{corr}(X_{i+t}, X_i) \leq a_t\ \forall\ t$, and $a_t \in [0,1]$, $\forall\ t$. Then, since

$$\sum_{t=1}^{\infty} a_t \leq \sum_{t=1}^{\infty} |\rho|^t = |\rho|/(1 - |\rho|) < \infty,$$

the sequence $\{X_n\}$ is asymptotic nonpositively correlated. □

[10] This is an example of a *stochastic process*, which we will revisit in our discussion of the general linear model in Chapter 8. "Stochastic process" means any collection of random variables $\{X_t; t \in T\}$, where $T$ is some index set that serves to order the random variables in the collection. Special cases of stochastic processes include a scalar random variable when $T = \{1\}$, an $n$-variate random vector when $T = \{1, 2, \ldots, n\}$, and a random sequence when $T = \{1, 2, 3, \ldots\}$.

**Theorem 5.28** (SLLN: Non-identically Distributed Case) *Let $\{X_n\}$ be a sequence of random variables such that* $\mathrm{E}X_i = \mu_i$, $\mathrm{var}(X_i) \leq b < \infty \; \forall \; i$, *and $\{X_n\}$ is asymptotic nonpositively correlated. Then* $\bar{X}_n - \bar{\mu}_n \xrightarrow{\text{as}} 0$.

**Proof** This follows directly from Theorem 3.7.2 in W. F. Stout (1974), Almost Sure Convergence. New York: Academic Press, p. 202. ■

The theorem indicates that a sequence of random variables will adhere to a SLLN if the variances of the random variables are bounded and if any positive correlation between random variables in the sequence eventually dissipates when the random variables are far enough apart in the sequence.

**Example 5.49** Recall Ex. 5.48, where it is known that $\{X_n\}$ is asymptotic, nonpositively correlated. Note further that since $\rho \in [0, 1)$, $\sigma_i^2 = \sigma^2 \sum_{j=1}^{i} \rho^{2(j-1)} < \infty \; \forall \; i$, where in fact $\sum_{j=1}^{\infty} \rho^{2(j-1)} = 1 + \rho^2 + \rho^4 + \cdots = 1/(1 - \rho^2)$, so that $\sigma_i^2 \leq 1/(1 - \rho^2) \; \forall \; i$. Since the $\sigma_i^2$'s are upper-bounded, the conditions of Theorem 5.28 are met, and it follows that $\bar{X}_n \xrightarrow{\text{as}} 0$, since $\mathrm{E}X_i = \mu = 0 \; \forall \; i$. □

Our final result on SLLNs concerns whether $\bar{X}_n - \bar{\mu}_n \xrightarrow{\text{as}} 0$ and $\bar{\mu}_n \to c$ together imply that $\bar{X}_n \xrightarrow{\text{as}} c$. The answer is yes, as stated in the next theorem.

**Theorem 5.29** $\bar{X}_n - \bar{\mu}_n \xrightarrow{\text{as}} 0$ *and* $\bar{\mu}_n \to c \Rightarrow \bar{X}_n \xrightarrow{\text{as}} c$.

**Proof** The proof is based on Theorems 5.16 and 5.17 and follows the approach of Theorem 5.23. Details are left to the reader. ■

## 5.5 Central Limit Theorems

Central limit theorems (CLTs) are concerned with the conditions under which sequences of random variables *converge in distribution* to known families of distributions. We will focus primarily on results concerning convergence in distribution of sequences of random variables $\{Y_n\}$ of the following form:[11]

$$Y_n = b_n^{-1}(S_n - a_n) \xrightarrow{\text{d}} N(0, \Sigma),$$

where $\{S_n\}$ is a sequence of scalar or vector random variables whose *n*th term is defined by $S_n = \sum_{i=1}^{n} X_i$, $\{X_n\}$ is a sequence of scalar or vector random variables, and $\{a_n\}$ and $\{b_n\}$ are suitably chosen sequences of real numbers, vectors, or matrices. A statement of conditions on $\{X_n\}$, $\{a_n\}$, and $\{b_n\}$ for which the convergence in distribution result holds true constitutes a **central limit theorem.**

[11] The reader who wishes to read about central limit theory in its most general form can examine Chapter 5 of R. G. Laha, and V. K. Rohatgi (1979), *Probability Theory.* New York: John Wiley.

As we had remarked in the introduction to the preceding section dealing with laws of large numbers, the reader may wonder why the particular problem concerning convergence in distribution defined above deserves such explicit attention. The answer lies in the fact that a large number of important parameter-estimation and hypothesis-testing procedures in econometrics and statistics are defined as functions of sums of random variables (note the $S_n$ term in the convergence problem above). Central limit theorems are then often useful for establishing asymptotic distributions for these procedures, as will be seen in specific examples in subsequent chapters.

In order to illustrate the general way in which the use of a CLT might arise in practice, suppose a CLT is applicable to a scalar random variable, say, as $Y_n = b_n^{-1}(S_n - a_n) \xrightarrow{d} Y \sim N(0,1)$. Then, since $S_n = g(Y_n, a_n, b_n) = b_n Y_n + a_n$, we can define an asymptotic distribution for $S_n$ using Def. 5.10 as $S_n \overset{a}{\sim} b_n Y + a_n$, or $S_n \overset{a}{\sim} N(a_n, b_n^2)$. Thus, for large $n$, $S_n = \sum_{i=1}^n X_i$ has an asymptotic distribution that is normal, with mean $a_n$ and variance $b_n^2$. Now suppose a particular statistical procedure is based on the random variable $W_n = h(S_n, c_n) = h(g(Y_n, a_n, b_n), c_n)$. Then $W_n \overset{a}{\sim} h(g(Y, a_n, b_n), c_n)$, so that for large $n$, an asymptotic distribution for $W_n$ is given by the distribution associated with the composite function $h \circ g$ of the standard normal random variable $Y$ or, equivalently, by the distribution associated with the function $h$ of the random variable $S_n$ under the assumption that $S_n \sim N(a_n, b_n^2)$. For example, if $W_n \overset{a}{\sim} h(g(Y, a_n, b_n), c_n) = c_n Y^2$, then $W_n \overset{a}{\sim}$ Gamma$(1/2, 2c_n)$ (since $Y^2 \sim \chi_1^2$, and then $c_n Y^2$ has a Gamma distribution with $\alpha = 1/2$ and $\beta = 2c_n$).

Defining asymptotic distributions for random variables of interest is most useful, and sometimes indispensable, when the *exact* distributions of the random variables are very difficult or impossible to derive. Furthermore, even if the exact distributions of random variables of interest can be defined, they may be very difficult to work with, whereas the asymptotic distribution may be relatively easy to analyze. For both of these reasons, central limit theorems figure prominently in the development of econometric and statistical theory and application.

We divide our presentation of CLTs into three subsections. The first subsection deals with the case of independent scalar random variables. The second subsection provides an introduction to the case where the scalar random variables in $\{X_n\}$ are *not* independent. In the final subsection, we present some CLT results relating to multivariate random variables.

### Independent Scalar Random Variables

We will examine three CLTs for independent random variables beginning with the simplest, but least general, iid case and ending with a CLT that presents necessary and sufficient conditions for $b_n^{-1}(S_n - a_n) \xrightarrow{d} N(0,1)$ when the $X_i$'s are independent but not necessarily identically distributed.

*Identically Distributed Case* We begin our examination of central limit theorems by presenting the simplest of all CLTs, the **Lindberg–Levy CLT.**

**Theorem 5.30** (Lindberg–Levy CLT) *Let $\{X_n\}$ be a sequence of iid random variables with* $\mathrm{E}X_i = \mu$ *and* $\mathrm{var}\,(X_i) = \sigma^2 \in (0,\infty)\ \forall\ i$. *Then,*

$$(n^{1/2}\sigma)^{-1}\left(\sum_{i=1}^{n} X_i - n\mu\right) = \frac{n^{1/2}\,(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathrm{d}} N(0,1).$$

**Proof** We prove the theorem for the case where the moment-generating function of $X_i$ exists. The proof can be made general by substituting characteristic functions in place of MGFs, and a general proof utilizing characteristic functions can be found in C. Rao, Statistical Inference, p. 127.

Let $Z_i = (X_i - \mu)/\sigma$, so that $\mathrm{E}Z_i = 0$ and $\mathrm{var}(Z_i) = 1$, and define $Y_n = (n^{1/2}\sigma)^{-1}(\sum_{i=1}^{n} X_i - n\mu) = n^{-1/2}\sum_{i=1}^{n} Z_i$. Then

$$M_{Y_n}(t) = \prod_{i=1}^{n} M_{Z_i}(t/n^{1/2}) = \left[M_{Z_1}(t/n^{1/2})\right]^n$$

since the $Z_i$'s are *independent* and *identically* distributed. Taking logarithms,

$$\lim_{n\to\infty} \ln M_{Y_n}(t) = \lim_{n\to\infty}\left[n \ln M_{Z_1}(t/n^{1/2})\right] = \lim_{n\to\infty}\left[\ln M_{Z_1}(t/n^{1/2})/n^{-1},\right]$$

which has the indeterminate form 0/0, so we can apply L'Hospital's rule. Letting $t_* = t/n^{1/2}$ and $\eta(t_*) = \ln M_{Z_1}(t_*)$,

$$\lim_{n\to\infty} \ln M_{Y_n}(t) = \lim_{n\to\infty}\left[\left(\frac{d\eta(t_*)}{dt_*}\right)\left(-\frac{t}{2n^{3/2}}\right)/(-n^{-2})\right] = \lim_{n\to\infty}\left[\frac{1}{2}\left[t\frac{d\eta(t_*)}{dt_*}/n^{-1/2}\right]\right],$$

which is still an indeterminate form since $d\eta(t_*)/dt_* \to d\eta(0)/dt_* = \mathrm{E}Z_1 = 0$ when $n \to \infty$. A second application of L'Hospital's rule yields

$$\lim_{n\to\infty} \ln M_{Y_n}(t) = (1/2)\lim_{n\to\infty}\left[t^2\frac{d^2\eta(t_*)}{dt_*^2}\right] = t^2/2$$

since

$$\lim_{n\to\infty}\frac{d^2\eta(t_*)}{dt_*^2} = \frac{d^2\eta(0)}{dt_*^2} = \frac{d^2M_{Z_1}(0)}{dt_*^2} - \left[\frac{dM_{Z_1}(0)}{dt_*}\right]^2 = \mathrm{E}Z_1^2 - (\mathrm{E}Z_1)^2 = \mathrm{var}(Z_1) = 1.$$

Thus,

$$\lim_{n\to\infty} M_{Y_n}(t) = \exp\left(\lim_{n\to\infty}\ln M_{Y_n}(t)\right) = e^{t^2/2},$$

which implies that $Y_n \xrightarrow{\mathrm{d}} N(0,1)$, by Theorem 5.2. ■

In order to enhance one's intuitive understanding of why the Lindberg–Levy CLT (LLCLT) holds, we now provide some additional rationale for the result, albeit at the expense of some degree of imprecision in the mathematical details. First note that under the conditions of the LLCLT, $\sum_{i=1}^{n} X_i$ is a random variable that has a mean of $n\mu$ and a variance of $n\sigma^2$. Since both $|n\mu|$ and $n\sigma^2$ diverge to $\infty$ (assuming $\mu \neq 0$) as $n$ increases, it is clear that some form of centering and scaling of $\sum_{i=1}^{n} X_i$ will be necessary for there to be any hope of convergence to some limiting distribution. By subtracting $n\mu$ and then dividing by $n^{1/2}\sigma$, one defines random variables $Y_n = (n^{1/2}\sigma)^{-1}(\sum_{i=1}^{n} X_i - n\mu) = (n^{1/2}\sigma)^{-1}\sum_{i=1}^{n} Z_i$ which have a mean zero and variance of 1 regardless of $n$. The random variables $Z_i = X_i - \mu, i = 1,\ldots,n$, are iid random variables with zero means and variances all equal to $\sigma^2$.

Now a key observation concerning an additional effect of the aforementioned centering and scaling: when $n \to \infty$, any effect of third- and higher-order moments of $Z_i$ on the moments of $Y_n$ are "centered and scaled away" so that all probability distributions for $Z_i$ that have the same mean and variance will lead to precisely the same moments for $Y_n$ as $n \to \infty$ (we assume that all moments of the $Z_i$'s exist). To see this, first consider the third moment of $Y_n$:

$$\mathrm{E}Y_n^3 = \left(n^{1/2}\sigma\right)^{-3} \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{E}(Z_i Z_j Z_k).$$

Since the $Z_i$'s are independent with zero means, it is only when $i = j = k$ that the expectation term is nonzero and equal to $\mu_3'$, the third moment of the distribution of the $Z_i$'s. But there are only $n$ of these terms, and thus $\mathrm{E}Y_n^3 = \sigma^{-3}n^{-3/2}(n\mu_3') \to 0$ as $n \to \infty$, and the third moment of $Y_n$ converges to zero regardless of $\mu_3'$. Following analogous logic applied to $\mathrm{E}Y_n^4$, which is defined in terms of a quadruple sum of $(Z_i Z_j Z_k Z_\ell)$ terms premultiplied by $(n^{1/2}\sigma)^{-4}$, it can be shown that $\mathrm{E}Y_n^4 = \sigma^{-4}n^{-2}[n\mu_4' + 3n(n-1)\sigma^4] \to 3$ regardless of $\mu_4'$, the fourth moment of $Z_i$. This type of argument can be continued ad infinitum to show that all higher-order moments of $Y_n$ converge to known constants and that higher order moments of the $Z_i$'s play no role in determining any of the moments of $Y_n$ when $n \to \infty$.

Now observe that the first four moments of $Y_n$, as defined in the previous paragraph, converge to the first four moments of the standard normal distribution, 0, 1, 0, and 3, respectively. Furthermore, all higher-order moments of $Y_n$ also converge to those of the standard normal distribution, which can in principle be verified one by one following the approach defined above, and in any case is implied by the proof of Theorem 5.30. While not true for all densities, members of the normal family of PDFs are uniquely identified by their moment sequences,[12] so that the moments of $Y_n$ are uniquely consistent with that of a standard normal density as $n \to \infty$ for all underlying probability distribution of the $X_i$'s having mean $\mu$ and variance $\sigma^2$. Thus, the centering and scaling of the $X_i$'s inherent in the definition of $Y_n$ remove any tendencies for higher-order moments of the $X_i$'s to cause the moments of $Y_n$ to deviate from those of a standard normal distribution as $n \to \infty$, leading to convergence of $Y_n$ to the $N(0, 1)$ limiting distribution.

[12] M. Kendall, and A. Stuart (1977), *The Advanced Theory of Statistics, Volume I.* New York: Macmillan, p. 115. Note that any random variable for which an MGF exists is such that its moment sequence uniquely identifies its probability distribution.

The establishment of the LLCLT provides an opportunity to revisit the relationship between *limiting distributions* and *asymptotic distributions*. Under the conditions of the LLCLT, $Y_n = (n^{1/2}\sigma)^{-1}(\sum_{i=1}^{n} X_i - n\mu) \xrightarrow{d} N(0,1)$. Then, following Def. 5.10, an asymptotic distribution for $S_n = \sum_{i=1}^{n} X_i$ can be defined as $S_n = (n^{1/2}\sigma)Y_n + n\mu \overset{a}{\sim} (n^{1/2}\sigma)Y + n\mu$, where $Y \sim N(0,1)$, so that $S_n \overset{a}{\sim} N(n\mu, n\sigma^2)$. Thus, the normal distribution with mean $n\mu$ and variance $n\sigma^2$ provides an approximation to the distribution of $S_n$ when $n$ is large. Regarding $\bar{X}_n$, note that the conditions of the Lindberg–Levy CLT imply $\bar{X}_n \xrightarrow{p} \mu$ by Khinchin's WLLN, so that $\bar{X}_n \xrightarrow{d} \mu$. Because the *limiting distribution* of $\bar{X}_n$ is *degenerate*, it is clear that the distribution provides no information about the variability of $\bar{X}_n$ for finite $n$. The *asymptotic distribution* for $\bar{X}_n$ is more useful in this regard. Noting that

$$\bar{X}_n = g(Y_n, n) = (\sigma/n^{1/2})Y_n + \mu,$$

it follows from Def. 5.10 that $\bar{X}_n \overset{a}{\sim} g(Y, n) = (\sigma/n^{1/2})Y + \mu$ for $Y \sim N(0,1)$, or $\bar{X}_n \overset{a}{\sim} N(\mu, \sigma^2/n)$. Thus, for large $n$, $\bar{X}_n$ is approximately normally distributed with mean $\mu$ and variance $\sigma^2/n$.

The following examples illustrate the application of the Lindberg–Levy CLT.

**Example 5.50** Approximating Binomial Probabilities via the Normal Distribution

Let $\{X_n\}$ be a sequence of iid Bernoulli-type random variables, i.e., $X_i \sim p^{x_i}(1-p)^{1-x_i} I_{\{0,1\}}(x_i)\ \forall\, i$ with $p \neq 0$ or 1. Then, by the Lindberg–Levy CLT,

$$\frac{\sum_{i=1}^{n} X_i - np}{n^{1/2}[p(1-p)]^{1/2}} \xrightarrow{d} N(0,1) \quad \text{and} \quad \sum_{i=1}^{n} X_i \overset{a}{\sim} N(np, np(1-p)).$$

Since $\sum_{i=1}^{n} X_i$ has a binomial distribution under the stated conditions, we have discovered an alternative to the Poisson density for approximating the binomial distribution for large $n$. Note that we are approximating the *discrete* binomial density with the aforementioned *continuous* normal density. It has been found in practice that such approximations are improved, especially when $n$ is not very large, by making a **continuity correction**, whereby each outcome, $x$, in the range of the discrete random variable is associated with the interval event $(x - (1/2), x + (1/2)]$ for the purpose of assigning probability via the asymptotic normal density. For example, if $n = 40$, $p = 1/2$, and $x = 20$, then since $\mu = np = 20$ and $\sigma^2 = np(1-p) = 10$, we have, using the normal asymptotic density $N(20, 10)$,

$$P(x = 20) \approx \int_{19.5}^{20.5} N(z; 20, 10)dz$$

$$\approx \int_{-.5/\sqrt{10}}^{.5/\sqrt{10}} N(z; 0, 1)dz = .1272 \quad \text{(from std. normal table).}$$

The actual probability assigned to the event $P(x = 20)$ by the binomial density is

$$P(x = 20) = \binom{40}{20}\left(\frac{1}{2}\right)^{40} = .1254. \qquad \square$$

**Example 5.51** Approximating $\chi^2$ Probabilities via the Normal Distribution

Let $\{X_n\}$ be a sequence of iid chi-square random variables with 1 degree of freedom, i.e., $X_i \sim \chi_1^2 \; \forall \; i$. By the additivity property of chi-square random variables, $\sum_{i=1}^n X_i \sim \chi_n^2$. Also, $\mathrm{E}X_i = 1$ and $\mathrm{var}(X_i) = 2 \; \forall \; i$ under the prevailing assumptions. Then, by the Lindberg–Levy CLT,

$$Y_n = \frac{\sum_{i=1}^n X_i - n}{(2n)^{1/2}} \xrightarrow{d} N(0,1) \quad \text{and} \quad \sum_{i=1}^n X_i \overset{a}{\sim} N(n, 2n).$$

We have thus discovered an approximation to the $\chi^2$ density function for large degrees of freedom. As an example of its use, note that (from tables of the $\chi^2$ distribution)

$$P(\chi_{30}^2 \leq 43.8) = .95.$$

We obtained our approximation of this probability by utilizing $Z_{30} \overset{a}{\sim} N(30, 60)$, and thus

$$P(z_{30} \leq 43.8) = P\left(\frac{z_{30} - 30}{(60)^{1/2}} \leq \frac{43.8 - 30}{(60)^{1/2}}\right) = P(z \leq 1.783) \quad (\text{where } Z \sim N(0,1))$$
$$= .9627. \qquad \square$$

The Lindberg–Levy CLT implies that any real-world experiment whose final outcome can be conceptualized as the result of a summation or average of the outcomes of a large number of iid random variables having a finite mean and variance can be treated as having approximately a normal distribution. Thus, for example, the total number of defective objects produced on an assembly line or the average miles per gallon achieved by a sample of Ford pickup trucks might be considered as approximately normally distributed to the extent that the independence and identical distribution assumptions hold true.

A natural question to ask in using asymptotic distributions is how large does $n$ have to be for the approximation to be good? Unfortunately, the answer depends on the characteristics of the true distribution underlying the sequence of random variables, and no general answer can be given. However, a number of inequalities have been developed that can be useful in answering the question if something is known about the moments of the underlying densities. Specifically, under the conditions of the Lindberg–Levy CLT, Van Beeck[13] has shown that the maximum absolute difference between $P(y_n \leq c)$

[13] P. Van Beeck (1972), An application of Fourier methods to the problem of sharpening the Berry-Esséen inequality. *Z. Wahrscheinlichkeits Theorie und Verw. Gebiete* 23, pp. 187–196.

for the *actual* density of $Y_n = n^{1/2}(\bar{X}_n - \mu)/\sigma$ and for that of the standard normal density is $.7975(\zeta_3/\sigma^3)n^{-1/2}$, where $\zeta_3 = E[|X - \mathrm{E}X|^3]$ and $\sigma$ refer to the common density of the $X_i$'s. As examples of the bound, if the $X_i$'s are iid Bernoulli or exponential random variables, then the bounds are respectively $.7975n^{-1/2}[1-2p(1-p)][p(1-p)]^{-1/2}$ and $.1653n^{-1/2}$. Then, for example, if $p = .5$ and $n = 1000$, the upper bounds on the errors when approximating $P(y_n \leq c)$ via the standard normal limiting distribution are .025 and .005, respectively. It should be emphasized that Van Beeck's result provides *omnibus* bounds that apply to *all* random variables having the prescribed moments, and as such the bound tends to be quite conservative, that is, the actual approximation errors are generally much smaller than the bound.

*Nonidentically Distributed Case* Although the Lindberg–Levy CLT is applicable in many experimental situations, it has the disadvantage of requiring that all of the random variables have the same mean, the same variance, and moreover, *the same probability distribution*. Various other central limit theorems can be constructed that utilize alternative conditions on the distributions of the random variables in the sequence $\{X_n\}$ (e.g., see Y. S. Chow and H. Teicher, *Probability Theory*, Chapter 9, and R. G. Laha and V. K. Rohatgi, *Probability Theory*, Chapter 5). The most general CLT for the case of independent random variables, which subsumes the LLCLT as a special case, is the **Lindberg CLT**.

***Theorem 5.31** (Lindberg's CLT) *Let $\{X_n\}$ be a sequence of independent random variables with $\mathrm{E}X_i = \mu_i$ and $\mathrm{var}(X_i) = \sigma_i^2 < \infty\ \forall\, i$. Define $b_n^2 = \sum_{i=1}^n \sigma_i^2$, $\bar{\sigma}_n^2 = n^{-1}\sum_{i=1}^n \sigma_i^2$, $\bar{\mu}_n = n^{-1}\sum_{i=1}^n \mu_i$, and let $f_i$ be the PDF of $X_i$. If $\forall\, \varepsilon > 0$,*

(*continuous case:*)
$$\lim_{n\to\infty} \frac{1}{b_n^2}\sum_{i=1}^n \int_{(x_i-\mu_i)^2 \geq \varepsilon b_n^2} (x_i - \mu_i)^2 f_i(x_i)dx_i = 0,$$

(*discrete case:*)
$$\lim_{n\to\infty} \frac{1}{b_n^2}\sum_{i=1}^n \sum_{\substack{(x_i-\mu_i)^2 \geq \varepsilon b_n^2 \\ f_i(x_i)>0}} (x_i - \mu_i)^2 f_i(x_i) = 0,$$

*then*

$$\frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2}} = \frac{n^{1/2}\left(\bar{X}_n - \bar{\mu}_n\right)}{\bar{\sigma}_n} \xrightarrow{\mathrm{d}} N(0,1).$$

**Proof** See Rohatgi, Mathematical Statistics, pp. 282–288 or Chow and Teicher, *Probability Theory*, pp. 291–293. ■

It can be shown that the limit conditions in the Lindberg CLT, known as the *Lindberg conditions*, imply that $\lim_{n\to\infty}\sigma_j^2/\sum_{i=1}^n \sigma_i^2 = 0\ \forall\, j$. That is, the contribution that each $X_j$ makes to the variance of $\sum_{i=1}^n X_i$ is negligible as $n \to \infty$. The Lindberg conditions can be difficult to verify in practice, and so we will present two useful special cases of the Lindberg CLT that rely on more easily verifiable conditions. The first special case essentially implies that if the random variables in the sequence $\{X_n\}$ are independent and bounded with probability 1, then $Y_n = n^{1/2}(\bar{X}_n - \bar{\mu}_n)/\bar{\sigma}_n \xrightarrow{d} N(0,1)$. It will be seen that the boundedness condition and $\sum_{i=1}^n \sigma_i^2 \to \infty$ imply the Lindberg condition.

**Theorem 5.32** (CLT for Bounded Random Variables) *Let $\{X_n\}$ be a sequence of independent random variables such that $P(|x_i| \le m) = 1\ \forall\, i$ for some $m \in (0,\infty)$, and suppose $\mathrm{E}X_i = \mu_i$ and $\mathrm{var}\,(X_i) = \sigma_i^2 < \infty\ \forall\, i$. If $\sum_{i=1}^n \mathrm{var}\,(X_i) = \sum_{i=1}^n \sigma_i^2 \to \infty$ as $n \to \infty$, then $n^{1/2}(\bar{X}_n - \bar{\mu}_n)/\bar{\sigma}_n \xrightarrow{d} N(0,1)$.*

**Proof** This follows from the Lindberg CLT by first noting that since $P(|x_i - \mu_i| \le 2m) = 1$ by the boundedness assumption, then $\forall\, \varepsilon > 0$ (for the continuous case—the discrete case is similar),

$$\int_{(x_i-\mu_i)^2 \ge \varepsilon b_n^2} (x_i - \mu_i)^2 f_i(x_i) dx_i \le 4m^2 \int_{(x_i-\mu_i)^2 \ge \varepsilon b_n^2} f_i(x_i) dx_i$$
$$\le 4m^2 P\left((x_i - \mu_i)^2 \ge \varepsilon b_n^2\right)$$
$$\le \frac{4m^2\sigma_i^2}{\varepsilon b_n^2}$$

where the last inequality results from Markov's inequality. Then

$$\frac{1}{b_n^2}\sum_{i=1}^n \int_{(x_i-\mu_i)^2 \ge \varepsilon b_n^2} (x_i - \mu_i)^2 f_i(x_i) dx_i \le \frac{1}{b_n^2}\sum_{i=1}^n \frac{4m^2\sigma_i^2}{\varepsilon b_n^2}$$
$$\le \frac{4m^2}{(\varepsilon b_n^2)}$$

since $b_n^2 = \sum_{i=1}^n \sigma_i^2$, so that if $b_n^2 \to \infty$ when $n \to \infty$, the right-hand side above has a zero limit $\forall\, \varepsilon > 0$, and the Lindberg condition is satisfied. ■

The following example illustrates the discovery of an asymptotic distribution for a simple form of the *least-squares estimator*, which we will examine in detail in Chapter 8.

**Example 5.52** Let the sequence $\{Y_n\}$ be defined by $Y_i = z_i\beta + \varepsilon_i$, where

**a.** $\beta$ is a real number,

**b.** $z_i$ is the $i$th element in the sequence of real numbers $\{z_n\}$ for which $n^{-1}\sum_{i=1}^n z_i^2 > a > 0\ \forall\, n$ and $|z_i| < d < \infty\ \forall\, i$, and

**c.** $\varepsilon_i$ is the $i$th element in a sequence of iid random variables $\{\varepsilon_n\}$ for which $\mathrm{E}\varepsilon_i = 0$, $\mathrm{var}(\varepsilon_i) = \sigma^2 \in (0,\infty)$, and $P(|\varepsilon_i| \le m) = 1\ \forall\, i$, where $m \in (0,\infty)$.

Find an asymptotic distribution for the least-squares estimator of $\beta$ defined by

$$\hat{\beta}_n = \sum_{i=1}^n z_i Y_i \Big/ \sum_{i=1}^n z_i^2.$$

**Answer**: We transform the problem into a form that allows both an application of a CLT and an application of Def. 5.10 to define an asymptotic distribution for $\hat{\beta}_n$. Note that

$$(\hat{\beta}_n - \beta) = \sum_{i=1}^{n} z_i(Y_i - z_i\beta) \Big/ \sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} z_i\varepsilon_i \Big/ \sum_{i=1}^{n} z_i^2,$$

so that

$$\left(\sum_{i=1}^{n} z_i^2\right)^{1/2} \frac{\left(\hat{\beta}_n - \beta\right)}{\sigma} = \sum_{i=1}^{n} \frac{z_i\varepsilon_i}{\left(\sigma^2 \sum_{i=1}^{n} z_i^2\right)^{1/2}} = \sum_{i=1}^{n} \frac{W_i}{\left(\sum_{i=1}^{n} \text{var}(W_i)\right)^{1/2}},$$

where $W_i = z_i\varepsilon_i$. The CLT of Theorem 5.32 is applicable to this function of $\hat{\beta}_n$. To see this, observe that $\text{E}W_i = 0$ and $\text{var}(W_i) = \sigma^2 z_i^2 \le \sigma^2 d^2 < \infty \ \forall\, i$. Also, $P(|z_i\varepsilon_i| \le dm) = P(|w_i| \le dm) = 1 \ \forall\, i$, and $\sum_{i=1}^{n} \text{var}(W_i) = \sigma^2 \sum_{i=1}^{n} z_i^2 \to \infty$, since $n^{-1} \sum_{i=1}^{n} z_i^2 > a > 0 \ \forall\, n$. Then, by Theorem 5.32,

$$\left(\sum_{i=1}^{n} z_i^2\right)^{1/2} \frac{\left(\hat{\beta}_n - \beta\right)}{\sigma} \xrightarrow{\text{d}} N(0, 1),$$

and by Def. 5.10,

$$\hat{\beta}_n \overset{\text{a}}{\sim} N\left(\beta, \sigma^2 \left(\sum_{i=1}^{n} z_i^2\right)^{-1}\right). \qquad \square$$

The **Liapounov CLT**, given next, relaxes the boundedness assumption of the previous CLT. In this case, a condition on the moments of the random variables in the sequence $\{X_n\}$ is used to imply the Lindberg conditions.

**Theorem 5.33** (Liapounov's CLT) *Let $\{X_n\}$ be a sequence of independent random variables such that* $\text{E}X_i = \mu_i$ *and* $\text{var}(X_i) = \sigma_i^2 < \infty \ \forall\, i$. *If, for some $\delta > 0$,*

$$\lim_{n \to \infty} \frac{\sum_{i=1}^{n} \text{E}|X_i - \mu_i|^{2+\delta}}{\left(\sum_{i=1}^{n} \sigma_i^2\right)^{1+\delta/2}} = 0,$$

*then*

$$\frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i}{\left(\sum_{i=1}^{n} \sigma_i^2\right)^{1/2}} = \frac{n^{1/2}(\bar{X}_n - \bar{\mu}_n)}{\bar{\sigma}_n} \xrightarrow{\text{d}} N(0, 1).$$

***Proof** This follows from the Lindberg CLT by first noting that, for $\delta > 0$ (for the continuous case—the discrete case is similar),

$$\begin{aligned}\int_{(x_i-\mu_i)^2 \geq \varepsilon b_n^2} (x_i - \mu_i)^2 f_i(x_i) dx_i &= \int_{(x_i-\mu_i)^2 \geq \varepsilon b_n^2} |x_i - \mu_i|^{\delta} |x_i - \mu_i|^{-\delta} (x_i - \mu_i)^2 f_i(x_i) dx_i \\ &\leq \left(\varepsilon b_n^2\right)^{-\delta/2} \int_{(x_i-\mu_i)^2 \geq \varepsilon b_n^2} |x_i - \mu_i|^{2+\delta} f_i(x_i) dx_i \\ &\leq \left(\varepsilon b_n^2\right)^{-\delta/2} \mathrm{E}|X_i - \mu_i|^{2+\delta},\end{aligned}$$

where the first inequality follows from the fact that over the range of integration, $(x_i - \mu_i)^2 \geq \varepsilon b_n^2$ implies that $|x_i - \mu_i|^{-\delta} \leq (\varepsilon b_n^2)^{-\delta/2}$, and the second inequality results from adding the *nonnegative* term

$$\left(\varepsilon b_n^2\right)^{-\delta/2} \int_{(x_i-\mu_i)^2 < \varepsilon b_n^2} |x_i - \mu_i|^{2+\delta} f_i(x_i) dx_i.$$

to the right-hand side of the first inequality. Then

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{b_n^2} \sum_{i=1}^{n} \int_{(x_i-\mu_i)^2 \geq \varepsilon b_n^2} (x_i - \mu_i)^2 f_i(x_i) dx_i &\leq \lim_{n\to\infty} \frac{1}{b_n^2} \sum_{i=1}^{n} \left(\varepsilon b_n^2\right)^{-\delta/2} \mathrm{E}|X_i - \mu_i|^{2+\delta} \\ &\leq \varepsilon^{-\delta/2} \lim_{n\to\infty} \sum_{i=1}^{n} \left[ \frac{\mathrm{E}|X_i - \mu_i|^{2+\delta}}{(b_n^2)^{1+\delta/2}} \right].\end{aligned}$$

The assumptions of the Liapounov CLT state that, for some $\delta > 0$,

$$\sum_{i=1}^{n} \frac{\mathrm{E}|X_i - \mu_i|^{2+\delta}}{(b_n^2)^{1+\delta/2}} \to 0 \quad \text{as } n \to \infty,$$

so that, $\forall\, \varepsilon > 0$ the Lindberg condition is met and the Liapounov CLT holds. ∎

An important implication of Liapounov's CLT is that $\sum_{i=1}^{n} X_i$ need not be a sum of *identically* distributed nor *bounded* random variables to have an asymptotic normal density. Under the conditions of the theorem, it follows that

$$S_n = \sum_{i=1}^{n} X_i \overset{a}{\sim} N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right) \quad \text{and} \quad \bar{X}_n \overset{a}{\sim} N\left(\bar{\mu}_n, n^{-1}\bar{\sigma}_n^2\right).$$

**Example 5.53** Let $\{X_n\}$ be a sequence of independent, uniformly distributed random variables such that

$$X_i \sim \frac{1}{2c_i} I_{[-c_i,c_i]}(x_i) \quad \forall\, i,$$

where $c_i \in [\tau, m]$, and $0 < \tau < m < \infty$. Then for $\delta > 0$, $\mathrm{E}|X_i - \mu_i|^{2+\delta} \leq m^{2+\delta}\ \forall\, i$, and $\mathrm{var}(X_i) = c_i^2/3 \geq \tau^2/3 > 0\ \forall\, i$. Letting $\delta = 1$ in Liapounov's CLT,

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} \mathrm{E}|X_i - \mu_i|^3}{\left[\sum_{i=1}^{n} \sigma_i^2\right]^{3/2}} \leq \lim_{n\to\infty} \left(\frac{\sqrt{3}\, m}{\tau}\right)^3 n^{-1/2} = 0,$$

so that (note $\mu_i = 0 \ \forall \ i$)

$$Y_n = \frac{\sum_{i=1}^n X_i}{\left[\sum_{i=1}^n (c_i^2/3)\right]^{1/2}} \xrightarrow{\mathrm{d}} N(0,1) \quad \text{and} \quad \bar{X}_n \overset{\mathrm{a}}{\sim} N\left(0, n^{-2}\sum_{i=1}^n\left(\frac{c_i^2}{3}\right)\right). \qquad \square$$

As an illustration of how either the CLT for bounded random variables or the Liapounov CLT might be applied in practice, consider a short-run production function of a firm and focus on the effect of labor input on production. There may exist a systematic engineering relationship between the quantity of labor applied to the complement of plant equipment and the expected output of the plant, say as $y = f(L)$, for any given time period of plant operation. However, it would undoubtedly be rare that the exact quantity of production expected in an engineering sense will actually be realized. Variations from the expected quantity could occur due to variations in the health, alertness, and general performance level of each of the various employees, the extent of machine failures in any given time period and their general performance level, the varying ability of management to schedule production efficiently on any given day, weather conditions if the production process is affected thereby, and so on. Viewing the overall deviation of total production from the engineering relationship as caused by the summation of a large number of random deviations with finite absolute upper bounds caused by various uncontrolled factors results in a production relationship of the form $Y = f(L) + \varepsilon$, where $\varepsilon$, and thus $Y$, has an asymptotic normal density (assuming the summation of the random deviations can be viewed as a sum of independent, although *not* necessarily identically distributed, random variables).

As in the case of the Lindberg–Levy CLT, bounds have been established on the maximum absolute deviation between the actual value of $P(y_n \leq c)$ for $Y_n = n^{1/2}(\bar{X}_n - \bar{\mu}_n)/\bar{\sigma}_n$, and the approximated value based on the standard normal distribution. Specifically, Zolotarev[14] has shown such an upper bound to be

$$.9051\left(\sum_{j=1}^n \xi_{3_j}\right)\left(\sum_{j=1}^n \sigma_{x_j}^2\right)^{-3/2},$$

where $\xi_{3_j} = \mathrm{E}(|X_j - \mathrm{E}X_j|^3)$ and $\sigma_{x_j}^2$ refer to the $j$th random variable $X_j$. Thus, if one knew the appropriate moments of the random variables $X_1,\ldots,X_n$, a bound on the approximation error could be ascertained. If such information were not available, the researcher knows only that as $n$ increases, the accuracy of the approximation improves. As we had remarked in our discussion of similar bounds for the LLCLT, these bounds tend to be conservative and the actual approximation errors tend to be significantly less than the value of the bound.

[14] M. Zolotarev (1967), "A sharpening of the inequality of Berry-Esséen." *Z. Wahrscheinlichkeits Theorie und Verw. Gebiete*, 8 pp. 332–342.

**Triangular Arrays* In analyzing the asymptotic properties of some types of econometric or statistical procedures, it is useful to be able to apply central limit theory to what is known as a **double array of random variables**. For our purposes, it will be sufficient to examine a special case of such a double array, called a **triangular array of random variables.**[15]

**Definition 5.17**
***Triangular array of random variables***

The ordered collection of random variables $\{X_{11}, X_{21}, X_{22}, X_{31}, X_{32}, X_{33}, \ldots, X_{nn}, \ldots\}$, or

$$\begin{array}{lllllll}
X_{11}; & & & & & & \\
X_{21} & X_{22}; & & & & & \\
X_{31} & X_{32} & X_{33}; & & & & \\
\vdots & \vdots & \vdots & \ddots & & & \\
X_{n1} & X_{n2} & X_{n3} & X_{n4} & \cdots & X_{nn}; & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

is called a triangular array of random variables and will be denoted by $\{X_{nn}\}$.

Central limit theorems applied to triangular arrays of random variables will be concerned with the limiting distributions of appropriately defined functions of the *row averages* $\bar{X}(n) = n^{-1}\sum_{i=1}^{n} X_{ni}$. Note that all of the CLTs examined so far have dealt with functions of averages of the type $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$, the $X_i$'s being elements of the sequence $\{X_n\}$. It is possible that $\bar{X}(n) = \bar{X}_n \; \forall \; n$, which would occur if $x_{ij} = x_j \, \forall \, i, j$, that is, all of the elements in any given *column* of the triangular array are *identical*. Thus, the CLT results obtained heretofore apply to this special case of a triangular array. However, the triangular array $\{X_{nn}\}$ is more general than a sequence $\{X_n\}$ in the sense that the random variables in a row of the array need *not* be the same as any random variables in any other row. Furthermore, $\bar{X}_n$ always involves *all* of the random variables in the sequence $\{X_n\}$ up to the $n$th element, while $\bar{X}(n)$ only involves the random variables *residing in the nth row of the triangular array*. The importance of this flexibility will become apparent when we analyze the asymptotic behavior of certain statistical procedures for which central limit theory can *only* be effectively applied in the context of triangular arrays of random variables.

*All* of the CLTs presented heretofore can be extended to the case of triangular arrays. We present here the extension of the Liapounov CLT. For additional details on such extensions, the reader can refer to the book by K. L. Chung cited in the proof of the theorem. Henceforth, we will let $\bar{\mu}(n) = n^{-1}\sum_{i=1}^{n} \mu_{ni}$ and $\bar{\sigma}^2(n) = n^{-1}\sum_{i=1}^{n} \sigma_{ni}^2$.

[15] A *double array* is one where the second subscript of the random variables in the $i$th row of the array ends with the value $k_i$, rather than with the value $i$ as in the case of the triangular array, and $k_n \rightarrow \infty$ as $n \rightarrow \infty$. See Serfling, Approximation Theorems, p. 31.

**Theorem 5.34** (Liapounov's CLT: Triangular Arrays) *Let* $\{X_{nn}\}$ *be a triangular array of random variables with independent random variables within rows. Let* $\mathrm{E}X_{ij} = \mu_{ij}$ *and* $\mathrm{var}(X_{ij}) = \sigma^2_{ij} < \infty \ \forall\, i,j$. *If, for some* $\delta > 0$,

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n \mathrm{E}|X_{ni} - \mu_{ni}|^{2+\delta}}{\left(\sum_{i=1}^n \sigma^2_{ni}\right)^{1+\delta/2}} = 0,$$

*then*

$$\frac{\sum_{i=1}^n X_{ni} - \sum_{i=1}^n \mu_{ni}}{\left(\sum_{i=1}^n \sigma^2_{ni}\right)^{1/2}} = n^{1/2}\left(\bar{X}(n) - \bar{\mu}(n)\right)/\bar{\sigma}(n) \xrightarrow{\mathrm{d}} N(0,1).$$

**Proof** See K. L. Chung (1974), *A Course in Probability Theory*, 2nd ed. New York: Academic Press, Section 7.2. ■

Note that the random variables *within* a row of the triangular array are assumed to be independent in the Liapounov CLT, but *no such assumption* is required for random variables in *different* rows. In fact, the random variables within a given row can be *arbitrarily dependent* on random variables in *other* rows.

The following example applies the Liapounov CLT for triangular arrays to establish the asymptotic normality of the least-squares estimator under more general conditions than those utilized in Ex. 5.52.

**Example 5.54** Let the sequence $\{Y_n\}$ be defined by $y_i = z_i\beta + \varepsilon_i$, and assume the conditions of Ex. 5.52, except replace the boundedness assumption $P(|\varepsilon_i| \le m) = 1 \ \forall\, i$ with the assumption that $\mathrm{E}|\varepsilon_i|^{2+\delta} \le m < \infty \ \forall\, i$ for some $\delta > 0$. Then the least-squares estimator $\hat{\beta}_n = \sum_{i=1}^n z_i Y_i / \sum_{i=1}^n z_i^2$ remains asymptotically normally distributed. To see this, note that

$$\left(\sum_{i=1}^n z_i^2\right)^{1/2} \frac{\left(\hat{\beta}_n - \beta\right)}{\sigma} = \sum_{i=1}^n z_i\varepsilon_i \Big/ \left(\sigma^2 \sum_{i=1}^n z_i^2\right)^{1/2} = \sum_{i=1}^n W_{ni},$$

where the random variables $W_{ni}$ are elements of a triangular array[16] for which $\mathrm{E}W_{ni} = 0$ and $\sigma^2_{ni} = \mathrm{var}(W_{ni}) = z_i^2/\sum_{i=1}^n z_i^2$. Because $\sum_{i=1}^n \sigma^2_{ni} = 1$, the limit condition of the Liapounov CLT for triangular arrays is met since, for $\delta > 0$,

$$\lim_{n\to\infty} \sum_{i=1}^n \mathrm{E}|W_{ni}|^{2+\delta} = \lim_{n\to\infty} \sum_{i=1}^n |z_i|^{2+\delta}\mathrm{E}|\varepsilon_i|^{2+\delta} \Big/ \left(\sigma^2 \sum_{i=1}^n z_i^2\right)^{1+\delta/2} \le \lim_{n\to\infty} \left[md^{2+\delta}/(\sigma^2 a)^{1+\delta/2}\right] n^{-\delta/2} = 0.$$

Therefore, $\sum_{i=1}^n W_{ni} \xrightarrow{\mathrm{d}} N(0,1)$, which then implies that

$$\hat{\beta} \overset{\mathrm{a}}{\sim} N\left(\beta, \sigma^2\left(\sum_{i=1}^n z_i^2\right)^{-1}\right).$$

□

[16] Note the $W_{ni}$, $i = 1,\ldots,n$, are independent because the $\varepsilon_i$'s are independent.

### *Dependent Random Variables

In this subsection we provide an introduction to the notion of defining CLTs when the random variables in the sequence $\{X_n\}$ exhibit some degree of dependence. Many different CLTs can be defined by allowing different types of dependencies among the random variables in $\{X_n\}$. CLTs for dependent random variables are generally much more complicated to state and prove than CLTs for independent random variables, and the level of mathematics involved is beyond the scope of our study. We will explicitly examine one useful CLT for a particular type of dependence called $m$-dependence. The reader can find additional reading on CLTs for the dependent random variable case in H. White, Asymptotic Theory, Chapters 3 and 5, and in R. J. Serfling (1968), "Contributions to central limit theory for dependent variables." *Ann. of Math. Stat.* 39 pp. 1158–1175.

**Definition 5.18** ***M-Dependence***

The sequence $\{X_n\}$ is said to exhibit $m$-dependence (or is said to be $m$-dependent) if, for $a_1 < a_2 < \cdots < a_k < b_1 < b_2 < \cdots < b_r$, $(X_{a_1}, X_{a_2}, \ldots, X_{a_k})$ is independent of $(X_{b_1}, X_{b_2}, \ldots, X_{b_r})$ whenever $b_1 - a_k > m$.

The definition states that $\{X_n\}$ is $m$-dependent if any two groups of random variables separated by more than $m$ positions in the sequence are independent of one another. Combining $m$-dependence with boundedness of the random variables in the sequence $\{X_n\}$ leads to the following CLT.

**Theorem 5.35** (CLT for Bounded $M$-Dependent Sequences) *Let $\{X_n\}$ be an $m$-dependent sequence of random scalars for which* $\mathrm{E}X_i = \mu_i$ *and* $P(|x_i| \leq c) = 1$ *for some* $c < \infty$ $\forall$ *i. Let* $\sigma^2_{*n} = \mathrm{var}\left(\sum_{i=1}^n X_i\right)$. *If* $n^{-2/3}\sigma^2_{*n} \to \infty$, *then*

$$(\sigma^2_{*n})^{-1/2}\left(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i\right) \xrightarrow{d} N(0,1).$$

**Proof** See R. G. Laha and V. K. Rohatgi, *Probability Theory*, p. 355, Section 5.5.31. ■

Regarding the variance condition stated in the theorem, note that if the $X_i$'s are *independent* and $\sigma_i^2 \geq b > 0 \ \forall\ i$, then $\sigma^2_{*n} \geq bn \to \infty$ at a rate of $n$, so that $n^{-2/3}\sigma^2_{*n} \geq bn^{1/3} \to \infty$ at a rate of $n^{1/3}$. If we let the $X_i$'s be *dependent*, then as long as the covariance terms in the definition of $\mathrm{var}(\sum_{i=1}^n X_i)$ do not collectively *decrease* the rate at which $\sigma^2_{*n}$ increases by a factor of $n^{1/3}$ or more, the variance condition of the CLT will hold. Thus, through restricting covariance terms, the variance condition of Theorem 5.35 places restrictions on the extent of the dependence that can exist between the $X_i$'s. Note that the restriction is effectively on *negative* covariances, since positive covariances actually increase the variance of $\sum_{i=1}^n X_i$.

**Example 5.55** Let the sequence $\{Y_n\}$ be defined by $Y_i = z_i\beta + \varepsilon_i$, and assume the conditions of Ex. 5.52, except replace the assumption $\mathrm{var}(\varepsilon_i) = \sigma^2 \in (0,\infty)$ with $\mathrm{var}(\varepsilon_i) = \sigma_i^2 \geq$

$\tau > 0 \;\forall\, i$, replace the assumption that the $\varepsilon_i$'s are iid with the assumption of $m$-dependence, and in addition assume that $z_i > 0 \;\forall\, i$ and $\text{cov}(\varepsilon_i, \varepsilon_j) \geq 0 \;\forall\, i \neq j$. From Theorem 5.35, it follows that

$$\hat{\beta} \overset{a}{\sim} N\left(\beta, \sigma_{*n}^2 \left(\sum_{i=1}^{n} z_i^2\right)^{-2}\right).$$

To see this, note that

$$\left(\sum_{i=1}^{n} z_i^2\right)\left(\hat{\beta}_n - \beta\right) = \sum_{i=1}^{n} z_i\varepsilon_i = \sum_{i=1}^{n} W_i,$$

where $\text{E}W_i = 0$ and $P(|z_i\varepsilon_i| \leq db) = P(|w_i| \leq c) = 1$ for $c = db < \infty \;\forall\, i$. Also,

$$\sigma_{*n}^2 = \text{var}\left(\sum_{i=1}^{n} W_i\right) = \sum_{i=1}^{n} \sigma_i^2 z_i^2 + \sum_{|i-j| \leq m, i \neq j}\sum z_i z_j \,\text{cov}(\varepsilon_i, \varepsilon_j)$$

since, by $m$-dependence, $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ when $|i-j| > m$. Because, $z_i z_j \,\text{cov}(\varepsilon_i, \varepsilon_j) \geq 0 \;\forall\, i$ and $j$, $\sigma_{*n}^2 \geq \sum_{i=1}^{n} \sigma_i^2 z_i^2 \geq na\tau$, so that $n^{-2/3}\sigma_{*n}^2 \geq n^{1/3}a\tau \to \infty$. Thus, by Theorem 5.35,

$$\left(\sum_{i=1}^{n} z_i^2\right)\frac{\left(\hat{\beta}_n - \beta\right)}{(\sigma_{*n}^2)^{1/2}} = \sum_{i=1}^{n} \frac{W_i}{(\sigma_{*n}^2)^{1/2}} \xrightarrow{d} N(0, 1),$$

so that $\hat{\beta}_n \overset{a}{\sim} N(\beta, \sigma_{*n}^2(\sum_{i=1}^{n} z_i^2)^{-2})$. □

### Multivariate Central Limit Results

The central limit theorems presented so far are applicable to sequences of random *scalars*. Central limit theorems can be defined for sequences of random *vectors*, in which case conditions are established to ensure that an appropriate (vector) function of the random sequence converges in distribution to a multivariate normal distribution. Due to a result discovered by H. Cramér and H. Wold,[17] and termed the **Cramér–Wold device**, questions of convergence in distribution for a multivariate random sequence can all be reduced to the question of convergence in distribution of sequences of random scalars, at least in principle. Thus, all of the central limit theorems discussed to this point remain highly relevant to the multivariate case.

**Theorem 5.36** (Cramér–Wold's Device) *The sequence of* $(k \times 1)$ *random vectors* $\{\mathbf{X}_n\}$ *converges in distribution to the random* $(k \times 1)$ *vector* $\mathbf{X}$ *iff* $\boldsymbol{\ell}'\mathbf{X}_n \xrightarrow{d} \boldsymbol{\ell}'\mathbf{X} \;\forall\, \boldsymbol{\ell} \in R^k$.

**Proof** Sufficiency will be motivated assuming the existence of MGFs. The general proof replaces MGFs with characteristic functions. From Theorem 5.2,

[17] H. Cramér and H. Wold (1936), "Some theorems on distribution functions." *J. London Math. Soc.*, 11, pp. 290–295.

$\boldsymbol{\ell}'\mathbf{X}_n \xrightarrow{d} \boldsymbol{\ell}'\mathbf{X} \Rightarrow M_{\boldsymbol{\ell}'\mathbf{X}_n}(t) \to M_{\boldsymbol{\ell}'X}(t)$ for $t \in (-h, h)$, $h > 0$, and $\forall\, \boldsymbol{\ell} \in R^k$. This implies

$$\mathrm{E}e^{t\boldsymbol{\ell}'\mathbf{X}_n} = \mathrm{E}e^{\mathbf{t}_*'\mathbf{X}_n} \to \mathrm{E}e^{\mathbf{t}_*'\mathbf{X}} = \mathrm{E}e^{t\boldsymbol{\ell}'\mathbf{X}}$$

$\forall\, \mathbf{t}_* = t\boldsymbol{\ell}$, since $\boldsymbol{\ell}$ can be chosen arbitrarily. But this is equivalent to $M_{\mathbf{X}_n}(\mathbf{t}_*) \to M_{\mathbf{X}}(\mathbf{t}_*)$, $\forall\, \mathbf{t}_*$, which implies $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ by the multivariate interpretation of Theorem 5.2. Necessity follows from Theorem 5.3 since $\boldsymbol{\ell}'\mathbf{x}$ is a continuous function of $\mathbf{x}$. The general proof based on characteristic functions can be found in V. Fabian and J. Hannan (1985), *Introduction to Probability and Mathematical Statistics*. New York: John Wiley, p. 144. ■

Note that in applying Theorem 5.36, $\boldsymbol{\ell}'\mathbf{X}_n \xrightarrow{d} \boldsymbol{\ell}'\mathbf{X}$ is *always* trivially true when $\boldsymbol{\ell} = [\mathbf{0}]$, and so the condition $\boldsymbol{\ell}'\mathbf{X}_n \xrightarrow{d} \boldsymbol{\ell}'\mathbf{X}$ need only be checked for $\boldsymbol{\ell} \neq [\mathbf{0}]$. We will be most concerned with convergence in distribution to members of the normal family of distributions. In this context, Theorem 5.36 implies that to establish convergence in distribution of the sequence of random $(k \times 1)$ vectors $\{\mathbf{X}_n\}$ to the random $(k \times 1)$ vector $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, it suffices to demonstrate that $\boldsymbol{\ell}'\mathbf{X}_n \xrightarrow{d} N(\boldsymbol{\ell}'\boldsymbol{\mu}, \boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell})\ \forall\, \boldsymbol{\ell} \in R^k$. We formalize this observation as a corollary to Theorem 5.36.

**Corollary 5.4** (Cramér–Wold Device for Normal Limiting Distributions) $\mathbf{X}_n \xrightarrow{d} N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ *iff* $\boldsymbol{\ell}'\mathbf{X}_n \xrightarrow{d} N(\boldsymbol{\ell}'\boldsymbol{\mu},\boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell})\ \forall\, \boldsymbol{\ell} \in R^k$.

The Cramér–Wold device can be used to define multivariate central limit theorems. The following is a multivariate extension of the Lindberg–Levy CLT.

**Theorem 5.37** (Multivariate Lindberg–Levy CLT) *Let* $\{\mathbf{X}_n\}$ *be a sequence of iid* $(k \times 1)$ *random vectors with* $\mathrm{E}\mathbf{X}_i = \boldsymbol{\mu}$ *and* $\mathbf{Cov}(\mathbf{X}_i) = \boldsymbol{\Sigma}\ \forall\, i$, *where* $\boldsymbol{\Sigma}$ *is a* $(k \times k)$ *positive definite matrix. Then*

$$n^{1/2}\left(n^{-1}\sum_{i=1}^{n}\mathbf{X}_i - \boldsymbol{\mu}\right) \xrightarrow{d} N([\mathbf{0}], \boldsymbol{\Sigma}).$$

**Proof** Examine $Z_i = \boldsymbol{\ell}'\mathbf{X}_i$, where $\boldsymbol{\ell} \neq [\mathbf{0}]$. Note that $\mu_z = \mathrm{E}Z_i = \mathrm{E}\boldsymbol{\ell}'\mathbf{X}_i = \boldsymbol{\ell}'\boldsymbol{\mu}$ and $\sigma_z^2 = \mathrm{var}(Z_i) = \mathrm{var}(\boldsymbol{\ell}'\mathbf{X}_i) = \boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell}\ \forall\, i$. Now, since $\{\mathbf{X}_n\}$ is a sequence of iid random vectors, then $\{Z_n\} = \{\boldsymbol{\ell}'\mathbf{X}_n\}$ is a sequence of iid random scalars, and applying the Lindberg–Levy CLT for random scalars to the iid sequence $\{Z_n\}$ results in

$$\frac{\sum_{i=1}^{n} Z_i - n\mu_z}{n^{1/2}\sigma_z} = \frac{\sum_{i=1}^{n}\boldsymbol{\ell}'\mathbf{X}_i - n\boldsymbol{\ell}'\boldsymbol{\mu}}{n^{1/2}(\boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell})^{1/2}} = \frac{\boldsymbol{\ell}'\left[\sum_{i=1}^{n}\mathbf{X}_i - n\boldsymbol{\mu}\right]}{n^{1/2}\,(\boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell})^{1/2}} = \frac{\boldsymbol{\ell}'n^{1/2}\left[n^{-1}\sum_{i=1}^{n}\mathbf{X}_i - \boldsymbol{\mu}\right]}{(\boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell})^{1/2}} \xrightarrow{d} N(0, 1).$$

Then, by Slutsky's theorem

$$\boldsymbol{\ell}'n^{1/2}\left[n^{-1}\sum_{i=1}^{n}\mathbf{X}_i - \boldsymbol{\mu}\right] \xrightarrow{d} N(0, \boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell}),$$

which holds for any choice of the vector $\boldsymbol{\ell} \neq [\mathbf{0}]$. By the Cramér–Wold device, we can conclude that

$$n^{1/2}\left[n^{-1}\sum_{i=1}^{n}\mathbf{X}_i - \boldsymbol{\mu}\right] \xrightarrow{d} N([\mathbf{0}], \boldsymbol{\Sigma}). \quad \blacksquare$$

It follows from the multivariate Lindberg–Levy CLT that $\bar{\mathbf{X}}_n \overset{a}{\sim} N(\boldsymbol{\mu}, n^{-1}\boldsymbol{\Sigma})$.

**Example 5.56** Shipments of CPU chips from two different suppliers are to be inspected before being accepted. Chips are randomly drawn, with replacement, from each shipment and are nondestructively tested in pairs, with $(X_{1i}, X_{2i})$ representing the outcome of the tests for pair $i$. An outcome of $x_{\ell i} = 1$ indicates a faulty chip, $x_{\ell i} = 0$ indicates a nondefective chip, and the joint density of $(X_{1i}, X_{2i})$ is given by

$$\mathbf{X}_i = \begin{bmatrix} X_{1i} \\ X_{2i} \end{bmatrix} \sim p_1^{x_{1i}}(1-p_1)^{1-x_{1i}} I_{\{0,1\}}(x_{1i}) p_2^{x_{2i}}(1-p_2)^{1-x_{2i}} I_{\{0,1\}}(x_{2i}),$$

where $p_i \in (0,1)$ for $i = 1, 2$. Note that

$$\mathrm{E}\mathbf{X}_i = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = p \quad \text{and} \quad \mathbf{Cov}(\mathbf{X}_i) = \boldsymbol{\Sigma} = \begin{bmatrix} p_1(1-p_1) & 0 \\ 0 & p_2(1-p_2) \end{bmatrix}.$$

With

$$\underset{(2\times 1)}{\bar{\mathbf{X}}_n} = n^{-1}\sum_{i=1}^{n}\begin{bmatrix} X_{1i} \\ X_{2i} \end{bmatrix},$$

it follows from the multivariate Lindberg–Levy CLT that

$$\mathbf{Z}_n = n^{1/2}[\bar{\mathbf{X}}_n - \mathbf{p}] \xrightarrow{d} \mathbf{Z} \sim N([\mathbf{0}], \boldsymbol{\Sigma})$$

and also that

$$\bar{\mathbf{X}}_n \overset{a}{\sim} N(\mathbf{p}, n^{-1}\boldsymbol{\Sigma}).$$

If one were interested in establishing an asymptotic distribution for the difference in the number of defectives observed in $n$ random pairs of CPUs from shipments 1 and 2, the random variable of interest would be $\mathbf{c}'(n\bar{\mathbf{X}}_n) = \sum_{i=1}^{n} X_{1i} - \sum_{i=1}^{n} X_{2i}$, where $\mathbf{c}' = [1 \ \ -1]$. Then

$$\mathbf{c}'(n\bar{\mathbf{X}}_n) = g(\mathbf{Z}_n, n) = \mathbf{c}'[n^{1/2}\mathbf{Z}_n + n\mathbf{p}],$$

so that, by Def. 5.10,

$$\mathbf{c}'(n\bar{\mathbf{X}}_n) \overset{a}{\sim} \mathbf{c}'\left[n^{1/2}\mathbf{Z} + n\mathbf{p}\right] \sim N(n\mathbf{c}'\mathbf{p}, n\mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}).$$

Thus, the asymptotic distribution for the difference in the number of defectives is

$$\mathbf{c}'(n\bar{\mathbf{X}}_n) \overset{a}{\sim} N(n(p_1 - p_2), n[p_1(1-p_1) + p_2(1-p_2)]). \quad \square$$

Another useful multivariate CLT concerns the case where the elements in the sequence $\{\mathbf{X}_n\}$ are independent but not necessarily identically distributed $(k \times 1)$ random vectors that exhibit uniform (i.e., across all $n$) absolute upper bounds with probability 1.

**Theorem 5.38** (Multivariate CLT: Independent Bounded Random Vectors) *Let $\{\mathbf{X}_n\}$ be a sequence of independent $(k \times 1)$ random vectors such that $p(|x_{1i}| \leq m, |x_{2i}| \leq m, \ldots, |x_{ki}| \leq m) = 1 \; \forall \; i$, where $m \in (0,\infty)$. Let $\mathrm{E}\mathbf{X}_i = \boldsymbol{\mu}_i$, $\mathbf{Cov}(\mathbf{X}_i) = \boldsymbol{\psi}_i$, and suppose that $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} \boldsymbol{\psi}_i = \boldsymbol{\psi}$, a finite, positive definite $(k \times k)$ matrix. Then*

$$n^{-1/2}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}_i) \xrightarrow{\mathrm{d}} N([\mathbf{0}], \boldsymbol{\psi}).$$

**Proof** Examine $Z_i = \boldsymbol{\ell}'\mathbf{X}_i$, where $\boldsymbol{\ell} \neq [\mathbf{0}]$. Note that $\mathrm{E}Z_i = \mathrm{E}\boldsymbol{\ell}'\mathbf{X}_i = \boldsymbol{\ell}'\boldsymbol{\mu}_i$ and $\mathrm{var}(Z_i) = \mathrm{var}(\boldsymbol{\ell}'\mathbf{X}_i) = \boldsymbol{\ell}'\boldsymbol{\psi}_i\boldsymbol{\ell}$. Since $\{\mathbf{X}_n\}$ is a sequence of independent random vectors, $\{Z_n\} = \{\boldsymbol{\ell}'\mathbf{X}_n\}$ is a sequence of independent random scalars. Furthermore, since outcomes of the vector $\mathbf{X}_n$ are contained within the closed and bounded rectangle $\times_{i=1}^{k}[-m, m]$ in $R^k$ with probability 1, then for any given nonzero vector of real numbers $\boldsymbol{\ell}$, outcomes of $Z_n = \boldsymbol{\ell}'\mathbf{X}_n$ exhibit an upper bound in absolute value with probability 1 uniformly $\forall \; n$. Thus, there exists a finite real number $\delta > 0$ such that $P(|z_i| \leq \delta) = 1 \; \forall \; i$. In addition, since $n^{-1}\sum_{i=1}^{n}\mathrm{var}(Z_i) = n^{-1}\sum_{i=1}^{n}\boldsymbol{\ell}'\boldsymbol{\psi}_i\boldsymbol{\ell} \to \boldsymbol{\ell}'\boldsymbol{\psi}\boldsymbol{\ell} > 0$, $\sum_{i=1}^{n}\mathrm{var}(Z_i) \to \infty$. It follows from the CLT of Theorem 5.32 that

$$\frac{\sum_{i=1}^{n} Z_i - \sum_{i=1}^{n}\boldsymbol{\ell}'\boldsymbol{\mu}_i}{\left(\sum_{i=1}^{n}\mathrm{var}(Z_i)\right)^{1/2}} = \frac{\boldsymbol{\ell}'\left(\sum_{i=1}^{n}\mathbf{X}_i - \sum_{i=1}^{n}\boldsymbol{\mu}_i\right)}{\left(\sum_{i=1}^{n}\boldsymbol{\ell}'\boldsymbol{\psi}_i\boldsymbol{\ell}\right)^{1/2}}$$

$$= \frac{\boldsymbol{\ell}'n^{-1/2}\left(\sum_{i=1}^{n}\mathbf{X}_i - \sum_{i=1}^{n}\boldsymbol{\mu}_i\right)}{\left(\sum_{i=1}^{n}\boldsymbol{\ell}'\left(\frac{1}{n}\boldsymbol{\psi}_i\right)\boldsymbol{\ell}\right)^{1/2}} \xrightarrow{\mathrm{d}} N(0, 1).$$

Premultiplying by the denominator term, and noting that

$$\lim_{n\to\infty}\left(\sum_{i=1}^{n}\boldsymbol{\ell}'\left(\frac{1}{n}\boldsymbol{\psi}_i\right)\boldsymbol{\ell}\right)^{1/2} = (\boldsymbol{\ell}'\boldsymbol{\psi}\boldsymbol{\ell})^{1/2},$$

Slutsky's theorem (Theorem 5.10) then results in

$$\boldsymbol{\ell}'n^{-1/2}\left(\sum_{i=1}^{n}\mathbf{X}_i - \sum_{i=1}^{n}\boldsymbol{\mu}_i\right) \xrightarrow{\mathrm{d}} N(0, \boldsymbol{\ell}'\boldsymbol{\psi}\boldsymbol{\ell}),$$

which holds for any choice of the real vector $\boldsymbol{\ell} \neq [\mathbf{0}]$. It follows by the Cramér–Wold device that

$$n^{-1/2}\left(\sum_{i=1}^{n}\mathbf{X}_i - \sum_{i=1}^{n}\boldsymbol{\mu}_i\right) \xrightarrow{\mathrm{d}} N([\mathbf{0}], \boldsymbol{\psi}).$$

■

Various other multivariate CLTs can be constructed using the Cramér–Wold device and CLTs for random scalars. In practice, one often relies on the Cramér–Wold device directly for establishing limiting distributions relating to statistical procedures of interest, and so we will not attempt to compile a list of additional multivariate CLTs here.

## 5.6 Asymptotic Distributions of Differentiable Functions of Asymptotically Normally Distributed Random Variables

In this section we examine results concerning the asymptotic distributions of differentiable functions of asymptotically normally distributed random variables. General conditions will be identified for which differentiable functions of asymptotically normally distributed random variables are themselves asymptotically normally distributed. The utility of these results in practice is that once the asymptotic distribution of $X_n$ is known, the asymptotic distributions of interesting functions of $X_n$ need not be derived anew. Instead, these asymptotic distributions can generally be defined by specifying the mean and the covariance matrix of a normal distribution according to well-defined and straightforwardly implemented formulas.

All of the results that we will examine in this section are based on *first-order* Taylor series expansions of the function $g(x)$ around a point $\mu$. We review the Taylor series expansion concept here, paying particular attention to the nature of the *remainder term*. Recall that $d(x,\mu) = [(x-\mu)'(x-\mu)]^{1/2}$ represents the distance between the points $x$ and $\mu$.

**Lemma 5.6** *First-Order Taylor Series Expansion and Remainder (Young's Form)*

Let $g: D \rightarrow R$ be a function having partial derivatives in a neighborhood of the point $\mu \in D$ that are continuous at $\mu$. Let $\mathbf{G} = [\partial g(\mu)/\partial x_1, \ldots, \partial g(\mu)/\partial x_k]$ be the gradient vector of $g(\mathbf{x})$ evaluated at the point $\mathbf{x} = \mu$. For $\mathbf{x} \in D$, define the *remainder term* $R(\mathbf{x})$ via

$$g(\mathbf{x}) = g(\mu) + \mathbf{G}(\mathbf{x} - \mu) + d(\mathbf{x}, \mu)R(\mathbf{x}),$$

with $R(\mu) = 0$. Then $R(\mathbf{x})$ is continuous at $\mathbf{x} = \mu$ and $\lim_{\mathbf{x}\rightarrow\mu} R(\mathbf{x}) = R(\mu) = 0$.

**Proof** Young's form of Taylor's theorem is not prevalent in calculus texts. G. H. Hardy (1952), *A Course of Pure Mathematics*, 10th ed. Cambridge, New York, The University Press, p. 278, proves the case where $x$ is a scalar. For a brief sketch of the proof in the multivariate case, first note that $R(\mathbf{x}) = (g(\mathbf{x}) - g(\mu) - \mathbf{G}(\mathbf{x} - \mu))/d(\mathbf{x}, \mu)$. Under the stated conditions of the theorem (partial derivatives exist in a neighborhood of $\mu$ and are continuous at $\mu$), the function $g(\mathbf{x})$ is *differentiable* at $\mu$, and the defining property of differentiability at $\mu$ is that $\lim_{\mathbf{x}\rightarrow\mu} R(\mathbf{x}) = 0$ (T. M. Apostol (1957), *Mathematical Analysis*. Cambridge, MA: Addison-Wesley, pp. 110 and 118). Thus, $R(\mathbf{x})$ is continuous at $\mathbf{x} = \mu$.

Our first result on asymptotic distributions of $g(\mathbf{X})$ concerns the case where $g(\mathbf{x})$ is a *scalar*-valued function. As will be common to all of the results we will examine, the principal requirement on the function $g(\mathbf{x})$ is that partial derivatives exist in a neighborhood of the point $\boldsymbol{\mu}$ and that they are continuous at $\boldsymbol{\mu}$ so that Lemma 5.6 can be utilized. In addition, we will also make assumptions relating to the nature of the asymptotic distribution of $\mathbf{X}$.

**Theorem 5.39** (Asymptotic Distribution of $g(\mathbf{X}_n)$—Scalar Function Case) *Let $\{\mathbf{X}_n\}$ be a sequence of $(k \times 1)$ random vectors such that $n^{1/2}(\mathbf{X}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{Z} \sim N([\mathbf{0}], \boldsymbol{\Sigma})$. Let $g(\mathbf{x})$ have first-order partial derivatives in a neighborhood of the point $\mathbf{x} = \boldsymbol{\mu}$ that are continuous at $\boldsymbol{\mu}$, and suppose the gradient vector of $g(\mathbf{x})$ evaluated at $\mathbf{x} = \boldsymbol{\mu}$, $\mathbf{G}_{(1\times k)} = [\partial g(\boldsymbol{\mu})/\partial x_1 \cdots \partial g(\boldsymbol{\mu})/\partial x_k]$, is not the zero vector. Then*

$$n^{1/2}(g(\mathbf{X}_n) - g(\boldsymbol{\mu})) \xrightarrow{d} N(0, \mathbf{G}\boldsymbol{\Sigma}\mathbf{G}') \quad \textit{and} \quad g(\mathbf{X}_n) \overset{a}{\sim} N(g(\boldsymbol{\mu}), n^{-1}\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}').$$

**Proof** Representing $g(\mathbf{X}_n)$ in terms of a first-order Taylor series expansion around the point $\boldsymbol{\mu}$ and the remainder term, as in Lemma 5.6, yields

$$g(\mathbf{X}_n) = g(\boldsymbol{\mu}) + \mathbf{G}(\mathbf{X}_n - \boldsymbol{\mu}) + d(\mathbf{X}_n, \boldsymbol{\mu})R(\mathbf{X}_n).$$

Multiplying by $n^{1/2}$ and rearranging terms obtain

$$n^{1/2}(g(\mathbf{X}_n) - g(\boldsymbol{\mu})) = \mathbf{G}\left[n^{1/2}(\mathbf{X}_n - \boldsymbol{\mu})\right] + n^{1/2}d(\mathbf{X}_n, \boldsymbol{\mu})R(\mathbf{X}_n).$$

The last term converges to zero in probability. To see this, first note by Slutsky's theorem that $\mathbf{X}_n - \boldsymbol{\mu} \xrightarrow{d} \operatorname{plim}(n^{-1/2})\mathbf{Z} = 0 \cdot \mathbf{Z} = [\mathbf{0}]$, so that $\operatorname{plim} \mathbf{X}_n = \boldsymbol{\mu}$ and then $\operatorname{plim} R(\mathbf{X}_n) = R(\boldsymbol{\mu}) = 0$ by the continuity of $R(\mathbf{x})$ at the point $\mathbf{x} = \boldsymbol{\mu}$. Also, by Theorem 5.3,

$$n^{1/2}d(\mathbf{X}_n, \boldsymbol{\mu}) = \left(\left[n^{1/2}(\mathbf{X}_n - \boldsymbol{\mu})\right]'\left[n^{1/2}(\mathbf{X}_n - \boldsymbol{\mu})\right]\right)^{1/2} \xrightarrow{d} (\mathbf{Z}'\mathbf{Z})^{1/2},$$

so that by Slutsky's theorem

$$n^{1/2}d(\mathbf{X}_n, \boldsymbol{\mu})R(\mathbf{X}_n) \xrightarrow{d} (\mathbf{Z}'\mathbf{Z})^{1/2} \cdot \operatorname{plim} R(\mathbf{X}_n) = (\mathbf{Z}'\mathbf{Z})^{1/2} \cdot 0 = 0$$

$$\xrightarrow{p} 0.$$

By continued applications of Slutsky's theorem, it follows that the limiting distribution of $n^{1/2}(g(\mathbf{X}_n) - g(\boldsymbol{\mu}))$ is the same as that of

$$\mathbf{G}[n^{1/2}(\mathbf{X}_n - \boldsymbol{\mu})] \xrightarrow{d} \mathbf{GZ} \sim N([0], \mathbf{G}\boldsymbol{\Sigma}\mathbf{G}').$$

The asymptotic distribution of $g(\mathbf{X}_n)$ is then as stated in the theorem. ■

**Example 5.57** Note from Ex. 5.50 that if $\{X_n\}$ is a sequence of iid Bernoulli-type random variables, then for $p \neq 0$ or 1, $n^{1/2}(\bar{X}_n - p) \xrightarrow{d} N(0, p(1-p))$. Consider using an outcome of $g(\bar{X}) = \bar{X}(1 - \bar{X})$ as an estimate of the variance $p(1-p)$ of the Bernoulli PDF, and define an asymptotic distribution for $g(\bar{X}_n)$.

**Answer**: Theorem 5.39 applies with $\mu = p$ and $\Sigma = \sigma^2 = p(1-p)$. Note that $dg(p)/d\bar{X} = 1 - 2p$, which is continuous in $p$ and is nonzero as long as $p \neq .5$. Also, $\sigma^2 \neq 0$ if $p \neq 0$ or 1. Then, for $p \neq 0$, .5, or 1, Theorem 5.39 implies that

$$\bar{X}_n(1-\bar{X}_n) \overset{a}{\sim} N(p(1-p), n^{-1}(1-2p)^2 p(1-p)).$$

An asymptotic density for $\bar{X}_n(1-\bar{X}_n)$ under the assumption $p = 1/2$ can be established using other methods (see Bickel and Doksum (1977), *Mathematical Statistics*. San Francisco: Holden-Day, p. 53); however, convergence is not to a normal distribution. Specifically, it can be shown that $n[\bar{X}_n(1-\bar{X}_n)-(1/4)] \xrightarrow{d} Z$, where $Z$ has the density of a $\chi^2_1$ random variable that has been multiplied by $(-1/4)$. If $p = 0$ or $p = 1$, the $X_i$'s are all degenerate random variables equal to 0 or 1, respectively, and the limiting density of $\bar{X}_n$ is then degenerate at 0 or 1 as well. □

By reinterpreting $\mathbf{g}(\mathbf{x})$ as a vector function and $\mathbf{G}$ as a Jacobian matrix, the conclusion of Theorem 5.39 regarding the asymptotic distribution of the *vector* function remains valid. The extension allows one to define the joint asymptotic distribution of the random vector $\mathbf{g}(\mathbf{X}_n)$.

**Theorem 5.40** (Asymptotic Distribution of $\mathbf{g}(\mathbf{X}_n)$—Vector Function Case) *Let $\{\mathbf{X}_n\}$ be a sequence of $(k \times 1)$ random vectors such that $n^{1/2}(\mathbf{X}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{Z} \sim N([0],\Sigma)$. Let $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}),\ldots,g_m(\mathbf{x}))'$ be an $(m \times 1)$ vector function $(m \leq k)$ having first-order partial derivatives in a neighborhood of the point $\mathbf{x} = \boldsymbol{\mu}$ that are continuous at $\boldsymbol{\mu}$. Let the Jacobian matrix of $\mathbf{g}(\mathbf{x})$ evaluated at $\mathbf{x} = \boldsymbol{\mu}$,*

$$\underset{m\times k}{\mathbf{G}} = \begin{bmatrix} \partial g_1(\boldsymbol{\mu})/\partial \mathbf{x}' \\ \vdots \\ \partial g_m(\boldsymbol{\mu})/\partial \mathbf{x}' \end{bmatrix} = \begin{bmatrix} \dfrac{\partial g_1(\boldsymbol{\mu})}{\partial x_1} & \cdots & \dfrac{\partial g_1(\boldsymbol{\mu})}{\partial x_k} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial g_m(\boldsymbol{\mu})}{\partial x_1} & \cdots & \dfrac{\partial g_m(\boldsymbol{\mu})}{\partial x_k} \end{bmatrix},$$

*have full row rank.*

*Then*

$$n^{1/2}(\mathbf{g}(\mathbf{X}_n) - \mathbf{g}(\boldsymbol{\mu})) \xrightarrow{d} N([0], \mathbf{G}\Sigma\mathbf{G}') \quad \textit{and} \quad \mathbf{g}(\mathbf{X}_n) \overset{a}{\sim} N(\mathbf{g}(\boldsymbol{\mu}), n^{-1}\mathbf{G}\Sigma\mathbf{G}').$$

**Proof** Following the proof of Theorem 5.39, a first-order Taylor series expansion applied to each coordinate function in $\mathbf{g}(\mathbf{X}_n)$ results in

$$n^{1/2}(\mathbf{g}(\mathbf{X}_n) - \mathbf{g}(\boldsymbol{\mu})) = \mathbf{G}[n^{1/2}(\mathbf{X}_n - \boldsymbol{\mu})] + n^{1/2}d(\mathbf{X}_n, \boldsymbol{\mu})\mathbf{R}(\mathbf{X}_n),$$

where $\mathbf{R}(\mathbf{X}_n) = (R_1(\mathbf{X}_n), \ldots, R_m(\mathbf{X}_n))'$ is now an $(m \times 1)$ vector of remainder terms. The approach used in the proof of Theorem 5.39 can be applied elementwise to conclude that $\text{plim}[n^{1/2}d(\mathbf{X}_n, \boldsymbol{\mu})\mathbf{R}(\mathbf{X}_n)] = [0]$. Then, by Slutsky's theorem, the limiting distribution of $n^{1/2}(\mathbf{g}(\mathbf{X}_n)-\mathbf{g}(\boldsymbol{\mu}))$ is the same as the limiting distribution of $\mathbf{G}[n^{1/2}(\mathbf{X}_n - \boldsymbol{\mu})]$, and

$$\mathbf{G}[n^{1/2}(\mathbf{X}_n - \boldsymbol{\mu})] \xrightarrow{d} \mathbf{G}\mathbf{Z} \sim N([0], \mathbf{G}\Sigma\mathbf{G}').$$

The asymptotic distribution of $\mathbf{g}(\mathbf{X}_n)$ follows directly. ■

**Example 5.58** Let $\{\hat{\beta}_n\}$ be a sequence of $(2 \times 1)$ random vectors such that $n^{1/2}(\hat{\beta}_n - \beta) \xrightarrow{d} N([0], \Sigma)$, where $\beta = [2\ 1]'$ and $\Sigma = \left[\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right]$. Find an asymptotic distribution for the vector function $g(\hat{\beta}) = (3\hat{\beta}[1]\hat{\beta}[2] \mid \hat{\beta}[2]/\hat{\beta}[1])'$.

**Answer**: All of the conditions of Theorem 5.40 are met, including the fact that

$$\underset{2\times 2}{G} = \begin{bmatrix} dg_1(\beta)/d\beta' \\ dg_2(\beta)/d\beta' \end{bmatrix} = \begin{bmatrix} 3\beta_2 & 3\beta_1 \\ -\beta_2/\beta_1^2 & 1/\beta_1 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ -1/4 & 1/2 \end{bmatrix}$$

has full row rank (note that the partial derivatives exist in an open rectangle containing $\beta$, and they are continuous at the point $\beta$). Then, since $g(\beta) = [6\ 1/2]'$ and $\mathbf{G}\Sigma\mathbf{G}' = \left[\begin{smallmatrix} 90 & 1.5 \\ 1.5 & .125 \end{smallmatrix}\right]$, it follows from Theorem 5.40 that

$$g(\hat{\beta}_n) = \begin{bmatrix} 3\hat{\beta}_n[1]\hat{\beta}_n[2] \\ \hat{\beta}_n[2]/\hat{\beta}_n[1] \end{bmatrix} \overset{a}{\sim} N\left(\begin{bmatrix} 6 \\ 1/2 \end{bmatrix}, n^{-1}\begin{bmatrix} 90 & 1.5 \\ 1.5 & .125 \end{bmatrix}\right).$$

A specific distribution is obtained once $n$ is specified. For example, if $n = 20$, then

$$g\left(\hat{\beta}_{20}\right) \overset{a}{\sim} N\left(\begin{bmatrix} 6 \\ 1/2 \end{bmatrix}, \begin{bmatrix} 4.5 & .075 \\ .075 & .00625 \end{bmatrix}\right). \qquad \square$$

The final result that we will examine concerning the asymptotic distribution of $\mathbf{g}(\mathbf{X}_n)$ generalizes the previous two theorems to cases for which $\mathbf{V}_n^{-1/2}(\mathbf{X}_n - \mu) \xrightarrow{d} N([\mathbf{0}], \mathbf{I})$, where $\{\mathbf{V}_n\}$ is a sequence of $(m \times m)$ positive definite matrices of real numbers such that $\mathbf{V}_n \rightarrow [0]$.[18] Note this case subsumes the previous cases upon defining $\mathbf{V}_n = n^{-1}\Sigma$, in which case $\mathbf{V}_n^{-1/2}(\mathbf{X}_n - \mu) = \Sigma^{-1/2}n^{1/2}(\mathbf{X}_n - \mu) \xrightarrow{d} N([\mathbf{0}], \mathbf{I})$ by Slutsky's theorem. The generalization allows additional flexibility in the means by which the asymptotic distribution of $\mathbf{X}_n$ is initially established and is especially useful in the context of the least squares estimator to be discussed in Chapter 8.

***Theorem 5.41** (Asymptotic Distribution of $\mathbf{g}(\mathbf{X}_n)$—Generalized) *Let $\{\mathbf{X}_n\}$ be a sequence of $(k \times 1)$ random vectors such that $\mathbf{V}_n^{-1/2}(\mathbf{X}_n - \mu) \xrightarrow{d} N([\mathbf{0}],\mathbf{I})$, where $\{\mathbf{V}_n\}$ is a sequence of $(m \times m)$ positive definite matrices for which $\mathbf{V}_n \rightarrow [\mathbf{0}]$. Let $\mathbf{g}(\mathbf{x})$ be an $(m \times 1)$ vector function satisfying the conditions of Theorem 5.40. If there exists a sequence of positive real numbers $\{a_n\}$ such that $\{[a_n\mathbf{G}\mathbf{V}_n\mathbf{G}']^{-1/2}\}$ is* O(1) *and $a_n^{1/2}(\mathbf{X}_n - \mu)$ is* $O_p(1)$, *then*

$$(\mathbf{G}\mathbf{V}_n\mathbf{G}')^{-1/2}[\mathbf{g}(\mathbf{X}_n) - \mathbf{g}(\mu)] \xrightarrow{d} N([\mathbf{0}], \mathbf{I})$$

*and*

$$\mathbf{g}(\mathbf{X}_n) \overset{a}{\sim} N(\mathbf{g}(\mu), \mathbf{G}\mathbf{V}_n\mathbf{G}').$$

[18] Recall that $\mathbf{V}_n^{1/2}$ is the symmetric square root matrix of $V_n$, and $\mathbf{V}_n^{-1/2}$ is the inverse of $\mathbf{V}_n^{1/2}$. The defining property of $\mathbf{V}_n^{1/2}$ is that $\mathbf{V}_n^{1/2}\mathbf{V}_n^{1/2} = \mathbf{V}_n$, while $\mathbf{V}_n^{-1/2}\mathbf{V}_n^{-1/2} = \mathbf{V}_n^{-1}$.

**Sketch of the Proof** Represent $\mathbf{g}(\mathbf{X}_n)$ in terms of a first-order Taylor series expansion plus remainder, as in the proof of Theorem 5.40, to obtain

$$[\mathbf{G}\mathbf{V}_n\mathbf{G}']^{-1/2}[\mathbf{g}(\mathbf{X}_n) - \mathbf{g}(\boldsymbol{\mu})] = [\mathbf{G}\mathbf{V}_n\mathbf{G}']^{-1/2}[\mathbf{G}(\mathbf{X}_n - \boldsymbol{\mu}) + d(\mathbf{X}_n, \boldsymbol{\mu})\mathbf{R}(\mathbf{X}_n)].$$

The first term to the right of the equality is such that

$$(\mathbf{G}\mathbf{V}_n\mathbf{G}')^{-1/2}\mathbf{G}(\mathbf{X}_n - \boldsymbol{\mu}) = (\mathbf{G}\mathbf{V}_n\mathbf{G}')^{-1/2}\mathbf{G}\mathbf{V}_n^{1/2}\mathbf{V}_n^{-1/2}(\mathbf{X}_n - \boldsymbol{\mu}) \xrightarrow{\mathrm{d}} N([\mathbf{0}], \mathbf{I}),$$

which follows from H. White, Asymptotic Theory, Lemma 4.23, p. 66, upon recognizing that $\{\mathbf{A}_n\} = \{(\mathbf{G}\mathbf{V}_n\mathbf{G}')^{-1/2}\mathbf{G}\mathbf{V}_n^{1/2}$ is an $O(1)$ sequence of matrices (note that $\mathbf{A}_n\mathbf{A}_n' = \mathbf{I}\ \forall\ n$) and $\mathbf{V}_n^{-1/2}(\mathbf{X}_n - \boldsymbol{\mu}) \xrightarrow{\mathrm{d}} N([\mathbf{0}], \mathbf{I})$.

The second term converges in probability to the zero matrix since $[\mathbf{G}\mathbf{V}_n\mathbf{G}']^{-1/2}d(\mathbf{X}_n, \boldsymbol{\mu}) = [a_n\mathbf{G}\mathbf{V}_n\mathbf{G}']^{-1/2}[a_n^{1/2}(\mathbf{X}_n - \boldsymbol{\mu})'(\mathbf{X}_n - \boldsymbol{\mu})a_n^{1/2}]^{1/2}$ is $O_p(1)$ and $\mathbf{R}(\mathbf{X}_n) \xrightarrow{\mathrm{p}} [\mathbf{0}]$. Then the convergence in distribution result of the theorem follows from Slutsky's theorem, and the asymptotic distribution of $\mathbf{g}(\mathbf{X}_n)$ follows subsequently. ■

## Key Words, Phrases, and Symbols

sequence

natural numbers, $N$

sequence of random variables

$Y \sim f(y)$, $Y$ has probability density $f(y)$, or $Y$ is distributed as $f(y)$

iid, independent and identically distributed

sequence of probability density functions

limit of a real number sequence

limit of the sequence $\{y_n\}$

converge to $y$

$\{y_n\}$ is bounded, unbounded

convergence of a sequence of functions $y_n \to y$, $\lim_{n\to\infty} y_n = y$

limiting function of $\{f_n\}$

order of magnitude of a sequence

$O(n^k)$, at most of order $n^k$

$o(n^k)$, of order smaller than $n^k$

convergence in distribution, $Y_n \xrightarrow{\mathrm{d}} Y$ or $Y_n \xrightarrow{\mathrm{d}} F$

converges in distribution to a constant, $Y_n \xrightarrow{\mathrm{d}} c$

limiting density of $\{Y_n\}$, limiting CDF of $\{Y_n\}$

limiting distribution

asymptotic distribution, $Z_n \overset{\mathrm{a}}{\sim} g(X, \Theta_n)$

convergence in probability, $Y_n \xrightarrow{\mathrm{p}} Y$

probability limit, plim

converges in probability to a constant

Slutsky's theorems

order of magnitude of a random sequence

order of magnitude in probability

$O_p(n^k)$, at most of order $n^k$ in probability

$o_p(n^k)$, of order smaller than $n^k$ in probability

bounded in probability

convergence in mean square, $Y_n \xrightarrow{\mathrm{m}} Y$

convergence in quadratic mean

almost-sure convergence, $Y_n \xrightarrow{\mathrm{as}} Y$

laws of large numbers

weak laws of large numbers, WLLN

Khinchin's WLLN

strong laws of large numbers, SLLN

Kolmogorov's inequality

Kolmogorov's SLLN

asymptotic nonpositively correlated

central limit theorems, CLTs

Lindberg-Levy's CLT

continuity correction

Lindberg CLT

Liapounov CLT

triangular array of random variables

$m$-dependent sequence

Cramér-Wold device

multivariate Lindberg-Levy CLT

symmetric matrix square root

asymptotic distributions of differentiable functions

## Problems

**1.** Given current technology, the production of active matrix color screens for notebook computers is a difficult process that results in a significant proportion of defective screens being produced. At one company the daily proportion of defective 9.5″ and 10.4″ screens is the outcome of a bivariate random variable, $X$, with joint density function $f(x_1, x_2; \alpha) = (\alpha x_1 + (2 - \alpha)x_2)I_{[0,1]}(x_1)I_{[0,1]}(x_2)$, where $\alpha \in (0, 2)$. The daily proportions of defectives are independent from day to day. A collection of $n$ iid outcomes of $X$ will be used to generate an estimate of the $(2 \times 1)$ vector of mean daily proportions of defectives, $\mu$, for the two types of screens being produced, as $\underset{(2\times 1)}{\bar{\mathbf{X}}_n} = \sum_{i=1}^{n} \mathbf{X}_i/n$, where $\mathbf{X}_i = \begin{bmatrix} X_{1i} \\ X_{2i} \end{bmatrix}$.

a. Does $\bar{\mathbf{X}}_n \xrightarrow{as} \boldsymbol{\mu}$? Does $\bar{\mathbf{X}}_n \xrightarrow{p} \boldsymbol{\mu}$? Does $\bar{\mathbf{X}}_n \xrightarrow{d} \boldsymbol{\mu}$?

b. Define an asymptotic distribution for the bivariate random variable $\bar{\mathbf{X}}_n$. If $\alpha = 1$ and $n = 200$, what is the approximate probability that $\bar{X}_n[1] > .70$, given that $\bar{X}_n[2] = .60$?

c. Consider using an outcome of the function $g(\bar{\mathbf{X}}_n) = \bar{X}_n[1]/\bar{X}_n[2]$ to generate an estimate of the relative expected proportions of defective 9.5″ and 10.4″ screens, $\mu_1/\mu_2$. Does $g(\bar{\mathbf{X}}_n) \xrightarrow{as} \mu_1/\mu_2$? Does $g(\bar{\mathbf{X}}_n) \xrightarrow{p} \mu_1/\mu_2$? Does $g(\bar{\mathbf{X}}_n) \xrightarrow{d} \mu_1/\mu_2$?

d. Define an asymptotic distribution for $g(\bar{\mathbf{X}}_n)$. If $\alpha = 1$ and $n = 200$, what is the approximate probability that the outcome of $g(\bar{\mathbf{X}}_n)$ will exceed 1?

**2.** Central limit theorems have important applications in the area of quality control. One such application concerns so-called control charts, and in particular, $\bar{X}$ charts, which are used to monitor whether the variation in the calculated mean levels of some characteristics of a production process are within acceptable limits. The actual chart consists of plotting calculated mean levels (vertical axis) over time (horizontal axis) on a graph that includes horizontal lines for the actual mean characteristic level of the process, $\mu$, and for *upper and lower control limits* that are usually determined by adding and subtracting two or more standard deviations, $\sigma_{\bar{X}}$, to the actual mean level. If, at a certain time period, the outcome of the calculated mean lies outside the control limits, the production process is considered to be no longer behaving properly, and the process is stopped for appropriate adjustments. For example, if a production process is designed to fill cans of soda pop to a mean level of 12 oz., if the standard deviation of the fill levels is .1, and if 100 cans of soda are randomly drawn from the packaging line to record fill levels and calculate a mean fill level $\bar{x}$, then the control limits on the daily calculated means of the filling process might be given by $12 \mp 3\,\text{std}(\bar{x}) = 12 \mp .03$.

a. Provide a justification for the $\bar{X}$ chart procedure described above based on asymptotic theory. Be sure to clearly define the conditions under which your justification applies.

b. Suppose that control limits are defined by adding and subtracting three standard deviations of $\bar{X}$ to the mean level $\mu$. In light of your justification of the control chart procedure in (a), what is the probability that the production process will be inadvertently stopped at a given time period, even though the mean of the process remains equal to $\mu$?

c. In the soda-can filling example described above, if the process were to change in a given period so that the mean fill level of soda cans became 12.05 oz. what is the probability that the control chart procedure would signal a shutdown in the production process in that period?

**3.** The lifetime of a certain computer chip that your company manufactures is characterized by the population distribution

$$f(z; \theta) = \frac{1}{\theta} e^{-z/\theta} I_{(0,\infty)}(z),$$

where $z$ is measured in thousands of hours. Let $(X_1, \ldots, X_n)$ represent iid random variables with the density $f(z; \theta)$. An outcome of the random variable $Y_n = (1 + \sum_{i=1}^{n} X_i)/n$ will be used to provide an estimate of $\theta$.

a. Is it true that $Y_n \xrightarrow{m} \theta$? Is it true that $Y_n \xrightarrow{p} \theta$?

b. Define an asymptotic distribution for $Y_n$.

c. Suppose $n = 100$ and $\theta = 10$. Use the asymptotic distribution you defined in (b) to approximate the probability that $Y_n \geq 15$.

**4.** In each case below, the outcome of some function, $T(X_{(n)})$, of $n$ iid random variables $X_{(n)} = (X_1, \ldots, X_n)$ is being considered for providing an estimate of some function of parameters, $q(\theta)$. Determine whether $\mathrm{E}T(X_{(n)}) = q(\theta)$, $\lim_{n\to\infty} \mathrm{E}T(X_{(n)}) = q(\theta)$, and $\text{plim}\, T(X_{(n)}) = q(\theta)$.

a. $X_i$'s $\sim$ iid Gamma$(\alpha, \beta)$ and $T(X_{(n)}) = \bar{X}_n$ is being used to estimate $q(\alpha, \beta) = \alpha\beta$.

b. $X_i$'s $\sim$ iid Gamma$(\alpha, \beta)$ and $T(X_{(n)}) = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2/(n-1)$ is being used to estimate $q(\alpha, \beta) = \alpha\beta^2$.

c. $X_i$'s $\sim$ iid Bernoulli$(p)$ and $T(X_{(n)}) = \bar{X}_n(1 - \bar{X}_n)$ is being used to estimate $q(p) = p(1 - p)$.

d. $X_i$'s $\sim$ iid $N(\mu, \sigma^2)$ and $T(X_{(n)}) = (\sum_{i=1}^n X_i - n^{1/2})/(n+1)$ is used to estimate $\mu$.

**5.** The daily number of customers entering a large grocery store who purchase one or more dairy products is given by the outcome of a binomial random variable $X_t$ with parameters $p$ and $n_t$ for day $t$. The number of customers who enter the grocery store on any given day, $n_t$, is itself an outcome of a random variable $N_t$ that has a discrete uniform distribution on the range $\{200, 201, \ldots, 300\}$. The $X_t$'s and the $N_t$'s are all independent. The local dairy products commission wants an estimate of the daily proportion of customers entering the store who purchase dairy products and wants you to use as an estimate an outcome of $\bar{X}_d = (1/d)\sum_{t=1}^d (X_t/N_t)$ where $d$ is the number of days.

a. Does $\bar{X}_d \xrightarrow{as} p$? Does $\bar{X}_d \xrightarrow{p} p$? Does $\bar{X}_d \xrightarrow{d} p$?

b. Define an asymptotic distribution for $\bar{X}_d$. If $p = .8$, $d = 300$, and $\sum_{i=1}^{300} n_t = 75{,}000$, what is the approximate probability that $\bar{X}_d \in (.78, .82)$?

**6.** Let $\{X_1, \ldots, X_n\}$ be iid random variables with $\sigma^2 < \infty$. We know from Khinchin's WLLN that

$$\bar{X} = n^{-1}\sum_{i=1}^{n} X_i \xrightarrow{p} u.$$

a. Find a functional relationship between $n$, $\sigma$, and $\varepsilon$ such that

$$P(\bar{x} \in (\mu - \varepsilon, \mu + \varepsilon)) \geq .99.$$

b. For what values of $n$ and $\sigma$ will an outcome of $\bar{X}$ be within $\mp .1$ of $\mu$ with probability $\geq .99$? Graph this relationship between the values of $n$ and $\sigma$.

c. If $\sigma = 1$, what value of $n$ will ensure that the outcome of $\bar{X}$ will be within $\mp .1$ of $\mu$ with probability $\geq .99$?

d. If $\sigma = 1$ *and* the $X_i$'s are normally distributed, what value of $n$ will ensure that the outcome of $\bar{X}$ will be within $\mp .1$ of $\mu$ with probability $= .99$?

**7.** Let the random variables in the sequence $\{X_n\}$ be iid with a gamma density having parameters $\alpha = 2$ and $\beta = 3$.

a. What is the probability density for $\bar{X}_n$?

b. What is an asymptotic density for $\bar{X}_n$?

c. Plot the actual versus asymptotic density for $\bar{X}_n$ when $n = 10$.

d. Repeat (c) for $n = 40$. Interpret the graphs in (c) and (d) in terms of asymptotic theory.

**8.** A pharmaceutical company claims that it has a drug that is 75 percent effective in generating hair growth on the scalps of balding men. In order to generate evidence regarding the claim, a consumer research agency conducts an experiment whereby a total of 1,000 men are randomly chosen and treated with the drug. Of the men treated with the drug, 621 experienced hair growth. Do the results support or contradict the company's claim?

**9.** A political candidate has hired a polling firm to assess her chances of winning a senatorial election in a large eastern state. She wants an estimate of the proportion of registered voters that would vote for her "if the election were held today." Registered voters are to be randomly chosen and interviewed, and their preferences recorded. The polling firm will use the outcome of $\bar{X}$ as an estimate of the proportion of voters in favor of the candidate. It is known that currently between 40 percent and 60 percent of the registered voters favor her in the election. She wants to have an estimate that is within 2 percentage points of the true proportion with probability $= .99$. How many registered voters must be interviewed, based on the asymptotic distribution of $\bar{X}$?

**10.** Let observations on the quantity supplied of a certain commodity be generated by $Y_i = x_i\beta + V_i$, where $|x_i| \in [a, b]\ \forall\ i$ are scalar observations on fixed prices, $\beta$ is an unknown slope coefficient, and the $V_i$'s are iid random variables having a mean of zero, a variance of $\sigma^2 \in (0, \infty)$, and $P(|V_i| \leq m) = 1\ \forall\ i$ ($a$, $b$, and $m$ are finite positive constants). Two functions of the $x_i$'s and $Y_i$'s are being considered for generating an estimate of the unknown value of $\beta$:

$$\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y} \quad \text{and} \quad \hat{\beta}_r = (\mathbf{x}'\mathbf{x} + k)^{-1}\mathbf{x}'\mathbf{Y},$$

where $k > 0$, $\mathbf{x}$ is an $(n \times 1)$ vector of observations on the $x_i$'s and $\mathbf{Y}$ is an $(n \times 1)$ vector of the corresponding iid $Y_i$'s.

a. Define the means and variances of the two estimators of $\beta$.

b. Is it true that $\lim_{n\to\infty} \mathrm{E}\hat{\beta} = \beta$ and/or $\lim_{n\to\infty} \mathrm{E}\hat{\beta}_r = \beta$?

c. Define the expected squared distances of the two estimators from $\beta$.

d. Which, if either, of the estimators converges in mean square to $\beta$?

e. Which, if either, of the estimators converges in probability to $\beta$?

f. Define asymptotic distributions for each of the estimators.

g. Under what circumstances would you prefer one estimator to the other for generating an estimate of the unknown value of $\beta$?

**11.** Let $X_1, \ldots, X_n$ be iid random variables having continuous uniform distributions of the form $f(z) = I_{(0,1)}(z)$.

a. Define an asymptotic distribution for $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$.

b. Using your result from (a), argue that $(\sum_{i=1}^{12} X_i) - 6 \approx Z \sim N(0, 1)$.

(This approximation is very accurate and is sometimes used for simulating $N(0, 1)$ outcomes using a uniform random-number generator.)

**12.** A company produces a popular beverage product that is distributed nationwide. The aggregate demand for the product during a given time period can be represented by

$$Q = \sum_{i=1}^{n} Q_i = \sum_{i=1}^{n} (\alpha_i - \beta_i p + V_i),$$

where $Q_i$ is quantity purchased by the $i$th consumer, $\alpha_i > 0$, $\beta_i > 0$, $\mathrm{E}V_i = 0$, $\mathrm{var}(V_i) \geq c > 0$, $P(|v_i| \leq m) = 1$ $\forall\, i$, and $c$ and $m$ are positive constants. It can be assumed that the quantities purchased by the various consumers are independent.

a. Define an asymptotic distribution for the aggregate quantity demanded, Q.

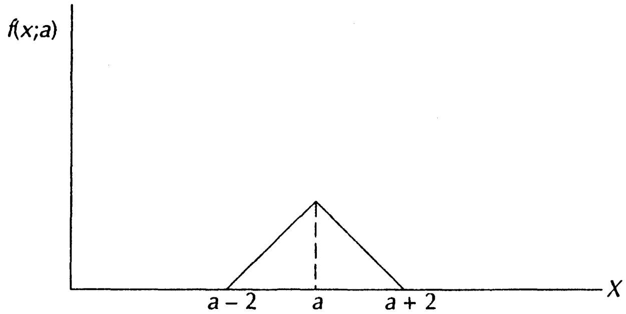
b. If $p = 2$, then $\mathrm{E}Q = 80$, and if $p = 5$, then $\mathrm{E}Q = 50$. If it costs the company \$2/unit to manufacture and distribute the product, and if $p = \$2.50$, what is the asymptotic distribution of aggregate company profit during the time period?

c. Define an interval around the mean of aggregate company profit that will contain the actual outcome of aggregate profit with (approximate) probability .95 when $p = \$2.50$.

**13.** The daily tonnage of garbage handled by the Enviro-Safe Landfill Co. is represented as the outcome of a random variable having some triangular distribution, as

$$X \sim f(x; a) = [(.5 - .25a) + .25x] I_{[a-2,a]}(x) + [(.5 + .25a) - .25x] I_{(a,a+2]}(x)$$

The distribution is represented graphically as follows:



Enviro-Safe is in the process of analyzing whether or not it needs to expand its facilities. It wants an estimate of the expected daily tonnage of garbage that it handles. It has collected 4 years of daily observations on tonnage handled ($n = 1,460$ observations) and provides you with the following summary statistic:

$$\sum_{i=1}^{n} x_i = 29,200.$$

You may treat the observations as outcomes of iid random variables.

a. Use $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ to provide an estimate of $a$, the expected tonnage of garbage handled.

b. Based on the LLCLT, define an asymptotic distribution for $\bar{x}$. You should be able to identify a *numerical value* for the variance of the asymptotic distribution.

c. Using the asymptotic distribution, how probable is it that the outcome of $\bar{x}$ will be within .05 tons of the *actual* expected value of daily garbage tonnage handled?

d. Use Van Beeck's inequality to provide an upper bound to the approximation error in the probability value that you assigned in part (c). Are you reasonably confident that you provided Enviro-Safe with an accurate "guess" of the expected daily tonnage? (Enviro-Safe management said that it would be satisfied if it could get an estimate that was "within ±1 ton of the actual expected daily tonnage.") Explain.

e. Using your estimate $\hat{a} = \bar{x}$ to estimate the density function, as $f(x; \hat{a})$, what is your estimate of the probability that tonnage handled by Enviro-Safe will exceed 21 tons on any given day?

**14.** A statistician wants to use iid outcomes from some exponential distribution

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta} I_{(0,\infty)}(x)$$

to generate an estimate of the variance of the exponential density, $\theta^2$. She wants to use the outcome of $\bar{X}_n^2$, where $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ to generate an estimate of $\theta^2$.

a. Does $\mathrm{E}\bar{X}_n^2 = \theta^2$? Does $\lim_{n\to\infty} \mathrm{E}\bar{X}_n^2 = \theta^2$?

b. Does $\operatorname{plim} \bar{X}_n^2 = \theta^2$?

c. Define an asymptotic distribution for $\bar{X}_n^2$.

**15.** We have shown that if $\{Y_n\}$ is a sequence of $\chi^2$ random variables, where $Y_n \sim \chi_n^2$, then $(Y_n - n)/\sqrt{2n} \xrightarrow{d} N(0,1)$. Since $Y_n \sim \chi_n^2$, we know that $P(y_{25} \leq 34.3816) = P(y_{50} \leq 63.1671) = P(y_{100} \leq 118.498) = .90$. Assign (approximate) probabilities to the three events using asymptotic distributions. How good are the approximations? Does $\operatorname{plim}(Y_n/n) = 1$? Why or why not?

**16.** Let $\{X_n\}$ be a sequence of random variables having binomial densities, where $X_n$ has a binomial density with parameters $n$ and $p$, i.e.,

$$X_n \sim \binom{n}{x} p^x (1-p)^{n-x} I_{\{0,1,2,\ldots,n\}}(x).$$

a. Show that $(X_n - np)/(\sqrt{n}(p(1-p))^{1/2}) \xrightarrow{d} N(0,1)$.

b. Define an asymptotic distribution for $X_n$. Use Van Beeck's inequality to provide a bound on the error in approximating the probability $P(x_n \leq c)$ using the asymptotic distribution for $X_n$. Calculate the numerical value of this bound when $p = .3$.

c. Using the binomial density for $X_n$ and letting $p = .3$, it follows that

| $n$ | $k$ | $p(x_n \leq k)$ |
|---|---|---|
| 15 | 6 | .8689 |
| 25 | 9 | .8106 |
| 100 | 34 | .8371 |

Assign (approximate) probabilities to the three events using asymptotic distributions based on your answer for (b). How good are the approximations?

d. In using the CLT to approximate probabilities of events for discrete random variables whose range consists of equally spaced points, it has been found that a **continuity correction** improves the accuracy of the approximation. In particular, letting

$R(X) = \{x_1, x_2, x_3, \ldots\}$, where $x_{i+1} - x_i = 2h > 0 \quad \forall\, i$, the continuity correction involves treating each elementary event $x_j$ for the discrete random variable $X$ as the interval event $(x_j - h, x_j + h]$ for the normal asymptotic distribution of the random variable $X$. For example, if $R(X) = \{0,1,2,3,\ldots\}$, then $P(x \in [1,2]) = \sum_{x=1}^2 f(x) \approx \hat{P}(x \in (.5, 2.5])$, where the latter (approximate) probability is assigned using the appropriate asymptotic normal distribution for $X$. Use the continuity correction to approximate the probabilities of the three events in (c).

**17.** The Nevada Gaming Commission has been directed to check the fairness of a roulette wheel used by the WINBIG Casino. In particular, a complaint was lodged stating that a "red" slot occurs more frequently than a "black" slot for the roulette wheel used by WINBIG, whereas red and black should occur with probability. 5 if the wheel is fair. The wheel is spun 100,000 times, and the number of red and black outcomes was recorded. The outcomes can be viewed as iid from some Bernoulli population distribution:

$$X \sim p^x(1-p)^{1-x} I_{\{0,1\}}(x)$$

for $p \in (0,1)$.
It was found that $\sum_{i=1}^n x_i = 49{,}873$, where $x_i = 1$ indicates that the $i$th spin resulted in a red outcome.

a. Use $\bar{X}_n$ to provide an estimate of $p$, the probability of observing a red outcome.

b. Define an asymptotic distribution for $\bar{X}_n$.

c. Using the outcome of $\bar{X}_n$ as an estimate of $p$ in the asymptotic distribution for $\bar{X}_n$, how probable is it that an outcome of $\bar{X}_n$ is within $\pm.005$ of the true probability? Use Chebyshev's inequality to argue that the estimate $\bar{X}_n$ for $p$ should be very accurate in the sense that outcomes of $\bar{X}_n$ are very close to $p$ with high probability.

d. Use Van Beeck's inequality to provide an upper bound to the approximation error that can occur in assigning probability to events like $P(\bar{X}_n \leq c)$ using the asymptotic distribution for $\bar{X}_n$. Your bound will unfortunately depend on the unknown value of $p$. Estimate a value for the bound using your outcome of $\bar{X}_n$ as an estimate of $p$.

e. Define an asymptotic distribution for $g(\bar{X}_n) = \bar{X}_n(1 - \bar{X}_n)$.

f. Compare the asymptotic distribution of the estimator in part (e) to the asymptotic distribution of the estimator $S^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2/(n-1)$. If the sample size were large, would you prefer one of the estimators of $p(1-p)$ over the other? Explain.

**18.** The Elephant Memory Chip Co. (EMC for short) instructs its resident statistician to investigate the operating-life characteristics of its new 4 megabyte memory chip in order to provide product information to potential buyers. The population distribution of operating lives can be specified as some exponential family distribution. The statistician intends to draw a random sample of 10,000 chips from EMC's production and apply a nondestructive test that will determine each chip's operating life. He then intends to use the outcome of the random sample to provide estimates of both the mean and the variance of the chip's operating life. He needs your help in answering a few statistical questions.

a. Letting $\theta$ represent the unknown parameter in the exponential population distribution, what is the distribution of the sample mean, $\bar{X}_n$? What is the mean and the variance of this distribution? Does $\operatorname{plim} \bar{X}_n = \theta$?

b. The outcome of the random sample resulted in the following two outcomes.

$$\bar{X}_n = 10.03702,$$

$$n^{-1}\sum_{i=1}^{n} X_i^2 = 199.09634.$$

Operating life is measured in 1,000-hour units.

(Side note: These are actual outcomes based on a *simulation* of the random sample outcome using a specific value of $\theta$.)

The statistician uses $\bar{X}_n$ to estimate the mean life, $\theta$, of the chips. He is considering using either $\bar{X}_n^2$ or $S_n^2 = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2/(n-1)$ to estimate the variance, $\theta^2$, of operating lives. He asks the following questions regarding the characteristics of $X_n^2$ and $S_n^2$: (show your work):

1. Does $E\bar{X}_n^2 = \theta^2$? Does $ES_n^2 = \theta^2$?
2. Does $\lim_{n\to\infty} E\bar{X}_n^2 = \theta^2$? Does $\lim_{n\to\infty} ES_n^2 = \theta^2$?
3. Does $\operatorname{plim} \bar{X}_n^2 = \theta^2$? Does $\operatorname{plim} S_n^2 = \theta^2$?
4. Define asymptotic distributions for $\bar{X}_n^2$ and $S_n^2$. Based on their asymptotic distributions, would you recommend the use of one random variable over the other for generating an estimate of $\theta^2$? Why or why not?
5. Calculate the outcomes of both $\bar{X}_n^2$ and $S_n^2$.

(Note: The *actual* value of $\theta = 10$, and thus the actual value of $\theta^2 = 100$.)

6. The statistician has an idea he wants you to react to. He doesn't like the fact that $E\bar{X}_n^2 \neq \theta$ (that *is* what you found—isn't it?). He wants to define a new random variable, $Y_n = a_n\bar{X}_n^2$, for an appropriate sequence of numbers $\{a_n\}$, so that $EY_n = \theta^2, \forall\, n$. Can he do it? How? If he (and you) can, then use the appropriate outcome of $Y_n$ to provide another estimate of $\theta^2$. Is it true that $\operatorname{plim} Y_n = \theta^2$?

**19.** Let $(X_1, \ldots, X_n)$ be a random sample from a Poisson population distribution. Derive the limiting distribution of

$$T = \frac{(\bar{X}_n - \mu)}{(S_n^2/n)^{1/2}},$$

where $S_n^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.

**20.** Liquid crystal displays (LCDs) that your wholesaling company is marketing for a large Japanese electronics firm are known to have a distribution of lifetimes of the following Gamma-distribution form:

$$f(z;\alpha) = \frac{1}{2^{\alpha}\Gamma(\alpha)} z^{\alpha-1} e^{-z/2} I_{(0,\infty)}(z),$$

where $z$ is measured in thousands of hours.

A set of $n$ iid outcomes of $Z$ will be used in an attempt to obtain information about the expected value of the lifetime of the LCD's.

a. Define the functional form of the joint density of the iid random variables, say $(X_1, \ldots, X_n)$, of LCD lifetimes.

b. What is the density function of the random variable $Y_n = \sum_{i=1}^{n} X_i$.

c. Supposing that $n$ were large, identify an asymptotic distribution for the random variable $Y_n$. (Note: since you don't know $\alpha$ at this point, your asymptotic distribution will depend on the unknown value of $\alpha$.)

d. If $\alpha$ were equal to 1/2, and the sample size was $n = 20$, what is the probability that $y_n \leq 31.4104$? Compare your answer to the approximate probability obtained using the asymptotic distribution you defined in (c).

e. If $\alpha$ were equal to 1/2, and the sample size was $n = 50$, what is the probability that $y_n \leq 67.5048$? Compare your answer to the approximate probability obtained using the asymptotic distribution you defined in (c).

# 6 Sampling, Sample Moments, Sampling Distributions, and Simulation

## 6.1 Introduction

Beginning with this chapter, we turn our attention toward concepts and procedures that are explicitly related to the problem of *statistical inference*. Prior to this point, our study of probability theory and its implications has essentially addressed questions of *deduction*: "Given a probability space, what can we deduce about the characteristics of outcomes of an experiment?" Our study of statistical inference will turn this question around: "Given the characteristics of outcomes of an experiment, what can we infer about the probability space?"

The term *statistical inference* refers to the inductive process of generating information about characteristics of a real-world population or process by analyzing a *sample* of objects or outcomes from the population or process. For example, a marketer may be interested in determining whether consumers with a certain sociodemographic profile (the *population*) would purchase a new product (the *characteristic*); an auditor would be interested in assessing the accuracy (the *characteristic*) of a firm's accounts (the *population*); and a quality-control engineer would have interest in determining whether commodities are being manufactured (the *process*) to within factory specifications (the *characteristic*). The statistical inference problems would involve analyses of samples of observations from the real-world population or process ultimately leading to inferences regarding the characteristics of interest. Figure 6.1 provides a schematic overview of the process of statistical inference.

It is apparent that for an analysis of a sample of objects or outcomes to lead to meaningful inferences about the characteristics of a real-world population or process, the sample of characteristics must be in some sense representative of the population or process characteristics. The linkage between the incidence of characteristics in the sample and in the real-world population or process is established by analyzing *random samples*. The objects or outcomes in a random sample are obtained using a selection procedure that establishes a connection between the incidence of characteristics in the population or process and the *probability* of observing the characteristics' outcomes in the sample. Then, based on random samples and the implications of probability theory, methods of statistical inference can be devised that generate inferences about the characteristics of a population or process with a degree of accuracy or representativeness that can be measured probabilistically.

In this chapter we examine the concept of random sampling and begin identifying methods of generating information on population or process characteristics from random sample observations. As organized in Figure 6.1, our coverage starts with the topic *Population or Process* and moves clockwise, ending with *Sample of Characteristic Outcomes*. Our discussion will also implicitly relate

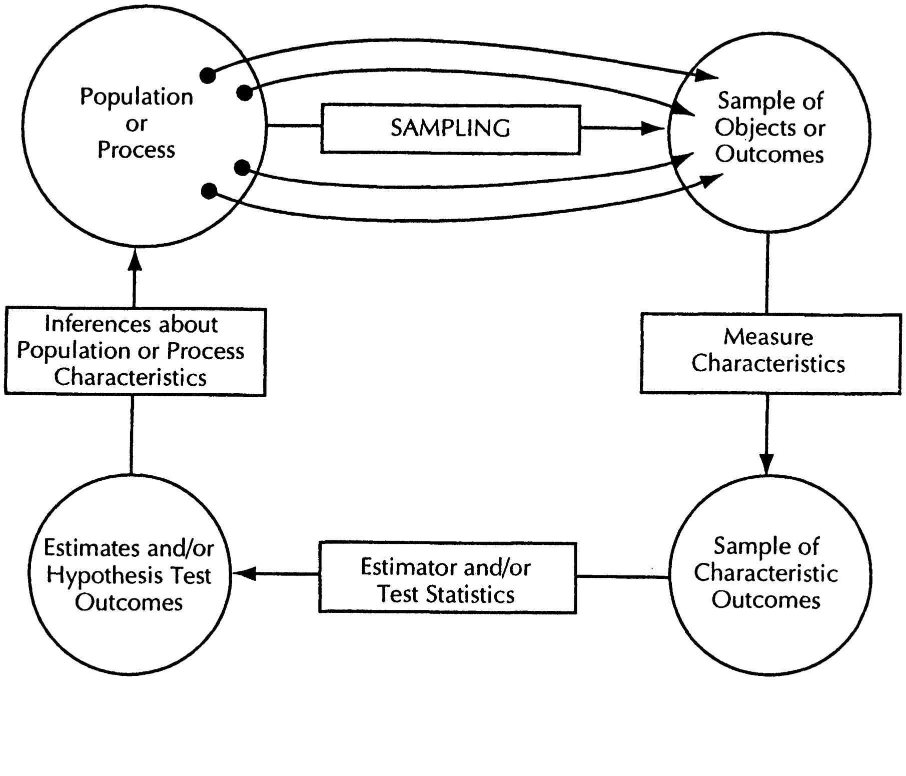


**Figure 6-1**
Overview of the statistical inference process.

to the remaining topics in the figure, but a detailed and focused analysis of estimation and hypothesis-testing concepts will not commence until the next chapter.

## 6.2 Random Sampling

The objective of statistical inference is to generate information about relevant characteristics of either the objects in some set of objects or the outcomes of some stochastic process. The term **population** will refer to any set of objects having associated characteristics that a researcher wishes to identify, enumerate, or generally obtain information about. The term **stochastic process** will refer to any collection of experiments or measurement activities whose outcome can be interpreted as the outcome of a collection of random variables[1] and whose characteristics are of interest to the researcher. (Henceforth, we will shorten *stochastic process* to simply *process* whenever the context is clear.) The purpose of random sampling is to obtain information about the characteristics of a population or process without having to examine each and every object or outcome relating to the population or process.

If a population has a finite number of objects, then an examination of each element of the population is conceivable, but there are reasons why the researcher might not wish to do so. An obvious reason would be if the measurement process were destructive, in which case measuring the characteristics of all of the population elements would lead to the population's destruction. This would be clearly undesirable if, for example, the objects in a population were manufactured by a firm that intended to offer them for sale (the tensile strength of steel beams is an example of a destructive measurement process, in which the beam is stressed to the breaking point). Another reason would be that the cost involved in evaluating each and every member of a population may be prohibitive. A related reason could be that the time available in which to perform thc analysis is insufficient for a complete enumeration of the population. Finally, the researcher may have no choice but to analyze a sample, instead of the whole, as would be the case for an ongoing manufacturing (or other) process having no stipulated or identifiable end.

If we were to attempt to obtain information about relevant characteristics of a population or process from the characteristics observed in a *sample* from that population or process, an obvious difficulty comes to mind. Namely, depending on which sample we choose, the incidence of the characteristics in the sample can differ from sample to sample, even though there exists only one fixed distribution of the characteristics in the population or process. It follows that we must expect, in general, that the existence or level of a characteristic in a given sample will not necessarily coincide with the existence or level of the

[1] More formally, the term *stochastic process* refers to any collection of random variables indexed by some index set $T$, i.e., $\{X_t, t \in T\}$ is a stochastic process. Recall Ex. 5.48 and the associated footnote.

characteristic in the population or process. The discrepancy between sample and population or process information is generically referred to as the **sampling error.**

As a concrete example of sampling error, suppose a shipment contained 100 objects, of which 30 were defective. Not knowing the number of defectives in the shipment, suppose an individual attempted to infer the proportion of the population of objects that were nondefective by observing the number of nondefectives in a sample of 10 objects from the shipment. It is unfortunately possible, depending on the particular set of 10 objects chosen, for the observed proportion of nondefectives in the sample to range between 0 and 1. Thus, the actual sampling error in inferring the proportion of nondefectives in the population would range between 0 and .7 in absolute value, although the researcher would clearly be unaware of this range, since she is unaware that the actual proportion of nondefectives is .7. Without any further information about how the sample of observations from the population was generated, nothing can be said about the reliability of the inference. Note that sample information in this context is essentially useless—the observed proportion of nondefectives will be between 0 and 1, and without any other information regarding the reliability of the inference, we know nothing more about the population than when we started the analysis.

In order for the researcher to be able to assess the reliability with which population or process characteristics are represented by sample information, it is necessary for the sample observations to have been generated in such a way that the researcher is able to establish some quantitative measures of confidence that the characteristics identified or enumerated in the sample accurately portray their incidence in the population or process. In particular, sample observations generated by a random sample will be useful for such inference. In essence, a random sample will refer to a sample whose observations can be interpreted as outcomes of a random vector having a probability distribution that ~~has been determined by the interaction between the distribution of character~~istics inherent in the population or process and the specific method governing the generation of sample observations. The linkage between the distribution of characteristics in the population or process and its influence on determining the form of the joint density function of random sample observations will be exploited to generate inference reliability measures.

We will examine three general methods for generating sample observations:

1. random sampling from a population distribution (which includes random sampling with replacement);
2. random sampling without replacement;
3. random sampling via a composite experiment.

### Random Sampling from a Population Distribution

In this case, the sample observations are defined by a collection of *independent and identical experiments*, and thus the sample observations can be conceptualized as the outcome of a collection of iid random variables. There are basically two generic types of random sampling from a population distribution. One type involves sampling from an existent, finite collection of objects, such as a shipment of personal computers or the citizens of a particular region of the country. The other type involves sampling from some ongoing stochastic process, such as a manufacturing process (e.g., measuring the net weight of potato chips in a bag of potato chips) or a market process (e.g., measuring stock market prices over time). In the former type of sampling, the ultimate objective is to obtain information about the finite collection of objects under study, while in the latter type the objective is to obtain information about the characteristics of the stochastic process under study (e.g., the expected weight of potato chips in a bag; the probability of producing a bag containing less than the advertised quantity of chips; the variance of stock prices).

When an existent, finite population is being sampled, random sampling from a population distribution is alternatively referred to as **random sampling with replacement**. In this case, there exists a finite set of $N$ objects having certain characteristics of interest to the researcher. The formal description of the sampling method is as follows.

**Definition 6.1**
***Random sampling with replacement***

1. An object is selected from the population in a way that gives all objects in the population an equal chance of being selected.
2. The characteristics level of the object selected is observed, and the object is returned to the population prior to any subsequent selection.
3. For a sample of size $n$, steps (1) and (2) are performed $n$ times.

Since all members of the population are equally likely to occur at each selection, the classical probability definition is appropriate for determining the probability of observing any level of the characteristics on a given selection, as

$$m(z) = \frac{\begin{bmatrix}\text{number of population objects}\\ \text{having characteristics level } z\end{bmatrix}}{N},$$

where $z$ will henceforth represent a scalar or vector, depending on the number of characteristics being measured for each selection, and $N$ represents the number of objects in the population. The density function, $m(z)$, is often referred to as the **population distribution** of the characteristics.

Since all of the selection experiments are independent, the probability of observing a sample of size $n$ exhibiting the collection of characteristic levels $(x_1, x_2, \ldots, x_n)$ is given by

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} m(x_i),$$

which defines the joint density function for $(X_1, \ldots, X_n)$, where $x_i$ is the characteristic(s) level associated with the $i$th observation.

**Example 6.1** Sampling for Defectives

Let a shipment of $N$ keyboards contain $J$ defective keyboards. (Note that in practice, $N$ will generally be known, while $J$ will be unknown, and information concerning $J$ would be the objective of statistical inference.) Let $z = 1$ denote a defective keyboard and $z = 0$ denote a nondefective keyboard. Then, if $n$ objects are sampled with replacement, and letting $p = J/N$, the *population distribution* of defective/nondefective keyboards is

$$m(z; p) = p^z(1-p)^{1-z} I_{\{0,1\}}(z),$$

where $p$ represents the proportion of defective keyboards in the population. The joint density governing the probabilities of observing outcomes $x_1, \ldots, x_n$ in the sample is given by

$$f(x_1, \ldots, x_n; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} I_{\{0,1\}}(x_i) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} I_{\{0,1\}}(x_i).$$

Note that the incidence of nondefectives in the population (the value of $p$) has a direct influence on the probabilities of events for the random sample outcomes. Thus, a probabilistic linkage is established between the incidence of characteristics in the population and in the sample. □

The outcome of random sampling with replacement can be interpreted as the outcome of an $n$-variate random variable $(X_1, \ldots, X_n)$ having the joint density function $f(x_1, \ldots, x_n)$, which will be referred to as the **joint density of the random sample**. The collection of random variables $(X_1, \ldots, X_n)$ is referred to as the **random sample**, whereas the outcome observed, $(x_1, \ldots, x_n)$, is referred to as the **outcome of the random sample**.

When the population from which we are sampling is not finite, random sampling from a population distribution will refer to an ongoing stochastic process in which all of the random variables whose outcomes are being measured are *independent* and *identically* distributed. In practice, this means that whatever underlying random mechanism determines the sample observations, it is unchanging from observation to observation, and observing a particular outcome for a given sample observation has no effect on the probability of outcomes of any other sample observations. Then each sample observation is the outcome of a random variable $Z \sim m(z)$, where $m(z)$ is the common PDF of characteristics' outcomes (also called the *population distribution*) associated with the stochastic process. Since all of the observation experiments are identical and independent, the PDF for the $n$ random variables $(X_1, \ldots, X_n)$ characterizing the outcomes of the $n$ observation experiments is given by

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} m(x_i).$$

As before, $(X_1, \ldots, X_n)$ is called the random sample, and $f(x_1, \ldots, x_n)$ is the joint density function of the random sample.

**Example 6.2** Sampling the Reliability of a Manufactured Product

The distribution of the operating lives until failure of halogen lamps produced by a domestic manufacturer, that is, the population distribution, is a given by a member of the gamma family of densities, as

$$m(z;\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} z^{\alpha-1} e^{-z/\beta} I_{(0,\infty)}(z).$$

The lamps are all produced using an identical manufacturing process, and $n$ lamps are arbitrarily selected from the production line for reliability testing. The $n$ measurements on operating life can be interpreted as the outcome of a random sample with joint density function

$$f(x_1,\ldots,x_n;\,\alpha,\beta) = \prod_{i=1}^{n} m(x_i;\alpha,\beta)$$

$$= \frac{1}{\beta^{n\alpha}(\Gamma(\alpha))^{n}} \prod_{i=1}^{n} x_i^{\alpha-1} e^{-\sum_{i=1}^{n} x_i/\beta} \prod_{i=1}^{n} I_{(0,\infty)}(x_i).$$

Note that the functional form of the population distribution, and in particular the actual values of the parameters $\alpha$ and $\beta$, will have a direct influence on the probabilities of events for random sample outcomes. Thus, a probabilistic linkage is established between the incidence of characteristics in the process and in the sample. □

### Random Sampling Without Replacement

Random sampling without replacement is relevant for a finite, existent population of objects, but it differs from random sampling with replacement in that once the characteristics of an object are observed, the object is removed from the population before another object is selected for observation. The sampling procedure is described as follows.

**Definition 6.2**
***Random sampling without replacement***

1. The first object is selected from the population in a way that gives all objects in the population an equal chance of being selected.
2. The characteristics level of the object is observed, but the object is not returned to the population.
3. An object is selected from the remaining objects in the population in a way that gives all remaining objects an equal chance of being selected, and step (2) is repeated. For a sample of size $n$, step (3) is performed $(n-1)$ times.

In this case, the sampling process can be characterized as a collection of $n$ experiments that are neither identical nor independent. In particular, the probability of observing characteristics level $x_i$ on the $i$th selection depends on what objects were observed and removed from the population in the preceding $(i-1)$ selections. For the first selection, the probability of observing characteristics level $x_1$ is given by

$$m(x_1) = \frac{\begin{bmatrix}\text{number of population objects}\\ \text{having characteristics level } x_1\end{bmatrix}}{N},$$

since all objects are equally likely to be selected. The density $m(x_1)$ can be thought of as the initial population distribution. On the $i$th selection, $i > 1$, the probability of observing characteristics level $x_i$ is conditioned on what was observed and removed from the population previously, so that

$$f(x_i \mid x_1,\ldots,x_{i-1}) = \frac{\begin{bmatrix}\text{number of objects remaining in the}\\ \text{population having characteristics level}\\ x_i \text{ after the } (i-1) \text{ selections } x_1,\ldots,x_{i-1}.\end{bmatrix}}{(N-i+1)},$$

for $N-i+1 > 0$. The joint density defining the probability of observing a sample of size $n$ that exhibits the collection of characteristics levels $(x_1,\ldots,x_n)$ can then be defined by[2]

$$\begin{aligned} f(x_1,\ldots,x_n) &= m(x_1)f(x_2 \mid x_1)f(x_3 \mid x_1,x_2)\cdots f(x_n \mid x_1,x_2,\ldots,x_{n-1}) \\ &= m(x_1)\prod_{i=2}^{n} f(x_i \mid x_1,\ldots,x_{i-1}). \end{aligned}$$

**Example 6.3** Sampling for Defectives

A shipment of $N$ LCD screens contains $J$ defectives. Suppose a random sample without replacement of size $n$ is drawn from the shipment. Let $x = 1$ denote that a screen is defective and $x = 0$ denote a nondefective. Then, assuming $N \geq n$ and $[(n-1)-(N-J)] \leq \sum_{i=1}^{n-1} x_i \leq J$,[3]

$$m(x_1) = \left(\frac{J}{N}\right)^{x_1}\left(\frac{(N-J)}{N}\right)^{1-x_1} I_{\{0,1\}}(x_1),$$

$$f(x_2 \mid x_1) = \left[\frac{J-x_1}{N-1}\right]^{x_2}\left[\frac{N-J-(1-x_1)}{N-1}\right]^{1-x_2}\prod_{i=1}^{2} I_{\{0,1\}}(x_i),$$

and in general, for the $n$th selection,

$$f(x_n \mid x_1,\ldots,x_{n-1}) = \left[\frac{J-\sum_{i=1}^{n-1} x_i}{N-(n-1)}\right]^{x_n}\left[\frac{N-J-\left(n-1-\sum_{i=1}^{n-1} x_i\right)}{N-(n-1)}\right]^{1-x_n}\prod_{i=1}^{n} I_{\{0,1\}}(x_i).$$

[2] This follows straightforwardly from the definitions of marginal and conditional density functions, as the reader should verify.

[3] These conditions, coupled with the condition that $x \in \{0, 1\}$, ensure that the denominators in the density expressions are positive and that the numerators are nonnegative.

The joint PDF for $(X_1, \ldots, X_n)$ is then given by

$$f(x_1, \ldots, x_n) = m(x_1)f(x_2 \mid x_1) \prod_{i=3}^{n} f(x_i \mid x_1, \ldots, x_{i-1}).$$

Note that the incidence of nondefectives in the population (the value of $J$) has a direct influence on the probabilities of events for the random sample outcomes. Thus, a probabilistic linkage is established between the incidence of characteristics in the population and in the sample. Compare this result to that from Ex. 6.1. □

As before, the outcome of the sampling process can be interpreted as the outcome of the $n$-variate random variable $(X_1, \ldots, X_n)$ having the joint PDF $f(x_1, \ldots, x_n)$, which will be called the **joint density of the random sample**. The collection $(X_1, \ldots, X_n)$ is referred to as the **random sample** while the outcome $(x_1, \ldots, x_n)$ is referred to as the **outcome of the random sample**.

### Sample Generated by a Composite Experiment

In this sampling method, observations on population or stochastic process characteristics are generated by the performance of $n$ experiments that are not independent and/or are not performed under identical conditions. In this case, the collection of experiments is called a **composite experiment**, with each of the $n$ experiments that lead to observations on each of the $X_i$'s being referred to as **component experiments**. The joint density for the sample outcomes is some density $f(x_1, \ldots, x_n)$ that is inherent to the collection of component experiments and their interrelationships. If the component experiments in the composite experiment happen to be independent, then the joint density of sample outcomes is given by $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i)$, where $f_i(x_i)$ is the probability density associated with the outcomes of the $i$th experiment.

**Example 6.4** Quantity Demanded Across Consumers

Let the quantity demanded of a commodity by a consumer in a given market be represented by

$$Q_i = g(p_i, y_i, \mathbf{z}_i) + V_i,$$

where $Q_i$ is the quantity demanded by consumer $i$, $p_i$ is the price of the commodity for consumer $i$, $y_i$ is the disposable income of consumer $i$, $\mathbf{z}_i$ is a vector of substitute/complement prices and sociodemographic characteristics for consumer $i$, and the $V_i$'s are independent but not necessarily identically distributed random variables whose outcomes represent deviations of $Q_i$ from $g(p_i, y_i, \mathbf{z}_i)$ caused by errors in utility optimization, lack of information, and/or inherent random human behavior. For the sake of exposition, assume that $V_i \sim N(0, \sigma_i^2)$. Then $(Q_1, \ldots, Q_n)$ is a random sample from a composite experiment relating to $n$ observations on a *demand process*. The $i$th component experiment consists of observing the demand by consumer $i$ for which $Q_i \sim N(q_i; g(p_i, y_i, \mathbf{z}_i), \sigma_i^2)$, and the joint density of the random sample is given by a product of nonidentical marginal densities as

$$f(q_1, q_2, \ldots, q_n; (p_i, y_i, \mathbf{z}_i, \sigma_i^2), i = 1, \ldots, n) = \prod_{i=1}^{n} N(q_i;\ g(p_i, y_i, \mathbf{z}_i), \sigma_i^2).$$

Note that the characteristics of the demand process directly influence the probabilities of events for random sample outcomes by their influence on the functional form of the joint density of the random sample. □

As before, the sample observations can be interpreted as an outcome $(x_1, \ldots, x_n)$ of the random sample $(X_1, \ldots, X_n)$, with the joint density of the random sample being $f(x_1, \ldots, x_n)$.

### Commonalities in Probabilistic Structure of Random Samples

The underlying rationale leading to the joint density function of a random sample is essentially identical in the cases of random sampling *from a population distribution* whether random sampling is from an existent finite population of objects or from an ongoing stochastic process. In both cases, an experiment whose probabilistic structure is characterized by the PDF, $m(z)$, is independently repeated $n$ times to obtain a random sample having a joint PDF defined by $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} m(x_i)$. In other words, in either case, the random sample can be thought of as a collection of iid random variables each having the PDF $m(z)$. In subsequent sections of this chapter, we will examine certain functions of random samples that have a number of properties that will be useful in statistical inference applications and that are derived from the fact that $X_1, \ldots, X_n$ are iid random variables. Since the iid property is shared by both types of random samples from a population distribution, any property of the function $g(X_1, \ldots, X_n)$ deduced from the fact that $X_1, \ldots, X_n$ are iid will apply regardless of whether random sampling is from a population that exists and is finite or is from an ongoing stochastic process.

There is also a commonality between how a random sample is generated via a composite experiment and how a random sample is generated by random sampling *without* replacement. In the latter case, observations are generated by a sequence of experiments that are neither independent nor performed under identical conditions, and this characterization of the experiments is subsumed under the general description of how a composite experiment can be performed. Note, however, that the composite experiment case is the broader concept, encompassing literally a myriad of different ways a series of component experiments can be interrelated, leading to a myriad of definitions for the joint density of the random sample. On the other hand, in random sampling without replacement, there is more structure to the definition of the joint density of the random sample, where the sampling procedure leads to a joint density definition based on the product of a collection of well-defined conditional density functions. In either case, it becomes somewhat difficult to establish *general* properties of functions of random samples, $g(X_1, \ldots, X_n)$, that are useful for purposes of statistical inference. In the case of random sampling without replacement, the mathematics involved in deriving properties of $g(X_1, \ldots, X_n)$ is generally more complex than in the case of random sampling from a population distribution—the iid assumption involved in the latter case introduces considerable simplifications. The sheer breadth of variations in the composite experiment case virtually relegates analyses of the properties of $g(X_1, \ldots, X_n)$ to analyses of special cases, and few generalizations to the entire class of composite experiments can be made.

In the remainder of our study, we will focus primarily on problems of statistical inference based on random sampling from a population distribution and on a specific composite experiment arising in the context of what is known as the general linear model. We will examine a few problems of interest involving random sampling without replacement from a finite population, but our analyses of these problems will be quite limited. For further reading on sampling without replacement that parallels some of the topics discussed in the remainder of this chapter related to sampling from a population distribution, and for a discussion of additional refinements to random sampling techniques, see M. Kendall and A. Stuart (1977), *The Advanced Theory of Statistics, Vol. 1*, 4th ed., New York: MacMillan, pp. 319–324, and W. G. Cochrane (1977), *Sampling Techniques*, 3rd ed., New York: Wiley.

### Statistics

In statistical inference, functions of the random sample will be used to map sample information into inferences regarding the relevant characteristics of a population or process. These functions, such as $T = t(X_1, \ldots, X_n)$, will be random variables whose probability densities depend on the joint density of the random sample. More specifically, inferential procedures will involve special functions known as **statistics**, defined as follows.

**Definition 6.3** ***Statistic***

> A real-valued function of observable random variables that is itself an observable random variable not depending on any unknown parameters.

By *observable* random variable, we simply mean a random variable whose numerical outcomes can actually be observed in the real world. Note the following example:

**Example 6.5** Let the outcome of a Beta-distributed random variable $X$ represent the proportion of a given day's telephone orders, received by the catalogue department of a large retail store, that are shipped the same day the order is received. Define the two random variables

$$Y = 100(X - .5) \quad \text{and} \quad W = a(X - b).$$

The random variable $Y$ is a *statistic* representing the number of percentage points above 50 percent that are shipped the same day. The random variable $W$ is *not* a statistic. It depends on the unknown values of $a$ and $b$, and until these values are specified, the random variable is *unobservable.* □

The reason for restricting our attention to statistics when attempting statistical inference is obvious. We cannot utilize a function of the random sample whose range elements are unobservable to make inferences about characteristics of the population or process from which we have sampled. In subsequent sections, we will examine a number of statistics that will be useful in statistical inference.

## 6.3 Empirical or Sample Distribution Function

There is a rather simple function of a random sample from a population distribution that can be used to provide an empirical characterization of the underlying population distribution from which a random sample is drawn. The function is called the **empirical distribution function** (EDF) (sometimes also referred to as the *sample distribution function*), and we will examine its definition and some of its properties in this section. After we have discussed the common measures used to judge the goodness of estimators in the next chapter, we will see that the EDF represents a useful estimator of the underlying population's cumulative distribution function. Furthermore, the EDF can be used to test hypotheses about the appropriate parametric family of distributions to which the population distribution belongs, as we will later examine in our discussion of hypothesis testing.

### EDF: Scalar Case

The EDF in the scalar case is defined as follows.

**Definition 6.4**
***Empirical distribution function, scalar case***

Let the scalar random variables $X_1, \ldots, X_n$ denote a random sample from some population distribution. Then the empirical distribution function is defined, for $t \in (-\infty, \infty)$, by

$$F_n(t) = n^{-1} \sum_{i=1}^{n} I_{(-\infty,t]}(X_i),$$

an outcome of which is defined by

$$\hat{F}_n(t) = n^{-1} \sum_{i=1}^{n} I_{(-\infty,t]}(x_i).$$

An outcome of the EDF can be defined alternatively using the size-of-set function $N(\cdot)$ as

$$\hat{F}_n(t) = \frac{N\left(\{x : x \leq t, x \in \{x_1, x_2, \ldots, x_n\}\}\right)}{n},$$

that is, $\hat{F}_n(t)$ equals the number of $x_i$'s in the outcome of the random sample that have values $\leq t$, divided by the sample size.

**Example 6.6** A random sample of size 10 from the population distribution of the yield per acre of a new wheat variety that a seed company has developed produced the following 10 outcomes of wheat yield, in bushels/acre:

$\{60, 71, 55, 50, 75, 78, 81, 78, 67, 90\}$.

Then the EDF is defined by

| $t$ | $\hat{F}_n(t)$ |
|---|---|
| $(-\infty, 50)$ | 0 |
| $[50, 55)$ | .1 |
| $[55, 60)$ | .2 |
| $[60, 67)$ | .3 |
| $[67, 71)$ | .4 |
| $[71, 75)$ | .5 |
| $[75, 78)$ | .6 |
| $[78, 81)$ | .8 |
| $[81, 90)$ | .9 |
| $[90, \infty)$ | 1.0 |

The graph of the EDF is given in Figure 6.2. □

Given Def. 6.4, it is apparent that the EDF is a random variable for each value of $t \in R$. In order to be able to assess the usefulness of the EDF in representing characteristics of the underlying population distribution, it will be informative to examine a number of important properties of the random variable $F_n(t)$. We begin by noting an important relationship between the binomial PDF and the PDF of $F_n(t)$.

**Theorem 6.1** *Let $F_n(t)$ be the EDF corresponding to a random sample of size n from a population distribution characterized by the CDF $F(t)$. Then the PDF of $F_n(t)$ is defined by*

$$P\left(\hat{F}_n(t) = \frac{j}{n}\right) = \begin{cases} \binom{n}{j}[F(t)]^j[1 - F(t)]^{n-j} & \textit{for } j \in \{0, 1, 2, \ldots, n\}, \\ 0 \quad \textit{otherwise}. \end{cases}$$

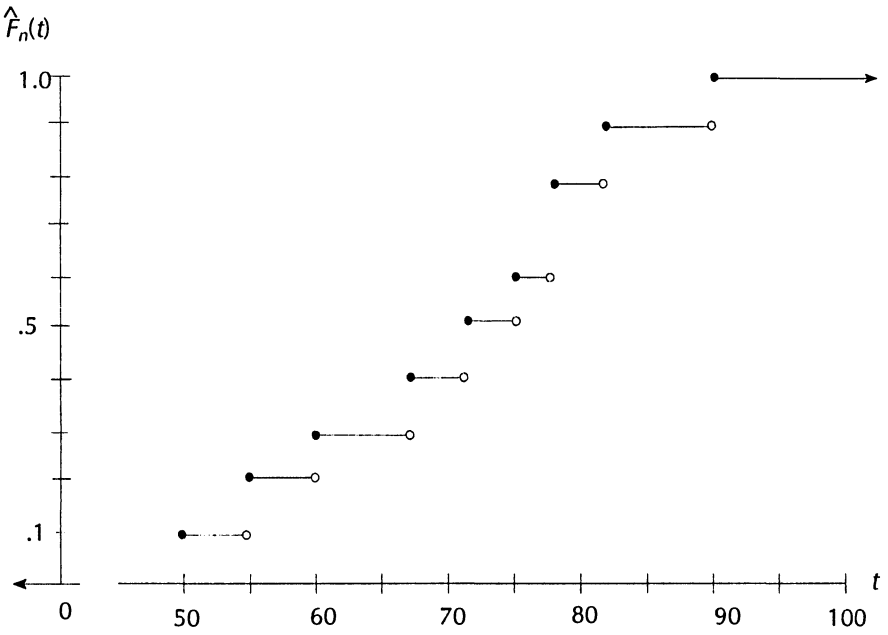


**Figure 6-2** EDF for wheat yields.

**Proof** From the definition of $F_n(t)$, it follows that

$$P\left(\hat{F}_n(t) = \frac{j}{n}\right) = P\left(\sum_{i=1}^{n} I_{(-\infty,t]}(x_i) = j\right).$$

Note that $Y_i = I_{(-\infty,t]}(X_i)$ is a Bernoulli random variable with $P(y_i = 1) = P(x_i \leq t) = F(t) = p$, and $P(y_i = 0) = P(x_i > t) = 1 - F(t) = 1 - p$, $\forall\, i = 1, \ldots, n$, and since $(X_1, \ldots, X_n)$ is a random sample from a population distribution, $Y_1, \ldots, Y_n$ are iid Bernoulli random variables. Then $\sum_{i=1}^{n} Y_i$ has a binomial density with parameters $n$ and $p = F(t)$, as

$$P\left(\sum_{i=1}^{n} I_{(-\infty,t]}(x_i) = j\right) = P\left(\sum_{i=1}^{n} y_i = j\right) = \binom{n}{j}[F(t)]^j[1 - F(t)]^{n-j}$$

for $j = 0, 1, 2, \ldots, n$, with all other values of $j$ assigned probability zero. ■

The implication of the preceding theorem is that for a *given* choice of $t \in (-\infty, \infty)$, $F_n(t)$ has the same probability distribution as the random variable $n^{-1}Z_n$, where $Z_n$ has a binomial distribution with parameters $n$ and $p = F(t)$. Now that we have discovered this probability distribution for $F_n(t)$, it is rather straightforward to derive the mean, variance, probability limit, and asymptotic distribution of $F_n(t)$:

**Theorem 6.2** *Let $F_n(t)$ be the EDF defined in Theorem 6.1. Then, $\forall\, t \in (-\infty,\infty)$,*

**a.** $\mathrm{E}F_n(t) = F(t)$,
**b.** $\mathrm{var}(F_n(t)) = n^{-1}[F(t)(1 - F(t))]$,

**c.** $\text{plim}\, F_n(t) = F(t)$,
**d.** $F_n(t) \overset{a}{\sim} N(F(t), n^{-1}[F(t)(1 - F(t))])$.

**Proof** Since $F_n(t)$ can be represented as $n^{-1}Z_n$, where $Z_n$ has a binomial density with $\mathrm{E}Z_n = nF(t)$ and $\text{var}(Z_n) = nF(t)(1 - F(t))$, properties (a) and (b) follow immediately. Property (c) follows from the fact that $F_n(t) \xrightarrow{m} F(t)$, since $\mathrm{E}F_n(t) = F(t)\ \forall\ n$ and $\text{var}(F_n(t)) \to 0$ as $n \to \infty$, which in turn implies $F_n(t) \xrightarrow{p} F(t)$. Property (d) follows from the fact that $F_n(t) = n^{-1}Z_n$ is the average of $n$ iid Bernoulli random variables, and it is known from the Lindberg–Levy CLT that this average has an asymptotic normal distribution with mean $F(t)$ and variance $n^{-1}[F(t)(1 - F(t))]$. ■

It will be seen from our discussion of estimator properties in Chapter 7 that the properties possessed by $F_n(t)$ will make it a good statistic to use in providing information about $F(t)$. In particular, the distribution of $F_n(t)$ is centered on $F(t)$; the variance, and thus, the "spread," of the distribution of $F_n(t)$ decreases as the sample size increases; and the probability that outcomes of $F_n(t)$ differ from $F(t)$ by any nonzero amount converges to zero as the sample size increases without bound.

The EDF can be used in a way analogous to the CDF to generate information on the probability that $x \in (a, b]$, namely, $\hat{P}_n(x \in (a, b]) = \hat{F}_n(b) - \hat{F}_n(a)$ provides an *empirical estimate* of the appropriate probability. Properties of the random variable $P_n(x \in (a, b]) = F_n(b) - F_n(a)$ generating these empirical estimates are established in the following theorem.

**Theorem 6.3** *Let $F_n(t)$ be the EDF defined in Theorem 6.1. Then $\forall\ t \in (-\infty, \infty)$, and for $a < b$,*

**a.** $\mathrm{E}[F_n(b) - F_n(a)] = F(b) - F(a)$,
**b.** $\text{var}\,(F_n(b) - F_n(a)) = n^{-1}[F(b) - F(a)]\,[1 - F(b) + F(a)]$,
**c.** $\text{plim}\,[F_n(b) - F_n(a)] = F(b) - F(a)$,
**d.** $F_n(b) - F_n(a) \overset{a}{\sim} N(F(b) - F(a),\ n^{-1}[F(b) - F(a)](1 - F(b) + F(a))$.

**Proof** Property (a) follows directly from applying the expectation operation to the linear combination of the two random variables and using Theorem 6.2, part (a). Property (c) follows from Theorem 6.2, part (c), and the fact that the probability limit of a sum is the sum of the probability limits.

To prove part (b), first note that our previous results concerning variances of linear combinations of random variables (Section 3.7) imply that

$$\text{var}(F_n(b) - F_n(a)) = \text{var}(F_n(b)) + \text{var}(F_n(a)) - 2\,\text{cov}(F_n(b), F_n(a)).$$

Values of the variance terms are known from Theorem 6.2, part (b). To find the value of the covariance term, first note that by definition,

$$\text{cov}(F_n(b), F_n(a)) = \text{cov}\left[n^{-1}\sum_{i=1}^{n} I_{(-\infty,b]}(X_i),\ n^{-1}\sum_{j=1}^{n} I_{(-\infty,a]}(X_j)\right]$$

$$= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{cov}(I_{(-\infty,b]}(X_i), I_{(-\infty,a]}(X_j))$$

$$= n^{-1}\,\text{cov}(I_{(-\infty,b]}(X_1)I_{(-\infty,a]}(X_1)).$$

The last equality follows from the fact that, for $i \neq j$, $I_{(-\infty,b]}(X_i)$ and $I_{(-\infty,a]}(X_j)$ are independent random variables since $X_i$ and $X_j$ are independent, which implies that the associated covariance terms are zero, and the $n$ remaining covariance terms are all identical and represented by the covariance term in the last equality involving the functions of $X_1$ (since $X_1, \ldots, X_n$ are iid). Then since $\text{cov}(Y, Z) = \text{E}YZ - \text{E}Y\text{E}Z$, by letting $Y = I_{(-\infty,b]}(X_1)$ and $Z = I_{(-\infty,a]}(X_1)$, we have that

$$\begin{aligned} cov(F_n(b), F_n(a)) &= n^{-1}[\text{E}I_{(-\infty,b]}(X_1)I_{(-\infty,a]}(X_1) - F(b)F(a)] \\ &= n^{-1}[F(a) - F(b)F(a)] \\ &= n^{-1}F(a)[1 - F(b)], \end{aligned}$$

where the next-to-last equality follows from the fact that $I_{(-\infty,b]}(x_1)I_{(-\infty,a]}(x_1) = I_{(-\infty,a]}(x_1)\ \forall\ x_1$, since $a < b$.

Having found the value of the covariance term, we finally have that

$$\begin{aligned} \text{var}(F_n(b) - F_n(a)) &= n^{-1}F(b)[1 - F(b)] + n^{-1}F(a)[1 - F(a)] - 2n^{-1}F(a)[1 - F(b)] \\ &= n^{-1}[F(b) - F(a)][1 - F(b) + F(a)]. \end{aligned}$$

To prove part (d), first note that $F_n(b)$ and $F_n(a)$ have an asymptotic bivariate normal distribution with mean vector $[F(b) \mid F(a)]'$ and covariance matrix

$$n^{-1}\Sigma = n^{-1}\left[\begin{array}{c|c} F(b)[1 - F(b)] & F(a)[1 - F(b)] \\ \hline F(a)[1 - F(b)] & F(a)[1 - F(a)] \end{array}\right],$$

which follows from applying the multivariate Lindberg–Levy CLT to the sample means of the iid Bernoulli random variables that define the random variables $F_n(b)$ and $F_n(a)$. Now define the function $g(F_n(b), F_n(a)) = F_n(b) - F_n(a)$, $g$ being a function of the two asymptotically normal random variables, and note that Theorem 5.40 is applicable. In particular,

$$\frac{\partial g}{\partial F_n(b)} = 1 \quad \text{and} \quad \frac{\partial g}{\partial F_n(a)} = -1,$$

so that $\mathbf{G} = [1 \quad -1]$ and thus $n^{-1}\mathbf{G}\Sigma\mathbf{G}' = n^{-1}[F(b) - F(a)][1 - F(b) + F(a)]$. Furthermore, it is clear that $g(F(b), F(a)) = F(b) - F(a)$. Then, Theorem 5.39 implies part (d) of the theorem. ■

The following example illustrates the use of the EDF to provide empirical estimates of probabilities.

**Example 6.7** Referring to the EDF in the wheat yield example (Ex. 6.6), provide an empirical calculation of the probability that (a) the wheat yield ≤ 76 bushels, and (b) the wheat yield > 62 bushels and ≤ 80 bushels.

**Answer:**

**a.** $\hat{P}(x \le 76) = \hat{F}_n(76) = .6.$
**b.** $\hat{P}(62 < x \le 80) = \hat{F}_n(80) - \hat{F}_n(62) = .8 - .3 = .5.$ □

It can be shown that the convergence of $F_n(t)$ to $F(t)$ in probability as $n \to \infty$ can be strengthened to almost-sure convergence, i.e., $\lim_{n\to\infty} F_n(t) = F(t)$ occurs with probability 1. Furthermore, it can be shown that $\lim_{n\to\infty} F_n(t) = F(t)$ occurs *simultaneously* $\forall\, t \in (-\infty, \infty)$ with probability 1. These results are given by the celebrated **Glivenko–Cantelli theorem,** which involves convergence to zero of the supremum (maximum if it exists) over all $t \in (-\infty, \infty)$ of the absolute difference between $F_n(t)$ and $F(t)$.

**Theorem 6.4** (Glivenko–Cantelli's Theorem) *Let* $D_n = \sup_t |F_n(t) - F(t)|$. *Then* $P(\lim_{n\to\infty} D_n = 0) = 1$.

**Proof** See V. Fabian and J. Hannan (1985), *Introduction to Probability and Mathematical Statistics*. New York: Wiley, pp. 80–82. ■

Note that the Glivenko–Cantelli theorem provides important additional information on the use of outcomes of $F_n(t)$ as a means of providing empirical representations of $F(t)$. In particular, the theorem implies that the sequence of *functions* $\{F_n\}$ converges as $n \to \infty$ to the *function F with probability 1*; this interpretation is supported by the fact that $F_n(t) \to F(t)\ \forall\, t$ with probability 1. Thus, for large enough $n$, the EDF represents a useful approximation to the CDF, $F$, *over its entire domain*, and it is the Glivenko–Cantelli theorem that provides rigorous justification for inferring the shape and functional form of $F$ from the shape and functional form of $F_n$.

### EDF: Multivariate Case

The empirical distribution function can be extended to the case where the random sample from the population distribution consists of a collection of multivariate random variables, as follows.

**Definition 6.5** ***Empirical distribution function, multivariate case***

Let the $(k \times 1)$ random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ denote a random sample from some population distribution. Then the empirical distribution function is defined for $\mathbf{t} = [t_1, \ldots, t_k]' \in R^k$ and $A(\mathbf{t}) = \times_{i=1}^{k}(-\infty, t_i]$ as

$$F_n(t) = n^{-1} \sum_{i=1}^{n} I_{A(\mathbf{t})}(\mathbf{X}_i),$$

an outcome of which is defined by

$$\hat{F}_n(t) = n^{-1} \sum_{i=1}^{n} I_{A(\mathbf{t})}(\mathbf{x}_i).$$

An outcome of the EDF can be defined alternatively using the size-of-set function $N(\cdot)$ as

$$\hat{F}_n(\mathbf{t}) = \frac{N(\{\mathbf{x} : \mathbf{x} \leq \mathbf{t}, \mathbf{x} \in \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}\})}{n},$$

i.e., $\hat{F}_n(t)$ equals the number of $\mathbf{x}_i$ *vectors* in the outcome of the random sample that have values $\leq$ the *vector* $\mathbf{t}$, divided by the sample size.

The properties of the EDF in the multivariate case parallel those of the EDF in the scalar case. In particular, the previous theorems apply to the multivariate case by simply reinterpreting $\mathbf{t}$, $\mathbf{a}$, and $\mathbf{b}$ as $(k \times 1)$ vectors instead of scalars in the statement of the theorems, and changing the condition $\forall\ t \in (-\infty, \infty)$ to $\forall\ \mathbf{t} \in R^k$. The proofs of Theorems 6.1, 6.2, and 6.3 in the multivariate case are analogous to the scalar case and are left to the reader.

## 6.4 Sample Moments and Sample Correlation

Based on a random sample from a population distribution, which we assume is the case throughout this section, statistics called *sample moments* can be defined that represent sample counterparts to the moments of the population distribution (henceforth called **population moments**). The sample moments have properties that make them useful for estimating the values of corresponding population moments. Sample moments also form the basis for the *method-of-moments* estimation procedure, which can be used to provide information on other characteristics of the population distribution besides moments. The method-of-moments procedure will be examined in Chapter 8.

The definitions of the various sample moments can all be unified through the use of the empirical distribution function concept. Specifically, the *empirical substitution principle* leads to the appropriate sample moment definitions.

**Definition 6.6**
***Empirical substitution principle***

Let $X_1, \ldots, X_n$ be a random sample from a population distribution having CDF $F$. Let $q = q(F)$ be any function of $F$. Then the **empirical substitution principle** representation of $q = q(F)$ is given by $\hat{q} = q(\hat{F}_n)$, where $\hat{F}_n$ is the EDF outcome associated with the random sample.

### Scalar Case

In order to use the empirical substitution principle to define sample moments when $(X_1, \ldots, X_n)$ is a collection of scalar random variables, first note that moments of the population distribution about the origin or mean can be expressed as functions of the CDF $F$. Specifically, letting $\mathrm{E}_F$ denote an expectation taken with respect to the probability distribution implied by $F$, we have $\mu'_r = \mathrm{E}_F(X^r)$ and $\mu_r = \mathrm{E}_F(X - \mathrm{E}_F(X))^r$, which are functions of $F$. Substituting the EDF outcome, $\hat{F}_n$, for $F$ when taking the expectations leads to the definition of sample moments about the origin and mean via the empirical substitution principle.

**Definition 6.7**
***Sample moments about the origin and mean***

Let the scalar random variables $X_1, \ldots, X_n$ be a random sample with EDF outcome $\hat{F}_n$. Then outcomes of the $r$th order sample moments about the origin and mean are defined as:

Sample moments about the origin: $m_r' = E_{\hat{F}_n}(X^r)$,

Sample moments about the mean: $m_r = E_{\hat{F}_n}(X - \mathrm{E}_{\hat{F}_n}(X))^r$.

In effect, when defining sample moments, one proceeds as if $X$ had the CDF $\hat{F}_n$, and then one calculates *population* moments associated with the population distribution defined by $\hat{F}_n$. Then the computational difference between population moments as defined in Section 3.4 and the sample moments defined here relates to which probability distribution is used in taking the expectations–the one implied by $F$ or by $\hat{F}_n$.

As it turns out, expectations taken with respect to the probability distribution represented by the EDF outcome $\hat{F}_n$ have a common mathematical definition, regardless of the form of the underlying CDF, $F$. To see this, first recall that in *either* the discrete or continuous case, $\hat{F}_n$ is a step function whose incremental value at each step equals the observed sample relative frequency of the random-variable outcome corresponding to the step (recall Figure 6.2). Since by its definition $\hat{F}_n$ can always be interpreted as a CDF for some *discrete* random-variable, the value of an incremental step can be interpreted as the probability assigned to the corresponding random-variable outcome by $\hat{F}_n$. In the case where an outcome value, say $x$, is observed only once in the outcome of a random sample of size $n$, it follows that the probability is assigned as $\hat{p}(x) = 1/n$. In general, the probability assigned by $\hat{F}_n$ to an outcome $x$ is given by

$$\hat{p}(x) = n^{-1} \sum_{i=1}^{n} I_{\{x\}}(x_i),$$

which is the relative frequency of the occurrence of outcome $x$ in the random sample outcome $(x_1, \ldots, x_n)$.

Now let $\hat{R}(X) = \{x : \hat{p}(x) > 0\}$ be the set of $x$ values assigned a positive density weighting by $\hat{p}(x)$, i.e., $\hat{R}(X)$ is the collection of *unique* values in the sample outcome $(x_1, \ldots, x_n)$. It follows that the expectation of $g(X)$ with respect to $\hat{p}(x)$, or equivalently with respect to the probability distribution implied by $\hat{F}_n$, is

$$\begin{aligned}\mathrm{E}_{\hat{F}_n} g(X) &= \Sigma_{x \in \hat{R}(X)} g(x)\hat{p}(x) = \sum_{x \in \hat{R}(X)} g(x) n^{-1} \sum_{i=1}^{n} I_{\{x\}}(x_i) \\ &= n^{-1} \Sigma_{x \in \hat{R}(X)} \sum_{i=1}^{n} g(x_i) I_{\{x\}}(x_i) = n^{-1} \sum_{i=1}^{n} g(x_i),\end{aligned}$$

where the third equality follows because only terms for which $x_i = x$ affect the value of the inner summation term. Then, by defining $g(X) = X^r$ or $g(X) =$

$(X - \mathrm{E}_{\hat{F}_N}(X))^r$ and taking expectations with respect to $\hat{F}_n$, we obtain the following alternative definition of sample moment outcomes.

**Definition 6.8** ***Sample moments about the origin and mean***

Assume the conditions of Def. 6.7. Then sample moment outcomes can be defined as

Sample moments about the origin: $m'_r = n^{-1} \sum_{i=1}^{n} x_i^r$,

Sample moments about the mean: $m_r = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^r$,

where $\bar{x}_n = m'_1 = n^{-1} \sum_{i=1}^{n} x_i$.

We emphasize that regardless of which representation, Def. 6.7 or Def. 6.8, is used in defining sample moment outcomes, all of the properties of population moments $\mu'_r$ and $\mu_r$ presented in Chapter 3 apply equally well to the outcomes of sample moments $m'_r$ and $m_r$. This is so because sample moment outcomes can be interpreted ultimately as (population) moments of a population distribution defined by the discrete CDF $\hat{F}_n$.

We now present a number of important properties of sample moments about the origin that are suggestive of the usefulness of sample moment outcomes for estimating the values of corresponding population moments.

**Theorem 6.5** (Properties of $M'_r$) *Let $M'_r = n^{-1} \sum_{i=1}^{n} X_i^r$ be the rth sample moment about the origin for a random sample $(X_1, \ldots, X_n)$ from a population distribution. Then, assuming the appropriate population moments exist,*

a. $\mathrm{E}M'_r = \mu'_r$,
b. $\mathrm{var}(M'_r) = n^{-1}[\mu'_{2r} - (\mu'_r)^2]$,
c. $\mathrm{plim}\, M'_r = \mu'_r$,
d. $(M'_r - \mu'_r)/[\mathrm{var}(M'_r)]^{1/2} \xrightarrow{d} N(0,1)$,
e. $M'_r \overset{a}{\sim} N(\mu'_r, \mathrm{var}(M'_r))$.

**Proof**

a. $\mathrm{E}M'_r = \mathrm{E}n^{-1} \sum_{i=1}^{n} X_i^r = n^{-1} \sum_{i=1}^{n} \mathrm{E}X_i^r = n^{-1} \sum_{i=1}^{n} \mu'_r = \mu'_r$, since $\mathrm{E}X_i^r = \mu'_r$ $\forall\, i$ because $(X_1, \ldots, X_n)$ is a random sample with iid elements.
b. $\mathrm{var}(M'_r) = \mathrm{var}(n^{-1} \sum_{i=1}^{n} X_i^r) = n^{-2}\, \mathrm{var} \sum_{i=1}^{n} X_i^r$, and note that the random variables $X_1^r, \ldots, X_n^r$ in the sum are iid since $X_i^r$ is the same real-valued function of $X_i$ $\forall\, i$, the $X_i$'s are identically distributed, and Theorem 2.9 applies.

   It follows from independence and the results on variances of linear combinations of random variables that

$$\mathrm{var}(M'_r) = n^{-2} \sum_{i=1}^{n} \mathrm{var}(X_i^r) = n^{-2} \sum_{i=1}^{n} [\mu'_{2r} - (\mu'_r)^2] = n^{-1}[\mu'_{2r} - (\mu'_r)^2].$$

**c.** Since $\mathrm{E}M'_r = \mu'_r \ \forall\, n$, and since $\mathrm{var}(M'_r) \to 0$ as $n \to \infty$, then Corollary 5.2 with $k = 1$ implies that $M'_r \xrightarrow{m} \mu'_r$ so that $\mathrm{plim}\, M'_r = \mu'_r$.[4]

**d. and e.** Since $(X_1^r, \ldots, X_n^r)$ are iid random variables with $\mathrm{E}X_i^r = \mu'_r$ and $\mathrm{var}(X_i^r) = \mu'_{2r} - (\mu'_r)^2 \ \forall\, i$, it follows upon substitution into the Lindberg–Levy central limit theorem that

$$Z_n = \frac{\sum_{i=1}^n X_i^r - n\mu'_r}{n^{1/2}[\mu'_{2r} - (\mu'_r)^2]^{1/2}} \xrightarrow{d} N(0,1).$$

Multiplying by 1 in the special form $n^{-1}/n^{-1}$ yields

$$Z_n = \frac{M'_r - \mu'_r}{n^{-1/2}[\mu'_{2r} - (\mu'_r)^2]^{1/2}} \xrightarrow{d} N(0,1),$$

which in turn implies by Def. 5.10 that

$$M'_r \overset{a}{\sim} N(\mu'_r, n^{-1}[\mu'_{2r} - (\mu'_r)^2]). \qquad \blacksquare$$

All told, the properties of $M'_r$ presented in Theorem 6.5 indicate that the expected value of a sample moment is equal to the value of the corresponding population moment, the variance of the sample moment monotonically decreases and converges to zero as $n \to \infty$, the sample moment converges in probability to the value of the corresponding population moment, and the sample moment is approximately normally distributed for large $n$. In the context of utilizing outcomes of $M'_r$ as estimates of $\mu'_r$, the properties indicate that the outcomes correctly estimate $\mu'_r$ *on average*, the *spread* of the estimates decreases and the outcomes of $M'_r$ become arbitrarily close to $\mu'_r$ with probability approaching 1 as the sample size increases, and the outcomes of $M'_r$ are approximately normally distributed around $\mu'_r$ for large enough sample sizes. The fact that outcomes of $M'_r$ are correct on average and become highly accurate individually with high probability as $n$ increases contributes to $M'_r$'s being a useful *estimator* of $\mu'_r$, as will be discussed further in Chapter 7. The fact that $M'_r$ is approximately normally distributed for large enough $n$ will facilitate testing hypotheses about the value of $\mu'_r$, to be discussed in Chapters 9 and 10.

The first-order sample moment about the origin is of particular importance in a number of point estimation and hypothesis-testing situations, and it is given a specific name and symbol.

**Definition 6.9**
***Sample mean***

Let $(X_1, \ldots, X_n)$ be a random sample. The **sample mean** is defined by

$$\bar{X}_n = n^{-1} \sum_{i=1}^n X_i = M'_1.$$

[4] An alternative proof, requiring only that moments up to the rth-order exist, can be based on Khinchin's WLLN. Although we will not use the property later, the reader can utilize Kolmogorov's SLLN to also demonstrate that $M'_r \xrightarrow{as} \mu'_r$ (see Chapter 5).

Based on sample moment properties, we know that $\mathrm{E}\bar{X}_n = \mu$, $\mathrm{var}(\bar{X}_n) = n^{-1}(\mu'_2 - \mu^2) = \sigma^2/n$, $\mathrm{plim}\,\bar{X}_n = \mu$, and $\bar{X}_n \overset{a}{\sim} N(\mu, \sigma^2/n)$. As a preview to a particular problem of statistical inference, note that $\bar{X}_n$ has properties that might be considered useful for estimating the population mean, $\mu$. In particular, the PDF of $\bar{X}_n$ is centered on $\mu$, and as the size of the random sample increases, the density of $\bar{X}_n$ concentrates within a small neighborhood of points around $\mu$ so that it becomes ever more improbable that an outcome of $\bar{X}_n$ would occur far from $\mu$. Thus, outcomes of $\bar{X}_n$ might be useful as estimates of the unknown value of a population mean.

The result on asymptotic normality extends to vectors of sample moments about the origin, in which case a multivariate, asymptotic, normal density is appropriate.

**Theorem 6.6** (Multivariate, Asymptotic Normality of Sample Moments About the Origin)

$$n^{1/2}\begin{bmatrix} M'_1 - \mu'_1 \\ \vdots \\ M'_r - \mu'_r \end{bmatrix} \xrightarrow{d} N\left(\underset{r\times 1}{[0]}, \underset{r\times r}{\Sigma}\right) \quad \textit{and} \quad \begin{bmatrix} M'_1 \\ \vdots \\ M'_r \end{bmatrix} \overset{a}{\sim} N\left((\mu'_1, \ldots, \mu'_r)', n^{-1}\Sigma\right),$$

*where the nonsingular covariance matrix* $\Sigma$ *has a typical* $(j,k)$ *entry* $\sigma_{jk} = \mu'_{j+k} - \mu'_j\mu'_k$.

**Proof** The proof relies on the multivariate version of the Lindberg-Levy central limit theorem. Let $\mathbf{Y}_i = (X_i^1, \ldots, X_i^r)'$. Since $X_1, \ldots, X_n$ is a random sample with iid elements, it follows from Theorem 2.9 that $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ are independent $(r \times 1)$ random vectors with $\mathrm{E}\mathbf{Y}_i = \boldsymbol{\mu}$ and $\mathbf{Cov}(\mathbf{Y}_i) = \Sigma$, $\forall\, i$, where $\boldsymbol{\mu} = (\mu'_1, \ldots, \mu'_r)'$. Then, given that $\Sigma$ is nonsingular, the multivariate Lindberg-Levy CLT applies, establishing the convergence in distribution result, which in turn implies the asymptotic density result.

The typical entry in $\Sigma$ is given by

$$\sigma_{jk} = \mathrm{cov}\left(X^j, X^k\right) = \mathrm{E}\left(X^j - \mu'_j\right)\left(X^k - \mu'_k\right)$$

$$= \mathrm{E}\left(X^{j+k} - \mu'_j X^k - \mu'_k X^j + \mu'_j\mu'_k\right) = \mu'_{j+k} - \mu'_j\mu'_k. \qquad \blacksquare$$

We will not study properties of sample moments *about the mean* in detail. Unlike the case of sample moments about the origin, the properties of higher-order sample moments about the mean become progressively more difficult to analyze. We will concentrate on properties of the second-order sample moment about the mean, called the *sample variance*. The reader interested in the general properties of sample moments about the mean can refer to R. Serfling (1980) *Approximation Theorems of Mathematical Statistics* New York: John Wiley, 1980, pp. 69–74.

**Definition 6.10**
***Sample variance***

Let $X_1, \ldots, X_n$ be a random sample of size $n$. The **sample variance** is defined as[5]

$$S_n^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = M_2.$$

Some important properties of the sample variance are presented in the following theorem.

**Theorem 6.7** (Properties of $S_n^2$) *Let $S_n^2$ be the sample variance for a random sample $(X_1,\ldots,X_n)$ from a population distribution. Then, assuming the appropriate population moments exist,*

**a.** $\mathrm{E}S_n^2 = ((n-1)/n)\sigma^2$,
**b.** $\mathrm{var}(S_n^2) = n^{-1}[((n-1)/n)^2\mu_4 - ((n-1)(n-3)/n^2)\sigma^4]$,
**c.** $\mathrm{plim}\, S_n^2 = \sigma^2$,
**d.** $n^{1/2}(S_n^2 - \sigma^2)\xrightarrow{d}N(0,\mu_4 - \sigma^4)$,
**e.** $S_n^2 \overset{a}{\sim} N(\sigma^2, n^{-1}(\mu_4 - \sigma^4))$.

**Proof** **a.**

$$\mathrm{E}S_n^2 = \mathrm{E}\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n} = \mathrm{E}\frac{\sum_{i=1}^n (X_i - \mu + \mu - \bar{X}_n)^2}{n}$$

$$= \mathrm{E}\frac{\left[\sum_{i=1}^n (X_i - \mu)^2\right]}{n} - \mathrm{E}\frac{\left[n(\mu - \bar{X}_n)^2\right]}{n} = \sigma^2 - \frac{\sigma^2}{n} = \left(\frac{n-1}{n}\right)\sigma^2.$$

**b.** The proof follows from expressing $\mathrm{var}(S_n^2) = \mathrm{E}(S_n^2 - \sigma^2)^2$ in terms of $X_i$'s and taking expectations. The proof is straightforward conceptually, but quite tedious algebraically. The details are left to the reader, or see R. G. Krutchkoff (1970), *Probability and Statistical Inference*, New York: Gordon and Breach, pp. 154–157.

**c.** Since $\mathrm{E}S_n^2 = ((n-1)/n)\sigma^2 \to \sigma^2$ as $n \to \infty$, and since $\lim_{n\to\infty}\mathrm{var}(S_n^2) = 0$, then Corollary 5.2 implies that $S_n^2 \xrightarrow{m} \sigma^2$, so that $\mathrm{plim}\, S_n^2 = \sigma^2$.

**d. and e.** First note that

$$nS_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n (X_i - \mu + \mu - \bar{X}_n)^2$$

$$= \sum_{i=1}^n (X_i - \mu)^2 + 2(\mu - \bar{X}_n)\sum_{i=1}^n (X_i - \mu) + n(\mu - \bar{X}_n^2),$$

[5] Some authors define the sample variance as $S_n^2 = \left(n/(n-1)\right)M_2$, so that $\mathrm{E}S_n^2 = \sigma^2$, which identifies $S_n^2$ as an *unbiased* estimator of $\sigma^2$ (see Section 7.2). However, this definition would be inconsistent with the aforementioned fact that $M_2$, and not $(n/(n-1))M_2$, is the second moment about the mean, and thus the *variance*, of the sample (empirical) distribution function, $\hat{F}_n$.

so that

$$n^{1/2}\left(S_n^2 - \sigma^2\right) = \frac{\sum_{i=1}^n (X_i - \mu)^2 - n\sigma^2}{n^{1/2}} + 2\left(\mu - \bar{X}_n\right) n^{-1/2} \sum_{i=1}^n (X_i - \mu)$$
$$+ n^{1/2}(\mu - \bar{X}_n)^2.$$

Of the three terms added together in the previous expression, the latter two converge in probability to zero. Regarding the third term, $\text{plim}(n^{1/2})\left(\mu - \bar{X}_n\right)^2 = \text{plim}(n^{1/2}o_p(n^{-1/2})) = 0$. In the second term, $n^{-1/2}\sum_{i=1}^n (X_i - \mu) \xrightarrow{d} N(0, \sigma^2)$ by the LLCLT, and $\text{plim}(\mu - \bar{X}_n) = \text{plim}\, \mu - \text{plim}\, \bar{X}_n = 0$. So by Slutsky's theorem, the second term converges in distribution, and thus in probability, to the constant 0.

In the first term, let $Y_i = (X_i - \mu)^2$, and note that $\text{E}Y_i = \sigma^2$ and $\text{var}(Y_i) = \mu_4 - \sigma^4$. Then since the $Y_i$'s are iid, the first term converges in distribution to $N(0, \mu_4 - \sigma^4)$ by the LLCLT and Slutsky's theorem. Thus, by Slutsky's theorem,

$$n^{1/2}\left(S_n^2 - \sigma^2\right) \xrightarrow{d} N(0, \mu_4 - \sigma^4), \quad \text{so that } S_n^2 \overset{a}{\sim} N\left(\sigma^2, n^{-1}(\mu_4 - \sigma^4)\right). \qquad \blacksquare$$

As another preview to a particular problem of statistical inference, note that $S_n^2$ has characteristics that might be considered useful for estimating the population variance, $\sigma^2$. Specifically, the distribution of $S_n^2$ becomes centered on $\sigma^2$ as $n \to \infty$, and as the size, $n$, of the random sample increases, the density of $S_n^2$ concentrates within a small neighborhood of points around $\sigma^2$ so that it becomes highly probable that an outcome of $S_n^2$ will occur close to $\sigma^2$.

**Example 6.8** Calculate outcomes of the sample mean and sample variance for the wheat yield data presented in Ex. 6.6.

**Answer**: The outcomes are respectively

$$\bar{x} = \sum_{i=1}^{10} \frac{x_i}{10} = 70.5 \quad \text{and} \quad s^2 = \sum_{i=1}^{10} \frac{(x_i - \bar{x})^2}{10} = 140.65. \qquad \square$$

### Multivariate Case

When the random sample consists of a collection of $k$-variate random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ (i.e., where $k$ characteristics are observed for each of the $n$ sample observations), one can define sample means, sample variances, or any $r$th-order sample moment for *each* of the $k$ entries in the $\mathbf{X}_i$-vectors. Furthermore, the concept of joint *sample* moments between pairs of entries in the $\mathbf{X}_i$ vectors becomes relevant in the multivariate case, and in particular one can define the notion of *sample covariance* and *sample correlation*.

The method of defining sample moments in the multivariate case is analogous to the approach in the scalar case—use the empirical substitution principle. All of the aforementioned moments can be defined as expectations taken with respect to the probability distribution defined by the (joint) EDF outcome,

$\hat{F}_n$, just as population moments can be defined using the population distribution represented by the (joint) CDF, $F$. Using analogous reasoning to the scalar case, the PDF implied by $\hat{F}_n$ is given by

$$\hat{p}(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} I_{\{\mathbf{x}\}}(\mathbf{x}_i),$$

i.e., the probability assigned to the *vector* outcome $\mathbf{x}$ is the relative frequency of the occurrence of $\mathbf{x}$ in the random sample of *vector* outcomes $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Then, following an argument analogous to the scalar case, expectations of $g(\mathbf{X})$ with respect to the discrete CDF $\hat{F}_n$ can be defined as

$$\mathrm{E}_{\hat{F}_n} g(\mathbf{X}) = \sum_{x \in \hat{R}(\mathbf{X})} \cdots \sum g(\mathbf{x})\hat{p}(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} g(\mathbf{x}_i).$$

By appropriate definitions of $g(\mathbf{X})$, one can define $r$th sample moments about the origin and mean for each entry in the vector $\mathbf{X}$, as well as covariances between entries in the vector $\mathbf{X}$. In particular, these function definitions would be respectively $g(\mathbf{X}) = X[j]^r$, $g(\mathbf{X}) = (X[j]-\bar{x}[j])^r$, and $g(\mathbf{X}) = (X[i]-\bar{x}[i])(X[j]-\bar{x}[j])$, where $\bar{x}[\ell] = \mathrm{E}_{\hat{F}_n} X[\ell]$, leading to the following definition of sample moments in the multivariate case.

**Definition 6.11**
***Sample Moments, Multivariate Case***

Let the $(k \times 1)$ vector random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from a population distribution. Then the following outcomes of sample moments can be defined for $j$ and $\ell \in \{1, 2, \ldots, k\}$:

Sample moments about the origin: $m'_r[j] = n^{-1} \sum_{i=1}^{n} x_i[j]^r$;

Sample means: $\bar{x}[j] = m'_1[j] = n^{-1} \sum_{i=1}^{n} x_i[j]$;

Sample moments about the mean: $m_r[j] = n^{-1} \sum_{i=1}^{n} (x_i[j] - \bar{x}[j])^r$;

Sample variances: $s^2[j] = m_2[j] = n^{-1} \sum_{i=1}^{n} (x_i[j] - \bar{x}[j])^2$;

Sample covariance: $s_{j\ell} = n^{-1} \sum_{i=1}^{n} (x_i[j] - \bar{x}[j])(x_i[\ell] - \bar{x}[\ell])$.

The properties of each of the sample moments about the origin and each of the sample variances are precisely the same as in the scalar case. The proofs are the same as in the scalar case upon utilizing a *marginal* probability distribution for each $X[j]$, $j = 1, \ldots, k$. The sample covariance has no counterpart in the scalar case, and we examine its properties below.

*Sample Covariance* For clarity of exposition, we will examine the case where $k = 2$ and distinguish the two random variables involved by $X$ and $Y$. A random sample of size $n$ will then be given by the $n$ two-tuples $((X_1, Y_1), \ldots, (X_n, Y_n))$, which can be interpreted as representing observations on two characteristics, $x_i$ and $y_i$, for each of $n$ outcomes of an experiment. Applications to cases where $k > 2$ are straightforwardly accomplished by applying the subsequent results to any pair of random variables in the $(k \times 1)$ vector $\mathbf{X}$.

**Theorem 6.8** (Properties of Sample Covariance) *Let $((X_1,Y_1), \ldots, (X_n,Y_n))$ be a random sample from a population distribution, and let $S_{XY}$ be the sample covariance between $X$ and $Y$. Then, assuming the appropriate population moments exist,*

**a.** $ES_{XY} = ((n-1)/n)\sigma_{XY}$,
**b.** $\text{var}(S_{XY}) = n^{-1}(\mu_{2,2} - (\mu_{1,1})^2) + o(n^{-1})$,
**c.** $\text{plim}\, S_{XY} = \sigma_{XY}$,
**d.** $S_{XY} \overset{a}{\sim} N(\sigma_{XY}, n^{-1}(\mu_{2,2} - (\mu_{1,1})^2))$.

**Proof** **a.** Examine the $i$th term in the sum, and note that

$$E(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = E\left[X_iY_i - \frac{1}{n}X_i\sum_{j=1}^{n} Y_j - \frac{1}{n}Y_i\sum_{j=1}^{n} X_j + \frac{1}{n^2}\sum_{j=1}^{n} X_j \sum_{j=1}^{n} Y_j\right].$$

Since $(X_1, Y_1), \ldots, (X_n, Y_n)$ are iid, $X_i$ is independent of $X_j$ and $Y_j$, for $i \neq j$, and $Y_i$ is independent of $X_j$ and $Y_j$ for $i \neq j$. Furthermore, since the $(X_i, Y_i)$'s have the same joint density function, and thus the same population moments $\forall\, i$, it follows that

$$\begin{aligned} E(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) &= \mu'_{1,1} - \frac{2}{n}\left[\mu'_{1,1} + (n-1)\mu_X\mu_Y\right] + \frac{1}{n^2}\left[n\mu'_{1,1} + n(n-1)\mu_X\mu_Y\right] \\ &= \left(\frac{n-1}{n}\right)(\mu'_{1,1} - \mu_X\mu_Y) = \left(\frac{n-1}{n}\right)\sigma_{XY}. \end{aligned}$$

Using this result, we find that

$$ES_{XY} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{n-1}{n}\right)\sigma_{XY} = \left(\frac{n-1}{n}\right)\sigma_{XY}.$$

**b.** The procedure used to derive the approximation is based on Taylor series expansions and can be found in M. Kendall, and A. Stuart, *The Advanced Theory of Statistics, Vol. 1*, 4th ed. pp. 246–250. An alternative motivation for the approximation is given below in the proof of part (d).

**c.** Since $ES_{XY} = ((n-1)/n)\sigma_{XY} \to \sigma_{XY}$ and $\text{var}(S_{XY}) \to 0$ as $n \to \infty$, it follows by Corollary 5.2 that $S_{XY} \longrightarrow \sigma_{XY}$, which in turn implies that $\text{plim}\, S_{XY} = \sigma_{XY}$.

**d.** First note that the sample covariance can alternatively be written as

$$S_{XY} = \left(\sum_{i=1}^{n}\frac{X_iY_i}{n} - \bar{X}_n\bar{Y}_n\right) = (M'_{1,1} - \bar{X}_n\bar{Y}_n),$$

where $M'_{1,1} = (1/n)\sum_{i=1}^{n} X_i Y_i$.[6]

Now examine the iid random vectors

$$\begin{bmatrix} X_i Y_i \\ X_i \\ Y_i \end{bmatrix}, \quad i = 1, \ldots, n,$$

and note that $\forall\, i$,

$$\mathrm{E}\begin{bmatrix} X_i Y_i \\ X_i \\ Y_i \end{bmatrix} = \begin{bmatrix} \mu'_{1,1} \\ \mu_X \\ \mu_Y \end{bmatrix},$$

$$\begin{aligned}\mathbf{\Sigma} = \mathbf{Cov}\begin{bmatrix} X_i Y_i \\ X_i \\ Y_i \end{bmatrix} &= \mathrm{E}\begin{bmatrix} X_i Y_i - \mu'_{1,1} \\ X_i - \mu_X \\ Y_i - \mu_Y \end{bmatrix}\begin{bmatrix} X_i Y_i - \mu'_{1,1} \\ X_i - \mu_X \\ Y_i - \mu_Y \end{bmatrix}' \\ &= \begin{bmatrix} \mu'_{2,2} - (\mu'_{1,1})^2 & \mu'_{2,1} - \mu_X \mu'_{1,1} & \mu'_{1,2} - \mu_Y \mu'_{1,1} \\ (\text{Symmetric}) & \sigma_X^2 & \sigma_{XY} \\ & & \sigma_Y^2 \end{bmatrix}.\end{aligned}$$

Then, by the multivariate Lindberg–Levy central limit theorem,

$$n^{1/2}\begin{bmatrix} M'_{1,1} - \mu'_{1,1} \\ \bar{X}_n - \mu_X \\ \bar{Y}_n - \mu_Y \end{bmatrix} \xrightarrow{d} N([\mathbf{0}], \mathbf{\Sigma}).$$

Now let $g(M'_{1,1}, \bar{X}_n, \bar{Y}_n) = M'_{1,1} - \bar{X}_n \bar{Y}_n$, and note by Theorem 5.39 that $g(\cdot)$ has an asymptotic normal distribution given by $g(\cdot) \overset{a}{\sim} N(\mu'_{1,1} - \mu_X \mu_Y,$ $n^{-1}\mathbf{G\Sigma G}' = N(\sigma_{XY}, n^{-1}\mathbf{G\Sigma G}')$, where $\mathbf{G}$ is the row vector of derivatives of $g$ with respect to its three arguments, evaluated at the expected values of $M'_{1,1}, \bar{X}_n, \bar{Y}_n$, i.e., $\mathbf{G}_{1\times 3} = [1 \quad -\mu_Y \quad -\mu_X]$. The variance of the asymptotic distribution of $g(\cdot)$ is then represented (after some algebraic simplification) as

$$\begin{aligned} n^{-1}\mathbf{G\Sigma G}' = n^{-1} \big[&\mu'_{2,2} - (\mu'_{1,1})^2 + 6\mu_X \mu_Y \mu'_{1,1} - 4\mu_X^2 \mu_Y^2 \\ &-2\mu_Y \mu'_{2,1} - 2\mu_X \mu'_{1,2} + \mu_y^2 \mu'_{2,0} + \mu_x^2 \mu'_{0,2}\big], \end{aligned}$$

and it can be shown by defining $\mu_{2,2}$ and $\mu_{1,1}$ in terms of moments about the origin that the preceding bracketed expression is identically equal to $\mu_{2,2} - (\mu_{1,1})^2$. Thus, $g(\cdot) \overset{a}{\sim} N(\sigma_{XY}, n^{-1}(\mu_{2,2} - (\mu_{1,1})^2))$. ■

[6] This is an example of a **joint sample moment about the origin**, the general definition being given by

$$M'_{r,s} = \frac{1}{n}\sum_{i=1}^{n} X_i^r Y_i^s.$$

The definition for the case of a **joint sample moment about the mean** replaces $X_i$ with $X_i - \bar{X}_i$, $Y_i$ with $Y_i - \bar{Y}_i$, and $M'_{r,s}$ with $M_{r,s}$.

As we have similarly argued in the case of the sample mean and sample variance, the sample covariance has properties that might be considered useful for estimating its population counterpart, the covariance $\sigma_{XY}$. In particular, the distribution of $S_{XY}$ becomes centered on $\sigma_{XY}$ as $n \to \infty$, and as the sample size $n$ increases, the density of $S_{XY}$ concentrates within a small neighborhood of $\sigma_{XY}$ so that it becomes highly probable that an outcome of $S_{XY}$ will occur close to $\sigma_{XY}$.

*Sample Correlation* Having defined the notion of a sample variance and sample covariance, we can make a rather natural definition of the *sample correlation* between random variables $X$ and $Y$ as follows.

**Definition 6.12**
***Sample correlation***

Let $((X_1, Y_1), \ldots, (X_n, Y_n))$ be a random sample from a population distribution. Then the **sample correlation** between $X$ and $Y$ is given by

$$R_{XY} = \frac{S_{XY}}{S_X S_Y},$$

where $S_X = (S_X^2)^{1/2}$ and $S_Y = (S_Y^2)^{1/2}$ are the **sample standard deviations** of $X$ and $Y$, respectively.

Regarding properties of $R_{XY}$, outcomes of the sample correlation are lower bounded by $-1$ and upper bounded by $+1$, analogous to their population counterparts.

**Theorem 6.9** *The outcomes of $R_{XY}$ are such that $r_{XY} \in [-1,1]$.*

**Proof** This follows directly from Theorems 3.31 and 3.32 upon recognizing that $s_{XY}$, $s_X^2$, and $s_Y^2$ can be interpreted as a population covariance and population variances associated with a population distribution defined by $\hat{F}_n$. ■

In an analogy to the population correlation, the numerical value of the sample correlation, $r_{XY}$, can be interpreted as a measure of the degree of *linear* association between given sample outcomes $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$. The motivation for this interpretation follows directly from Theorems 3.35 and 3.36 upon recognizing that $\hat{F}_n$ can be thought of as defining a population distribution for $X$ and $Y$ to which the theorems can be subsequently applied. Expectations in the statements of the theorems are then interpreted as expectations based on $\hat{F}_n$, and all references to moments are then interpreted in the context of sample moments.

With the preceding interpretation of Theorem 3.35 in mind, and recalling Figure 3.12, $r_{XY} = 1$ implies that $y_i = a_1 + bx_i$, for $i = 1, \ldots, n$, where $a_1 = \bar{y} - b\bar{x}$ and $b = (s_Y/s_X)$. Thus, the outcomes $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$ have a perfect positive linear relationship. If $r_{XY} = -1$, then $y_i = a_2 - bx_i$, for $i = 1, \ldots, n$, where $a_2 = \bar{y} + b\bar{x}$, so the $y_i$'s and $x_i$'s have a perfect negative linear relationship.

To interpret the meaning of $r_{XY} \in (-1,1)$, examine Theorem 3.36 and its subsequent discussion applied to the discrete CDF $\hat{F}_n$ and the associated random sample outcome $(x_1, y_1), \ldots, (x_n, y_n)$. The best predictor of the $y_i$'s in terms of a linear function of the associated $x_i$'s is thus given by $\hat{y}_i = a + bx_i$, where $a = \bar{y} - b\bar{x}$ and $b = s_{XY}/s_X^2$. In the current context, *best* means the choice of $a$ and $b$ that minimizes $\mathrm{E}_{\hat{F}_n} d^2(Y, \hat{Y}) = n^{-1} \sum_{i=1}^{n} (y_i - (a + bx_i))^2$, which has a strictly monotonically increasing functional relationship with the distance $\left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right]^{1/2}$ between the vector $(y_1, \ldots, y_n)$ and the vector $(\hat{y}_1, \ldots, \hat{y}_n)$. (This is analogous to the least-squares criterion that will be discussed in Section 8.2.) Then, since

$$\mathrm{E}_{\hat{F}_n} d^2(Y, \hat{Y}) = s_Y^2 \left[1 - r_{XY}^2\right]$$

(recall the discussion following Theorem 3.36), it follows that the closer $r_{XY}$ is to either $-1$ or 1, the smaller the distance is between the outcomes $y_1, \ldots, y_n$ and the best linear prediction of these outcomes based on $x_1, \ldots, x_n$. Therefore, $r_{XY}$ is a measure of the degree of *linear* association between sample outcomes $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$ in the sense that the larger $|r_{XY}|$ is, the smaller the distance is between $(y_1, \ldots, y_n)$ and $(\hat{y}_1, \ldots, \hat{y}_n)$, with $\hat{y}_i = a + bx_i$. Also, recall from the discussion of Theorem 3.36 that $r_{XY}^2$ has an interpretation as the proportion of the (sample) variance in the $y_i$'s that is explained by the $\hat{y}_i$'s. Thus, the closer $|r_{XY}|$ is to 1, the more the sample variance in the $y_i$'s is explained by $a + bx_i$, $i = 1, \ldots, n$. The arguments are completely symmetric in the $y_i$'s and $x_i$'s, and a reversal of their roles leads to an interpretation of $r_{XY}$ (or $r_{YX}$) as a measure of the distance between $(x_1, \ldots, x_n)$ and $(\hat{x}_1, \ldots, \hat{x}_n)$ with $\hat{x}_i = a + by_i$ ($a$ and $b$ suitably redefined). Also, $r_{XY}^2$ (or $r_{YX}^2$) is the proportion of the sample variance in the $x_i$'s explained by $a + by_i$, $i = 1, \ldots, n$.

Besides its use as a measure of linear association between the sample *outcomes*, one might also inquire as to the relationship between $R_{XY}$ and its population counterpart, $\rho_{XY}$. Like the case of the sample variance or the sample covariance, it is not true that the expectation of the sample correlation equals its population counterpart, i.e., in general, $\mathrm{E}R_{XY} \neq \rho_{XY}$. Furthermore, general expressions for the mean and variance of $R_{XY}$ are quite complicated to state and derive.[7] We will concentrate here on establishing the probability limit and asymptotic distribution of $R_{XY}$.

**Theorem 6.10** *Let $(X_i, Y_i)$, $i = 1, \ldots, n$, be a random sample from a population distribution, and let $R_{XY}$ be the sample correlation between $X$ and $Y$. Then*

**a.** $\operatorname{plim} R_{XY} = \rho_{XY}$,

**b.** $R_{XY} \overset{a}{\sim} N(\rho_{XY}, n^{-1}\boldsymbol{\tau}'\boldsymbol{\Sigma}\boldsymbol{\tau})$,

*with $\boldsymbol{\tau}$ and $\boldsymbol{\Sigma}$ defined ahead.*

[7] See Kendall and Stuart, *Advanced Theory, Vol. I*, pp. 246–251, for an approach based on Taylor series expansions that can be used to approximate moments of the sample correlation.

**Proof** **a.** Note that $R_{XY} = S_{XY}/(S_X S_Y)$ is a continuous function of $S_{XY}$, $S_X$, and $S_Y$ for all $S_X > 0$ and $S_Y > 0$ and, in particular, is continuous at the values $\sigma_{XY} = \operatorname{plim} S_{XY}$, $\sigma_X = \operatorname{plim} S_X$ and $\sigma_Y = \operatorname{plim} S_Y$. It follows from Slutsky's theorem that

$$\operatorname{plim} R_{XY} = \operatorname{plim}\left(\frac{S_{XY}}{S_X S_Y}\right) = \frac{\operatorname{plim} S_{XY}}{(\operatorname{plim} S_X \cdot \operatorname{plim} S_Y)} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$
$$= \rho_{XY}.$$

**b.** The proof follows the approach used by Serfling, *Approximation Theorems*, pp. 125–126. Define the $(5 \times 1)$ vector $\mathbf{W} = (\bar{X}, \bar{Y}, (1/n)\sum_{i=1}^n X_i^2, (1/n)\sum_{i=1}^n Y_i^2, (1/n)\sum_{i=1}^n X_i Y_i)'$, so that the sample correlation can be expressed as

$$R_{XY} = g(\mathbf{W}) = \frac{W_5 - W_1 W_2}{(W_3 - W_1^2)^{1/2}(W_4 - W_2^2)^{1/2}}.$$

Note that $\mathbf{Z}_i = (X_i, Y_i, X_i^2, Y_i^2, X_i Y_i)'$, $i = 1, \ldots, n$, are iid random vectors, so that an application of the multivariate Lindberg–Levy CLT to the $\mathbf{Z}_i$'s implies that

$$n^{1/2}(\mathbf{W} - \mathrm{E}\mathbf{W}) \xrightarrow{d} N([\mathbf{0}], \boldsymbol{\Sigma}),$$

because $n^{-1}\sum_{i=1}^n \mathbf{Z}_i = \bar{\mathbf{Z}} = \mathbf{W}$, where $\boldsymbol{\Sigma}$ is the covariance matrix of any $\mathbf{Z}_i$.

A direct application of Theorem 5.39 to $R_{XY} = g(\mathbf{W})$ yields the statement in the theorem, where

$$\tau_1 = \frac{\partial g(\mathrm{E}\mathbf{W})}{\partial w_1} = \left(\frac{\rho_{XY}\mu'_{1,0}}{\sigma_X^2}\right) - \left(\frac{\mu'_{0,1}}{\sigma_X \sigma_Y}\right),$$

$$\tau_2 = \frac{\partial g(\mathrm{E}\mathbf{W})}{\partial w_2} = \left(\frac{\rho_{XY}\mu'_{0,1}}{\sigma_Y^2}\right) - \left(\frac{\mu'_{1,0}}{\sigma_X \sigma_Y}\right),$$

$$\tau_3 = \frac{\partial g(\mathrm{E}\mathbf{W})}{\partial w_3} = \frac{-\rho_{XY}}{(2\sigma_X^2)},$$

$$\tau_4 = \frac{\partial g(\mathrm{E}\mathbf{W})}{\partial w_4} = \frac{-\rho_{XY}}{(2\sigma_Y^2)},$$

$$\tau_5 = \frac{\partial g(\mathrm{E}\mathbf{W})}{\partial w_5} = (\sigma_X \sigma_Y)^{-1}.$$

■

As we have noted for the other statistics we have examined in this section, the sample correlation can be useful for estimating its population counterpart, $\rho_{XY}$, at least in large samples. In particular, as $n \to \infty$, the density of $R_{XY}$ concentrates within a small neighborhood of $\rho_{XY}$ so that as $n$ increases, it becomes highly probable that an outcome of $R_{XY}$ will occur close to $\rho_{XY}$.

**Example 6.9** Sample Means, Variances, Covariance and Correlation Matrices

A stock analyst has 15 observations on daily average prices of three common stocks:

| Stock 1 | Stock 2 | Stock 3 |
|---|---|---|
| 1.38 | 1.66 | 4.85 |
| 3.45 | 5.95 | 2.26 |
| 4.80 | 3.02 | 4.41 |
| 4.68 | 7.08 | 4.61 |
| 9.91 | 7.55 | 9.62 |
| 6.01 | 9.49 | 9.34 |
| 8.13 | 9.43 | 8.35 |
| 8.64 | 11.96 | 10.02 |
| 12.54 | 11.32 | 13.97 |
| 11.20 | 12.09 | 10.37 |
| 15.20 | 11.85 | 13.75 |
| 13.52 | 16.98 | 14.59 |
| 15.77 | 16.81 | 16.80 |
| 16.26 | 18.03 | 18.64 |
| 18.21 | 17.07 | 16.95 |

She wants to calculate summary statistics of the data consisting of sample means, sample variances, sample covariances, and sample correlations. Calculate these statistics, providing the variances and covariances in the form of a *sample covariance matrix*, and the sample correlations in terms of a *sample correlation matrix*. To what extent are there linear relationships between pairs of stock prices?

**Answer**: Letting **x** represent the $(15 \times 3)$ matrix of stock prices, the three sample means are given by

$$\underset{(3\times 1)}{\bar{\mathbf{x}}} = \frac{1}{15}\sum_{i=1}^{15}\mathbf{x}[i,\cdot]' = \begin{bmatrix} 9.98 \\ 10.69 \\ 10.57 \end{bmatrix}.$$

The **sample covariance matrix**, containing the respective sample variances on the diagonal and the sample covariances between stock $i$ and stock $j$ prices in the $(i, j)$th entry, $i \neq j$, is

$$\underset{(3\times 3)}{\widehat{\mathrm{Cov}}(\mathbf{X})} = \frac{1}{15}\sum_{i=1}^{15}(\mathbf{x}[i,\cdot]' - \bar{\mathbf{x}})(\mathbf{x}[i,\cdot]' - \bar{\mathbf{x}})'$$

$$= \begin{bmatrix} 25.402 & 22.584 & 23.686 \\ & 24.378 & 22.351 \\ \text{(symmetric)} & & 24.360 \end{bmatrix}.$$

The **sample correlation matrix**, containing the sample correlations between stock $i$ and stock $j$ prices in the $(i, j)$th entry, $i \neq j$, is

$$\widehat{\mathrm{Corr}}(\mathbf{X}) = \begin{bmatrix} s_{x_1} & & \\ & s_{x_2} & \\ & & s_{x_3} \end{bmatrix}^{-1} \widehat{\mathrm{Cov}}(\mathbf{X}) \begin{bmatrix} s_{x_1} & & \\ & s_{x_2} & \\ & & s_{x_3} \end{bmatrix}^{-1} = \begin{bmatrix} 1 & .908 & .952 \\ & 1 & .917 \\ \text{(symmetric)} & & 1 \end{bmatrix}.$$

The sample correlations indicate that sample observations have a pronounced tendency to follow a *positively sloped* linear relationship between pairs of stock prices. The linear relationship between stock 1 and stock 3 prices is the most pronounced with $(.952)^2 \times 100 = 90.631$ percent of the variance in either of the stock prices explained by a linear function of the other stock price. □

## 6.5 Properties of $\bar{X}_n$ and $S_n^2$ When Random Sampling from a Normal Distribution

Additional properties of $\bar{X}_n$ and $S_n^2$ that are useful in applications are obtained when random sampling is from a normal population distribution. In particular, $\bar{X}_n$ and $S_n^2$ are then independent random variables, $\bar{X}_n$ is normally distributed, $nS_n^2/\sigma^2$ has a $\chi^2$ distribution with $(n - 1)$ degrees of freedom and, thus, $S_n^2$ has a Gamma distribution. We examine these properties in subsequent theorems. It can also be shown that $n^{1/2}(\bar{X}_n - \mu)/\hat{\sigma}_n$, where $\hat{\sigma}_n^2 = (n/(n-1))S_n^2$, has a "$t$-distribution," but we defer our examination of this property until Section 6.7, where we will examine the $t$-distribution in detail.

We begin by stating a theorem on the independence of linear and quadratic forms in normally distributed random variables that has applicability in a variety of situations.

**Theorem 6.11** *Let* $\mathbf{B}$ *be a* $(q \times n)$ *matrix of real numbers, let* $\mathbf{A}$ *be an* $(n \times n)$ *symmetric matrix of real numbers having rank* $p$*, and let* $\mathbf{X}$ *be an* $(n \times 1)$ *random vector such that* $\mathbf{X} \sim N(\boldsymbol{\mu}_\mathbf{X}, \sigma^2\mathbf{I})$*. Then* $\mathbf{BX}$ *and* $\mathbf{X'AX}$ *are independent if* $\mathbf{BA} = [0]$.[8]

**Proof** Since $\mathbf{A}$ is symmetric, it can be diagonalized by pre- and postmultiplication using the matrix of characteristic vectors of $\mathbf{A}$ as

$$\mathbf{P'AP} = \boldsymbol{\Lambda} = \left[\begin{array}{ccc|c} \lambda_1 & & & \\ & \ddots & & [0] \\ & & \lambda_p & \\ \hline & [0] & & [0] \end{array}\right],$$

where $\mathbf{P}$ is the $(n \times n)$ matrix of characteristic vectors stored columnwise, and the $\lambda_i$'s are the nonzero characteristic roots of $\mathbf{A}$. It is assumed without loss of

[8] The theorem can be extended to the case where $\mathbf{X} \sim N(\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma})$, in which case the condition for independence is that $\mathbf{B\Sigma A} = [0]$.

generality that the columns of $\mathbf{P}$ have been ordered so that the first $p$ columns correspond to characteristic vectors associated with the $p$ nonzero characteristic roots. Let $\mathbf{BA} = [\mathbf{0}]$, so that $\mathbf{BPP'AP} = [\mathbf{0}]$, since $\mathbf{P}$ is orthogonal (i.e., $\mathbf{PP'} = \mathbf{P'P} = \mathbf{I}$). Let $\mathbf{C} = \mathbf{BP}$, so that $\mathbf{C\Lambda} = \mathbf{BPP'AP} = [\mathbf{0}]$. Partitioning $\mathbf{C}$ and $\mathbf{\Lambda}$ appropriately, we have that

$$\mathbf{C\Lambda} = \left[\begin{array}{c|c} \underset{(q\times p)}{\mathbf{C}_1} & \underset{q\times(n-p)}{\mathbf{C}_2} \end{array}\right] \left[\begin{array}{c|c} \underset{p\times p}{\mathbf{D}} & \underset{p\times(n-p)}{[\mathbf{0}]} \\ \hline \underset{(n-p)\times p}{[\mathbf{0}]} & \underset{(n-p)\times(n-p)}{[\mathbf{0}]} \end{array}\right] = \underset{(q\times n)}{[\mathbf{0}]},$$

where $\mathbf{D} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix}$ is the diagonal matrix of nonzero characteristic roots of $\mathbf{A}$.

The above matrix equation implies that $\mathbf{C}_1\mathbf{D} = [\mathbf{0}]$, and since $\mathbf{D}$ is invertible, we have that $\mathbf{C}_1 = [\mathbf{0}]$. Thus

$$\mathbf{C} = \left[\begin{array}{c|c} \underset{(q\times p)}{[\mathbf{0}]} & \underset{q\times(n-p)}{\mathbf{C}_2} \end{array}\right].$$

Now define $\mathbf{Z}_{(n\times 1)} = \mathbf{P'X} \sim N(\mathbf{P'}\boldsymbol{\mu}_{\mathbf{X}}, \sigma^2\mathbf{I})$ (recall that $\mathbf{P'P} = \mathbf{I}$), and note that $(Z_1, \ldots, Z_n)$ are independent random variables. Since $\mathbf{X'AX} = \mathbf{Z'P'APZ} = \mathbf{Z'\Lambda Z} = \sum_{i=1}^{p} \lambda_i Z_i^2 = g_1(Z_1, \ldots, Z_p)$ and $\mathbf{BX} = \mathbf{BPZ} = \mathbf{CZ} = \mathbf{C}_2\begin{bmatrix} Z_{p+1} \\ \vdots \\ Z_n \end{bmatrix} = g_2(Z_{p+1}, \ldots, Z_n)$, and because $(Z_1, \ldots, Z_p)$ and $(Z_{p+1}, \ldots, Z_n)$ are independent, then by Theorem 2.9, $\mathbf{X'AX}$ and $\mathbf{BX}$ are independent. ■

We use the preceding theorem to prove a theorem that establishes the distribution of the random variable $nS_n^2/\sigma^2$, which we will later find to have important applications in testing hypotheses about the value of $\sigma^2$.

**Theorem 6.12** *If $\bar{X}_n$ and $S_n^2$ are the sample mean and sample variance, respectively, of a random sample of size n from a normal distribution with mean $\mu$ and variance $\sigma^2$, then*

**a.** *$\bar{X}_n$ and $S_n^2$ are independent,*

**b.** $(nS_n^2/\sigma^2) \sim \chi^2_{n-1}$.

**Proof** **a.** In the context of Theorem 6.11, let $\mathbf{B}_{(1\times n)} = [n^{-1} \ldots n^{-1}]$, so that $\mathbf{BX} = \bar{X}_n$. Also, let

$$\underset{(n\times n)}{\mathbf{H}} = \begin{bmatrix} \mathbf{B} \\ \mathbf{B} \\ \vdots \\ \mathbf{B} \end{bmatrix} = \begin{bmatrix} n^{-1} & \ldots & n^{-1} \\ \vdots & \ddots & \vdots \\ n^{-1} & \ldots & n^{-1} \end{bmatrix} \quad \text{so that } (\mathbf{I} - \mathbf{H})\mathbf{X} = \begin{bmatrix} X_1 - \bar{X}_n \\ \vdots \\ X_n - \bar{X}_n \end{bmatrix},$$

implying $nS_n^2 = \mathbf{X}'(\mathbf{I}-\mathbf{H})'(\mathbf{I}-\mathbf{H})\mathbf{X} = \mathbf{X}'(\mathbf{I}-\mathbf{H})\mathbf{X}$ since $\mathbf{I}-\mathbf{H}$ is symmetric and $(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) = \mathbf{I}-\mathbf{H}$, i.e., $\mathbf{I}-\mathbf{H}$ is idempotent. Then letting $\mathbf{A} = n^{-1}(\mathbf{I}-\mathbf{H})$, we have that $\mathbf{X}'\mathbf{A}\mathbf{X} = S_n^2$.

It follows from Theorem 6.11 that $\bar{\mathbf{X}}_n = \mathbf{B}\mathbf{X}$ and $S_n^2 = \mathbf{X}'\mathbf{A}\mathbf{X}$ are independent since $\mathbf{B}\mathbf{A} = n^{-1}\mathbf{B}(\mathbf{I}-\mathbf{H}) = n^{-1}(\mathbf{B}-\mathbf{B}) = [0]$.

**b.** From the proof of part (a), it follows that

$$\frac{nS_n^2}{\sigma^2} = \frac{1}{\sigma^2}(\mathbf{X}'(\mathbf{I}-\mathbf{H})'(\mathbf{I}-\mathbf{H})\mathbf{X}).$$

Let $\boldsymbol{\mu}_\mathbf{X} = (\mu\ \mu\ \ldots\ \mu)'$, and note that $(\mathbf{I}-\mathbf{H})\boldsymbol{\mu}_\mathbf{X} = \boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu}_\mathbf{X} = [0]$. Therefore,

$$\frac{nS_n^2}{\sigma^2} = \frac{1}{\sigma^2}(\mathbf{X}-\boldsymbol{\mu}_\mathbf{X})'(\mathbf{I}-\mathbf{H})'(\mathbf{I}-\mathbf{H})(\mathbf{X}-\boldsymbol{\mu}_\mathbf{X}) = \frac{1}{\sigma^2}(\mathbf{X}-\boldsymbol{\mu}_\mathbf{X})'(\mathbf{I}-\mathbf{H})(\mathbf{X}-\boldsymbol{\mu}_\mathbf{X}),$$

since $\mathbf{I}-\mathbf{H}$ is symmetric and idempotent. Note that $\mathrm{tr}[\mathbf{I}-\mathbf{H}] = n-1$, which implies that $(n-1)$ characteristic roots have value 1 because a symmetric idempotent matrix has rank equal to its trace, and its characteristic roots are a collection of 1's and 0's with the number of 1's equal to the rank of the matrix. Diagonalizing $(\mathbf{I}-\mathbf{H})$ by its orthogonal characteristic vector matrix $\mathbf{P}$ then yields

$$\mathbf{P}'(\mathbf{I}-\mathbf{H})\mathbf{P} = \left[\begin{array}{c|c} \mathbf{I} & [0] \\ \hline [0] & 0 \end{array}\right] = \boldsymbol{\Lambda},$$

where $\mathbf{I}$ is an $(n-1)$-dimensional identity matrix. Therefore, $\mathbf{I}-\mathbf{H} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$, and then $(nS_n^2/\sigma^2) = (1/\sigma^2)(\mathbf{X}-\boldsymbol{\mu}_\mathbf{X})'\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'(\mathbf{X}-\boldsymbol{\mu}_\mathbf{X}) = \mathbf{Z}'\boldsymbol{\Lambda}\mathbf{Z}$ where $\mathbf{Z} = (1/\sigma)\mathbf{P}'(\mathbf{X}-\boldsymbol{\mu}_x) \sim N([0],\mathbf{I})$ since $\mathbf{P}'\mathbf{P} = \mathbf{I}$. Finally, $(nS_n^2/\sigma^2) = \sum_{i=1}^{n-1}\mathbf{Z}_i^2 \sim \chi^2_{n-1}$, given the definition of $\boldsymbol{\Lambda}$, i.e., we have the sum of squares of $(n-1)$ iid standard normal random variables. ■

It follows from part (b) of Theorem 6.12 that $S_n^2$ has a Gamma density, as stated and proved in the following theorem.

**Theorem 6.13** *Under the assumptions of Theorem 6.12, $S_n^2 \sim$ Gamma with $\alpha = (n-1)/2$ and $\beta = (2\sigma^2/n)$.*

**Proof** Let $Y = nS_n^2/\sigma^2$. Then it is known from Theorem 6.12.b that $M_Y(t) = (1-2t)^{-(n-1)/2}$. Note that $S_n^2 = ((\sigma^2/n))Y$. Then $M_{S_n^2}(t) = \mathrm{E}\exp(S_n^2 t) = \mathrm{E}\exp((\sigma^2/n)Yt) = \mathrm{E}\exp(Yt_*)$ where $t_* = (\sigma^2 t/n)$. But since $M_Y(t_*) = \mathrm{E}\exp(Yt_*) = (1-2t_*)^{-(n-1)/2}$, it follows that $M_{S_n^2}(t) = (1-2(\sigma^2 t/n))^{-(n-1)/2}$, which is associated with the Gamma density having $\alpha = (n-1)/2$ and $\beta = (2\sigma^2/n)$. ■

Since $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ is a linear combination of iid $N(\mu,\sigma^2)$ random variables $X_i \sim N(\mu,\sigma^2)$, $\bar{X}_n$ is also normally distributed, as indicated in the following theorem.

**Theorem 6.14** *Under the assumptions of Theorem 6.12, $\bar{X}_n \sim N(\mu,\sigma^2/n)$.*

**Proof** Let $\mathbf{A}_{(1\times n)} = (1/n \ldots 1/n)$ and $b = 0$ in Theorem 4.9. Then since $\mathbf{X}_{(n\times 1)} \sim N(\boldsymbol{\mu}_{\mathbf{X}}, \sigma^2\mathbf{I})$, where $\mathbf{X}_{(n\times 1)} = (X_1, \ldots, X_n)'$ and $\boldsymbol{\mu}_{\mathbf{X}_{(n\times 1)}} = (\mu, \ldots, \mu)'$, then $\bar{\mathbf{X}} = \mathbf{A}\mathbf{X} \sim N(\mathbf{A}\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{A}\sigma^2\mathbf{I}\mathbf{A}') = N\left(\mu, \sigma^2/n\right)$. ■

**Example 6.10** An inspector from the state's Department of Weights and Measures is investigating a claim that an oil company is underfilling quarts of motor oil that are sold at retail. The filling process is such that the actual volumes of oil placed in quart cans can be interpreted as being normally distributed. According to state specifications, a can is considered legally full if it contains $\geq 31.75$ ounces of oil. The company claims that its filling process has a mean of 32.25 and a standard deviation of .125, and that therefore the probability that a can is legally full exceeds .9999.

The inspector randomly samples 200 quarts of oil produced by the oil company and finds that $\sum_{i=1}^{200} x_i = 6{,}462$ and $\sum_{i=1}^{200} x_i^2 = 208{,}791$. Calculate the sample mean and sample variance. Assuming the *company's* claims to be true, do the outcomes of $\bar{X}$ and $S^2$ appear to be unusual in any way? Does the outcome support the claim against the company?

**Answer:** $\bar{x} = \sum_{i=1}^{200} x_i/200 = 32.31$ and $s^2 = \sum_{i=1}^{200}(x_i - \bar{x})^2/200 = (\sum_{i=1}^{200} x_i^2 - 200\bar{x}^2)/200 = .0189$. *Given* the company's claim, $\bar{X}_n \sim N(32.25, .78125 \times 10^{-4})$ and $(200)S^2/(.015625) \sim \chi^2_{199}$. If the company's claim were true, then the event $\bar{x} \in [\mu - 4\sigma_{\bar{X}}, \mu + 4\sigma_{\bar{X}}] = [32.2146, 32.2854]$ would occur with probability $>$ .9999. The particular outcome of $\bar{X}$ would seem to be unusually *high* compared to the claimed mean, suggesting that $\mu$ might actually be *higher* than the company claims (if the claimed standard deviation is correct). Regarding the sample variance, and *given* the company's claim, $\mathrm{E}S^2 = (199/200)\sigma^2 = .015547$ and the event $s^2 \in [.01265, .01875]$ would occur with probability .95.[9] Then $s^2 = .0189$ appears to be somewhat high compared to the claimed variance.

Overall, there is strong evidence to suggest that the filling process does not have the characteristics claimed by the company, but the evidence given above does not support the claim that the cans are being *underfilled*. We will examine this issue further after we derive the $t$-distribution in Section 6.7. □

## 6.6 Sampling Distributions: Deriving Probability Densities of Functions of Random Variables

The statistics we examined in the preceding sections are useful in a number of statistical inference problems, but one needs to be concerned with a much larger variety of functions of random samples to adequately deal with the variety of inference problems that arise in practice. Furthermore, in order to assess

[9] This interval and associated probability was obtained by noting that if $Y \sim \chi^2_{199}$, then $P(161.83 \leq y \leq 239.96) = .95$, which was obtained via numerical integration of the $\chi^2_{199}$ density, leaving .025 probability in both the right and left tails of the density. Using the relationship $S^2 \sim (.015625/200)Y$ then leads to the stated interval.

the adequacy of statistical procedures, it will be necessary to identify probability spaces for functions of random samples that are proposed as estimators or as hypothesis-testing statistics in order to evaluate their probability characteristics. In particular, we will be interested in deriving the functional forms of the PDFs associated with functions of a random sample, where such PDFs will be referred to as **sampling densities** or **sampling distributions**.

### MGF Approach

If $Y = g(X_1, \ldots, X_n)$ is a function of interest, then one can attempt to derive the moment-generating function of $Y = g(X_1, \ldots, X_n)$, i.e., $M_{g(\cdot)}(t) = \mathrm{E}\exp(g(X_1, \ldots, X_n)t)$, and identify the density function characterized by the moment-generating function. If $X_1, \ldots, X_n$ is a random sample, one would refer to the density of $Y$ as the sampling distribution of $Y$. Of course, **g** can be a vector function, in which case we would employ the multivariate moment-generating function concept in the hope of identifying the joint density function of the multivariate random variable $Y$.

**Example 6.11** What is the sampling distribution of the sample mean $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ when the random sample $(X_1, \ldots, X_n)$ is drawn from a population distribution represented by the exponential family of densities?

**Answer**: Using the MGF approach, we attempt to find $\mathrm{E}\exp(\bar{X}_n t) = \mathrm{E}\exp((\sum_{i=1}^n X_i)t/n)$. We know that the population distribution is given by $m(z) = (1/\theta(e^{-z/\theta}I_{(0,\infty)}(z)$, so that the joint density of the random sample is given by

$$f(x_1, \ldots, x_n) = \frac{1}{\theta^n}\exp\left(\frac{-\sum_{i=1}^n x_i}{\theta}\right)\prod_{i=1}^n I_{(0,\infty)}(x_i).$$

Then

$$\mathrm{E}\exp\left(\left(\sum_{i=1}^n X_i\right)t/n\right) = \prod_{i=1}^n \mathrm{E}\exp\left(\frac{X_i t}{n}\right) = \prod_{i=1}^n \left(1-\theta\frac{t}{n}\right)^{-1} = \left(1-\theta\frac{t}{n}\right)^{-n},$$

for $t < n/\Theta$, which is recognized as the MGF of a Gamma density with $\alpha = n$ and $\beta = \Theta/n$. Thus, the *sampling distribution* of $\bar{X}$ is gamma with the aforementioned parameters. □

### CDF Approach

One might also attempt to derive the sampling distribution of $Y = g(X_1, \ldots, X_n)$ by identifying the cumulative distribution function of $Y$ and then using the correspondence between cumulative distribution functions and density functions to derive the latter. The usefulness of this approach depends on how difficult it is to derive the cumulative distribution function for a particular problem. The advantages of this approach over the MGF approach are that it is applicable even if the MGF does not exist, and it circumvents the problem of the researcher's not knowing which density function corresponds to a particular MGF when the MGF does exist.[10]

**Example 6.12** Refer to Ex. 6.11, and suppose we want the sampling distribution of $Y = g(X_1, \ldots, X_n) = \bar{X}_n^2$. Note that

$$F(c) = \begin{cases} P[y \le c] = P\left[\bar{x}_n^2 \le c\right] = P\left[\bar{x}_n \in [-c^{1/2}, c^{1/2}]\right] = \int_0^{c^{1/2}} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} dx & \text{for } c > 0, \\ 0 \quad \text{otherwise,} & \end{cases}$$

since $\bar{X}_n$ is Gamma distributed. The following lemma regarding differentiating an integral with respect to its bounds will be useful here and in other applications of the CDF approach for defining the PDF of $g(X)$.

**Lemma 6.1**
***Leibniz's rules***

Let $w_1(c)$ and $w_2(c)$ be functions that are differentiable at $c$, and let $h(x)$ be continuous at $x = w_1(c)$ and $x = w_2(c)$ and integrable. Then

$$\frac{d \int_{w_1(c)}^{w_2(c)} h(x) dx}{dc} = h(w_2(c)) \frac{dw_2(c)}{dc} - h(w_1(c)) \frac{dw_1(c)}{dc}.$$

*Special cases*:

**a.** Let $w_1(c) = k \;\forall\; c$, where $k$ is a constant. Then

$$\frac{d \int_{k}^{w_2(c)} h(x) dx}{dc} = h(w_2(c)) \frac{dw_2(c)}{dc},$$

and the function $h(x)$ need not be continuous at $k$. Note the result is still valid if $k$ is replaced by $-\infty$.

**b.** Let $w_2(c) = k \;\forall\; c$, where $k$ is a constant. Then

$$\frac{d \int_{w_1(c)}^{k} h(x) dx}{dc} = -h(w_1(c)) \frac{dw_1(c)}{dc},$$

and the function need not be continuous at $k$. Note that the result is still valid if $k$ is replaced by $\infty$.

(Adapted from D. V. Widder (1961), *Advanced Calculus*, 2nd ed., Englewood Cliffs, NJ: Prentice-Hall, pp. 350–353, and R. G. Bartle (1976), *The Elements of Real Analysis*, 2nd ed., New York: John Wiley, pp. 245–246.)

[10] Although, as we have mentioned previously, the density function can always be identified in principle by an integration problem involving the MGF in the integrand.

In the case at hand, $w_2(c) = c^{1/2}$, and $w_1(c) = 0$, so that Lemma 6.1 implies that the sampling distribution of $g(X_1, \ldots, X_n) = \bar{X}_n^2$ is given by

$$\frac{dF(c)}{dc} = f(c) = \begin{cases} \dfrac{1}{\beta^\alpha \Gamma(\alpha)} c^{(1/2)(\alpha-1)} \exp\left[\dfrac{-c^{1/2}}{\beta}\right]\left(\dfrac{1}{2}\right) c^{-1/2} & \\ \quad = \dfrac{1}{2\beta^\alpha \Gamma(\alpha)} c^{(\alpha/2)-1} \exp\left[\dfrac{-c^{1/2}}{\beta}\right] & \text{for } c > 0, \\ 0 \quad \text{otherwise.} & \end{cases}$$ □

### Equivalent Events Approach (Discrete Case)

In the case of discrete random variables, a third, conceptually straightforward, approach for deriving the density function of functions of random variables is available. In fact, without calling attention to it, we have already used the procedure, called the **equivalent-events approach**, in proving a theorem concerning the expectation of functions of random variables. Specifically, if $y = g(x)$ is a real-valued function of $x$, then $P_Y(y) = P_X(A_y)$, $A_y = \{x : y = g(x), x \in R(X)\}$, because the elementary event $y$ for $Y$ is equivalent to the event $A_y$ for $X$ in the sense that $y$ occurs iff $A_y$ occurs. It follows that the density function of $Y$ can be obtained from the density function of $X$ for scalar $X$ and $Y$ as

$$h(y) = \sum_{\{x : g(x) = y, x \in R(X)\}} f(x).$$

The extension to the case of multivariate random variables and vector functions is straightforward. In particular, interpreting **y** and/or **x** as vectors, we have that

$$h(\mathbf{y}) = \sum_{\{\mathbf{x}:\mathbf{g}(\mathbf{x})=\mathbf{y}, \mathbf{x}\in R(\mathbf{X})\}} \cdots \sum f(\mathbf{x}).$$

In either the scalar or multivariate case, if $\mathbf{y} = \mathbf{g}(\mathbf{x})$ is invertible, so that $\mathbf{x} = \mathbf{g}^{-1}(\mathbf{y})\ \forall\ \mathbf{x} \in R(\mathbf{X})$, then $h(\mathbf{y}) = f(\mathbf{g}^{-1}(\mathbf{y}))$.

**Example 6.13** Your company ships videocassettes in lots of 100. As part of your quality-control program, you assure potential buyers that no more than 1 videocassette will be defective in each lot of 100. In reality, the probability that a cassette manufactured by your company is defective is .001, and whether a given cassette is defective is independent of whether or not any other cassette is defective. Then the PDF for the number of defectives in a lot of your cassettes is given by the binomial density function

$$f(x) = \begin{cases} \dbinom{100}{x} (.001)^x (.999)^{n-x} & \text{for } x \in \{0, 1, \ldots, 100\}, \\ 0 \quad \text{otherwise.} & \end{cases}$$

Define a new random variable, $Z$, such that $z = 1$ represents the event that your quality-control claim is valid (i.e., $x = 0$ or $1$) while $z = 0$ represents the event that your claim is invalid (i.e., $x = 2$ or more) on a lot of your videocassettes. Define the density function of $Z$.

**Answer**: The density function of $Z$ is given by

$$h(z) = \sum_{\{x:g(x)=z, x\in R(X)\}} f(x),$$

where $g(x) = I_{\{0,1\}}(x)$. Then, referring to the binomial density function defined above,

$$h(1) = \sum_{x=0}^{1} f(x) = .995 \quad \text{and} \quad h(0) = \sum_{x=2}^{100} f(x) = .005$$

so that $h(z) = .995I_{\{1\}}(z) + .005I_{\{0\}}(z)$. □

### Change of Variables (Continuous Case)

Another very useful procedure for deriving the PDF of functions of *continuous* random variables is available if the functions involved are continuously differentiable and if $Y = g(X)$ admits an inverse function $X = g^{-1}(Y)$. We first examine the case where both $Y$ and $X$ are *scalar* random variables. We introduce the term **support of the density**, meaning the set of domain elements of a PDF that are assigned positive density values, i.e., $\{x : f(x) > 0\}$ is the support of the density $f(x)$, or simply, the **support of $f(x)$**.

**Theorem 6.15** (Change-of-Variables Technique (Univariate and Invertible)) *Suppose the continuous random variable $X$ has PDF $f(x)$. Let $g(x)$ be continuously differentiable with $dg/dx \neq 0\ \forall\ x$ in some open interval, $\Delta$, containing the support of $f(x)$, $\Xi$. Also, let the inverse function $x = g^{-1}(y)$ be defined $\forall\ y \in \Psi = \{y{:}y = g(x), x \in \Xi\}$. Then the PDF of $Y = g(X)$ is given by*

$$h(y) = f(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right| \quad \textit{for } y \in \Psi, \quad \textit{with } h(y) = 0 \textit{ elsewhere.}$$

**Proof** If $dg/dx \neq 0$ and is a continuous function, then $g$ is either a monotonically increasing or monotonically decreasing function, i.e., either $dg/dx > 0$ or $dg/dx < 0$, respectively, $\forall x \in \Delta$. (Note that $dg/dx$ cannot $> 0$ for some $x$ and $< 0$ for other $x$'s, since continuity of $dg/dx$ would then necessarily require that $dg/dx = 0$ at some point.)

**a.** Case where $dg/dx > 0$: In this case, $P(y \leq b) = P(g(x) \leq b) = P(x \leq g^{-1}(b))$, $\forall\ b \in \Psi$, since $x = g^{-1}(y)$ is monotonically increasing for $y \in \Psi$ (see Figure 6.3). Thus the cumulative distribution function for $Y$ can be defined for all $b \in \Psi$ by

$$H(b) = P(y \leq b) = P(x \leq g^{-1}(b)) = \int_{-\infty}^{g^{-1}(b)} f(x)dx.$$

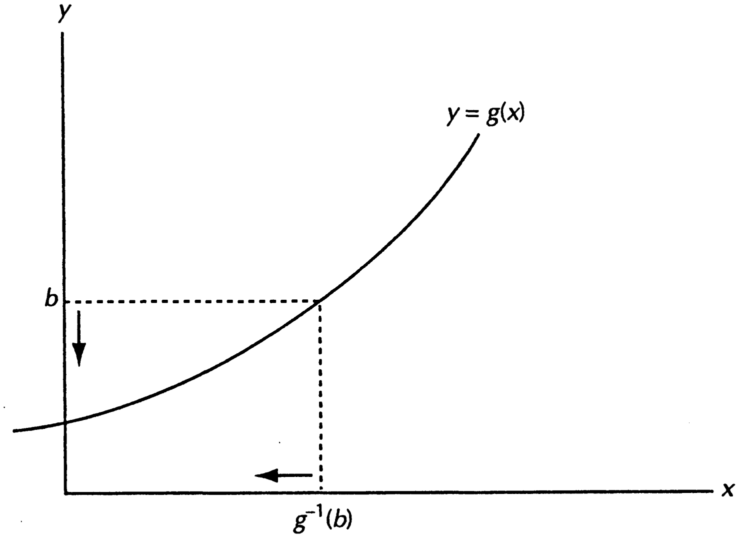


**Figure 6-3**
Monotonically increasing function.

Then using Lemma 6.1, we can derive the density function for $Y$ by differentiation as

$$h(b) = \frac{dH(b)}{db} = \begin{cases} \dfrac{d\int_{-\infty}^{g^{-1}(b)} f(x)dx}{db} = f(g^{-1}(b))\dfrac{dg^{-1}(b)}{db} & \text{for } b \in \Psi, \\ 0 \quad \text{otherwise.}^{11} & \end{cases}$$

**b.** Case where $dg/dx < 0$: In this case, $P(y \le b) = P(g(x) \le b) = P(x \ge g^{-1}(b))$, $\forall\, b \in \Psi$, since $x = g^{-1}(y)$ is monotonically decreasing for $y \in \Psi$ (see Figure 6.4). Thus the cumulative distribution function for $Y$ can be defined for all $b \in \Psi$ by

$$H(b) = P(y \le b) = P(x \ge g^{-1}(b)) = \int_{g^{-1}(b)}^{\infty} f(x)dx.$$

Then, using Lemma 6.1, we can derive the density function for $Y$ by differentiation as

$$h(b) = \frac{dH(b)}{db} = \begin{cases} \dfrac{d\int_{g^{-1}(b)}^{\infty} f(x)dx}{db} = -f(g^{-1}(b))\dfrac{dg^{-1}(b)}{db} & \text{for } b \in \Psi, \\ 0 \quad \text{otherwise,} & \end{cases}$$

[11] Here, and elsewhere, we are suppressing a technical requirement that $f$ be continuous at the point $g^{-1}(b)$, so that we can invoke Lemma 6.1 for differentiation of the cumulative distribution function. Even if $f$ is discontinuous at $g^{-1}(b)$, we can nonetheless *define* $h(b)$ as indicated above, since a density function can be redefined arbitrarily at a finite number of points of discontinuity without affecting the assignment of probabilities to any events.

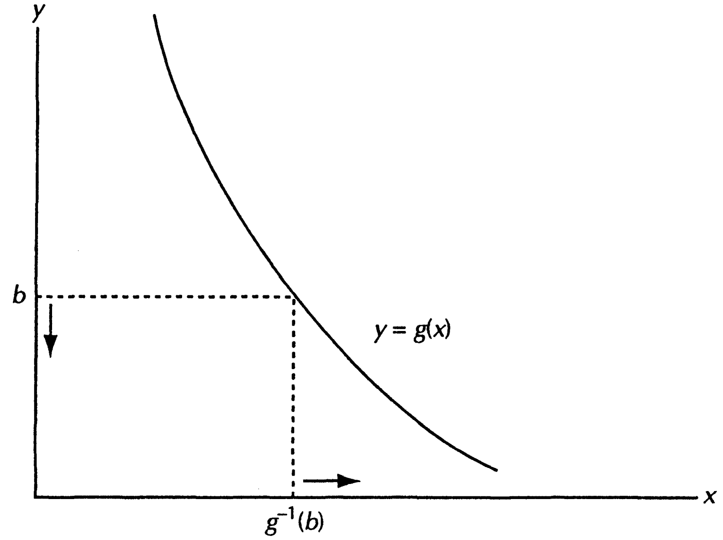


**Figure 6-4** Monotonically decreasing function.

Note that in this latter case, since $(dg^{-1}(b)/db) < 0$, we can write $h(b)$ alternatively as

$$h(b) = \begin{cases} f(g^{-1}(b))\left|\dfrac{dg^{-1}(b)}{db}\right| & \text{for } b \in \Psi, \\ 0 \quad \text{otherwise,} \end{cases}$$

which is also an alternative representation of the previous case where $dg/dx > 0$. ■

**Example 6.14** The Cobb–Douglas production function for a certain commodity is given by

$$Q = \beta_0 \prod_{i=1}^{k} x_i^{\beta_i} e^W = \beta_0 \prod_{i=1}^{k} x_i^{\beta_i} V,$$

where $W \sim N(0, \sigma^2)$ and $V = e^W$. What is the PDF of $V$?

**Answer**: $v = e^w$ is a monotonically increasing function of $w$ for which $dv/dw = e^w > 0 \; \forall \; w$. The inverse function is given by $w = \ln(v) \; \forall \; v > 0$, and $dw/dv = v^{-1}$. Then Theorem 6.15 applies, and the distribution of $V$ can be defined as

$$h(v) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\,\sigma v} \exp\left(-\dfrac{[\ln(v)]^2}{2\sigma^2}\right) & \text{for } v > 0, \\ 0 \quad \text{elsewhere} \end{cases}$$

The density function is called a **log-normal distribution**. □

The change-of-variables approach can be generalized to certain cases where the function defined by $y = g(x)$ does *not* admit an inverse function. The procedure essentially applies to functions that are *piecewise invertible*, as defined below.

**Theorem 6.16** (Change-of-Variables Technique (Univariate and Piecewise Invertible)) *Suppose the continuous random variable $X$ has PDF $f(x)$. Let $g(x)$ be continuously differentiable with $dg(x)/dx \neq 0$ for all but perhaps a finite number of x's in an open interval $\Delta$ containing the support of $f(x)$, $\Xi$. Let $\Xi$ be partitioned into a collection of disjoint intervals $D_1, \ldots, D_n$ for which $g:D_i \to R_i$ has an inverse function $g_i^{-1}:R_i \to D_i \; \forall \, i$.*[12] *Then the probability density of $Y = g(X)$ is given by*

$$h(y) = \begin{cases} \displaystyle\sum_{i \in I(y)} f\left(g_i^{-1}(y)\right) \left|\frac{dg_i^{-1}(y)}{dy}\right| & \textit{for } y \in \Psi = \{y : y = g(x), x \in \Xi\}, \\ 0 \quad \textit{elsewhere,} \end{cases}$$

*where $I(y) = \{i\colon \exists x \in D_i \text{ such that } y = g(x), i = 1,\ldots,n\}$ and $(dg_i^{-1}(y)/dy) \equiv 0$ whenever it would otherwise be undefined.*[13]

**Sketch of the Proof** The CDF of $Y = g(X)$ can be represented as

$$P(y \leq b) = \sum_{i=1}^{n} P(x : g(x) \leq b, x \in D_i).$$

Note the following possibilities for $P(x : g(x) \leq b, x \in D_i)$:

| | $P(x : g(x) \leq b, x \in D_i)$ | |
|---|---|---|
| **Case** | ***g*** **monotonically increasing on** $D_i$ | ***g*** **monotonically decreasing on** $D_i$ |
| 1. $b < \min_{x \in D_i} g(x)$ | 0 | 0 |
| 2. $\min_{x \in D_i} g(x) \leq b \leq \max_{x \in D_i} g(x)$ | $\int_{\min(D_i)}^{g_i^{-1}(b)} f(x)\,dx$ | $\int_{g_i^{-1}(b)}^{\max(D_i)} f(x)dx$ |
| 3. $b > \max_{x \in D_i} g(x)$ | $P(x_i \in D)$ | $P(x_i \in D)$ |

where $\min(D_i)$ and $\max(D_i)$ refer to the minimum and maximum values in $D_i$.[14] In either case 1 or 3, $dP(x : g(x) \leq b, x \in D_i)/db = 0$, which corresponds to the case where $\nexists x$ such that $g(x) = b$. In case 2, Leibniz's rule for differentiating the integrals yields $f(g^{-1}(b))[dg^{-1}(b)/db]$ and $-f(g^{-1}(b))[dg^{-1}(b)/db]$ in the monotonically increasing and decreasing cases, respectively, both of which can

[12] These properties define a function that is **piecewise invertible** on the domain $\cup_{i=1}^{n} D_i$.

[13] Note that $I(y)$ is an index set containing the indices of all of the $D_i$ sets that have an element whose image under the function $g$ is the value $y$.

[14] If max and min do not exist, they are replaced with sup and inf.

be represented by $f(g^{-1}(b))|dg^{-1}(b)/db|$. Then the density of $Y$ can be defined as $h(b) = dP(y \leq b)/db$ for $b \in \Psi$, with $|(dg_i^{-1}(y)/dy)|$ arbitrarily set to zero in the finite number of instances where $dg(b)/dx = 0$, for $b \in D_i$ implies $(dg_i^{-1}(g(b))/dy)$ does not exist, $i = 1, \ldots, n$. ■

The following example illustrates the use of this more general change-of-variables technique.

**Example 6.15** The manufacturing process for a certain product is sensitive to deviations from an optimal ambient temperature of 72° F. In a poorly air-conditioned plant owned by the Excelsior Corporation, the average daily temperature is uniformly distributed in the range [70, 74], and the deviations from 72° F are represented by the outcome of $X \sim f(x) = .25I_{[-2,2]}(x)$. The percentage of production lost on any given day due to temperature deviations can be represented by the outcome of $Y = g(X) = X^2$. Derive the probability density for the percentage of production lost on any given day.

**Answer**: First note that Theorem 6.15 does not apply since $y = x^2$ is such that $dy/dx = 0$ at $x = 0$, which is in the support of $f(x)$, and $y = x^2$ does not admit an inverse function $\forall\, y \in \Psi = [0, 4]$ (see Figure 6.5).

However, we can utilize Theorem 6.16 to derive the density function of $Y$. The function $g(x) = x^2$ is continuously differentiable $\forall\, x$, and $dg(x)/dx = 2x \neq 0$ $\forall\, x$ except $x = 0$. The sets $D_1 = [-2, 0)$ and $D_2 = [0, 2]$ are two disjoint intervals whose union equals the support of $f(x)$. Also, $g : D_1 \to (0, 4]$ has inverse function $g_1^{-1} : (0, 4] \to D_1$ defined by $g_1^{-1}(y) = -y^{1/2}$ for $y \in (0, 4]$, while $g : D_2 \to [0, 4]$ has inverse function $g_2^{-1} : [0, 4] \to D_2$ defined by $g_2^{-1}(y) = y^{1/2}$ for $y \in [0, 4]$.[15] Finally, note that $dg_1^{-1}(y)/dy = -(1/2)y^{-1/2}$ and $dg_2^{-1}(y)/dy = (1/2)y^{-1/2}$ are defined $\forall\, y \in$

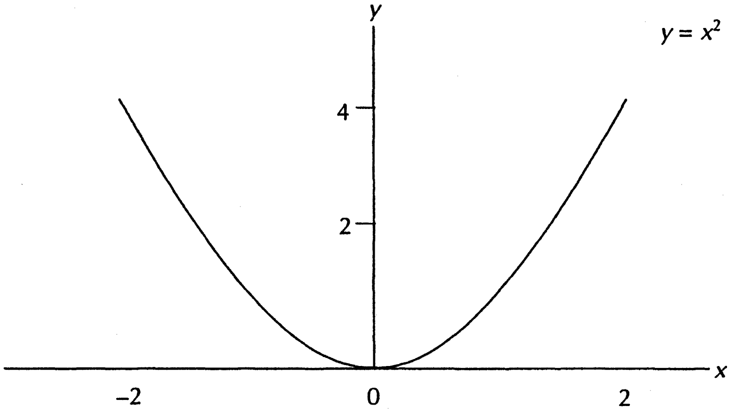


**Figure 6-5**
Function with no inverse.

[15] We are continuing to use the convention that $y^{1/2}$ refers to the positive square root of $y$, so that $-y^{1/2}$ refers to the negative square root.

$(0, 4]$. Then, from Theorem 6.16, it follows that

$$h(y) = \begin{cases} \frac{1}{4} I_{[-2,2]}(-y^{1/2}) \left| -.5y^{-.5} \right| + \frac{1}{4} I_{[-2,2]}(y^{1/2}) \left| .5y^{-.5} \right| \\ \qquad = \frac{1}{8} I_{[0,4]}(y) y^{-.5} + \frac{1}{8} I_{[0,4]}(y) y^{-.5} = \frac{1}{4} y^{-.5} I_{[0,4]}(y) \quad \text{for } y \in (0, 4], \\ 0 \quad \text{elsewhere.} \end{cases} \qquad \square$$

The change-of-variables technique can be extended to the multivariate case. We examine the case where the inverse function exists.

**Theorem 6.17** (Change of Variables Technique (Multivariate and Invertible)) *Suppose the continuous* $(n \times 1)$ *random vector* $\mathbf{X}$ *has joint PDF* $f(\mathbf{x})$. *Let* $\mathbf{g}(\mathbf{x})$ *be a* $(n \times 1)$ *real-valued vector function that is continuously differentiable* $\forall$ $\mathbf{x}$ *vector in some open rectangle of points,* $\Delta$, *containing the support of* $f(\mathbf{x})$, $\Xi$. *Assume the inverse vector function* $\mathbf{x} = \mathbf{g}^{-1}(\mathbf{y})$ *exists,* $\forall\, \mathbf{y} \in \Psi = \{\mathbf{y}{:}\mathbf{y} = \mathbf{g}(\mathbf{x}), \mathbf{x} \in \Xi\}$. *Furthermore, let*

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial g_1^{-1}(\mathbf{y})}{\partial y_1} & \cdots & \dfrac{\partial g_1^{-1}(\mathbf{y})}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial g_n^{-1}(\mathbf{y})}{\partial y_1} & \cdots & \dfrac{\partial g_n^{-1}(\mathbf{y})}{\partial y_n} \end{bmatrix},$$

*called the* **Jacobian matrix,** *be such that* $\det(\mathbf{J}) \neq 0$ *with all partial derivatives in* $\mathbf{J}$ *being continuous* $\forall\, \mathbf{y} \in \Psi$. *Then the joint density of* $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ *is given by*

$$h(\mathbf{y}) = \begin{cases} f\left(g_1^{-1}(\mathbf{y}), \ldots, g_n^{-1}(\mathbf{y})\right) \left| \det(\mathbf{J}) \right| & \textit{for } \mathbf{y} \in \Psi, \\ 0 \quad \textit{otherwise,} \end{cases}$$

*where* $|\det(\mathbf{J})|$ *denotes the absolute value of the determinant of the Jacobian.*

**Proof** The proof is quite complex and is best left to a more advanced course of study. The proof is based on the change-of-variables approach in multivariate integration problems. See T. Apostol (1974), *Mathematical Analysis*, 2nd ed., Reading, MA: Addison-Wesley, pp. 421. ■

We defer an illustration of the use of this theorem until the next section, where we will use the procedure to derive both the $t$-density and the $F$-density. The reader should note that in Theorem 6.17, there are as many coordinate functions in the vector function $\mathbf{g}$ as there are elements in the random variable vector $\mathbf{X}$. In practice, if the researcher is interested in establishing the density of fewer than $n$ random variables defined as real-valued functions of the $(n \times 1)$ vector $\mathbf{X}$, she is obliged to define "auxiliary" random variables to obtain an invertible vector function having $n$ coordinate functions, and then later integrate out the auxiliary random variables from the derived $n$-dimensional joint density to arrive at the marginal density of interest. This approach will be illustrated in the next section.

A generalization of Theorem 6.17 to noninvertible functions along the lines of Theorem 6.16 is possible. See Mood, Graybill, and Boes (1974), *Introduction to the Theory of Statistics*. New York: McGraw-Hill, p. 209, for one such generalization.

## 6.7 *t*-and *F*-Densities

In this section we consider two important PDFs relating to ratios of functions involving normally distributed random variables. These ratios and PDFs will figure prominently in the construction of hypothesis-testing procedures when random sampling is from a normal probability distribution.

### *t*-Density

**Theorem 6.18** *Let $Z \sim N(0,1)$, let $Y \sim \chi^2_v$, and let $Z$ and $Y$ be independent random variables. Then $T = Z/(Y/v)^{1/2}$ has the* **t-density** *with v degrees of freedom, defined as*

$$f(t; v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma(v/2)\sqrt{\pi v}}\left(1 + \frac{t^2}{v}\right)^{-\left(\frac{v+1}{2}\right)}.$$

**Proof** Define the $(2 \times 1)$ vector function $\mathbf{g}$ as

$$\begin{bmatrix} t \\ w \end{bmatrix} = \begin{bmatrix} g_1(z,y) \\ g_2(z,y) \end{bmatrix} = \begin{bmatrix} \dfrac{z}{(y/v)^{1/2}} \\ y \end{bmatrix},$$

where $g_2(z, y)$ is an auxiliary function of $(z, y)$ defined to allow the use of Theorem 6.17. Note that $\mathbf{g}$ is continuously differentiable $\forall\, z$ and $\forall\, y > 0$, which represents an open rectangle of $(z, y)$ points containing the support of the joint density of $(Z, Y)$ (note that the joint density of $(Z, Y)$ in this case is the product of a standard normal density and a $\chi^2$ density, which has support $(-\infty, \infty) \times (0, \infty)$). The inverse vector function $\mathbf{g}^{-1}$ is defined by

$$\begin{bmatrix} z \\ y \end{bmatrix} = \begin{bmatrix} g_1^{-1}(t,w) \\ g_2^{-1}(t,w) \end{bmatrix} = \begin{bmatrix} t\left(\frac{w}{v}\right)^{1/2} \\ w \end{bmatrix}$$

$\forall\, (t, w) \in \Psi = \{(t, w) : t \in (-\infty, \infty), w > 0\}$. The Jacobian of the inverse function is thus

$$J = \left[\begin{array}{c|c} \left(\frac{w}{v}\right)^{1/2} & \frac{t}{2}\left(\frac{w}{v}\right)^{-1/2}\left(\frac{1}{v}\right) \\ \hline 0 & 1 \end{array}\right],$$

where the elements of the Jacobian are continuous functions of $(t, w)$ and $|\det(\mathbf{J})| = |(w/v)^{1/2}| = (w/v)^{1/2} \neq 0 \;\forall\; (t, w) \in \Psi$. Given the density assumptions concerning $Z$ and $Y$,

$$f(z, y) = m_Z(z) m_Y(y) = \frac{1}{(2\pi)^{1/2}} \exp(-z^2/2) \frac{1}{2^{v/2}\Gamma(v/2)} y^{(v/2)-1} \exp(-y/2) I_{(0,\infty)}(y).$$

Then, by Theorem 6.17, the joint density of $(T, W)$ is given by

$$\begin{aligned} h(t, w) &= \frac{1}{(2\pi)^{1/2}} \exp\left(-(1/2)t^2 \left(\frac{w}{v}\right)\right) \frac{1}{2^{v/2}\Gamma(v/2)} w^{(v/2)-1} \exp\left(-\frac{w}{2}\right) I_{(0,\infty)}(w) \left(\frac{w}{v}\right)^{1/2} \\ &= \frac{1}{\Gamma(v/2)(\pi v)^{1/2} 2^{(v+1)/2}} w^{(v-1)/2} \exp\left(-\frac{w}{2}\left(1 + \frac{t^2}{v}\right)\right) I_{(0,\infty)}(w). \end{aligned}$$

Our interest centers on the density of $T$, and so we require the marginal density of $T$, $f_T(t; v) = \int_0^\infty h(t, w) dw$. Making the substitution $p = (w/2)(1 + (t^2/v))$ in the integral, so that $w = 2p/(1 + (t^2/v))$ and $dw = 2/(1 + (t^2/v))dp$, yields

$$\begin{aligned} f_T(t; v) &= \int_0^\infty \frac{1}{\Gamma(v/2)(\pi v)^{1/2} 2^{(v+1)/2}} \left[\frac{2p}{1 + \frac{t^2}{v}}\right]^{(v-1)/2} e^{-p} \left[\frac{2}{1 + \frac{t^2}{v}}\right] dp \\ &= \frac{1}{\Gamma(v/2)(\pi v)^{1/2}} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2} \int_0^\infty p^{(v-1)/2} e^{-p} dp \\ &= \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma(v/2)(\pi v)^{1/2}} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}, \end{aligned}$$

where we have used the fact that the integral in the next-to-last expression is the definition of the Gamma function $\Gamma((v + 1)/2)$. The density is known as the $t$-density with $v$ degrees of freedom, the *degrees of freedom* referring to the denominator $\chi^2_v$ random variable in the $t$-ratio. ■

The preceding theorem facilitates the derivation of the probability distribution of $T = n^{1/2}(\bar{X}_n - \mu)/\hat{\sigma}_n$ when random sampling is from a population distribution where $\hat{\sigma}_n = (n/(n - 1))^{1/2} S_n$. We will later find that the random variable $T$ has important applications in testing hypotheses about the value of the population mean, $\mu$.

**Theorem 6.19** *Under the assumptions of Theorem 6.12 and defining $\hat{\sigma}_n = (n/(n - 1))^{1/2} S_n$,*

$$T = \frac{n^{1/2}(\bar{X}_n - \mu)}{\hat{\sigma}_n} \sim \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma(v/2)(\pi v)^{1/2}} \left[1 + \frac{t^2}{v}\right]^{-(v+1)/2}$$

*where $v = n - 1$.*

**Proof** The proof is based on the fact that $T$ is defined as a standard normal random variable divided by the square root of a $\chi^2_v$ random variable that has been divided by its degrees of freedom $v = n - 1$ (in this case) and the two random variables are independent, so Theorem 6.18 applies. Details are left to the reader. (Hint: Divide numerator and denominator by $\sigma$.) ■

**Example 6.16** Recall Ex. 6.10, regarding the question of underfilling of cans. By Theorem 6.19, we know that $T = n^{1/2}(\bar{X}_n - \mu)/\hat{\sigma}_n$ has a $t$-distribution with 199 degrees of freedom when $n = 200$. *Assuming* the company's claim of $\mu = 32.25$ to be true, the outcome of $T$ in this case is 6.1559. The probability of obtaining an outcome $\geq 6.1559$ is $P(t \geq 6.1559) = 2.022 \times 10^{-9}$ (obtained by integrating the aforementioned $t$-distribution on a computer). The evidence suggests that the company's claim is suspect—its claimed value of $\mu$ may in fact be *too low*, and there is no support for the claim that the company is underfilling its packages. □

We present some properties of the $t$-distribution below.

**Family Name: *t*-Family**

*Parameterization* $v \in \Omega = \{v : v \text{ is a positive integer}\}$

*Density Definition* See Theorem 6.18.

*Moments* $\mu = 0$ for $v > 1$, $\sigma^2 = v/(v-2)$ for $v > 2$, $\mu_3 = 0$ for $v > 3$

*MGF* Does not exist

As Figure 6.6 shows, the graph of the $t$-density is symmetric about zero, and when compared to the standard normal density, the $t$-density has a smaller mode at zero and has fatter tails. However, as $v \to \infty$, $T_v \xrightarrow{d} N(0, 1)$.

**Theorem 6.20** *Let $T_v = Z/(Y_v/v)^{1/2}, Z \sim N(0,1)$, $Y_v \sim \chi^2_v$, and let $Z$ and $Y_v$ be independent, so that $T_v$ has the t-density with $v$ degrees of freedom. Then, as $v \to \infty$, $T_v \xrightarrow{d} N(0,1)$.*

**Proof** Since $Y_v \sim \chi^2_v$, then $\mathrm{E}Y_v = v$ and $\mathrm{var}(Y_v) = 2v$, and thus $\mathrm{E}(Y_v/v) = 1$, and $\sigma^2 = var(Y_v/v) = 2v^{-1}$. It follows that $(Y_v/v) \xrightarrow{m} 1$, so that $\mathrm{plim}(Y_v/v) = 1$. Also note that since $Z \sim N(0, 1)\ \forall\ v$, it follows trivially that $Z \xrightarrow{d} N(0, 1)$. Then, by Slutsky's theorems, $T_v = (Y_v/v)^{-1/2} Z \xrightarrow{d} 1 \cdot Z \sim N(0, 1)$. ■

The convergence of the $t$-density to the standard normal density is rapid, and for $v > 30$, the standard normal density provides an excellent approximation to the $t$-density. Tables of integrals of the $t$-density are widely available, generally giving the value of $c$ for which

$$P[t_v \geq c] = \int_c^\infty f_T(t; v)dt = \alpha$$

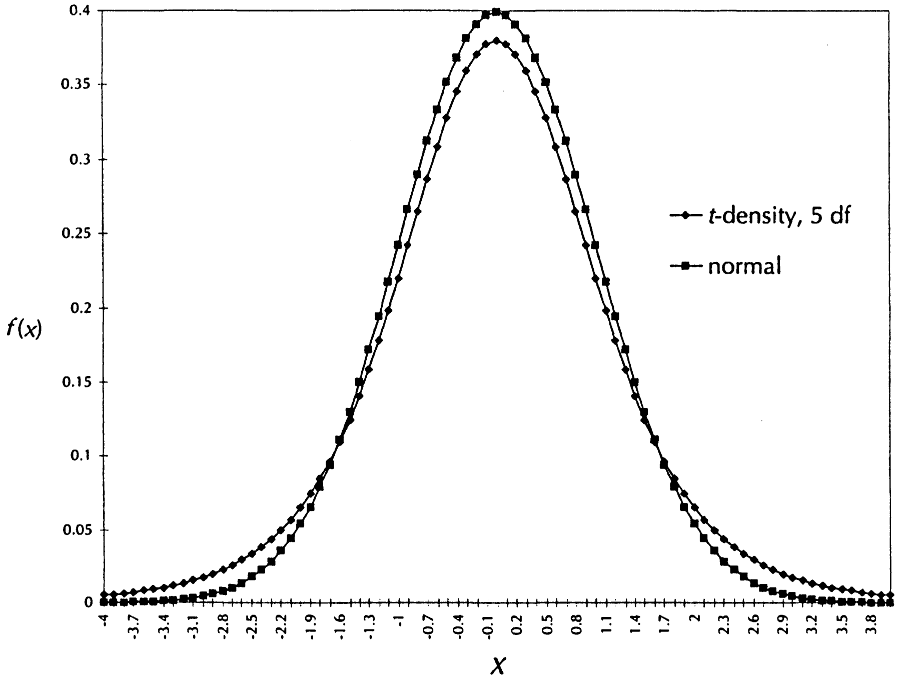


**Figure 6-6**
Standard normal and *t*-density.

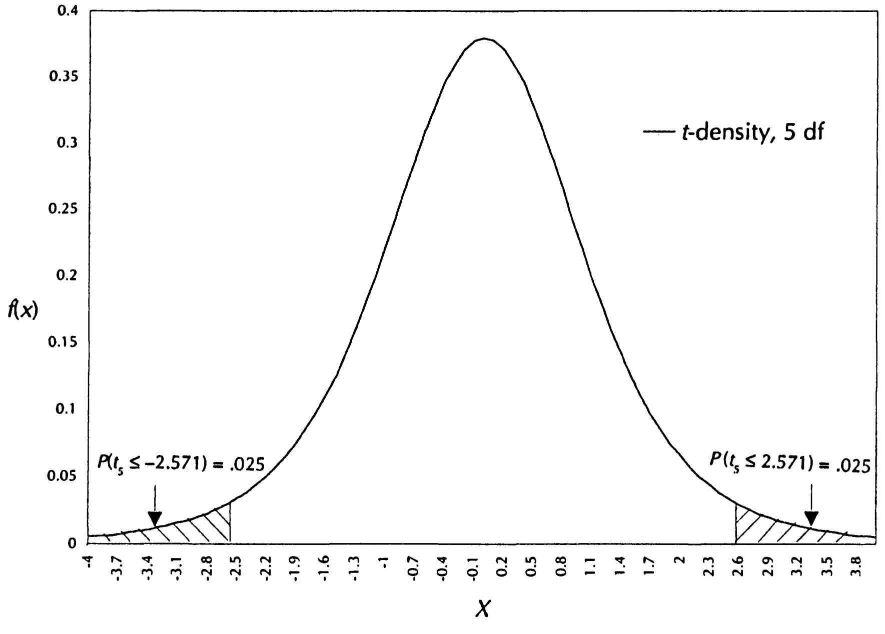


**Figure 6-7**
*t*-density symmetry.

for specific choices of $\alpha$ such as .01, .025, and .05, and for selected choices of $v$. The symmetry of the $t$-density implies that if $c$ is such that $P[t_v \geq c] = \alpha$, then $-c$ is such that $P[t_v \leq -c] = \alpha$ (see Figure 6.7). If more detail is required than what is available in the table of the $t$-density, computer programs exist for numerically integrating the $t$-density (e.g., IMSLIB (routine MDTD), SAS, SSP, Shazam, and GAUSS).

### $F$-Density

**Theorem 6.21** *Let $Y_1 \sim \chi^2_{v_1}$, let $Y_2 \sim \chi^2_{v_2}$, and let $Y_1$ and $Y_2$ be independent. Then $F = (Y_1/v_1)/(Y_2/v_2)$ has the* **F-density** *with $v_1$ numerator and $v_2$ denominator degrees of freedom, defined as*

$$m(f; v_1, v_2) = \frac{\Gamma\left(\frac{v_1+v_2}{2}\right)}{\Gamma\left(\frac{v_1}{2}\right)\Gamma\left(\frac{v_2}{2}\right)} \left(\frac{v_1}{v_2}\right)^{v_1/2} f^{(v_1/2)-1} \left(1 + \frac{v_1}{v_2} f\right)^{-(1/2)(v_1+v_2)} I_{(0,\infty)}(f).$$

**Proof** Define the $(2 \times 1)$ vector function $\mathbf{g}$ as

$$\begin{bmatrix} f \\ w \end{bmatrix} = \begin{bmatrix} g_1(y_1, y_2) \\ g_2(y_1, y_2) \end{bmatrix} = \begin{bmatrix} \dfrac{y_1/v_1}{y_2/v_2} \\ y_2 \end{bmatrix},$$

where $g_2(y_1, y_2)$ is an auxiliary function defined to allow the use of Theorem 6.17. Note that $\mathbf{g}$ is continuously differentiable $\forall\, y_1 > 0$ and $\forall\, y_2 > 0$, which represents an open rectangle of $(y_1, y_2)$ points containing the support of the joint density of $(Y_1, Y_2)$ (the support being $\times_{i=1}^2 (0, \infty)$. The inverse function $\mathbf{g}^{-1}$ is defined by

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} g_1^{-1}(f, w) \\ g_2^{-1}(f, w) \end{bmatrix} = \begin{bmatrix} \dfrac{v_1 f w}{v_2} \\ w \end{bmatrix}$$

$\forall\, (f, w) \in \Psi = \{(f, w) : f > 0, w > 0\}$. The elements of the Jacobian matrix are continuous functions of $(f, w)$, and the absolute value of the determinant of the Jacobian of the inverse function is such that $|\det(\mathbf{J})| = v_1 w / v_2 \neq 0\ \forall\, (f, w) \in \Psi$. Then since $Y_1$ and $Y_2$ have independent $\chi^2$-densities with $v_1$ and $v_2$ degrees of freedom, respectively, the joint density of $(F, W)$ is given by Theorem 6.17 as

$$h(f, w) = \frac{\left(\frac{v_1}{v_2}\right)^{v_1/2} f^{(v_1/2)-1} w^{((v_1+v_2)/2)-1}}{2^{(v_1+v_2)/2}\Gamma(v_1/2)\Gamma(v_2/2)} \exp\left(-\frac{w}{2}\left(1 + \frac{v_1 f}{v_2}\right)\right) \cdot I_{(0,\infty)}(f) I_{(0,\infty)}(w).$$

Substituting for $w$ using $p = (w/2)(1 + (v_1 f / v_2))$, and then integrating out $p$ finally yields the marginal density of $F$, as stated in the theorem. ■

We present some useful properties of the $F$-distribution below.

**Family Name: *F*-Family**

*Parameterization* $(v_1, v_2) \in \Omega = \{(v_1, v_2) : v_1 \text{ and } v_2 \text{ are positive integers}\}$

*Density Definition* See Theorem 6.21.

*Moments*

$$\mu = \frac{v_2}{v_2 - 2} \quad \text{for } v_2 > 2, \qquad \sigma^2 = \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)} \quad \text{for } v_2 > 4,$$

$$\mu_3 = \left(\frac{v_2}{v_1}\right)^3 \frac{8v_1(v_1 + v_2 - 2)(2v_1 + v_2 - 2)}{(v_2 - 2)^3(v_2 - 4)(v_2 - 6)} > 0 \quad \text{for } v_2 > 6.$$

*MGF* Does not exist

The graph of the *F*-density is skewed to the right, and for $v_1 > 2$ it has the typical shape shown in Figure 6.8.

The mode of the density occurs at $F^* = v_2(v_1 - 2)/v_1(v_2 + 2)$.

If $v_1 = 2$, the density is monotonically decreasing and approaches an intercept on the vertical axis equal to $\tau = 2v_2^{-1}\Gamma(1 + v_2/2)/\Gamma(v_2/2)$ as $F \to 0$ (see Figure 6.9).

If $v_1 = 1$, then the density is monotonically decreasing and approaches $\infty$ as $F \to 0$ (see Figure 6.10).

Values of $c$ for which $P[F \geq c] = \alpha$ for selected values of $v_1$, $v_2$, and $\alpha$ are available in tables of the *F*-distribution. However, choices of $\alpha$ are very limited in these tables, generally being .05 and .01. Computer programs are available for integrating the *F*-density more generally (e.g., IMSLIB (routine MDFD), SAS, SSP, Shazam, or GAUSS).

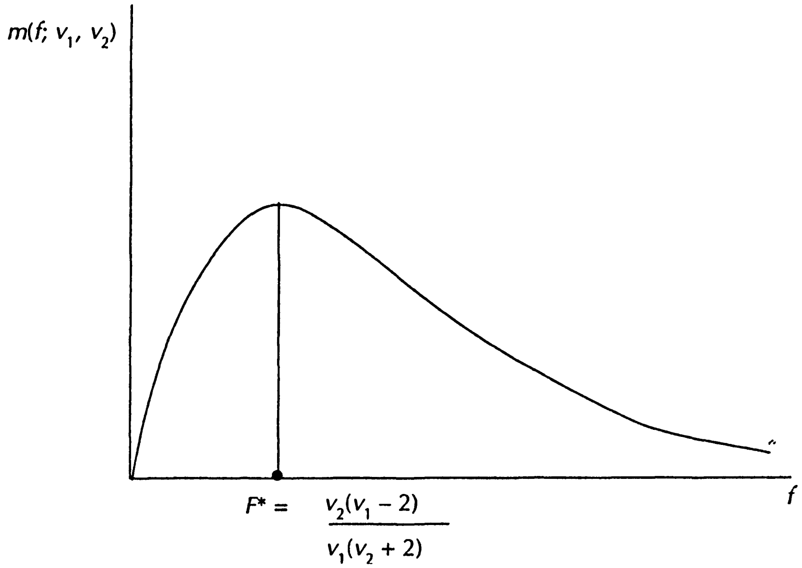


**Figure 6-8**
*F*-density for $v_1 > 2$.

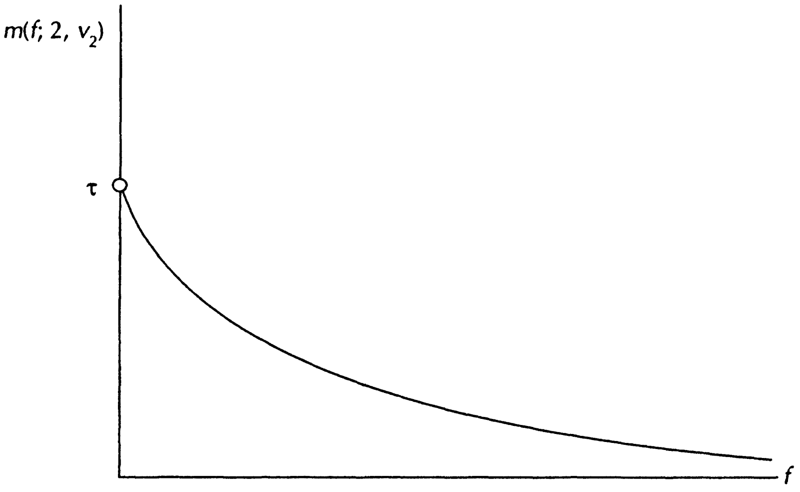


**Figure 6-9**
$F$-density for $v_1 = 2$.

It is interesting to note that if $T$ has a $t$-density with $v$ degrees of freedom, then $T^2$ has an $F$-density with 1 and $v$ degrees of freedom. The reader should verify this result, which follows directly from definitions. The reader can also verify that by letting $v_2 \rightarrow \infty$ while holding $v_1$ constant, $v_1 F \xrightarrow{d} \chi^2_{v_1}$ (a proof based on Slutsky's theorems, similar to the approach used in Theorem 6.20 can be constructed). Finally, note that if $F_{v_1,v_2}$ denotes a random variable that has the $F$-distribution with $v_1$ (numerator) and $v_2$ (denominator) degrees of freedom, then $(F_{v_1,v_2})^{-1}$ (i.e., the *reciprocal* of $F_{v_1,v_2}$ has the $F$-distribution with $v_2$ numerator and $v_1$ denominator degrees of freedom (note the reversal of degrees of freedom), which follows immediately by definition. Therefore, if the value of $c$ is such that $P[F_{v_2,v_1} \geq c] = \alpha$, then $P[(F_{v_2,v_1})^{-1} \leq c^{-1}] = P[F_{v_1,v_2} \leq c^{-1}] = \alpha$, which allows

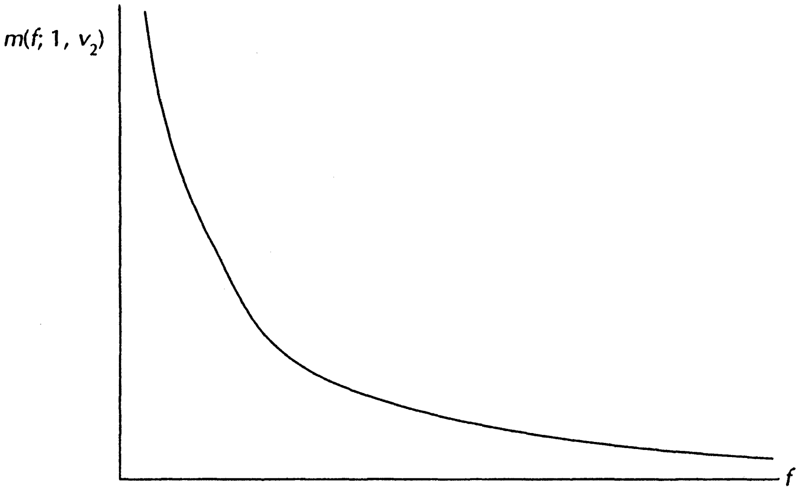


**Figure 6-10**
$F$-density for $v_1 = 1$.

one to construct lower-tail events having probability $\alpha$ from upper-tail events having probability $\alpha$.

**Example 6.17** Suppose it is desired to find the value of $b$ for which $P(F_{2,4} \leq b) = .05$ were true. From the tables of the $F$-density, one can obtain the result that $P(F_{4,2} \geq 19.25) = .05$ (notice the reversed order of the numerator and denominator degrees of freedom). It follows that $P(F_{2,4} \leq .0519) = .05$, where $b = (19.25)^{-1} = .0519$. □

## 6.8 Random Sample Simulation and the Probability Integral Transformation

There is a special real-valued function of a continuous random variable $X$, called the *probability integral transformation*, which is useful for simulating the outcomes of a random sample of size $n$ from a (continuous) probability distribution, $f(x)$. In addition, the function is useful in a certain goodness-of-fit test that we will examine later. The function of interest is defined as follows.

**Definition 6.13** ***Probability integral transformation***

Let $X$ be a continuous random variable having the cumulative distribution function $F$. Then $Y = F(X)$ is called the **probability integral transformation** of $X$.

**Example 6.18** If $X \sim (1/\theta)e^{-x/\theta}I_{(0,\infty)}(x)$, then the CDF of $X$ is given by $F(x) = (1 - e^{-x/\theta})I_{(0,\infty)}(x)$. The probability integral transformation of $X$ is then $Y = (1 - e^{-X/\theta})I_{(0,\infty)}(X)$. □

There is an important relationship between the probability integral transformation and the continuous uniform PDF, which is identified in the following theorem.

**Theorem 6.22** *Let $Y = F(X)$ be the probability integral transformation of the continuous random variable $X \sim f(x)$. Then $Y$ is uniformly distributed on the interval $(0,1)$, i.e., $Y \sim I_{(0,1)}(y)$.*

**Proof** a. *Case where $F$ is strictly increasing and continuously differentiable on an open interval containing the support of $f(x)$.* Since $F$ is strictly increasing, $dF/dx = f(x) > 0 \; \forall \; x$ in the support of $f(x)$, and $x = F^{-1}(y)$ defines the inverse function $\forall \; y \in (0,1)$. Then, by the change-of-variables technique (Theorem 6.15), the density function for $Y$ is given by

$$h(y) = \begin{cases} f(F^{-1}(y))\left|\dfrac{dF^{-1}(y)}{dy}\right| = f(x)/f(x) = 1 & \forall y \in (0,1), \\ 0 \text{ elsewhere} & \end{cases}$$

since $dF^{-1}(y)/dy = [dF(x)/dx]^{-1} = (f(x))^{-1}$ for $y = F(x)$, so that $Y$ is uniformly distributed on the interval $(0,1)$.

**b.** *Case where X has MGF $M_X(t)$:*

$$M_Y(t) = \mathrm{E}e^{Yt} = \mathrm{E}e^{F(X)t} = \int_{-\infty}^{\infty} e^{F(x)t} f(x)dx.$$

Make a change of variables using $y = tF(x)$, so that $dy/dx = tf(x)$ and $dx/dy = [tf(x)]^{-1}$, yielding

$$M_Y(t) = \int_o^t \frac{e^y}{t} dy = \frac{e^t - 1}{t} \quad \text{for } t \neq 0,$$

which by the MGF uniqueness theorem identifies a uniform distribution on (0,1) (recall the MGF of the uniform density presented in Section 4.2). A general proof of the theorem can be constructed using the characteristic function in place of the MGF above. ■

**Example 6.19** Recall Ex. 6.18. The density of $Y = g(X) = F(X)$ can be found using the change-of-variables approach. In particular, $\partial g/\partial x = (1/\theta)e^{-x/\theta} > 0 \ \forall \ x \in (0, \infty)$, the inverse function $x = -\theta \ln(1 - y)$ is defined and continuously differentiable $\forall \ y \in (0, 1)$, and $|dg^{-1}(y)/dy| = \theta/(1 - y)$. Then,

$$h(y) = \begin{cases} \dfrac{1}{\theta} \exp\left(\dfrac{\theta \ln(1 - y)}{\theta}\right) \dfrac{\theta}{1 - y} = 1 & \text{for } y \in (0, 1), \\ 0 \quad \text{elsewhere,} \end{cases}$$

so that $Y \sim I_{(0,1)}(y)$. □

The preceding theorem indicates how a random variable having any continuous CDF, $F$, can be transformed into a random variable having a uniform probability distribution on the interval (0,1). We now examine the converse to the result of Theorem 6.22—a uniformly distributed random variable on the interval (0,1) can be transformed into a random variable having any continuous CDF, $F$. In particular, if the continuous CDF $F$ admits an inverse function, then it can be shown that $X = F^{-1}(Y)$ has the CDF $F$ if $Y \sim I_{(0,1)}(y)$. Even if $F^{-1}$ does not exist, one can define a function of $Y$ that involves $F$ and that defines a random variable that has the CDF $F$. The details are provided in the next theorem.

**Theorem 6.23** *Let $Y \sim I_{(0,1)}(y)$, and let F be a continuous CDF. Then $X_* = \min_x\{x{:}F(x) \geq Y\}$ has the CDF F. If F has an inverse on (0,1), then $X_* = F^{-1}(Y)$ has the CDF F.*

**Proof**

**a.** *Case where F has an inverse on* (0,1). Since $F^{-1}$ exists, and since $F$ is continuous and strictly increasing, $\{x : F(x) \leq y\} = \{x : x \leq F^{-1}(y)\}$. The CDF of $X_*$ can be represented as $G(c) = P(x_* \leq c) = P(F^{-1}(y) \leq c) = P(y \leq F(c)) = F(c)$, since $Y$ is uniformly distributed on (0,1). Therefore, $X_*$ has the CDF $F$.

**b.** *General.* Let $x_* = \min_x\{x : F(x) \geq y\}$. Then $P(x_* \leq c) = P(y \leq F(c)) = F(c)$, since $Y \sim I_{(0,1)}(y)$, and thus $X_*$ has the CDF $F$. ■

Both of the preceding theorems have important applications. As a preview to how Theorem 6.22 might be useful in practice, suppose we are dealing with a random sample $(X_1, \ldots, X_n)$ and we hypothesize that the population distribution from which the sample was drawn is characterized by the CDF $F$. Necessarily, then, by Theorem 6.22, if our hypothesized distribution were the correct one, $Y_i = F(X_i)$, $i = 1, \ldots, n$, would constitute a random sample from a uniform population distribution on the interval $(0, 1)$. Then if the outcomes of the random sample, i.e., $y_i = F(x_i)$, $i = 1, \ldots, n$, did not exhibit behavior that would be "appropriate" for a random sample from the uniform probability distribution, our hypothesized CDF, $F$, would be suspect. We will examine how "appropriate" behavior is evaluated in this case when we examine the question of hypothesis testing.

Theorem 6.23 suggests a procedure that can be used for simulating outcomes of random samples of size $n$ from any continuous probability distribution function, $F$. Specifically, one begins with a source of outcomes of independent, uniformly distributed, random variables on the interval $(0,1)$. Computer programs are readily available that generate independent outcomes of $Y \sim I_{(0,1)}(y)$ using numerical techniques. These computer programs are commonly referred to as "uniform random-number generators" and are available in SAS, SSP, IMSLIB, Shazam, and GAUSS software, for example. Theorem 6.23 indicates that the outcome of a random sample of size $n$ from the desired population CDF $F$ can then be obtained by calculating

$$x_{*_i} = \min_x\{x : F(x) \geq y_i\}, \qquad i = 1, \ldots, n,$$

where $(y_1, \ldots, y_n)$ are the independent outcomes of $Y \sim I_{(0,1)}(y)$ generated via the computer. In many cases of interest, the function $F$ has an inverse function $F^{-1}$, so that the random sample can be represented simply as

$$x_{*i} = F^{-1}(y_i), \quad i = 1, \ldots, n.$$

**Example 6.20** Suppose we wish to simulate a random sample of size 5 from an exponential population distribution of computer memory chip lifetimes, where $\theta = 100$, so that $X \sim (1/100)e^{-x/100} I_{(0,\infty)}(x)$ and $F(x) = (1 - e^{-x/100}) I_{(0,\infty)}(x)$. We use the result in Theorem 6.23, and note that for $y \in (0, 1)$, $F$ has an inverse function defined by $x = -100 \ln(1 - y)$. Then, using five independent outcomes of $Y \sim I_{(0,1)}(y)$ generated by computer, as $\{.127, .871, .464, .922, .761\}$, we simulate a random sample outcome of 5 chip lifetimes as $\{13.582, 204.794, 62.362, 255.105, 143.129\}$. □

The preceding approach has wide applicability in simulating random samples from continuous probability distributions, but sometimes the procedure can be numerically complex if the inverse function is impossible to solve for explicitly, or if the solution to $\min_x\{x : F(x) \geq y\}$ is difficult to obtain. Other approaches exist for simulating outcomes of random samples, and the interested reader can consult the book by M. Johnson (1987), *Multivariate Statistical Simulation*. New York: John Wiley.

To this point we have not discussed a method for generating random samples from a population distribution when the distribution is discrete. In principle, the procedure in this case is relatively straightforward and, as in the continuous case, utilizes outcomes of a random variable that is uniformly distributed on $(0,1)$. Specifically, let the range of the discrete random variable $X$ be given by $R(X) = \{x_1, x_2, x_3, \ldots\}$, where $x_1 < x_2 < x_3 < \ldots$, and let the density function of $X$ be $f(x)$. Define the function $g(y)$ as

$$g(y) = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{bmatrix} \quad \text{if} \quad \begin{bmatrix} 0 < y \leq F(x_1) \\ F(x_1) < y \leq F(x_2) \\ F(x_2) < y \leq F(x_3) \\ \vdots \end{bmatrix}.$$

By the duality between CDFs and PDFs, this definition of $g(y)$ implies that $P(g(y) = x_i) = f(x_i)$ if $Y \sim I_{(0,1)}(y)$ since $P(y \in (F(x_{i-1}), F(x_i)]) = F(x_i) - F(x_{i-1}) = f(x_i)$. Thus $g(y)$, for $Y \sim I_{(0,1)}(y)$, is a discrete random variable with PDF $f$.

**Example 6.21** In order to simulate a random sample of consumer buying behavior regarding a given product, assume that the appropriate probability distribution from which random sampling would occur in this case is given by the Bernoulli distribution with $p = .5$. Then define

$$g(y) = \begin{cases} 0 & \text{if } 0 < y \leq .5, \\ 1 & \text{if } .5 < y \leq 1, \end{cases}$$

where 1 denotes a purchase, 0 denotes no purchase. Utilizing the numbers $\{.217, .766, .822, .402, .674\}$ generated via a computerized uniform random-number generator for outcomes of $Y$, we then calculate a simulated random sample of consumer purchasing decisions to be $\{0, 1, 1, 0, 1\}$. □

## 6.9 Order Statistics

Situations arise in practice where the relative magnitudes of observations are of primary interest. For example, the *largest* value in a sample of observations may be of particular interest, such as if a business is planning capacity in order to successfully service the demands of customers over a number of operating periods. The *smallest* observation may also be of interest, such as in the manufacture of a consumer electronics product that will properly function only as long as the most short-lived critical electronics component in its circuitry. One might also be interested in the median value of sample observations as a measure of central tendency of the observations, or the sample range of observations as a measure of spread. All of these statistics are examples of *order statistics* or functions of order statistics, which also play an important role in nonparametric hypothesis testing, to be introduced in Chapter 10.

The order statistics corresponding to a random sample $(X_1, \ldots, X_n)$ are simply the $X_i$'s arranged in order of increasing magnitude. The formal definition is as follows.

**Definition 6.14** ***Order statistics***

Let $\mathrm{y} = \mathrm{SORT}(x_1, \ldots, x_n)$ be the $n \times 1$ vector function whose value is the $n \times 1$ vector $[x_1, \ldots, x_n]'$ sorted from the lowest to the highest value. Then the order statistics $\mathbf{X}_o = [X_{[1]}, \ldots, X_{[n]}]'$ corresponding to the random sample $X = [X_1, \ldots, X_n]'$ are defined as $\mathbf{X}_o = \mathrm{SORT}(X)$. The random variable $X_{[k]}$ is called the kth **order statistic.**

Note that the order statistics are indeed *statistics*, since they are defined as (vector) functions of the random sample. It is apparent from the definition that outcomes of order statistics satisfy the inequalities $x_{[1]} \le x_{[2]} \le \ldots \le x_{[n]}$.

We will examine the sampling distribution of order statistics under the assumption that we are random sampling *from a population distribution*, i.e., the $X_i$'s are iid. In this case, the following result is available concerning the sampling distribution of the $k$th order statistic.

**Theorem 6.24** (Sampling distribution of $X_{[k]}$) *Let $(X_1, \ldots, X_n)$ be a random sample from a population distribution with CDF F, and let $X_{[k]}$ be the kth order statistic corresponding to the random sample. Then the CDF of $X_{[k]}$ is given by*

$$F_{X_{[k]}}(b) = \sum_{j=k}^{n} \binom{n}{j} F(b)^j [1 - F(b)]^{n-j}.$$

**Proof** For a given value of $b$, define the random variable $Y_i = I_{(-\infty,b]}(X_i)$. Note that $Y_i$ has a Bernoulli distribution with $p = P(y_i = 1) = P(x_i \le b) = F(b)$. Since the $Y_i$'s are iid, it follows that $\sum_{i=1}^{n} Y_i$ has the binomial distribution with parameters $n$ and $p$.

Now note the equivalence of the following events:

$$\{(x_1, \ldots, x_n) : x_{[k]} \le b\} = \left\{(x_1, \ldots, x_n) : \sum_{i=1}^{n} I_{(-\infty,b]}(x_i) \ge k\right\}.$$

The event to the left of the equality corresponds to the situation where the $k$th largest outcome in $(x_1, \ldots, x_n)$ is less than or equal to $b$. This event can happen iff *at least* $k$ outcomes in $(x_1, \ldots, x_n)$ are less than or equal to $b$, which is the event to the right of the equality. Hence, the events are equivalent. It then follows from event equivalence and the binomial distribution of $\sum_{i=1}^{n} Y_i$ that

$$P(x_{[k]} \le b) = P\left(\sum_{i=1}^{n} y_i \ge k\right)$$

$$= \sum_{j=k}^{n} \binom{n}{j} F(b)^j [1 - F(b)]^{n-j}. \qquad \blacksquare$$

The CDF identified in Theorem 6.24 simplifies considerably in two cases of particular relevance in applications—the smallest and the largest order statistic.

**Corollary 6.1** (Sampling Distributions of $X_{[1]}$ and $X_{[n]}$) *Assume the conditions of Theorem 6.24. Then* $F_{X_{[1]}}(b) = 1 - [1 - F(b)]^n$ *and* $F_{X_{[n]}}(b) = F(b)^n$.

**Proof** From Theorem 6.24, we have

$$F_{X_{[1]}}(b) = \sum_{j=1}^{n} \binom{n}{j} F(b)^j [1 - F(b)]^{n-j} = 1 - [1 - F(b)]^n,$$

where the second equality follows from the fact that $\sum_{j=0}^{n} \binom{n}{j} F(b)^j [1-F(b)]^{n-j} = 1$ because we are summing a binomial density over all the values in its support. Also,

$$F_{X_{[n]}}(b) = \sum_{j=n}^{n} \binom{n}{j} F(b)^j (1 - F(b))^{n-j} = [F(b)]^n. \qquad \blacksquare$$

**Example 6.22** The waiting time between customer arrivals at the pharmacy department in a variety store is given by the exponential PDF $f(x) = .2e^{-.2x} I_{(0,\infty)}(x)$, where $x$ is measured in minutes. In a random sample of 10 customers, what is the probability that the smallest waiting time will be greater than 2 minutes? What is the probability that the largest waiting time will be no greater than 5 minutes?

**Answer**: The CDF of the exponential population distribution is given by $F(b) = 1 - e^{-.2b}$. Then

$$P(x_{[1]} > 2) = 1 - P(x_{[1]} \leq 2) = 1 - F_{X_{[1]}}(2)$$

$$= \left[1 - \left(1 - e^{-.2(2)}\right)\right]^{10} = .0183,$$

$$P(x_{[n]} \leq 5) = F_{X_{[n]}}(5) = \left[1 - e^{-.2(5)}\right]^{10} = .0102. \qquad \square$$

The PDFs of the order statistics can be found using the duality between CDFs and PDFs. If the population distribution is discrete with support $x_1 < x_2 < \ldots < x_n$, say, then the PDF of the $k$th order statistic can be defined in the usual way as

$$f_{X_{[k]}}(x_i) = F_{X_{[k]}}(x_i) - F_{X_{[k]}}(x_{i-1}) \quad \text{for } i \geq 2,$$

with $f_{X_{[k]}}(x_1) = F_{X_{[k]}}(x_1)$. In the continuous case, the PDF is found after differentiation of the CDF, as well as some algebraic manipulation, to be

$$f_{X_{[k]}}(x) = \frac{dF_{X_k}(x)}{dx} = \frac{n!}{(k-1)!(n-k)!} f(x) F(x)^{k-1} [1 - F(x)]^{n-k}.$$

In order to assign probabilities to events involving outcomes of the sample range, i.e., $X_{[n]} - X_{[1]}$, or to make a joint probability statement concerning the outcomes of the largest and smallest observations in a random sample, the joint sampling distribution of $(X_{[1]}, X_{[n]})$ is needed. Furthermore, if the sample size $n$ is an even number and one wishes to assign probabilities to events involving the **sample median**, defined as $\widehat{\text{med}}(X) = \left(X_{[k]} + X_{[k+1]}\right)/2$ for $k = n/2$, the joint sampling distribution of $(X_{[k]}, X_{[k+1]})$ is needed. (When $n$ is odd, the sample median is defined as $\widehat{\text{med}}(X) = X_{[k]}$, where $k = (n+1)/2$, and so Theorem 6.24 covers this case.) The sampling distribution for any pair of order statistics is given as follows.

**Theorem 6.25** (Sampling Distribution of $(X_{[k]},X_{[\ell]})$) *Let $X_{[k]}$ and $X_{[\ell]}$, $k < \ell$, be the kth and $\ell$th order statistics corresponding to the random sample $X = (X_1,\ldots,X_n)$ from a population distribution with CDF F and PDF f. Then the joint CDF of $(X_{[k]},X_{[\ell]})$ is given by*

$$F_{X_{[k]},X_{[\ell]}}(b_k, b_\ell) = \begin{cases} F_{X_{[\ell]}}(b_\ell) & \text{for } b_k \geq b_\ell, \\ \displaystyle\sum_{i=k}^{n} \sum_{j=\max\{0,\ell-i\}}^{n-i} \frac{n!}{i!j!(n-i-j)!} F(b_k)^i [F(b_\ell) - F(b_k)]^j [1 - F(b_\ell)]^{n-i-j} & \text{for } b_k < b_\ell. \end{cases}$$

**Proof** Given $k < \ell$, it follows by definition of the order statistics that $b_k \geq b_\ell$ implies $\{x : x_{[\ell]} \leq b_\ell\} \subset \{x : x_{[k]} \leq b_k\}$, so that

$$F_{X_{[k]},X_{[\ell]}}(b_k, b_\ell) = P(x_{[k]} \leq b_k, x_{[\ell]} \leq b_\ell) = P(x_{[\ell]} \leq b_\ell) = F_{x_{[\ell]}}(b_\ell),$$

proving the first part of the definition of the CDF.

When $b_k < b_\ell$, note that the event $\{x_{[k]} \leq b_k,\ x_{[\ell]} \leq b_\ell\}$ corresponds to the event that *at least* $k$ of the random sample outcomes $x_1, \ldots, x_n$ are less than or equal to $b_k$ and *at least* $\ell$ are less than or equal to $b_\ell$. Defining the index set $I = \{(i,j) : \max\{0, \ell - i\} \leq j \leq n - i;\ k \leq i \leq n;\ i \text{ and } j \text{ are integers}\}$, we can then represent the event as

$$\{x_{[k]} \leq b_k, x_{[\ell]} \leq b_\ell\} = \bigcup_{(i,j)\in I} \{\text{exactly } i\ x_i\text{'s} \leq b_k,\ \text{ exactly } j\ x_i\text{'s such that } b_k < x_i \leq b_\ell\}.$$

Now note that each of the *disjoint* events involved in the union operation can be assigned probability via the *multinomial distribution*, where the outcome of each $X_i$ in the random sample is categorized into one of three types, $x_i \leq b_k$, $x_i \in (b_k, b_\ell]$, and $x_i > b_\ell$ which occur with probabilities $F(b_k)$, $F(b_\ell) - F(b_k)$, and $1 - F(b_\ell)$, respectively. Then directly applying the multinomial distribution (Section 4.1) and summing the probabilities of all of the disjoint events in the union operation yield the second part of the definition of the CDF. ■

As indicated following Theorem 6.24, the (joint) PDF of $(X_{[k]}, X_{[\ell]})$ can be obtained through the duality between CDFs and PDFs. In the discrete case, the PDF would be obtained by appropriate differencing of the CDF, while in the continuous case, the CDF would be differentiated (recall Theorems 2.4 and 2.5). Details are left to the reader.

The CDF simplifies considerably for the case of the joint distribution of the **extreme values**[16] $(X_{[1]}, X_{[n]})$.

**Corollary 6.2** (Sampling Distribution of $(X_{[1]},X_{[n]})$) *Let $k = 1$ and $\ell = n$ in Theorem 6.25. Then by the binomial theorem,*

$$F_{X_{[1]},X_{[n]}}(b_1, b_n) = \begin{cases} F(b_n)^n & \text{for } b_1 \geq b_n, \\ F(b_n)^n - [F(b_n) - F(b_1)]^n & \text{for } b_1 < b_n. \end{cases}$$

**Example 6.23** Revisit Ex. 6.22 regarding waiting time between customer arrivals. In a sample of 10 customer arrivals, what is the probability that the minimum waiting time ≤ 4 minutes and the maximum waiting time ≤ 8 minutes?

**Answer**:

$$\begin{aligned} F_{X_{[1]},X_{[n]}}(4, 8) &= F(8)^{10} - [F(8) - F(4)]^{10} \\ &= \left[1 - e^{-.2(8)}\right]^{10} - \left[e^{-.2(4)} - e^{-.2(8)}\right]^{10} = .1049 \end{aligned}$$

□

Sampling distributions for functions of order statistics, such as the sample range and the sample median when $n$ is even, can be pursued using the change-of-variables approach presented in Section 6.6. The multinomial logic of the proof of Theorem 6.25 can be extended to derive joint densities of three or more order statistics. For additional details on properties and functions of order statistics, see M. Kendall and A. Stuart, *Advanced Theory*, Vol. 1, Chapter 14, and the references therein.

## Key Words, Phrases, and Symbols

population
stochastic process
sampling error
random sample
random sampling with replacement
population distribution
joint density of the random sample
outcome of the random sample
random sampling without replacement
random sampling from a population distribution
composite experiment
component experiments
statistic
empirical distribution function (EDF)
Glivenko–Cantelli theorem
population moment
empirical substitution principle
$r$th sample moment about the origin, $M_r'$
$r$th sample moment about the mean, $M_r$
sample mean, $\bar{X}_n$
sample variance, $S_n^2$
sample covariance, $S_{XY}$
joint sample moment about the origin
joint sample moment about the mean

[16] The largest and smallest order statistics are sometimes alternatively referred to as **extreme values**.

sample correlation, $R_{XY}$
sample standard deviation
sample covariance matrix
sample correlation matrix
sampling density or sampling distribution of a function of a random sample
MGF approach
CDF approach
Equivalent-events approach
change-of-variables technique
support of the density
support of $f(x)$
log-normal distribution
Jacobian matrix
$t$-density
$F$-density
probability integral transformation
simulation
order statistics
sample median
extreme values

## Problems

**1.** Let $X$ be a random sample of size $n$ from an $N(\mu, \sigma^2)$ population distribution representing the weight, in ounces, of cereal placed in cereal boxes for a certain brand and type of breakfast cereal. Define $\hat{\sigma}$ as in Theorem 6.19.

a. Show that the random variable $T = n^{1/2}(\bar{X} - \mu)/\hat{\sigma}$ has the $t$-distribution with $(n - 1)$ degrees of freedom.

b. Let $n = 25$. What is the probability that the *random interval* $(\bar{X} - 2.06\hat{\sigma}/n^{1/2}, \bar{X} + 2.06\hat{\sigma}/n^{1/2})$ will have an outcome that contains the value of $\mu$? (This random interval is an example of a **confidence interval**—in this case for the population mean $\mu$. See Section 10.6.)

c. Suppose that $\bar{x} = 16.3$ and $s^2 = .01$. Define a *confidence interval* that is designed to have a .90 probability of generating an outcome that contains the value of the population mean weight of cereal placed in the cereal boxes. Generate a confidence-interval outcome for the mean weight.

**2.** Let $X$ and $Y$ be two independent random samples of sizes $n_x$ and $n_y$, respectively, from two normal population distributions that do not necessarily have the same means or variances. The two distributions refer to the miles per gallon achieved by two half-ton pickup trucks produced by two rival Detroit manufacturers. Define $\hat{\sigma}$ as in Theorem 6.19.

a. Show that the random variable $F = (\hat{\sigma}_X^2/\sigma_X^2)/(\hat{\sigma}_Y^2/\sigma_Y^2)$ has the $F$-distribution with $(n_x - 1)$ numerator and $(n_y - 1)$ denominator degrees of freedom.

b. Let $n_x = 21$ and $n_y = 31$. What is the probability that the *random interval* $(.49(\hat{\sigma}_Y^2/\hat{\sigma}_X^2), 1.93(\hat{\sigma}_Y^2/\hat{\sigma}_X^2))$ will have an outcome that contains the value of the ratio of the variances $\sigma_Y^2/\sigma_X^2$? (This random interval is another example of a confidence interval—in this case for the ratio of the population variances $\sigma_Y^2/\sigma_X^2$.)

c. Suppose that $s_x^2 = .25$ and $s_y^2 = .04$. Define a confidence interval that is designed to have a .98 probability of generating an outcome that contains the value of the ratio of population variances associated with the miles per gallon achieved by the two pickup trucks. Generate a confidence-interval outcome for the ratio of variances.

**3.** Let $X$ be a random sample of size $n$ from an $N(\mu, \sigma^2)$ population distribution representing the yield per acre, in pounds, of a new strain of hops used in the production of premium beer.

a. Justify that the random interval $(nS^2/\chi_\alpha^2, nS^2/\chi_{1-\alpha}^2)$ will have an outcome that contains the value of the population variance $\sigma^2$ with probability $(1 - 2\alpha)$, where $\chi_\delta^2$ is a number for which $P(x > \chi_\delta^2) = \delta$ when $X$ has a $\chi^2$-distribution with $(n - 1)$ degrees of freedom.

b. Suppose that $s^2 = 9$ and $n = 20$. Based on your answer to (a), define a confidence interval that is designed to have a .95 probability of generating an outcome that contains the value of the population variance of hop yields, $\sigma^2$. Generate a confidence-interval outcome for the variance.

**4.** The shipping and receiving department of a large toy manufacturer is contemplating two strategies for sampling incoming parts deliveries and estimating the proportion, $p$, of defective parts in a shipment. The two strategies are differentiated on the basis of whether random sampling will be *with* or *without* replacement. In each case, a sample mean will be calculated and used as an estimate of the proportion of defective parts in the shipment. The department wants to use the strategy that will generate estimates that are smallest in terms of expected squared distance from $p$.

a. Compare the means and variances of the sample mean under both sampling strategies. Which strategy should be used?

b. Describe conditions under which there will be little difference between the two methods in terms of expected squared distance from the true proportion of defectives.

c. Do the sample means converge in probability to $p$ in each case? Do they converge in mean square? Explain.

d. If a shipment contains 250 parts, of which 10 percent are defective, and if a random sample of 50 will be taken from the shipment, calculate the percentage reduction in expected squared distance that can be obtained by using the better strategy.

**5.** GenAg, Inc., a genetics engineering laboratory specializing in the production of better seed varieties for commercial agriculture, is analyzing the yield response to fertilizer application for a new variety of barley that it has developed. GenAg has planted 40 acres of the new barley variety and applied a different fixed level of fertilizer to each one-acre plot. In all other respects, the cultivation of the crop was identical. The GenAg scientists maintain that the relationship between observed levels of yield, in bushels per acre, and the level of fertilizer applied, in pounds per acre, will be a quadratic relationship as $Y_j = \beta_0 + \beta_1 f_j + \beta_2 f_j^2 + V_j$, where $f_j$ is the level of fertilizer applied to the $j$th one-acre plot, the $\beta$'s are fixed parameters, and the $V_j$'s are iid random variables with some continuous probability density function for which $\mathrm{E}V_j = 0$ and $\mathrm{var}(V_j) = \sigma^2$.

a. Given GenAg's assumptions, is $(Y_1, \ldots, Y_{40})$ a random sample from a population distribution or from a composite experiment? Explain.

b. Express the mean and variance of the sample mean $\bar{Y}_{40}$ as a function of the parameters and $f_j$ variables. If the sample size could be increased without bound, would $\bar{Y}_n$ converge in probability to some constant? Explain.

c. Is it true that $(\bar{Y}_n - (\beta_0 + \beta_1 n^{-1} \sum_{i=1}^{n} f_i + \beta_2 n^{-1} \sum_{i=1}^{n} f_j^2)) \xrightarrow{p} 0$? If so, interpret the meaning of this result. Based on your analysis to this point, does it appear that an outcome of $\bar{Y}_{40}$ will produce a meaningful estimate of any characteristic of the yield process?

d. Suppose that the 40 one-acre plots were all contiguous on a given 40-acre plot of land. Might there be reasons for questioning the assumption that the $V_j$'s are iid? What would the outcome of $V_j$ represent in GenAg's representation of the yield process? Presuming that the $V_j$'s were not iid, would this change your answer to (a)?

**6.** The Always Ready Battery Co. has developed a new "Failsafe" battery that incorporates a small secondary battery that becomes immediately functional upon failure of the main battery. The operating life of the main battery is a Gamma-distributed random variable as $X_1 \sim$ Gamma(3,1), where $X_1$ is measured in years. The operating life of the secondary battery is also Gamma distributed as $X_2 \sim$ Gamma(2,1). The operating lives of the main and secondary batteries are independent.

a. Let $Y_1 = X_1 + X_2$ represent the total operating life of the Failsafe battery, and let $Y_2 = X_1/(X_1 + X_2)$ represent the proportion of total operating life that is contributed by the main battery. Derive the joint probability distribution of $(Y_1, Y_2)$.

b. Are $Y_1$ and $Y_2$ independent random variables?

c. Define the marginal densities of $Y_1$ and $Y_2$. To which specific families of densities do these marginal densities belong?

d. What is the expected proportion of total operating life contributed by the main battery? What is the probability that the secondary battery contributes more than 50 percent of the total operating life of the Failsafe battery?

e. What is the expected total operating life of the Failsafe battery?

**7.** The seasonal catch of a commercial fishing vessel in a certain fishery in the southern hemisphere can be represented by $Q = c(\mathbf{z})V$, where $\mathbf{z}$ is a vector of characteristics of the vessel relating to tonnage, length, number of crew members, holding tank size, etc., $c(\mathbf{z})$ represents maximum fishing capacity of the boat; $Q$ represents the tons of fish caught; and $V \sim \theta v^{\theta-1} I_{[0,1]}(v)$ represents the proportion of fishing capacity realized.

a. Derive the density function of seasonal catch.

b. If $\theta = 10$ and $c(\mathbf{z}) = 5{,}000$, what is the expected value of seasonal catch?

**8.** A company markets its line of products directly to consumers through telephone solicitation. Salespersons are given a base pay that depends on the number of documented phone calls made plus incentive pay for each phone call that results in a sale. It can be assumed that the number of phone calls that result in sales is a binomial random variable with parameters $p$ (probability of sale) and $n$ (number of phone calls). The base pay is \$0.50 per call, and the incentive pay is \$5.00 per sale.

a. Derive the probability distribution of pay received by a salesperson making $n$ calls in a day.

b. Given that 100 calls are made in a day and $p = .05$, what is the expected pay of the salesperson? What is the probability that the pay will be \$50 or less?

**9.** The daily quantity of a commodity that can be produced using a certain type of production technology is given by the outcome of the random variable Q, defined as

$$Q = 10x_1^{.35}x_2^{.5}V,$$

where $Q$ is measured in tons/day, $x_1$ represents units of labor per day, $x_2$ represents units of capital per day, and $V = e^{\varepsilon}$, with $\varepsilon \sim N(0, \sigma^2)$.

a. Derive the probability density function of the random variable $V$. (What you will have derived is a PDF that is a member of the "log-normal" family of densities. In *general*, if $X \sim N(\mu, \sigma^2)$, $Y = e^X \sim$ log-normal, with mean $= \exp(\mu + \sigma^2/2)$, variance $= \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2)$, and $\mu'_r = \exp[r\mu + 1/2r^2\sigma^2]$).

b. Derive the density of $Q$ if $x_1 = x_2 = 2$.

c. Define the expected value of $Q$ in terms of the levels of $x_1$, $x_2$, and $\sigma^2$. What is the expected value of $Q$ if $x_1 = x_2 = 2$ and $\sigma^2 = 1$?

d. The above production technology is used in 1,600 plants in a country where the economy is centrally planned. All of the plants are required to use labor and capital at the levels $x_1 = 7.24579$ and $x_2 = 4$. Assume that $\text{var}(\varepsilon) = .25$.

An economist says that the aggregate daily production function

$$Q^* = \sum_{i=1}^{1600} Q_i = 10x_1^{.35}x_2^{.5}\sum_{i=1}^{1600} V_i$$

is such that aggregate daily production, $Q^*$, can be considered to be approximately *normally* distributed. Do you agree? Justify or refute the economist's proposition. You may assume that the $V_i$'s, and hence the $Q_i$'s, are independent random variables.

e. Define the appropriate Berry-Esséen inequality bound on the approximation error when applying the LLCLT to part (d).

**10.** The daily price, $p$, and quantity sold, $q$, of ground beef produced by the Red Meat Co. can be represented by outcomes of the bivariate random variable $(P, Q)$ having bivariate density function

$$f(p, q) = 2pe^{-pq}I_{[.5,1]}(p)I_{(0,\infty)}(q),$$

where $p$ is measured in dollars and $q$ is measured in thousands of pounds.

a. Derive the probability density function for $R = PQ$, where outcomes of $R$ represent daily revenue from ground beef sales. (Hint: define $W = P$ as an "auxiliary" random variable, and use the change-of-variables approach.)

b. What is the expected value of daily revenue? What is the probability that daily revenue exceeds \$1,000?

**11.** Let $X = (X_1, \ldots, X_{26})$ and $Y = (Y_1, \ldots, Y_{31})$ represent two independent random samples from two normal population distributions. Let $S_X^2$ and $S_Y^2$ represent the sample variances associated with the two random samples, and let $\bar{X}$ and $\bar{Y}$ represent the respective sample means. Define $\hat{\sigma}$ as in Theorem 6.19.

a. What is the value of $P(|(\bar{x} - \text{E}\bar{X})/(\hat{\sigma}_X^2/26)^{1/2}| \leq 1.316)$?

b. What is the value of $P(26s_X^2/\sigma_X^2 > 37.652)$?

c. What is the value of $P(s_X^2 > 6.02432)$, assuming that $\sigma_X^2 = 4$?

d. What is the value of $P(s_Y^2 > 1.92s_X^2)$, assuming that $\sigma_X^2 = \sigma_Y^2$?

e. Find the value of $c$ for which the following probability statement is true:

$$P\left(\frac{\sigma_Y^2\hat{\sigma}_X^2}{\sigma_X^2\hat{\sigma}_Y^2} \leq c\right) = .05.$$

**12.** The daily price, $p$, and daily quantity sold, $q$, of pink salmon produced by the AJAX Fish Packing Co. can be represented by outcomes of the bivariate random variable $(P, Q)$ with density function

$$f(p, q) = qe^{-q(1+p)}I_{(0,\infty)}(p)I_{(0,\infty)}(q),$$

where $p$ is measured in dollars, and $q$ is measured in thousands of pounds.

a. Derive the density function for $R = PQ$, where outcomes of $R$ represent daily revenue from fish sales. (Hint: Define $W = Q$ as an auxiliary random variable, and use the change-of-variables approach.)

b. What is the expected value of daily revenue? What is the probability that daily revenue exceeds \$300?

**13.** The probability that a customer entering an electronics store will make a purchase is equal to $p = .15$, and customers' decisions whether to purchase electronics equipment are independent random variables.

a. Simulate the buying behavior of ten customers entering the store using the following ten outcomes from a Uniform(0,1) computer random-number generator:

$$(.4194, .3454, .8133, .1770, .5761, .6869, .5394, .5098, .4966, .5264).$$

b. Calculate the sample mean and sample variance, and compare them to the appropriate population mean and variance.

**14.** Under the conditions of the previous problem:

a. Of the first 10 customers who enter the store on 10 consecutive days, simulate the daily number of customers that make a purchase. Use the following 10 outcomes from a Uniform(0,1) computer random-number generator:

$$(.0288, .7936, .8807, .4055, .6605, .3188, .6717, .2329, .1896, .8719).$$

b. Calculate the sample mean and sample variance, and compare them to the appropriate population mean and variance.

**15.** The number of times that a copy machine malfunctions in a day is the outcome of a Poisson process with $\lambda = .1$.

a. Simulate the operating behavior of the copy machine regarding the daily number of malfunctions over a 10 day period using the following 10 outcomes from a Uniform(0,1) computer random-number generator:

$$(.5263, .8270, .8509, .1044, .6216, .9214, .1665, .5079, .1715, .1726).$$

b. Calculate the sample mean and sample variance, and compare them to the appropriate population mean and variance.

**16.** The length of time that a 4 megabyte PC computer memory module operates until failure is the outcome of an exponential random variable with mean $EX = 3.25$, where $x$ is measured in 100,000 hour units.

a. Simulate the operating lives of 10 memory modules using the following 10 outcomes from a Uniform(0,1) computer random-number generator:

$$(.2558, .5938, .1424, .9476, .5748, .8641, .0968, .5839, .3201, .1577).$$

b. Calculate the sample mean and sample variance, and compare them to the appropriate population mean and variance.

**17.** The monthly proportion of purchases paid by check to a large grocery store that are returned because of insufficient funds can be viewed as the outcome of a Beta(1,20) distribution.

a. Simulate 12 monthly proportions of returned checks using the following 24 outcomes from a uniform(0,1) computer random-number generator:

$$(.6829, .4283, .0505, .7314, .8538, .6762, .6895, .9955, .2201, .9144, .3982, .9574, .0801, .6117, .3706, .2936, .2799, .3900, .7533, .0113, .5659, .9063, .5029, .6385).$$

(Hint: In a previous problem you have proven that $Y = X_1/(X_1 + X_2)$ has a Beta distribution with parameters $(a, b)$ if $(X_1, X_2)$ are independent Gamma-distributed random variables with parameters $(a, \beta)$ and $(b, \beta)$, respectively.)

b. Calculate the sample mean and sample variance, and compare them to the appropriate population mean and variance.

**18.** The daily closing price for a certain stock issue on the NYSE can be represented as the outcome of $Y_t = Y_{t-1} + V_t$, where $y_t$ is the value of the stock price on day $t$, and $V_t \sim N(0,4)$. (This is an example of a stochastic process known as a **random walk.**)

*a. Use the change-of-variables approach to verify that if $(U_1, U_2)$ are independent and identically distributed Uniform(0,1) random variables, then

$$V_1 = [-2\ln(U_1)]^{.5}\cos(2\pi U_2)$$

and

$$V_2 = [-2\ln(U_1)]^{.5}\sin(2\pi U_2)$$

are independent and identically distributed $N(0,1)$ random variables.

b. Simulate ten days worth of stock prices $(y_1, \ldots, y_{10})$ using $y_0 = 50$, the result in (a), and the following 10 outcomes from a Uniform(0,1) computer random-number generator:

(.9913, .4661, .1018, .0988, .4081, .3422, .1585, .6351, .0634, .4931).

c. Are you simulating a random sample from a population distribution or a random sample from a composite experiment? Calculate the sample mean and sample variance and compare them to whatever characteristics of the *random-walk* process that you feel is appropriate.

**19.** A random sample of the gas mileage achieved by 20 domestic compact automobiles resulted in the following outcome:

(25.52, 24.90, 22.24, 22.36, 26.62, 23.46, 25.46, 24.98, 25.82, 26.10, 21.59, 22.89, 27.82, 22.40, 23.98, 27.77, 23.29, 24.57, 23.97, 24.70).

a. Define and graph the empirical distribution function.

b. What is the estimated probability that gas mileage will exceed 26 miles per gallon?

c. What is the estimated probability that gas mileage will be between 24 and 26 miles per gallon?

d. Acting as if the EDF is the true CDF of gas mileages, calculate the expected value of gas mileage. Is the value you calculated equal to the sample mean? Why or why not?

**20.** The time between work-related injuries at the Imperial Tool and Die Co. during a given span of time resulted in the following 20 observations, where time was measured in weeks:

(9.68, 6.97, 7.08, .50, 6.71, 1.13, 2.20, 9.98, 4.63, 7.59, 3.99, 3.26, .92, 3.07, 17.96, 4.69, 1.80, 8.73, 18.13, 4.02).

a. Define and graph the empirical distribution function.

b. What is the estimated probability that there will be at least 8 weeks between work-related injuries?

c. What is the estimated probability that there will be between 4 and 8 weeks between work-related injuries?

d. Acting as if the EDF is the true CDF of time between work-related injuries, calculate the expected value of time between injuries. Is the value you calculated equal to the sample mean? Why or why not?

**21.** A realtor randomly samples homeowners who have purchased homes in the last two years and records their income, $y$, and home purchase price, $p$ (the population is large enough that one can consider this a random sample with replacement):

| Income | Price | Income | Price |
|---|---|---|---|
| 21,256 | 49,412 | 37,589 | 74,574 |
| 97,530 | 170,249 | 137,557 | 232,097 |
| 24,759 | 56,856 | 67,598 | 124,309 |
| 18,369 | 45,828 | 83,198 | 144,103 |
| 35,890 | 73,703 | 46,873 | 92,600 |
| 38,749 | 80,050 | 24,897 | 61,763 |
| 57,893 | 110,658 | 36,954 | 77,971 |

a. Calculate the sample covariance between income and home price.

b. Calculate the sample correlation between income and home price.

c. Calculate the linear function of income of the form $\hat{p} = a+by$ that is minimum distance from the home price observations.

d. Discuss the extent to which there is a linear relationship between income and home price.

**22.** The proportion of the work force of a large manufacturing firm that takes at least one day's sick leave in a given work week is assumed to be the outcome of a random variable whose PDF is well represented by a uniform distribution on the interval [0,.10].

a. In a random sample of 8 work weeks, what is the probability that the maximum proportion of workers who take sick leave is less than or equal to .05?

b. What is the probability that the sum of all 8 sample observations will be between .25 and .75?

**23.** A large aircraft manufacturer produces a passenger jet having a navigation component consisting of three sequentially functioning redundant navigation systems that will allow the jet to be properly controlled so long as at least one of the systems remains operational. The operating life of each of the systems is the outcome of a random variable having the exponential PDF

$$f(x) = .1e^{-.1x}I_{[0,\infty)}(x),$$

where $x$ is measured in thousands of hours, and the operating lives are independent of one another.

a. What is the probability that the navigation component will continue to function for at least 20,000 hours?

b. In an economizing mode, the management is considering a redesign of the jet that will reduce the navigation component from three redundant systems to two. How does this affect the probability of the event in (a)?

# 7 Elements of Point Estimation Theory

## 7.1 Introduction

The problem of point estimation examined in this chapter is concerned with the estimation of the values of unknown parameters, or functions of parameters, that represent characteristics of interest relating to the probability space of some collection of economic, sociological, biological, or physical experiments. The outcomes generated by the collection of experiments are assumed to be outcomes of a random sample with some joint probability density function $f(x_1, \ldots, x_n; \Theta)$. The random sample need *not* be from a population distribution, that is, it is not necessarily the case that $X_1, \ldots, X_n$ are iid. The objective of point estimation will be to utilize functions of the random sample outcome to generate good (in some sense) estimates of the unknown characteristics of interest.

The types of estimation problems that will be examined in this (and the next) chapter are also referred to as problems of **parametric estimation**, as opposed to nonparametric estimation problems. A problem of *parametric* estimation is concerned with estimating the values of *parameters* or *functions of parameters*. In particular, the objective is to utilize a sample outcome $[x_1, \ldots, x_n]'$ of $\mathbf{X} = [X_1, \ldots, X_n]'$ to estimate the unknown value $\Theta_*$ or $\mathbf{q}(\Theta_*)$, where $\Theta_*$ denotes the value of the parameter vector associated with the joint PDF that *actually* determines the probabilities of events for the random sample outcome. That is, $\Theta_*$ is the value of $\Theta$ such that $\mathbf{X} \sim f(\mathbf{x}; \Theta_*)$ is a *true statement*, and for this reason $\Theta_*$ is often referred to as the **true value of** $\Theta$, and we can then also speak of $\mathbf{q}(\Theta_*)$ as being the **true value of** $\mathbf{q}(\Theta)$ and $f(\mathbf{x};\Theta_*)$ as being the **true PDF of X**. Some examples of functions of $\Theta_*$ that might be of interest when sampling from a population distribution $f(z;\Theta_*)$ include

1. $q_1(\Theta_*) = \mathrm{E}(Z) = \int_{-\infty}^{\infty} z f(z;\Theta_*)dz$ (mean),
2. $q_2(\Theta_*) = \mathrm{E}(Z - \mathrm{E}Z)^2 = \int_{-\infty}^{\infty}(z - \mathrm{E}Z)^2 f(z;\Theta_*)dz$ (variance),
3. $q_3(\Theta_*)$ defined implicitly by $\int_{-\infty}^{q_3(\Theta_*)} f(z;\Theta_*)dz = .5$ (median),
4. $q_4(\Theta_*) = \int_a^b f(z;\Theta_*)dz$ $(P(z \in [a,b]))$.

The method used to solve a parametric estimation problem will generally depend on the degree of specificity with which one can define the family of candidates for the true PDF of the random sample, **X**. The situation for which the most statistical theory has been developed, both in terms of the actual procedures used to generate point estimates and in terms of the evaluation of the properties of the procedures, is the **distribution-specific case**. In this case, the candidates, $f(x_1,\ldots,x_n;\Theta)$, are assumed at the outset to belong to a *specific parametric family* of PDFs (e.g., normal, Gamma, binomial). Application of the celebrated *maximum likelihood* estimation procedure (see Chapter 8) relies on the candidates for the distribution of $X$ being members of a specific collection of density functions that are indexed, and fully algebraically specified, by the values of $\Theta$.

In the distribution-nonspecific, or **distribution-free**, case, a specific functional definition of the potential PDFs for **X** is *not* assumed, although some assumptions about the lower-order moments of $f(\mathbf{x};\Theta)$ are generally made. In any case, it is often still possible to generate useful point estimates of various characteristics of the probability space of **X** that are conceptually functions of parameters, such as moments, quantiles, and probabilities, even if the specific parametric family of PDFs for **X** is not specified. For example, useful point estimates (in a number of respects) of the parameters in the so-called general linear model representation of a random sample from a composite experiment can be generated with only a few general assumptions regarding the lower-order moments of $f(x_1,\ldots,x_n;\Theta)$ without any assumptions that the density is of a *specific* parametric form (see Section 8.2).

Distribution-free methods of parametric estimation have an advantage of being applicable to a wide range of sampling distributions since they are defined in a distribution-nonspecific context that inherently subsumes many different functional forms for $f(\mathbf{x};\Theta)$. However, it is usually the case that superior methods of estimating $\Theta_*$ or $\mathbf{q}(\Theta_*)$ exist if a parametric family of PDFs for **X** can be specified. Put another way, the more information one has about the form of $f(\mathbf{x};\Theta)$ at the outset, the more precisely one can estimate $\Theta_*$ or $\mathbf{q}(\Theta_*)$ (in most cases). We will encounter examples of both distribution-specific and distribution-free parametric estimation in the remainder of our study.

There is an alternative method of inference, called **nonparametric estimation**, that is used when estimating characteristics of the probability space of **X** that are not directly concerned with parameters. For example, a researcher may be interested in obtaining an estimate of the CDF associated with a population distribution under study. This type of estimation problem can be conceptualized without reference to any parameters whatsoever. The researcher may wish to estimate $F(z)$, where $(X_1, \ldots, X_n)$ is an iid random sample from the population distribution $F(z)$, and no mention is made, nor required, regarding *parameters* of the CDF. We have already examined a method for estimating the CDF in the case where the random sample is from a population distribution, namely, the empirical distribution function, $F_n$, provides an estimate of $F$. We leave the general study of nonparametric estimation to a more advanced course of study; interested readers should refer to M. Puri and P. Sen (1985), *Nonparametric Methods in General Linear Models*. New York: John Wiley; F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel (1986), *Robust Statistics*. New York: John Wiley; and J. Pratt and J. Gibbons (1981), *Concepts of Nonparametric Theory*. New York: Springer-Verlag.[1]

## 7.2 Statistical Models

A problem of point estimation begins with either a fully or partially specified *statistical model* for the random sample $\mathbf{X} = [X_1, \ldots, X_n]'$ whose outcome $\mathbf{x} = [x_1, \ldots, x_n]'$ constitutes the observed data being analyzed in a real-world problem of statistical inference.

**Definition 7.1**
***Statistical model***

A **statistical model** for a random sample $\mathbf{X}$ consists of a parametric functional form, $f(\mathbf{x}; \Theta)$, for the sampling density of $\mathbf{X}$ together with a parameter space, $\Omega$, that defines the set of potential candidates for the true sampling density of $\mathbf{X}$ as $\{f(\mathbf{x}; \Theta), \Theta \in \Omega\}$.

The statistical model defines the probabilistic and parametric context in which point estimation proceeds. Once the statistical model has been specified, interest centers on estimating the true values of some (or all) of the parameters, or on estimating the true values of some functions of the parameters of the problem. The specific objectives of any point estimation problem depend on the needs of the researcher, who will identify which quantities are to be estimated.

The case of distribution-specific parametric estimation of $\Theta_*$ or $q(\Theta_*)$ is associated with a fully specified statistical model in which a specific parametric family of PDFs is represented by $\{f(\mathbf{x}; \Theta), \Theta \in \Omega\}$. For example, a fully specified statistical model for a random sample of the miles per gallon achieved by 25 randomly chosen trucks from the assembly line of a Detroit manufacturer might be defined as $\{\prod_{i=1}^{25} N(x_i; \mu, \sigma^2), (\mu, \sigma^2) \in \Omega\}$, where $\Omega = (0, \infty) \times (0, \infty)$.

[1] There is not universal agreement on the meaning of the terms *parametric*, *nonparametric*, and *distribution-free*. Sometimes nonparametric and distribution-free are used synonymously, although the case of distribution-free *parametric* estimation is pervasive in econometric work. See J. D. Gibbons (1982) *Encyclopedia of Statistical Sciences*, Vol. 4. New York: Wiley, pp. 400–401.

In the distribution-free case, a specific functional form for $f(\mathbf{x}; \Theta)$ is not defined and $\Omega$ may or may not be fully specified. For example, in the preceding truck mileage example, a partially specified statistical model would be $\{f(\mathbf{x}; \mu, \sigma^2), (\mu, \sigma^2) \in \Omega\}$, where $\Omega = (0, \infty) \times (0, \infty)$ and $f(\mathbf{x}; \mu, \sigma^2)$ is *some continuous PDF*. In this latter case, the statistical model allows for the possibility that $f(\mathbf{x}; \mu, \sigma^2)$ is *any* continuous PDF having a mean of $\mu$ and variance of $\sigma^2$, with both $\mu$ and $\sigma^2$ positive, e.g., normal, Gamma, or uniform PDFs would be potential candidates.

In specifying a statistical model, the researcher presumably identifies the parametric family based on a combination of experience, consideration of the real-world characteristics of the experiments involved, theoretical considerations, past analyses of similar problems, an attempt at a reasonably robust approximation to the probability distribution, and/or pragmatism. The degree of detail with which the parametric family of densities is specified can vary from problem to problem. In certain situations there will be great confidence in a detailed choice of parametric family. For example, suppose we are interested in estimating the proportion, $p$, of defective manufactured items in a shipment of $N$ items. If a random sample with replacement of size $n$ is taken from the shipment (population) of manufactured items, then

$$(X_1, \ldots, X_n) \sim f(x_1, \ldots, x_n; p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \prod_{i=1}^n I_{\{0,1\}}(x_i)$$

unequivocally represents the parametric family of densities characterizing the joint density of the random sample, and interest centers on estimating the unknown value of the parameter $p$.

On the other hand, there will be situations in which the specification of the parametric family is quite tentative. For example, suppose one were interested in estimating the average operating life of a certain brand of hard-disk based on outcomes of a random sample of hard-disk lifetimes. In order to add some mathematical structure to the estimation problem, we might represent the $i$th random variable in the random sample of lifetimes $(X_1, \ldots, X_n)$ as $X_i = \mu + V_i$, where $\mu$ represents the unknown mean of the population distribution of lifetimes, an outcome of $X_i$ represents the actual lifetime observed for the $i$th hard disk sampled, and the corresponding outcome of $V_i$ represents the deviation of $x_i$ from $\mu$. Since $(X_1, \ldots, X_n)$ is a random sample from the population distribution, it follows that $\mathrm{E}X_i = \mu$ and $\mathrm{var}(X_i) = \sigma^2 \forall i$ can be assumed, so that $\mathrm{E}V_i = 0$ and $\mathrm{var}(V_i) = \sigma^2 \forall i$ can also be assumed. Moreover, it is then legitimate to assume that $(X_1, \ldots, X_n)$ and $(V_1, \ldots, V_n)$ are each a collection of iid random variables. To this point, we have already specified that the parametric family of distributions associated with $\mathbf{X}$ is of the form $\prod_{i=1}^n m(x_i; \Theta)$, where the density $m(z; \Theta)$ has mean $\mu$ and variance $\sigma^2$ (what is the corresponding specification for $V$?).

Now, what parametric functional specification of $m(x_i; \Theta)$ can be assumed to contain the specific density that represents the actual probability distribution of $X_i$ or $V_i$? (Note, of course, that specifying a parametric family for $V_i$ would imply a corresponding parametric family for $X_i$, and vice versa.) One general specification would be the collection of all continuous joint density functions $f(x_1, \ldots, x_n; \Theta)$ for which $f(x_1, \ldots, x_n; \Theta) = \prod_{i=1}^{n} m(x_i; \Theta)$ with $EX_i = \mu$ and $\text{var}(X_i) = \sigma^2, \forall\, i$. The advantage of such a general specification of density family is that we have great confidence that the actual density function of **X** is contained within the implied set of potential PDFs, which we will come to see as an important component of the specification of any point estimation problem. In this particular case, the general specification of the statistical model actually provides sufficient structure to the point estimation problem for a useful estimate of mean lifetime to be generated (for example, the least-squares estimator can be used to estimate $\mu$—see Chapter 8). We will see that one disadvantage of very general specifications of the statistical model is that the interpretation of the properties of point estimates generated in such a general context is also usually not as specific or as detailed as when the density family can be defined with greater specificity.

Consider a more detailed specification of the statistical model of hard-disk operating lives. If we feel that lifetimes are symmetrically distributed around some point, $\mu$, with likelihoods of lifetimes declining the more distant the measurement is from $\mu$, we might consider the normal parametric family for the distribution of $V_i$. It would, of course, follow that $X_i$ would then also be normally distributed, and thus the normal distribution could serve only as an *approximation* since negative lifetimes are impossible. Alternatively, if we felt that the distribution of lifetimes was skewed to the right, the Gamma parametric family provides a rich source of density shapes, and we might specify that the $X_i$'s have some Gamma density, and thus the $V_i$'s would have the density of a Gamma-type random variable that has been shifted to the left by $\mu$ units. Hopefully, the engineering staff could provide some guidance regarding the most defensible parametric family specification to adopt. In cases where there is considerable doubt concerning the appropriate parametric family of densities, tests of hypotheses concerning the adequacy of a given parametric family specification can be performed. Some such tests will be examined in Chapter 10. In some problem situations, it may not be possible to provide any more than a general specification of the density family, in which case the use of distribution-free methods of parametric estimation will be necessary.

Given that a parametric functional form is specified to characterize the joint density of the random sample, a parameter space, $\Omega$, must also be identified to complete the statistical model. There are often natural choices for the parameter space. For example, if the Bernoulli family were specified, then $\Omega = \{p : p \in [0, 1]\}$, or if the normal family were specified, then $\Omega = \{(\mu, \sigma) : \mu \in (-\infty, \infty), \sigma > 0\}$. However, if only a general definition of the parametric family of densities is specified at the outset of the point estimation problem, the specification of the parameter space for the parametric family will then also be general and often incomplete. For example, a parameter space specification for the aforementioned point estimation problem involving hard-disk lifetimes could be $\Omega = \{(\mu, \sigma^2, \Theta_o) : \mu \geq 0, \sigma^2 \geq 0\}$. In this case, since the specific algebraic form of $f(x_1, \ldots, x_n; \Theta)$ is not specified, we can only state that the mean and variance of hard-disk lifetimes are nonnegative, and we leave any other unknown parameters of the density function, denoted by $\Theta_o$, unrestricted.

Regardless of the level of detail with which $\Omega$ is specified, there are two important assumptions regarding the specification of $\Omega$ that are made in the context of a point estimation problem. First, it is assumed that $\Omega$ contains the true value of $\Theta$, so the statistical model given by $\{f(x : \Theta), \Theta \in \Omega\}$ can be assumed to contain the true sampling distribution for the random sample under study. Put another way, in the context of a point estimation problem, the set $\Omega$ is assumed to represent the entire collection of possible values for $\Theta_*$. The relevance of this assumption in the context of point estimation is perhaps obvious—if our objective in point estimation is to estimate the value of $\Theta_*$ or $\mathbf{q}(\Theta_*)$, we do not want to preclude $\Theta_*$ or $\mathbf{q}(\Theta_*)$ from the set of potential estimates. Note that, in practice, this may be a tentative assumption that is subjected to statistical test for verification or refutation (see Chapter 10).

The second assumption on $\Omega$ concerns the notion of the *identifiability* of the parameter vector $\Theta$. As we alluded to in our discussion of parametric families of densities in Chapter 4, parameterization of density families is *not* unique. Any invertible transformation of $\Theta$, say $\lambda = \mathbf{h}(\Theta)$, defines an alternative parameter space $\Lambda = \{\lambda : \lambda = \mathbf{h}(\Theta), \Theta \in \Omega\}$ that can be used to specify an alternative statistical model for $\mathbf{X}$ that contains the same PDF candidates as the statistical model based on $\Omega$, i.e.,

$$\{f(\mathbf{x}; \mathbf{h}^{-1}(\lambda)), \lambda \in \Lambda\} = \{f(\mathbf{x}; \Theta), \Theta \in \Omega\}.$$

Defining $m(\mathbf{x}; \lambda) \equiv f(\mathbf{x}; \mathbf{h}^{-1}(\lambda))$, the alternative statistical model could be written as $\{m(\mathbf{x}; \lambda), \lambda \in \Lambda\}$. The analyst is free to choose whatever parameterization appears to be most natural or useful in the specification of a statistical model, so long as the parameters in the chosen parameterization are *identified*. In stating the definition of parameter identifiability, we use the terminology **distinct PDFs** to refer to PDFs that assign *different* probabilities to at least one event for $\mathbf{X}$.

**Defintion 7.2**
***Parameter identifiability***

Let $\{f(\mathbf{x}; \Theta), \Theta \in \Omega\}$ be a statistical model for the random sample $\mathbf{X}$. The parameter vector $\Theta$ is said to be **identified** or **identifiable** iff $\forall\, \Theta_1$ and $\Theta_2 \in \Omega$, $f(\mathbf{x}; \Theta_1)$ and $f(\mathbf{x}; \Theta_2)$ are *distinct* if $\Theta_1 \neq \Theta_2$.

The importance of parameter identifiability is related to the ability of random sample outcomes to provide discriminatory information regarding the choice of $\Theta \in \Omega$ to be used in estimating $\Theta_*$. If the parameter vector in a statistical model is *not* identified, then two or more different values of the parameter vector $\Theta$, say $\Theta_1$ and $\Theta_2$, are associated with precisely the same sampling distribution for $\mathbf{X}$. In this event, random-sample outcomes cannot possibly be used to discriminate between the values of $\Theta_1$ and $\Theta_2$ since the probabilistic behavior of $\mathbf{X}$ under either possibility is indistinguishable. We thus insist on parameter identifiability in a point estimation problem so that different values of $\Theta$ are associated with different probabilistic behavior of the outcomes of the random sample.

**Example 7.1** Parameter Identifiability

The yield per acre of tomatoes on 20 geographically dispersed parcels of irrigated land is thought to be representable as the outcomes of

$$Y_i = \beta_0 + \beta_1 T_i + V_i, \quad i = 1, \ldots, 20,$$

where $\beta_0$ and $\beta_1 > 0$, and $(T_i, V_i)$, $i = 1, \ldots, 20$, are iid outcomes of a bivariate normal population distribution with mean vector $[\mu_T, 0]'$, for $\mu_T > 0$, and diagonal covariance matrix with diagonal entries $\sigma_T^2$ and $\sigma_V^2$. The outcome $y_i$ represents bushels/acre on parcel $i$, and $t_i$ is the season average temperature measured at the growing site. The statistical model for the random sample $\mathbf{Y} = [Y_1, \ldots, Y_{20}]'$ is subsequently specified as

$$\left\{ \prod_{i=1}^{20} N\left(y_i; \beta_0 + \beta_1 \mu_T, \beta_1^2 \sigma_T^2 + \sigma_V^2\right), (\beta_0, \beta_1, \mu_T, \sigma_T^2, \sigma_v^2) \in \Omega \right\},$$

where $\Omega = \times_{i=1}^{5}(0, \infty)$. Is the parameter vector $[\beta_0, \beta_1, \mu_T, \sigma_T^2, \sigma_V^2]'$ identified?
**Answer**: No. Define $\mu = \beta_0 + \beta_1 \mu_T$ and $\sigma^2 = \beta_1^2 \sigma_T^2 + \sigma_V^2$, and examine the statistical model for $\mathbf{Y}$ given by

$$\left\{ \prod_{i=1}^{20} N(y_i; \mu, \sigma^2), (\mu, \sigma^2) \in \Lambda \right\},$$

where $\Lambda = \times_{i=1}^{2}(0, \infty)$. Note that any choice of positive values for $\beta_0, \beta_1, \mu_T, \sigma_T^2$, and $\sigma_V^2$ that result in given positive values for $\mu$ and $\sigma^2$ result in precisely the same sampling distribution for $\mathbf{Y}$ (there is an infinite set of such choices for each value of the vector $[\mu, \sigma^2]'$). Thus, the original parameter vector is not identified. Note that the parameter vector $[\mu, \sigma^2]'$ in the latter statistical model for $\mathbf{Y}$ *is* identified since the sampling distributions associated with two different positive values of the vector $[\mu, \sigma^2]'$ are distinct. □

Once the researcher has identified the characteristics of the sampling distribution that are to be estimated, the problem becomes one of devising a procedure to generate estimates of the relevant characteristics. We will presume that the only information the researcher has that relates to the characteristics is represented by the information contained in the sample outcome. If additional *prior* information is available regarding the likely values of characteristics, the Bayesian approach to inference, which combines both sample information and prior information into the estimation procedure, might be considered. We will not examine the Bayesian approach here. The interested reader can consult A. Zellner, (1971), *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley, and G. E. P. Box and G. C. Tiao (1973), *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley, for further information.

We pause here to introduce a convention regarding the interpretation of phrases such as **estimating** $\Theta$ or **estimating** $\mathbf{q}(\Theta)$, or **an estimate of** $\Theta$ **(or of** $\mathbf{q}(\Theta)$**)**. Since $\Theta$ is simply a parameter vector that indexes a family of density functions and that can assume a range of alternative values (those specified in $\Omega$), the reader might wonder what such phrases could possibly mean. That is, *what* are we estimating if we are estimating, say, $\Theta$? The phrases are used as a shorthand or abbreviated way of stating that one is *estimating the true value of* $\Theta$, or *estimating the true value of* $\mathbf{q}(\Theta)$, or that one has *an estimate of the true value of* $\Theta$ *(or of* $\mathbf{q}(\Theta)$*)*. There is widespread use of such phrases in the statistics and econometrics literature, and we will make frequent use of such phrases in this book as well. In general, one must rely on the context of the discussion to be sure whether $\Theta$ or $\mathbf{q}(\Theta)$ refers to the quantity being estimated or merely to the indexing parameter of a family of joint density functions.

## 7.3 Estimators and Estimator Properties

### Estimators

Point estimation is concerned with estimating $\Theta$ or $\mathbf{q}(\Theta)$ from knowledge of the outcome $\mathbf{x} = [x_1, \ldots, x_n]'$ of a random sample $\mathbf{X}$. It follows from this basic description of point estimation that *functions* are at the heart of the estimation problem: The inputs or domain elements are sample outcomes, $\mathbf{x}$, and the outputs or range elements are estimates of $\Theta$ or $\mathbf{q}(\Theta)$. More formally, estimates will be generated via some function of the form $\mathbf{t} : R(\mathbf{X}) \rightarrow R(\mathbf{t})$, where $R(\mathbf{t})$ is the range of $\mathbf{t}$ defined as $R(\mathbf{t}) = \{\mathbf{t} : \mathbf{t} = \mathbf{t}(\mathbf{x}), \mathbf{x} \in R(\mathbf{X})\}$. Note that $R(\mathbf{t})$ represents the set of all possible estimates of $\Theta$ or $\mathbf{q}(\Theta)$ that can be generated as outcomes of $\mathbf{t}(\mathbf{X})$. We will always tacitly assume that $\mathbf{t}(\mathbf{X})$ is an observable random variable, and hence a *statistic,* so that estimates are observable and empirically informative.

Henceforth, when the function $\mathbf{t} : R(\mathbf{X}) \rightarrow R(\mathbf{t})$ represented by $\mathbf{t} = \mathbf{t}(\mathbf{x})$ is being utilized to generate estimates of $\mathbf{q}(\Theta)$, we will refer to the random variable $\mathbf{T} = \mathbf{t}(\mathbf{X})$ as an **estimator** for $\mathbf{q}(\Theta)$, and $\mathbf{q}(\Theta)$ will be referred to as the **estimand.** An outcome, $\mathbf{t} = \mathbf{t}(\mathbf{x})$, of the estimator will be referred to as an **estimate** of $\mathbf{q}(\Theta)$. We formalize these three terms in the following definitions:

**Definition 7.3**
***Point estimator and estimand***

A statistic or vector of statistics, $\mathbf{T} = \mathbf{t}(\mathbf{X})$, whose outcomes are used to estimate the value of a scalar or vector function, $\mathbf{q}(\Theta)$, of the parameter vector, $\Theta$, is called a **point estimator,** with $\mathbf{q}(\Theta)$ called the **estimand.**[2]

[2] Note, as always, that the function $\mathbf{q}$ can be the identity function $\mathbf{q}(\Theta) \equiv \Theta$, in which case we could be referring to estimating $\Theta$ itself. Henceforth, it will be understood that since $\mathbf{q}(\Theta) \equiv \Theta$ is a possible choice of $\mathbf{q}(\Theta)$, all discussion of estimating $\mathbf{q}(\Theta)$ could be referring to estimating $\Theta$ itself.

**Definition 7.4**
***Point Estimate***

An observed outcome of an estimator is called a **point estimate.**

Figure 7.1 contains a schematic overview of the general context of the point estimation problem to this point.

### Estimator Properties

Since there is literally an *uncountably infinite* set of possible functions of $\mathbf{X}$ that are potential estimators of $\mathbf{q}(\Theta)$, a fundamental problem in point estimation is the choice of a "good" estimator. In order to rank the efficacy of estimators and/or to choose the *optimal* estimator of $\mathbf{q}(\Theta)$, an objective function that establishes an appropriate measure of "goodness" must be defined.

A natural measure to use in ranking estimators would seem to be the distance between outcomes of $\mathbf{t}(\mathbf{X})$ and $\mathbf{q}(\Theta)$, which is a direct measure of how *close* estimates are to what is being estimated. In the current context, this distance measure would be defined by

$$d(\mathbf{t}(\mathbf{x}), \mathbf{q}(\Theta)) = ([\mathbf{t}(\mathbf{x}) - \mathbf{q}(\Theta)]'[\mathbf{t}(\mathbf{x}) - \mathbf{q}(\Theta)])^{1/2},$$

which specializes to $|t(\mathbf{x}) - q(\Theta)|$ when $k = 1$. However, upon closer examination, this closeness measure has an obvious practical flaw for comparing alternative functions for estimating $\mathbf{q}(\Theta)$—the estimate that would be preferred depends on the true value of $\mathbf{q}(\Theta)$, which is unknown (otherwise there would be no point estimation problem). This problem is clearly not the fault of the particular closeness measure chosen, since any reasonable measure of closeness between the two values $\mathbf{t}(\mathbf{x})$ and $\mathbf{q}(\Theta)$ would depend on where $\mathbf{q}(\Theta)$ actually is in $R^k$ vis-a-vis where $\mathbf{t}(\mathbf{x})$ is located. Thus, comparing alternative functions for estimating $\mathbf{q}(\Theta)$ on the basis of the closeness to $\mathbf{q}(\Theta)$ of an actual estimate $\mathbf{t}(\mathbf{x})$ is

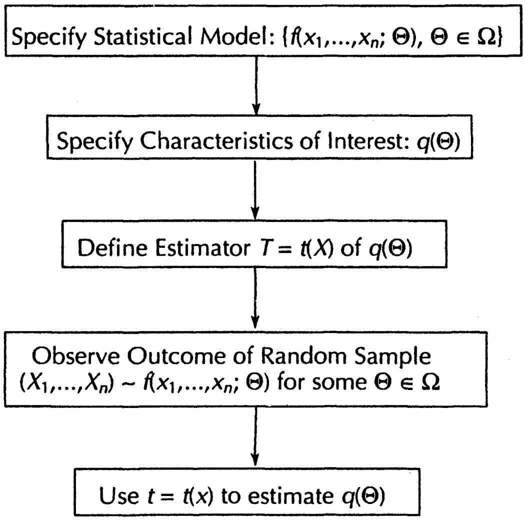

Specify Statistical Model: $\{f(x_1,\ldots,x_n;\Theta), \Theta \in \Omega\}$
Specify Characteristics of Interest: $q(\Theta)$
Define Estimator $T = t(X)$ of $q(\Theta)$
Observe Outcome of Random Sample $(X_1,\ldots,X_n) \sim f(x_1,\ldots,x_n;\Theta)$ for some $\Theta \in \Omega$
Use $t = t(x)$ to estimate $q(\Theta)$

**Figure 7-1**
General point estimation procedure.

not tractable—we clearly need additional criteria with which to judge whether $\mathbf{t}(\mathbf{X})$ generates "good" estimates of $\mathbf{q}(\Theta)$.

Various criteria for judging the usefulness of a given estimator $\mathbf{t}(\mathbf{X})$ for estimating $\mathbf{q}(\Theta)$ have been presented in the literature.[3] The measures evaluate and rank estimators in terms of closeness of estimates to $\mathbf{q}(\Theta)$ in an *expected* or *probabilistic* sense. Note that since $\mathbf{t}(\mathbf{X})$ is a function of $\mathbf{X}$, and thus a random variable, a sampling distribution (i.e., the probability distribution of $\mathbf{t}(\mathbf{X})$) is defined on $R(\mathbf{T})$ that is derived from the probability distribution of the random sample, $\mathbf{X} = [X_1, \ldots, X_n]'$. Roughly speaking, the fact that the distribution of $\mathbf{X}$ depends on $\Theta$ will generally result in the sampling distribution of $\mathbf{t}(\mathbf{X})$ depending on $\Theta$ as well, and this latter dependence in turn can lead to changes in location, spread, and/or shape of the distribution of $\mathbf{t}(\mathbf{X})$ as $\Theta$ changes. If the sampling distribution of $\mathbf{t}(\mathbf{X})$ changes with $\Theta$ in a way that keeps the spread of potential estimates generated by $\mathbf{t}(\mathbf{X})$ narrowly focused on $\mathbf{q}(\Theta)$ so that outcomes of $\mathbf{t}(\mathbf{X})$ occur near $\mathbf{q}(\Theta)$ with high probability under all contingencies for $\Theta \in \Omega$, (see Figure 7.2), then the function $\mathbf{T}$ would be useful for generating estimates of $\mathbf{q}(\Theta)$. We turn our attention to specific estimator properties that have been used to measure whether these objectives have been achieved.

Figure 7.3 presents an overview of the basic evaluation scheme most often adopted in econometric and statistical practice for ranking estimators of $\mathbf{q}(\Theta)$. We will examine the components of this evaluation scheme in detail. (Discussion of additional ranking criteria can be found in Lehmann.[4]) The paths of evaluation are differentiated on the basis of whether one is evaluating properties of the exact sampling distribution of the estimator for finite sample sizes (the *finite sample properties* path on the left in Figure 7.3) or whether one is dealing with approximations to sampling distribution properties based on asymptotic distributions and convergence in probability considerations (the *asymptotic properties* path). It is generally preferable to rank estimators based on properties of their exact sampling distributions, but this is not always possible. In some cases, the finite sample properties are intractable to analyze, or else they do not exist at all because appropriate population moments do not exist. In these cases, which arise often in advanced statistical and econometric practice,

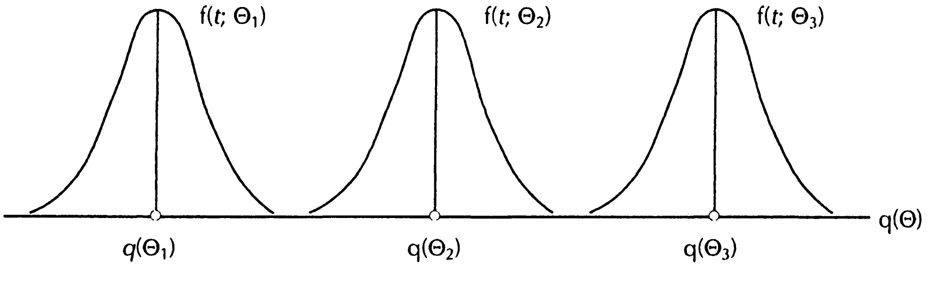


**Figure 7-2**
Estimator PDFs for various values of $\Theta$.

[3] A concise review and comparison of a number of alternative criteria is given by T. Amemiya (1994), *Introduction to Statistics and Econometrics*. Cambridge, MA: Harvard University Press, pp. 118–121.

[4] E. L. Lehmann (1983), *Theory of Point Estimation*. New York: John Wiley.

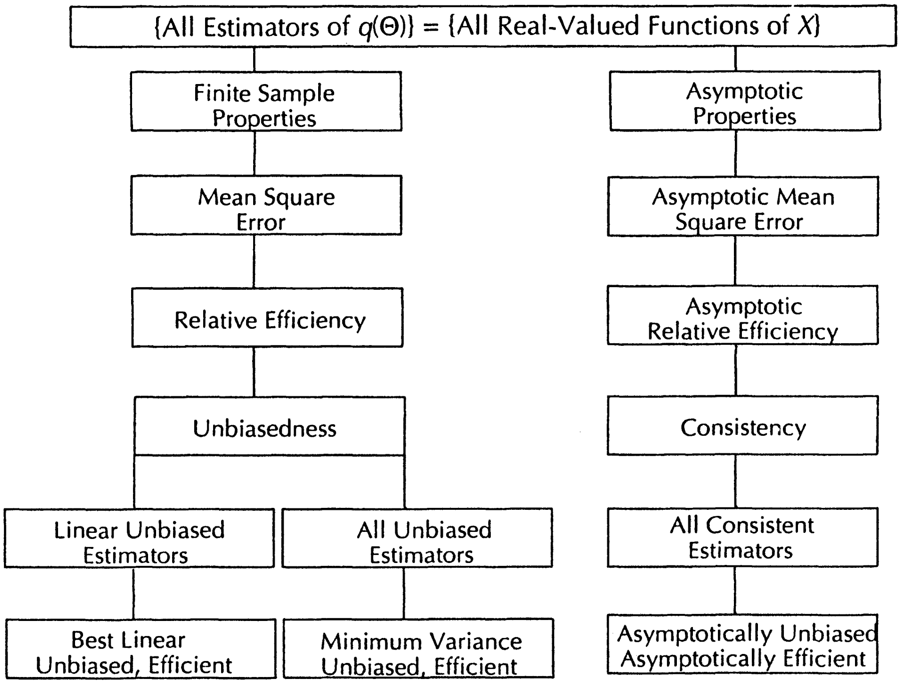

{All Estimators of $q(\Theta)$} = {All Real-Valued Functions of $X$}
Finite Sample Properties
Mean Square Error
Relative Efficiency
Unbiasedness
Linear Unbiased Estimators
All Unbiased Estimators
Best Linear Unbiased, Efficient
Minimum Variance Unbiased, Efficient
Asymptotic Properties
Asymptotic Mean Square Error
Asymptotic Relative Efficiency
Consistency
All Consistent Estimators
Asymptotically Unbiased Asymptotically Efficient

**Figure 7-3**
Evaluation scheme for estimators of $\mathbf{q}(\Theta)$.

the analyst must rely on asymptotic properties to guide his or her estimator choice. We will examine each of the evaluation paths in turn.

In discussing estimator properties, we will sometimes utilize a $\Theta$-subscript such as $\mathrm{E}_\Theta(\cdot)$, $P_\Theta(\cdot)$, or $\mathrm{var}_\Theta(\cdot)$ to emphasize that expectations or probabilities are being calculated using a particular value of $\Theta$ for the parameter vector of the underlying probability distribution. In cases where the parametric context of expectations and probabilities is clear or does not need to be distinguished, the subscript $\Theta$ will not be explicitly displayed.

### Finite Sample Properties

*Mean Square Error and Relative Efficiency* The term *mean square error* (MSE) is an alternative term for the *expected squared distance* between an estimator $\mathbf{T} = \mathbf{t}(\mathbf{X})$, and the estimand $\mathbf{q}(\Theta)$. When $T$ and $q(\Theta)$ are scalars, the following definition applies.

**Definition 7.5**
***Mean square error (scalar case)***

> The **mean square error** of an estimator $T$ of $q(\Theta)$ is defined as $\mathrm{MSE}_\Theta(T) = \mathrm{E}_\Theta d^2(T, q(\Theta)) = \mathrm{E}_\Theta(T - q(\Theta))^2 \forall\, \Theta \in \Omega$.

The MSE criterion accounts for both the degree of spread in the sampling distribution of $T$ as well as the degree to which the central tendency of $T$'s distribution deviates from $q(\Theta)$. We will make this notion precise upon defining the concept of bias, as follows.

**Definition 7.6**
***Estimator bias***

> The **bias** of an estimator $T$ of $q(\Theta)$ is defined as $\text{bias}_\Theta(T) = \mathrm{E}_\Theta(T - q(\Theta))$, $\forall\, \Theta \in \Omega$. The **bias vector** of an estimator $\mathbf{T}$ of $\mathbf{q}(\Theta)$ is defined as $\mathbf{Bias}_\Theta(\mathbf{T}) = \mathrm{E}_\Theta(\mathbf{T} - \mathbf{q}(\Theta))$, $\forall\, \Theta \in \Omega$.

The MSE of a scalar estimator $T$ can be decomposed into the sum of the variance of $T$ and the squared bias of $T$, as

$$\begin{aligned}\mathrm{MSE}_\Theta(T) &= \mathrm{E}_\Theta\,(T - \mathrm{E}_\Theta T + \mathrm{E}_\Theta T - q(\Theta))^2 = \mathrm{E}_\Theta(T - \mathrm{E}_\Theta T)^2 + [\mathrm{E}_\Theta T - q(\Theta)]^2\\ &= \mathrm{var}_\Theta(T) + [\mathrm{bias}_\Theta(T)]^2.\end{aligned}$$

The MSE criterion thus penalizes an estimator for having a high variance, a high bias, or both. It also follows that the MSE criterion allows a tradeoff between variance and bias in the ranking of estimators. In the final analysis, it is the expected squared distance between $T$ and $q(\Theta)$ implied by $\mathrm{var}_\Theta(T)$ and $\mathrm{bias}_\Theta(T)$, and not the variance and bias per se, that determines an estimator's relative ranking via MSE.

In the multivariate case, the MSE criterion is generalized through the use of the *mean square error matrix.*

**Definition 7.7**
***Mean square error matrix***

> The **mean square error matrix** of the estimator $\mathbf{T}$ of the $(k \times 1)$ vector $\mathbf{q}(\Theta)$ is defined by $\mathbf{MSE}_\Theta(\mathbf{T}) = \mathrm{E}_\Theta(\mathbf{T} - \mathbf{q}(\Theta))(\mathbf{T} - \mathbf{q}(\Theta))' \forall\, \Theta \in \Omega$.

To appreciate the information content of $\mathbf{MSE}_\Theta(\mathbf{T})$, first note that the diagonal of the MSE matrix contains the MSE of the estimator $T_i$ for $q_i(\Theta)$, $i = 1, \ldots, k$, since the typical diagonal entry in $\mathbf{MSE}_\Theta(\mathbf{T})$ is $\mathrm{E}_\Theta(T_i - q_i(\Theta))^2$. More generally, let $\mathbf{c}$ be any $(k \times 1)$ vector, and examine the MSE of $\mathbf{c}'\mathbf{T} = \sum_{i=1}^{k} c_i T_i$ as an estimator of $\mathbf{c}'\mathbf{q}(\Theta) = \sum_{i=1}^{k} \mathrm{c}_i q_i(\Theta)$,

$$MSE_\Theta(\mathbf{c}'\mathbf{T}) = \mathrm{E}_\Theta(\mathbf{c}'\mathbf{T} - \mathbf{c}'\mathbf{q}(\Theta))^2 = \mathrm{E}_\Theta \mathbf{c}'[\mathbf{T} - \mathbf{q}(\Theta)][\mathbf{T} - \mathbf{q}(\Theta)]'\mathbf{c} = \mathbf{c}'\mathbf{MSE}_\Theta(\mathbf{T})\mathbf{c}.$$

Thus, the MSEs of every possible linear combination of the $T_i$'s as estimators of the corresponding linear combination of the $q_i(\Theta)$'s can be obtained from the MSE matrix. Note further that the *trace* of the MSE matrix defines the *expected squared distance* of the *vector* estimator $\mathbf{T}$ from the *vector* estimand $\mathbf{q}(\Theta)$, as

$$\mathrm{tr}\mathbf{MSE}_\Theta(\mathbf{T}) = \mathrm{tr}\mathrm{E}_\Theta[\mathbf{T} - \mathbf{q}(\Theta)][\mathbf{T} - \mathbf{q}(\Theta)]' = \mathrm{E}[\mathbf{T} - \mathbf{q}(\Theta)]'[\mathbf{T} - \mathbf{q}(\Theta)] = \mathrm{E}_\Theta d^2(\mathbf{T}, \mathbf{q}(\Theta)).$$

This is the direct vector analogue to the measure of the closeness of $T$ to $q(\Theta)$ that is provided by the MSE criterion in the scalar case.

The MSE matrix can be decomposed into variance and bias components, analogous to the scalar case. Specifically, $\mathbf{MSE}(\mathbf{T})$ is equal to the sum of the covariance matrix of $\mathbf{T}$ and the outer product of the bias vector of $\mathbf{T}$ as

$$\begin{aligned}\mathbf{MSE}(\mathbf{T}) &= \mathrm{E}[\mathbf{T} - \mathrm{E}\mathbf{T} + \mathrm{E}\mathbf{T} - \mathbf{q}(\Theta)][\mathbf{T} - \mathrm{E}\mathbf{T} + \mathrm{E}\mathbf{T} - \mathbf{q}(\Theta)]'\\ &= \mathbf{Cov}(\mathbf{T}) + \mathbf{Bias}(\mathbf{T})\mathbf{Bias}(\mathbf{T})'.\end{aligned}$$

The outer product of the bias vector forms a $(k \times k)$ matrix that is called the **bias matrix**.

In the case of a scalar $q(\Theta)$, estimators with smaller MSEs are preferred. Note, however, that since the true $\Theta$ is unknown (or else there is no point estimation problem to begin with), one must consider the performance of an estimator for all possible contingencies for the true value of $\Theta$, which is to say, for all $\Theta \in \Omega$. It is quite possible, and often the case, that an estimator will have lower MSEs than another estimator for some values of $\Theta \in \Omega$ but not for others. These considerations lead to the concepts of *relative efficiency* and *relatively more efficient.*

**Definition 7.8**
***Relative efficiency (scalar case)***

Let $T$ and $T^*$ be two estimators of a scalar $q(\Theta)$. The **relative efficiency** of $T$ with respect to $T^*$ is given by

$$\mathrm{RE}_\Theta(T, T^*) = \frac{\mathrm{MSE}_\Theta(T^*)}{\mathrm{MSE}_\Theta(T)} = \frac{\mathrm{E}_\Theta(T^* - q(\Theta))^2}{\mathrm{E}_\Theta(T - q(\Theta))^2}, \quad \forall \Theta \in \Omega.$$

$T$ is **relatively more efficient** than $T^*$ if $\mathrm{RE}_\Theta(T, T^*) \geq 1 \forall \Theta \in \Omega$ and $> 1$ for some $\Theta \in \Omega$.

In comparing two estimators of $q(\Theta)$, if $T$ is *relatively more efficient* than $T^*$, then there is no value of $\Theta$ for which $T^*$ is preferred to $T$ on the basis of MSE, and for one or more values of $\Theta$, $T$ is preferred to $T^*$. In this case, it is evident that $T^*$ can be discarded as an estimator of $q(\Theta)$, and thus $T^*$ is considered an *inadmissible* estimator of $q(\Theta)$.

**Definition 7.9**
***Estimator admissibility***

Let $T$ be an estimator of $q(\Theta)$. If there exists another estimator of $q(\Theta)$ that is *relatively more efficient* than $T$, then $T$ is called **inadmissible** for estimating $q(\Theta)$. Otherwise, $T$ is called **admissible**.

It is evident that in the search for good estimators of $q(\Theta)$ in terms of MSE, the analyst need not consider any estimators that are *inadmissible.*

**Example 7.2** Suppose $(X_1, \ldots, X_n)$ is a random sample from a Bernoulli population distribution, where $x_i$ represents whether $(x_i = 1)$ or not $(x_i = 0)$ the $i$th customer contacted by telephone solicitation purchases a product. Consider two estimators for the unknown proportion, $p$, of the consumer population who will purchase the product:

$$T = \bar{X} = n^{-1}\sum_{i=1}^{n} X_i \quad \text{and} \quad T^* = (n+1)^{-1}\sum_{i=1}^{n} X_i = \left(\frac{n}{n+1}\right)\bar{X}.$$

Which estimator, if either, is the preferred estimator of $p$ on the basis of MSE given that $n = 25$? Does either estimator render the other inadmissible?

**Answer**: Note that $\mathrm{bias}(T) = \mathrm{E}\bar{X} - p = 0$, $\mathrm{bias}(T^*) = \mathrm{E}((n/(n+1))\bar{X}) - p = -p/(n+1) = -p/26$, $\mathrm{var}(T) = p(1-p)/n = p(1-p)/25$, and $\mathrm{var}(T^*) = np(1-p)/(n+1)^2 =$

$p(1-p)/27.04$. Then the MSEs of the two estimators are given by

$$\mathrm{MSE}(T) = \frac{p(1-p)}{25}$$

and

$$\mathrm{MSE}(T^*) = \frac{p(1-p)}{27.04} + \frac{p^2}{676}.$$

Examine the MSE of $T^*$ relative to the MSE of $T$, as

$$\mathrm{RE}_p(T, T^*) = \frac{\mathrm{MSE}_p(T^*)}{\mathrm{MSE}_p(T)} = .9246 + .0370p/(1-p).$$

Since the ratio depends on the value of $p$, which is unknown, we must consider all of the possible contingencies for $p \in [0, 1]$. Note that the ratio is monotonically increasing in $p$, taking its smallest value of .9246 when $p = 0$, and diverging to infinity as $p \to 1$. The ratio of MSEs equals 1 when $p = .6708$. Thus, without constraints on the potential values of $p$, neither estimator is preferred to the other on the basis of MSE, and thus neither estimator is rendered inadmissible by the other. □

In contrast to the scalar case, a myriad of different MSE comparisons is possible when $\mathbf{q}(\Theta)$ is a $(k \times 1)$ vector. First of all, there are $k$ individual MSE comparisons that can be made between corresponding entries in the two estimators $\mathbf{T}^*$ and $\mathbf{T}$. One could also compare the expected squared distances of $\mathbf{T}^*$ and $\mathbf{T}$ from $\mathbf{q}(\Theta)$, which is equivalent to comparing the sums of the mean-square errors of the entries in $\mathbf{T}^*$ and $\mathbf{T}$. Furthermore, one could contemplate estimating linear combinations of the entries in $\mathbf{q}(\Theta)$ via corresponding linear combinations of the entries in $\mathbf{T}^*$ and $\mathbf{T}$, so that MSE comparisons between the estimators $\ell'\mathbf{T}^*$ and $\ell'\mathbf{T}$ for $\ell'\mathbf{q}(\Theta)$ are then of interest. *All* of the preceding MSE comparisons are accounted for simultaneously in the following *strong mean square error* (SMSE) criterion.

**Definition 7.10**
***Strong mean square error superiority***

Let $\mathbf{T}^*$ and $\mathbf{T}$ be two estimators of the $(k \times 1)$ vector $\mathbf{q}(\Theta)$. $\mathbf{T}^*$ is **strong mean square error superior** to $\mathbf{T}$ iff $\mathbf{MSE}_\Theta(\mathbf{T}^*) - \mathbf{MSE}_\Theta(\mathbf{T})$ is negative semidefinite $\forall\, \Theta \in \Omega$ and unequal to the zero matrix for some $\Theta \in \Omega$.

If $\mathbf{T}^*$ is SMSE superior to $\mathbf{T}$, it follows directly from Def. 7.10 that $MSE_\Theta(T_i^*) \leq MSE_\Theta(T_i)\ \forall i$ and $\forall\, \Theta \in \Omega$ since if $\mathbf{MSE}_\Theta(\mathbf{T}^*) - \mathbf{MSE}_\Theta(\mathbf{T})$ is negative semidefinite, the matrix difference necessarily has nonpositive diagonal entries.[5] It follows subsequently that

$$\mathrm{E}_\Theta d^2(\mathbf{T}^*, \mathbf{q}(\Theta)) = \sum_{i=1}^{k} MSE_\Theta(T_i^*) \leq \sum_{i=1}^{k} MSE_\Theta(T_i) = \mathrm{E}_\Theta d^2(\mathbf{T}, \mathbf{q}(\Theta)) \quad \forall\, \Theta \in \Omega.$$

Furthermore, in terms of estimating $\boldsymbol{\ell}'\mathbf{q}(\Theta)$,

$$MSE_\Theta(\boldsymbol{\ell}'T^*) = \boldsymbol{\ell}'\mathbf{MSE}_\Theta(\mathbf{T}^*)\boldsymbol{\ell} \leq \boldsymbol{\ell}'\mathbf{MSE}_\Theta(\mathbf{T})\boldsymbol{\ell} = MSE_\Theta(\boldsymbol{\ell}'\mathrm{T}) \quad \forall\, \Theta \in \Omega \text{ and } \forall\, \boldsymbol{\ell}.$$

Thus in the sense of *all* of the MSE comparisons mentioned previously, $\mathbf{T}^*$ is *at least as good as* $\mathbf{T}$.

The fact that $\mathbf{MSE}_\Theta(\mathbf{T}^*) - \mathbf{MSE}_\Theta(\mathbf{T})$ is negative semidefinite *and* unequal to the zero matrix for some $\Theta \in \Omega$ implies that some of the weak inequalities ($\leq$) in the aforementioned MSE comparisons become strong inequalities ($<$) for some $\Theta$. To see this, note that a nonzero, negative semidefinite, symmetric matrix necessarily has one or more *negative* diagonal entries.[6] Therefore, $MSE_\Theta(T_i^*) < MSE_\Theta(T_i)$ for some $\Theta$ and $i$, so that $\mathrm{E}_\Theta d^2(\mathbf{T}^*, \mathbf{q}(\Theta)) < \mathrm{E}_\Theta d^2(\mathbf{T}, \mathbf{q}(\Theta))$ for some $\Theta$ and $MSE_\Theta(\boldsymbol{\ell}'\mathrm{T}^*) < MSE_\Theta(\boldsymbol{\ell}'T)$ for some $\Theta$ and $\boldsymbol{\ell}$. Thus, $\mathbf{T}^*$ is superior to $\mathbf{T}$ for at least some MSE comparisons in addition to being no worse for any of the MSE comparisons. We now have the following multivariate analogues to the notions of relative efficiency and admissibility.

**Definition 7.11**
***Relative efficiency and admissibility with respect to SMSE***

Let $\mathbf{T}^*$ and $\mathbf{T}$ be estimators of the $(k \times 1)$ vector $\mathbf{q}(\Theta)$. If $\mathbf{T}^*$ is SMSE superior to $\mathbf{T}$, then $\mathbf{T}^*$ is said to be **relatively more efficient** than $\mathbf{T}$. If there exists an estimator that is relatively more efficient than $\mathbf{T}$, then $\mathbf{T}$ is said to be **inadmissible.** Otherwise, $\mathbf{T}$ is said to be **admissible.**

As in the scalar case, the analyst need not consider any estimators of $\mathbf{q}(\Theta)$ that are inadmissible when searching for good estimators of $\mathbf{q}(\Theta)$ in terms of MSE.[7]

In either the scalar or multivariate case, a natural question to ask is whether an optimal estimator exists that has the smallest MSE or MSE matrix among all estimators of $\mathbf{q}(\Theta)$. We might call such an estimator most efficient, or simply efficient. In general, no such estimator exists.To clarify the problem involved, consider the scalar case, and note that the degenerate estimator $T^* = t^*(X) = \Theta_o$ would certainly have minimum mean-square error for estimating $\Theta$ if mean-square error were evaluated *at the point* $\Theta = \Theta_o$, i.e., $\mathrm{MSE}_\Theta(T^*) = 0$ for $\Theta = \Theta_o$.

[5] By definition, $\mathbf{A}$ is negative semidefinite iff $\boldsymbol{\ell}'\mathbf{A}\boldsymbol{\ell} \leq 0 \forall \boldsymbol{\ell}$. Then the $i$th diagonal entry of $\mathbf{A}$ must be $\leq 0$ since this entry can be defined by $\boldsymbol{\ell}'\mathbf{A}\boldsymbol{\ell}$ with $\boldsymbol{\ell}$ being a zero vector except for a 1 in the $i$th position.

[6] A nonzero matrix has at least unit rank. The rank of a negative semidefinite symmetric matrix is equal to the number of negatively valued eigenvalues, and all eigenvalues of a negative semidefinite matrix $\leq 0$. The trace of a negative semidefinite symmetric matrix is equal to the sum of its eigenvalues. Since all diagonal entries in a negative semidefinite matrix are $\leq 0$, it follows that a nonzero negative semidefinite symmetric matrix must have one or more negative diagonal entries.

[7] Some analysts use a *weak mean-square error* (WMSE) criterion that relates only to expected squared distance considerations. $\mathbf{T}^*$ is WMSE superior to $\mathbf{T}$ iff $\mathrm{E}_\Theta d^2(\mathrm{T}^*, \mathrm{q}(\Theta)) \leq \mathrm{E}_\Theta d^2(\mathrm{T}, \mathrm{q}(\Theta)) \forall \Theta \in \Omega$, and $<$ for some $\Theta \in \Omega$. Relative efficiency and admissibility can be defined in the context of WMSE superiority and are left to the reader.

Since a similar degenerate estimator could be defined for each $\Theta \in \Omega$, then for a given estimator to have minimum mean-square error *for every potential value of* $\Theta$ (i.e., uniformly in $\Theta$), it would be necessary that $\text{MSE}_\Theta(T) = 0 \; \forall \; \Theta \in \Omega$, which would imply that $\text{var}_\Theta(T) = 0 \; \forall \; \Theta \in \Omega$, and thus that $P_\Theta(t(x) = \Theta) = 1 \; \forall \; \Theta \in \Omega$. In order to construct an estimator $T$ that satisfied the condition $P_\Theta(t(x) = \Theta) = 1 \; \forall \; \Theta \in \Omega$, it would be necessary to be able to identify the true value of $\Theta$ directly upon observing the sample outcome, $x$. This essentially requires that the range of the random sample be dependent on the value of $\Theta$, denoted as $R_\Theta(X)$, in such a way that the sets $R_\Theta(X)$, $\Theta \in \Omega$, are all mutually exclusive, i.e., $R_{\Theta'}(X) \cap R_{\Theta''}(X) = \emptyset$ for $\Theta' \neq \Theta''$. Then, upon observing $x$, one would only need to identify the set $R_\Theta(X)$ to which $x$ belonged, and $\Theta$ would be immediately known. This is rarely, if ever, possible in practice, and so adopting a *minimum mean-square error criterion* for choosing an estimator of $q(\Theta)$ is not feasible. A similar argument leads to the conclusion that there is in general no estimator of a $(k \times 1)$ vector $\mathbf{q}(\mathbf{\Theta})$ whose MSE matrix is smallest among the MSE matrices of all estimators of $\mathbf{q}(\mathbf{\Theta})$.[8]

While there generally does not exist an estimator that has a uniformly (i.e., for all $\mathbf{\Theta} \in \Omega$) minimum MSE or MSE matrix relative to *all* other estimators of $\mathbf{q}(\mathbf{\Theta})$, it is often possible to find an optimal estimator if one restricts the type of estimators under consideration. Two such restrictions that have been widely used in practice are *unbiasedness* and *linearity*, which we examine in the next two subsections.

*Unbiasedness* The property of unbiasedness refers to the balancing point or expectation of an estimator's probability distribution being equal to the estimand.

**Definition 7.12**
***Unbiased estimator***

An estimator $\mathbf{T}$ is said to be an **unbiased estimator** of $\mathbf{q}(\mathbf{\Theta})$ iff $\text{E}_\mathbf{\Theta}\mathbf{T} = \mathbf{q}(\mathbf{\Theta}) \; \forall \, \mathbf{\Theta} \in \Omega$. Otherwise, the estimation is said to be **biased.**

As in the case of MSE criteria, it is important to appreciate the significance of the condition $\forall \, \mathbf{\Theta} \in \Omega$ in the above definition. In the context of the point estimation problem, we have assumed that the true value of $\mathbf{\Theta}$, say $\mathbf{\Theta}_*$, is some element of the specified parameter space, $\Omega$, but we do not know which one. Thus, the property of unbiasedness is stated for *all possible contingencies* regarding the potential values for the true value of $\mathbf{\Theta}$. Due to the condition $\forall \, \mathbf{\Theta} \in \Omega$, the requirement for unbiasedness essentially means that $\text{E}\mathbf{T} = \mathbf{q}(\mathbf{\Theta})$ *regardless* of which value of $\mathbf{\Theta} \in \Omega$ is the true value. Thus, for $\mathbf{T}$ to be unbiased, its density function must be balanced on the point $\mathbf{q}(\mathbf{\Theta})$ *no matter what the true value of* $\mathbf{\Theta}$ is. Evidently, the density function of $\mathbf{T}$ must be related to $\mathbf{\Theta}$ in a special way for unbiasedness to be obtained. Whether or not $\mathbf{T}$ has the unbiasedness property depends on the functional definition of $\mathbf{T}$ and, in particular,

[8] By "smallest MSE matrix," we mean that $\text{MSE}_\mathbf{\Theta}(\mathbf{T}) - \text{MSE}_\mathbf{\Theta}(\mathbf{T}^*)$ is a *negative semidefinite matrix* for all estimators $\mathbf{T}^*$ of $\mathbf{q}(\mathbf{\Theta})$ and for all $\mathbf{\Theta}$.

on how the function translates the density function of $\mathbf{X} \sim f(x_1, \ldots, x_n; \boldsymbol{\Theta})$ into the density function of $\mathbf{T} \sim f(\mathbf{t}; \boldsymbol{\Theta})$.

An unbiased estimator has the intuitively appealing property of being equal to $\mathbf{q}(\boldsymbol{\Theta})$ *on the average*, the phrase having two useful interpretations. First, since the expectation operation is itself a weighted average of the outcomes of $\mathbf{T}$, then the outcomes of $\mathbf{T}$ have a weighted average equal to $\mathbf{q}(\boldsymbol{\Theta})$. Alternatively, if we were to repeatedly and independently observe outcomes of the random sample $\mathbf{X}$, and thus repeatedly generate estimates of $\mathbf{q}(\boldsymbol{\Theta})$ using corresponding outcomes of the vector $\mathbf{T}$, then the simple average of all of the observed estimates would converge in probability (and, in fact, converge almost surely) elementwise to $\mathbf{q}(\boldsymbol{\Theta})$ by Khinchin's WLLN (or by Kolmogorov's SLLN in the case of almost-sure convergence), provided only that $\mathbf{q}(\boldsymbol{\Theta})$ is finite.

We provide the following example of an unbiased estimator of a parameter.

**Example 7.3** Let the population distribution of hard-disk operating lives for a certain brand of hard disk be a member of the exponential family of densities $f(z; \theta) = (1/\theta)e^{-z/\theta}I_{(0,\infty)}(z)$, $\theta \in \Omega = (0, \infty)$. Let $T = n^{-1}\sum_{i=1}^{n} X_i$ be an estimator of the expected operating life of the hard disk, $\theta$, where $(X_1, \ldots, X_n)$ is a random sample from the exponential population distribution. Then $T$ is an unbiased estimator of $\theta$, since $\mathrm{E}X_i = \theta \; \forall \, i$, which implies that $\mathrm{E}T = \mathrm{E}n^{-1}\sum_{i=1}^{n} X_i = n^{-1}\sum_{i=1}^{n} \mathrm{E}X_i = \theta$ *regardless* of the value of $\theta > 0$. Thus, for example, if the true value of $\theta$ were 2, then $\mathrm{E}T = 2$, or if the true value of $\theta$ were 100, then $\mathrm{E}T = 100$. □

*MVUE, BLUE, and Efficiency* The unbiasedness criterion ensures only that an estimator will have a density that has a central tendency or balancing point of $\mathbf{q}(\boldsymbol{\Theta})$. However, it is intuitively clear that we also desire that the density not be too spread out around this balancing point for fear that an estimate could be generated that was a significant distance from $\mathbf{q}(\boldsymbol{\Theta})$ with high probability. Graphically, we would prefer the estimator $T$ to the estimator $T^*$ in Figure 7.4, where both of these estimators are unbiased estimators of $q(\Theta)$.

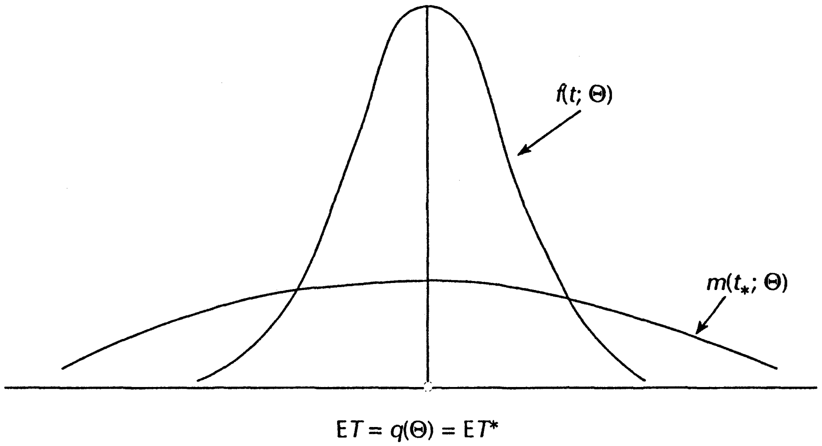


**Figure 7-4**
The densities of two unbiased estimators of $q(\Theta)$

The foregoing considerations motivate the objective that an estimator have minimum variance, or minimum covariance matrix if $T$ is a vector, among all unbiased estimators of $\mathbf{q}(\Theta)$. Since $\mathbf{Bias}_\Theta(\mathbf{T}) = [0]$ for all estimators in the unbiased class of estimators, $\text{MSE}_\Theta(T) = \text{var}_\Theta(T)$ or $\mathbf{MSE}_\Theta(\mathbf{T}) = \mathbf{Cov}_\Theta(\mathbf{T})$, and we can thus view the objective of minimizing $\text{var}(T)$ or $\mathbf{Cov}(\mathbf{T})$ equivalently as searching for the estimator with the smallest MSE or smallest MSE matrix *within the class of unbiased estimators.* In the following definition, we introduce the notation $\mathbf{A} \preceq \mathbf{B}$ to indicate that **matrix A is smaller than matrix B** by a negative semidefinite matrix, i.e., $\mathbf{A} - \mathbf{B} =$ negative semidefinite matrix.

**Definition 7.13** ***Minimum variance unbiased estimator (MVUE)***

An estimator $\mathbf{T}$ is said to be a minimum-variance unbiased estimator of $\mathbf{q}(\Theta)$ iff $\mathbf{T}$ is an unbiased estimator of $\mathbf{q}(\Theta)$, and

**a.** (scalar case) $\text{var}_\Theta(T) \leq \text{var}_\Theta(T^*)\ \forall\, \Theta \in \Omega$ and $\forall\, T^* \in \text{U}_{q(\Theta)}$;

**b.** (vector case) $\mathbf{Cov}_\Theta(T) \preceq \mathbf{Cov}_\Theta(\mathbf{T}^*)\ \forall\, \Theta \in \Omega$ and $\forall\, \mathbf{T}^* \in \text{U}_{\mathbf{q}(\Theta)}$;

where $\text{U}_{\mathbf{q}(\Theta)}$ is the set of all unbiased estimators of $\mathbf{q}(\Theta)$.[9]

Definition 7.13 implies that an estimator is a MVUE if the estimator is unbiased and if there is no other unbiased estimator that has a smaller variance or covariance matrix for any $\Theta \in \Omega$. Drawing direct analogies to the discussion of MSE criteria, a MVUE, $\mathbf{T}$, is such that $\text{MSE}_\Theta(T_i) = \text{var}_\Theta(T_i) \leq \text{var}_\Theta(T_i^*) = \text{MSE}_\Theta(T_i^*)\ \forall\, \Theta \in \Omega$ and $\forall\, i$, where $\mathbf{T}^*$ is any estimator *in the unbiased class of estimators*. Furthermore, $\text{E}_\Theta d^2(\mathbf{T}, \mathbf{q}(\Theta)) \leq \text{E}_\Theta d^2(\mathbf{T}^*, \mathbf{q}(\Theta))\ \forall\, \Theta \in \Omega$ and $\text{MSE}_\Theta(\boldsymbol{\ell}'\mathbf{T}) = \text{var}_\Theta(\boldsymbol{\ell}'\mathbf{T}) \leq \text{var}_\Theta(\boldsymbol{\ell}'\mathbf{T}^*) = \text{MSE}_\Theta(\boldsymbol{\ell}'\mathbf{T}^*)\ \forall\, \Theta \in \Omega$ and $\forall\, \boldsymbol{\ell}$. Thus, within the class of unbiased estimators, a MVUE of $\mathbf{q}(\Theta)$ is at least as good as any other estimator of $\mathbf{q}(\Theta)$ in terms of all of the types of MSE comparisons that we have discussed previously. If $\mathbf{T}$ is a MVUE for $\mathbf{q}(\Theta)$, then $\mathbf{T}$ is said to be **efficient**.

Unfortunately, without the aid of theorems that facilitate the discovery of MVUES, finding a MVUE of $\mathbf{q}(\Theta)$ can be quite challenging even when the point estimation problem appears to be quite simple. Study the following example, which illustrates the general concepts involved.

**Example 7.4** Find a MVUE of the parameter p using a random sample of size 2 from the Bernoulli population distribution $f(z; p) = p^z(1-p)^{1-z}I_{\{0,1\}}(z)$.

**Answer:** First, the range of the random sample is $\{(0,0), (0,1), (1,0), (1,1)\}$, which represents the domain of the estimator function $T = t(\mathbf{X})$. For $t(\mathbf{X})$ to be in the unbiased class, we must have that

$$\text{E}t(\mathbf{X}) = t(0,0)(1-p)^2 + t(0,1)(1-p)p + t(1,0)p(1-p) + t(1,1)p^2 = p \quad \forall p \in [0,1].$$

[9] This is alternatively referred to in the literature by the term uniformly minimum variance unbiased estimator (UMVUE), where the adverb "uniformly" is used to emphasize the condition "$\forall\, \Theta \in \Omega$." In our usage of the terms, MVUE and UMVUE will be interchangeable.

If $p = 1$, then

$$\mathrm{E}t(\mathbf{X}) = t(1,1)p^2 = p \quad \text{iff } t(1,1) = 1.$$

If $p = 0$, then

$$\mathrm{E}t(\mathbf{X}) = t(0,0)(1-p)^2 = p \quad \text{iff } t(0,0) = 0.$$

If $p \in (0,1)$, and using $t(1,1) = 1$ and $t(0,0) = 0$,

$$\mathrm{E}t(\mathbf{X}) = t(0,1)(1-p)p + t(1,0)p(1-p) + p^2 = p,$$

which implies that $t(0,1)(1-p)+t(1,0)(1-p) = 1-p$, so that $t(0,1)+t(1,0) = 1$. Now consider the variance of $t(\mathbf{X})$, which by definition can be written as

$$\begin{aligned}\mathrm{var}(t(\mathbf{X})) &= (t(0,0)-p)^2(1-p)^2 + (t(0,1)-p)^2(1-p)p \\ &\quad + (t(1,0)-p)^2p(1-p) + (t(1,1)-p)^2p^2 \\ &= 2p^2(1-p)^2 + (t(0,1)-p)^2(1-p)p + (t(1,0)-p)^2p(1-p),\end{aligned}$$

where we have used $\mathrm{E}t(\mathbf{X}) = p$, $t(0,0) = 0$, and $t(1,1) = 1$ since $t(\mathbf{X})$ is to be unbiased. Also, because of the unbiasedness condition, we can substitute $t(0,1) = 1 - t(1,0)$ into the variance expression to obtain

$$\mathrm{var}(t(\mathbf{X})) = 2p^2(1-p)^2 + (1-p-t(1,0))^2(1-p)p + (t(1,0)-p)^2p(1-p).$$

The first-order condition for a minimum of the variance is given by

$$\frac{d\,\mathrm{var}\,t(\mathbf{X})}{dt(1,0)} = -2(1-p)^2p + 2t(1,0)p(1-p) + 2t(1,0)p(1-p) - 2p^2(1-p) = 0,$$

which implies that $4p(1-p)t(1,0) = 2(1-p)^2p + 2p^2(1-p)$, so that $t(1,0) = (1/2)$, which then implies that $t(0,1) = (1/2)$.

We have thus defined the function $T = t(\mathbf{X})$ that represents a MVUE of $p$ by associating an appropriate outcome of $T$ with each random sample outcome. The preceding results can be represented collectively as $t(x_1, x_2) = (1/2)(x_1 + x_2) = \bar{x}$. □

A number of general theorems that can often be used to simplify the search for a MVUE will be presented in Section 7.5.

For purposes of simplicity and tractability, as well as for cases where little can be assumed about the statistical model other than conditions on low-order moments, attention is sometimes restricted to estimators that are unbiased and that have minimum variance or covariance matrix among all unbiased estimators that are *linear* functions of the sample outcome. Such an estimator is called a BLUE or MVLUE, as indicated in the following definition.

**Definition 7.14**
***Best linear unbiased estimator (BLUE) or minimum-variance linear unbiased estimator (MVLUE)***

An estimator $\mathbf{T}$ is said to be a BLUE or MVLUE of $\mathbf{q}(\Theta)$ iff

1. $\mathbf{T}$ is a linear (vector) function, $\mathbf{T} = \mathbf{t}(\mathbf{X}) = \mathbf{AX} + \mathbf{b}$, of the random sample $\mathbf{X} = [X_1, \ldots, X_n]'$,
2. $\mathrm{E}_\Theta \mathbf{T} = \mathbf{q}(\Theta) \; \forall \, \Theta \in \Omega$ ($\mathbf{T}$ is unbiased),
3. $\mathbf{T}$ has minimum variance or covariance matrix among all unbiased estimators that are also linear functions of the random sample $\mathbf{X} \; \forall \, \Theta \in \Omega$.

A BLUE estimator of $\mathbf{q}(\Theta)$ is also referred to as an **efficient** estimator *within the class of linear unbiased estimators*. The following is an example identifying the BLUE or MVLUE of the mean of *any* population distribution with a finite mean and variance.

**Example 7.5** BLUE of Population Mean

Let $(X_1, \ldots, X_n)$ be a random sample from some population distribution $f(z; \Theta)$ of consumer incomes having a finite mean $\mu = q_1(\Theta)$ and variance $\sigma^2 = q_2(\Theta)$. What is the BLUE of the mean of the population distribution of consumer incomes?

**Answer**: We are examining linear estimators, and thus $t(\mathbf{X}) = \sum_{i=1}^n a_i X_i + b$. For $T$ to be unbiased, we require that $\sum_{i=1}^n a_i = 1$ and $b = 0$, since $\mathrm{E}T = \mathrm{E}[\sum_{i=1}^n a_i X_i + b] = [\sum_{i=1}^n a_i \mathrm{E} X_i] + b = \mu[\sum_{i=1}^n a_i] + b = \mu$, which holds for all potential $\mu$ iff $\sum_{i=1}^n a_i = 1$ and $b = 0$. The variance of $T$ is simply $\sigma^2 \sum_{i=1}^n a_i^2$ since $(X_1, \ldots, X_n)$ is a random sample from $f(z; \Theta)$. Thus, to find the BLUE, we must solve the following minimization problem:

$$\min_{a_1,\ldots,a_n} \sigma^2 \sum_{i=1}^n a_i^2 \qquad \text{such that} \sum_{i=1}^n a_i = 1.$$

The Lagrangian form of this minimization problem is given by

$$L = \sigma^2 \sum_{i=1}^n a_i^2 - \lambda \left[ \sum_{i=1}^n a_i - 1 \right],$$

and the first-order conditions are

$$\frac{\partial L}{\partial a_i} = 2\sigma^2 a_i - \lambda = 0, \quad i = 1, \ldots, n,$$

and

$$\frac{\partial L}{\partial \lambda} = 1 - \sum_{i=1}^n a_i = 0.$$

The first $n$ conditions imply $a_1 = a_2 = \ldots = a_n$, since $a_i = \lambda/2\sigma^2 \; \forall i$, and then $\sum_{i=1}^n a_i = 1$ requires that $a_i = 1/n$, $i = 1, \ldots, n$. Thus, $t(\mathbf{X}) = \sum_{i=1}^n a_i X_i + b = n^{-1} \sum_{i=1}^n X_i = \bar{X}$, so that the sample mean is the BLUE (or MVLUE) of the mean of a population distribution having a finite mean and variance. (The reader should check that the second-order conditions for a minimum are in fact met.) □

In addition to estimating the means of population distributions, a prominent BLUE arises in the context of least-squares estimation of the parameters of a general linear model, which we will examine in Chapter 8.

## Asymptotic Properties

When finite sample properties are intractable or else inapplicable due to the nonexistence of the appropriate expectations, the analyst may have to rely on asymptotic properties to rank the efficiency of estimators. In addition, asymptotic properties may be of more fundamental interest if the analyst is interested in assessing the effects on estimator properties of an ever-increasing number of sample observations.

Asymptotic properties of estimators are essentially equivalent in concept to the finite sample properties presented heretofore, except that asymptotic properties are based on the *asymptotic distributions of estimators* rather than estimators' exact finite sampling distributions. In particular, asymptotic analogues to MSE, relative efficiency, unbiasedness, and minimum-variance unbiasedness can be defined with reference to asymptotic distributions of estimators. However, a problem of nonuniqueness of asymptotic properties can arise because of the inherent nonuniqueness of asymptotic distributions.

To clarify the difficulties that can arise when using asymptotic distributions as a basis for asymptotic properties, let $T_n$ denote an estimator of the scalar $q(\Theta)$ based on $n$ sample observations, and suppose $b_n^{-1}(T_n - q(\Theta)) \xrightarrow{d} N(0, 1)$. Then one might consider defining asymptotic properties of $T_n$ in terms of the asymptotic distribution $N(q(\Theta), b_n^2)$. However, by Slutsky's theorem it follows that $(n/(n-k))^{1/2} b_n^{-1}(T_n - q(\Theta)) \xrightarrow{d} N(0, 1)$ for a fixed value of $k$ since $(n/(n-k))^{1/2} \to 1$, so that an alternative asymptotic distribution could be $T_n \overset{a}{\sim} N(q(\Theta), ((n-k)/n)b_n^2)$, which would lead to different asymptotic properties. The difficulty is that the centering and scaling required to achieve a limiting distribution is not unique, leading to both nonunique asymptotic distributions and nonunique asymptotic properties derived from them.

There are two basic ways of addressing the aforementioned nonuniqueness problem when dealing with asymptotic properties. One approach, which we will mention only briefly, is to rank estimators on the basis of limits of asymptotic property comparisons so as to expunge any arbitrary scaling or centering from the comparison. For example, referring to the previous illustration of nonuniqueness, let the asymptotic distribution of $T_n$ be $N(q(\Theta), b_n^2)$, and let $T_n^*$ have the asymptotic distribution $N(q(\Theta), ((n-k)/n)b_n^2)$. **Asymptotic MSEs** based on these asymptotic distributions can be calculated as $\text{AMSE}_\Theta(T_n) = \text{E}_\Theta(T_n - q(\Theta))^2 = b_n^2$ and $\text{AMSE}_\Theta(T_n^*) = \text{E}_\Theta(T_n^* - q(\Theta))^2 = ((n-k)/n)b_n^2$. Then the **asymptotic relative efficiency** (ARE) of $T_n$ with respect to $T_n^*$ could be represented as

$$\text{ARE}_\Theta(T_n, T_n^*) = \frac{\text{AMSE}_\Theta(T_n^*)}{\text{AMSE}_\Theta(T_n)} = \frac{n-k}{n}.$$

Using the ARE, one would be led to the conclusion that $T_n^*$ is **asymptotically relatively more efficient** than $T_n$, which in the context of the previous illustration of nonuniqueness would be absurd since $T_n$ and $T_n^*$ are the same estimator. However, $\lim_{n\to\infty}$ ARE $= 1$, which would lead to the conclusion that $T_n$ and $T_n^*$ are equally preferable on the basis of asymptotic MSE considerations, which of course is the correct conclusion in the current context. The limit operation removes the arbitrary scaling of the asymptotic variance. To operationalize this approach for general applications requires extensions of limit notions (limit superiors, or lim sups) which we will leave for future study (see L. Schmetterer (1974), *Introduction to Mathematical Statistics*. New York: Springer-Verlag, pp. 335–342).

An alternative approach for avoiding nonuniqueness of asymptotic properties is to restrict the use of asymptotic properties to classes of estimators for which the problem will not occur. For our purposes, it will suffice to examine the consistent asymptotically normal (CAN) class of estimators (for other possibilities, see E. Lehmann, *Point Estimation*, pp. 347–348).

Prior to identifying the CAN class of estimators, we examine the property of consistency.

### *Consistency*

**Definition 7.15**
***Consistent estimator***

$\mathbf{T}_n$ is said to be a **consistent estimator** of $\mathbf{q}(\Theta)$ iff $\text{plim}_\Theta \mathbf{T}_n = \mathbf{q}(\Theta)\ \forall\, \Theta \in \Omega$.

A consistent estimator is an estimator that converges in probability (element wise if $\mathbf{T}_n$ is a vector) to what is being estimated. Thus, for large enough $n$ (i.e., for large enough sample size), there is a high probability that the outcome of $T_n$ will be in the interval $(q(\Theta) - \varepsilon, q(\Theta) + \varepsilon)$ for arbitrarily small $\varepsilon > 0$ regardless of the value of $\Theta$. Relatedly, the sampling density of $T_n$ concentrates on the true value of $q(\Theta)$ as the sample size → ∞ if $T_n$ is a consistent estimator of $q(\Theta)$. Consistency is clearly a desirable property of an estimator, since it ensures that increasing sample information will ultimately lead to an estimate that is essentially certain to be arbitrarily close to the estimand $q(\Theta)$.

Since $\mathbf{T}_n \xrightarrow{m} \mathbf{q}(\Theta)$ implies $\mathbf{T}_n \xrightarrow{p} \mathbf{q}(\Theta)$, we can state *sufficient conditions for consistency* of $\mathbf{T}_n$ in terms of unbiasedness and in terms of variance convergence to zero. Specifically, if $\mathbf{T}_n$ is unbiased, or if the bias vector converges to zero as $n \to \infty$ and if $\text{var}(T_n) \to 0$ as $n \to \infty$ (or $\mathbf{Cov}(\mathbf{T}_n) \to [0]$ as $n \to \infty$ if $\mathbf{T}_n$ is a vector), then $\mathbf{T}_n$ is a consistent estimator of $\mathbf{q}(\Theta)$ by mean-square convergence.

**Example 7.6** Let $(X_1, \ldots, X_n)$ be a random sample from a population distribution, $f(z; \Theta)$, of the incomes of buyers of new Cadillacs, where $EZ = q_1(\Theta) = \mu$ and $\text{var}(Z) = q_2(\Theta) = \sigma^2 < \infty$. The sample mean $t_n(\mathbf{X}) = \bar{X}_n$ is a consistent estimator of the mean income, $\mu$, of Cadillac buyers since $\bar{X}_n$ is unbiased, and its variance $(\sigma^2/n) \to 0$ as $n \to \infty$. □

We note for future reference that a sequence of estimators can be consistent for $\mathbf{q}(\Theta)$ *without* $\mathrm{E}\mathbf{T}_n \to \mathbf{q}(\Theta)$ even if $\mathrm{E}\mathbf{T}_n$ exists $\forall n$. This at first seems counter intuitive, since if $\mathbf{T}_n$ is consistent, its density collapses on $\mathbf{q}(\Theta)$ as $n \to \infty$. Note the following counterexample.

**Example 7.7** Let the density of $T_n$ be defined as $f(t_n; \Theta) = (1-n^{-1/2})I_{\{\Theta\}}(t_n)+n^{-1/2}I_{\{n\}}(t_n)$. Note that as $n \to \infty$, $\lim_{n\to\infty} P[|t_n - \Theta| < \varepsilon] = 1$, for any $\varepsilon > 0$, and $T_n$ is consistent for $\Theta$. However, since $ET_n = \Theta(1 - n^{-1/2}) + n(n^{-1/2}) = \Theta(1 - n^{-1/2}) + n^{1/2}$, then as $n \to \infty$, $ET_n \to \infty$. □

The divergence of the expectation in Ex. 7.7 is due to the fact that the density function of $T_n$, although collapsing to the point $\Theta$ as $n \to \infty$, was not collapsing at a *fast enough rate* for the expectation to converge to $\Theta$. In particular, the density weighting assigned to the outcome $n$ in defining the expectation went to zero at a rate slower than $n$ went to infinity as $n \to \infty$, causing the divergence. A sufficient condition for the statement $T_n \xrightarrow{p} q(\Theta) \Rightarrow \lim_{n\to\infty} ET_n = q(\Theta)$ to be true is that $ET_n^2$ exists and is bounded $\forall n$, i.e., $ET_n^2 \leq m < \infty \; \forall n$ (Rao, *Statistical Inference*, p. 121). Note that this condition does not hold in Ex. 7.7.

*Class of Consistent Asymptotically Normal (CAN) Estimators and Asymptotic Properties* The class of *consistent asymptotically normal* estimators of $\mathbf{q}(\Theta)$ is defined in the statistical literature to be the collection of all estimators of $\mathbf{q}(\Theta)$ for which $n^{1/2}(\mathbf{T}_n - \mathbf{q}(\Theta)) \xrightarrow{d} N([\mathbf{0}], \Sigma_\mathrm{T})$, where $\Sigma_\mathrm{T}$ is a positive definite covariance matrix that may depend on the value of $\Theta$. We will allow this dependence to be implicit rather than utilize notation such as $\Sigma_\mathrm{T}(\Theta)$. Note that consistency of $\mathbf{T}_n$ follows immediately, since by Slutsky's theorem $n^{-1/2}[n^{1/2}(\mathbf{T}_n - \mathbf{q}(\Theta))] = \mathbf{T}_n - \mathbf{q}(\Theta) \xrightarrow{d} 0 \cdot \mathbf{Z} = [\mathbf{0}]$, where $\mathbf{Z} \sim N([\mathbf{0}], \Sigma_\mathrm{T})$, which implies $\mathbf{T}_n - \mathbf{q}(\Theta) \xrightarrow{p} [\mathbf{0}]$ or $\mathbf{T}_n \xrightarrow{p} \mathbf{q}(\Theta)$. The CAN class contains a large proportion of the estimators used in empirical work.

Because all of the estimators in the CAN class utilize precisely the same sequence of centering (i.e., $\mathbf{q}(\Theta)$ is subtracted from $\mathbf{T}_n$) and scaling (i.e., $\mathbf{T}_n - \mathbf{q}(\Theta)$ is multiplied by $n^{1/2}$), the problem of nonuniqueness of asymptotic distributions and properties does not arise. Asymptotic versions of MSEs, MSE matrices, bias vectors, variances, and covariance matrices can be defined via expectations taken with respect to the unique asymptotic distribution of estimators, where $\mathbf{T}_n \overset{a}{\sim} N(\mathbf{q}(\Theta), n^{-1}\Sigma_\mathrm{T})$. In particular, letting the prefix A denote an asymptotic property, and letting $\mathrm{E}_A$ denote an expectation taken with respect to an asymptotic distribution, we have within the CAN class, $\forall\,\Theta \in \Omega$,

$$\mathbf{AMSE}(\mathbf{T}_n) = \mathbf{ACov}(\mathbf{T}_n) = \mathrm{E}_{\mathrm{A}}(\mathbf{T}_n - \mathbf{q}(\Theta))(\mathbf{T}_n - \mathbf{q}(\Theta))' \quad \text{(multivariate)},$$

$$= \mathrm{Avar}(T_n) = \mathrm{E}_{\mathrm{A}}(T_n - q(\Theta))^2 \quad \text{(scalar)},$$

and

$$\mathbf{ABIAS}(\mathbf{T}_n) = \mathrm{E}_{\mathrm{A}}(\mathbf{T}_n - \mathbf{q}(\Theta)) = [0].$$

The last property above indicates that a CAN estimator of $\mathbf{q}(\Theta)$ is necessarily **asymptotically unbiased**. We pause to note that there is a lack of consensus in the literature regarding the definition of *asymptotic unbiasedness*, and Ex. 7.7 is useful for illustrating the issues involved. Some statisticians define asymptotic unbiasedness of an estimator sequence in terms of the limit of the expected values of the estimators in the sequence, where $\lim_{n\to\infty} \mathrm{E}\mathbf{T}_n = \mathbf{q}(\Theta)\ \forall\,\Theta \in \Omega$ characterizes an **asymptotically unbiased estimator**. Under this definition, the estimator in Ex. 7.7 would *not* be asymptotically unbiased, but rather would be **asymptotically biased**. It is clear that this definition of asymptotic unbiasedness requires that the expectations in the sequence exist, as they do in Ex. 7.7. Within the CAN class, the two definitions of asymptotic unbiasedness will coincide if the second-order moments of the estimators in the sequence $\{\mathbf{T}_n\}$ are bounded, since then $\lim_{n\to\infty} \mathrm{E}\mathbf{T}_n = \mathbf{q}(\Theta) = \mathrm{E}_{\mathrm{A}}(\mathbf{T}_n)$ (recall our remark on convergence of expectations following Ex. 7.7). Otherwise, the definitions may refer to different concepts of unbiasedness, as Ex. 7.7 demonstrates. Having alerted the reader to the alternative definition, we will continue to define asymptotic properties exclusively in terms of asymptotic distributions.

Given the preceding definition of asymptotic properties, we can now define the meaning of *asymptotic relative efficiency* and *asymptotic admissibility* uniquely within the context of CAN estimators.

**Definition 7.16**
***Asymptotic relative efficiency and admissibility***

Let $\mathbf{T}_n$ and $\mathbf{T}_n^*$ be CAN estimators of $\mathbf{q}(\Theta)$ such that $n^{1/2}(\mathbf{T}_n - \mathbf{q}(\Theta)) \xrightarrow{d} N([0], \Sigma_{\mathbf{T}})$ and $n^{1/2}(\mathbf{T}_n^* - \mathbf{q}(\Theta)) \xrightarrow{d} N([0], \Sigma_{\mathbf{T}^*})$.

a. If $T_n$ and $T_n^*$ are scalars, then the **asymptotic relative efficiency** of $T_n$ with respect to $T_n^*$ is given by

$$\mathrm{ARE}_\Theta(T_n, T_n^*) = \frac{\mathrm{AMSE}_\Theta(T_n^*)}{\mathrm{AMSE}_\Theta(T_n)} = \frac{\Sigma_{T^*}}{\Sigma_T}\ \forall\,\Theta \in \Omega.$$

$T_n$ is **asymptotically relatively more efficient** than $T_n^*$ if $\mathrm{ARE}_\Theta(T_n, T_n^*) \geq 1\ \forall\,\Theta \in \Omega$ and $> 1$ for some $\Theta \in \Omega$.

b. $\mathbf{T}_n$ is asymptotically relatively more efficient than $\mathbf{T}_n^*$ iff $\Sigma_{\mathbf{T}} - \Sigma_{\mathbf{T}^*}$ is negative semidefinite $\forall\,\Theta \in \Omega$ and $\Sigma_{\mathbf{T}} - \Sigma_{\mathbf{T}^*} \neq [0]$ for some $\Theta \in \Omega$.

c. If there exists an estimator that is asymptotically relatively more efficient than $\mathbf{T}_n$, then $\mathbf{T}_n$ is **asymptotically inadmissible**. Otherwise $\mathbf{T}_n$ is **asymptotically admissible**.

A discussion of the meaning of ARE and asymptotic admissibility, as well as all of the other asymptotic properties presented to this point, would be completely analogous to the discussion presented in the finite sample case, except now all interpretations would be couched in terms of approximations based on asymptotic distributions. We leave it to the reader to draw the analogies.

**Example 7.8** Recall Ex. 7.3 regarding the estimation of the expected operating lives of hard disks, $\theta$, using a random sample from an exponential population distribution. As an alternative estimator of $\theta$, consider the following function of the random sample:

$$T_n^* = t_n^*(\mathbf{X}) = \left[\frac{1}{2}\left(n^{-1}\sum_{i=1}^{n} X_i^2\right)\right]^{1/2} = (M_2'/2)^{1/2}.$$

Recall that in the case of the exponential probability distribution, $\mu_2' = \mathrm{E}X_i^2 = 2\theta^2$. Since $M_2'$ is the second-order sample moment based on a random sample from a probability distribution, we know that $M_2' \xrightarrow{p} 2\theta^2$ and that

$$\frac{n^{1/2}(M_2' - \mu_2')}{[\mu_4' - (\mu_2')^2]^{1/2}} = \frac{n^{1/2}(M_2' - 2\theta^2)}{[20\theta^4]^{1/2}} \xrightarrow{d} N(0, 1),$$

where $\mu_4' = [(d^4(1 - \theta t)^{-1}/dt^4)]_{t=0} = 24\theta^4$.

Now note that $T_n^*$ is a continuous function of $M_2'$, so that $\operatorname{plim} T_n^* = \operatorname{plim}(M_2'/2)^{1/2} = (\operatorname{plim}(M_2')/2)^{1/2} = \theta$ by Theorem 5.5. Therefore, $T_n^*$ is a consistent estimator of $\theta$. Furthermore, $n^{1/2}(M_2' - \theta)$ has a normal limiting distribution and is thus a CAN estimator. To see this, recall Theorem 5.39 on the asymptotic distribution of functions of asymptotically normal random variables, where in this application, $T_n^*$ is a function of the asymptotically normal random variable $M_2' \overset{a}{\sim} N(2\theta^2, 20\theta^4/n)$. Since

$$G = \left[\frac{dT_n^*}{dM_2'}\right]_{M_2'=2\theta^2} = \left[\frac{1}{4}(M_2'/2)^{-1/2}\right]_{M_2'=2\theta^2} = (4\theta)^{-1},$$

which $\neq 0\ \forall \theta > 0$ and is thus of "full row rank," it follows that

$$n^{1/2}(T_n^* - \theta) \xrightarrow{d} N(0, G[20\theta^4/n]G') = N(0, 1.25\theta^2).$$

In comparing $T_n^*$ with $\bar{X}_n$ as estimators of $\theta$, it is now clear that although both are consistent and asymptotically normal estimators of $\theta$, $\bar{X}_n$ is *asymptotically more efficient* than $T_n^*$, since in comparing the asymptotic variances of the limiting distributions of $n^{1/2}(\bar{X}_n - \theta)$ and $n^{1/2}(T_n^* - \theta)$, we have $\theta^2 < 1.25\theta^2$. □

*Asymptotic Efficiency* At this point it would seem logical to proceed to a definition of *asymptotic efficiency* in terms of a choice of estimator in the CAN class that has the smallest asymptotic variance or covariance matrix $\forall\, \Theta \in \Omega$ (compare to Def. 7.13). Unfortunately, LeCam[10] has shown that such an estimator

[10] L. LeCam, (1953), "On some asymptotic properties of maximum likelihood estimates and related Bayes estimates." *University*

does not exist without further restrictions on the class of estimators. In particular, LeCam effectively showed that, for *any* CAN estimator, one can always define an alternative estimator that has a smaller variance or covariance matrix for at least one $\Theta \in \Omega$. The implication of this result is that one cannot define an achievable lower bound to the asymptotic variances or covariance matrices of CAN estimators and so no asymptotically optimal estimator exists.

On the other hand, LeCam also showed that under mild regularity conditions, there does exist a lower bound to the asymptotic variance or covariance matrix of a CAN estimator that holds for all $\Theta \in \Omega$ except on a set of $\Theta$ values having *Lebesque measure zero* (this is called the *Cramér–Rao lower bound,* which will be discussed in Section 7.5). Note that the Lebesque measure of a set of $\Theta$ values can be thought of as the volume of the set within the $k$-dimensional parameter space. A set having Lebesque measure zero is a set with zero volume in $k$-space, e.g., a collection of isolated points, or a set of points having dimension less than $k$ (such as a square and its interior in a three-dimensional space, or a line in two-dimensional space). A set of Lebesque measure zero is a nonstochastic analogue to a set having probability zero, and such a set is thus practically irrelevant relative to its complement. It is thus meaningful to speak of a lower bound on the asymptotic variance or covariance matrix of a CAN estimator of $\mathbf{q}(\Theta)$ that holds *almost everywhere* in the parameter space (i.e., except for a set of Lebesque measure zero), and then a search for an estimator that achieves this bound becomes meaningful as well.

At this point we will state a general definition of asymptotic efficiency for CAN estimators. In Section 7.5, we will be much more precise about the functional form of the asymptotic covariance matrix of an asymptotically efficient estimator.

**Definition 7.17**
***Asymptotic efficiency***

If $\mathbf{T}_n$ is a CAN estimator of $\mathbf{q}(\Theta)$ having the smallest asymptotic covariance matrix among all CAN estimators $\forall \Theta \in \Omega$, except on a set of Lebesque measure zero, $\mathbf{T}_n$ is said to be asymptotically efficient.

As a final remark, it is possible to remove the qualifier "except on a set of Lebesque measure zero" if the CAN class of estimators is further restricted so that only estimators that converge *uniformly* to the normal distribution are considered. Roughly speaking, *uniform* convergence of a function sequence $F_n(x)$ to $F(x)$ requires that the rate at which convergence occurs is uniform across all $x$ in the domain of $F(x)$, unlike ordinary convergence (recall Def. 5.7), which allows the rate to be different for each $x$. The restricted class of estimators is called the consistent uniformly asymptotically normal (CUAN) class, and within the CUAN class it is meaningful to speak of an estimator that literally has the smallest asymptotic covariance matrix. The interested reader can con-

*of California Publications in Statistics,* 1:277–330.

sult C. R. Rao (1963), "Criteria of estimation in large samples." *Sankhya*, Series A, pp. 189–206, for further details.

## 7.4 Sufficient Statistics

Sufficient statistics for a given estimation problem are a collection of statistics or, equivalently, a collection of functions of the random sample, that summarize or represent all of the information in a random sample that is useful for estimating any $\mathbf{q}(\Theta)$. Thus, in place of the original random sample outcome, it is *sufficient* to have observations on the sufficient statistics to estimate any $\mathbf{q}(\Theta)$. Of course, the random sample itself must then be considered a collection of $n$ sufficient statistics, but an objective in defining sufficient statistics is to reduce the number of functions of the random sample needed to represent all of the sample information relevant for estimating $\mathbf{q}(\Theta)$. If a small collection of sufficient statistics can be found for a given statistical model, then for defining estimators of $\mathbf{q}(\Theta)$ it is sufficient to consider only functions of the smaller set of sufficient statistic outcomes as opposed to functions of all $n$ outcomes of the original random sample. In this way the sufficient statistics allow a data-reduction step to occur in a point estimation problem. Relatedly, it will be shown that the search for estimators of $\mathbf{q}(\Theta)$ having the MVUE property or small MSEs can always be restricted to functions of the smallest collection of sufficient statistics. Finally, if the sufficient statistics have a special property, referred to as *completeness*, then an explicit procedure utilizing the complete sufficient statistics is available that is often useful in defining MVUEs. We begin by presenting a more rigorous definition of sufficient statistics. Unless explicitly indicated otherwise, the scalar $\Theta$ can be replaced by the vector $\Theta$ throughout this section.

**Definition 7.18**
***Sufficient statistics***

Let $(X_1,\ldots,X_n) \sim f(x_1,\ldots,x_n;\Theta)$ be a random sample, and let $(S_1,\ldots,S_r) = (s_1(X_1,\ldots,X_n),\ldots,s_r(X_1,\ldots,X_n))$ be $r$ statistics. The $r$ statistics are said to be **sufficient statistics** for $f(x;\Theta)$ iff

$$f(x_1,\ldots,x_n;\Theta|s_1,\ldots,s_r) = h(x_1,\ldots,x_n),$$

i.e., the conditional density of $\mathbf{X}$, given $\mathbf{s} = [s_1,\ldots,s_r]'$, does not depend on the parameter $\Theta$.[11]

[11] Note that the conditional density function referred to in this definition is *degenerate* in the general sense alluded to in footnote 20 of Chapter 3. That is, since $(x_1,\ldots,x_n)$ satisfies the $r$ restrictions $s_i(x_1,\ldots,x_n) = s_i$, for $i = 1,\ldots,r$, by virtue of the event being conditioned upon, the arguments $x_1,\ldots,x_n$ of the conditional density are not all free to vary, but rather are functionally related. If one wished to utilize the conditional density for actually calculating *conditional* probabilities of events for $(X_1,\ldots,X_n)$, and if the random variables were continuous, then line integrals would be required, as discussed previously in Chapter 3 concerning the use of degenerate densities. This technical problem is of no concern in our current discussion of sufficient statistics since we will have no need to actually calculate conditional probabilities from the conditional density.

An intuitive interpretation of Def. 7.18 is that once the outcomes of the $r$ sufficient statistics are observed, there is no additional information on $\Theta$ in the sample outcome. The definition also implies that *given* the function values $\mathbf{s}(\mathbf{x}) = \mathbf{s}$, no other function of $\mathbf{X}$ provides any additional information about $\Theta$ other than that obtained from the outcomes $\mathbf{s}$. To motivate these interpretations, first note that the conditional density function $f(\mathbf{x}; \Theta|\mathbf{s})$ can be viewed as representing the probability distribution of all of the various ways in which random sample outcomes, $\mathbf{x}$, occur so as to generate the conditional value of $\mathbf{s}$. This is because the event being conditioned on requires that $\mathbf{x}$ satisfy $\mathbf{s}(\mathbf{x}) = \mathbf{s}$. Definition 7.18 states that if $\mathbf{S}$ is a vector of sufficient statistics, then $\Theta$ is a ghost in $f(\mathbf{x}; \Theta|\mathbf{s})$. That is, the conditional density function really does not depend on the value of $\Theta$ since $f(\mathbf{x}; \Theta|\mathbf{s}) = h(\mathbf{x})$. It follows that the probabilistic behavior of the various ways in which $\mathbf{x}$ results in $\mathbf{s}(\mathbf{x}) = \mathbf{s}$ *has nothing to do with* $\Theta$—it is independent of $\Theta$. Thus, analyzing the various ways in which a given value of $\mathbf{s}$ can occur, or examining additional functions of $X$, cannot possibly provide any additional information about $\Theta$ since the behavior of the outcomes of $X$, conditioned on the fact that $\mathbf{s}(\mathbf{x}) = \mathbf{s}$, is totally unrelated to $\Theta$.

**Example 7.9** Let $(X_1, \ldots, X_n)$ be a random sample from a Bernoulli population distribution representing whether phone call solicitations to potential customers result in a sale, so that

$$f(\mathbf{x}; p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \prod_{i=1}^n I_{\{0,1\}}(x_i),$$

where $p \in \Omega = (0, 1)$, $x_i = 1$ denotes a sale on the $i$th call, and $x_i = 0$ denotes no sale on the $i$th call. In this case, $\sum_{i=1}^n X_i$, representing the total number of sales in the sample, is a sufficient statistic for $f(x; p)$. To see that this is true, first note that the appropriate conditioning event in the context of Def. 7.18 would be $s(\mathbf{x}) = \sum_{i=1}^n x_i = s$, i.e., the total number of sales equals the value $s$. It follows from the definition of conditional probability that the conditional density function can be defined as[12]

$$f(\mathbf{x}; p|s) = \frac{P(x_1, \ldots, x_n, s(\mathbf{x}) = s)}{P(s(\mathbf{x}) = s)}.$$

The denominator probability is given directly by

$$P(s(\mathbf{x}) = s) = \binom{n}{s} p^s (1-p)^{n-s} I_{\{0,1,\ldots,n\}}(s)$$

because $s(\mathbf{X}) = \sum_{i=1}^n X_i$ is the sum of iid Bernoulli random variables, which we know to have a binomial distribution. The numerator probability is defined by

[12] The reader may wonder why we define the conditional density "from the definition of conditional probability," instead of using the rather straightforward methods for defining conditional densities presented in Section 2.6. The problem is that here we are conditioning on an event that involves *all* of the random variables $X_1, \ldots, X_n$, whereas in Chapter 2 we were dealing with the usual case where the event being conditioned upon involves only a subset of the random variables $X_1, \ldots, X_n$ having fewer than $n$ elements.

an appropriate evaluation of the joint density of the random sample, as

$$P(x_1,\ldots,x_n, s(\mathbf{x}) = s) = f(\mathbf{x}; p) I_{\{\sum_{i=1}^n x_i = s\}}(\mathbf{x})$$
$$= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}\left(\prod_{i=1}^n I_{\{0,1\}}(x_i)\right) I_{\{\sum_{i=1}^n x_i = s\}}(\mathbf{x}),$$

which is directly the probability of $x_1,\ldots,x_n$ *and* $s(\mathbf{x}) = \sum_{i=1}^n x_i = s$. Using the preceding functional representations of the numerator and denominator probabilities in the ratio defining the conditional density function, and using the fact that $s = \sum_{i=1}^n x_i$, we obtain (after appropriate algebraic cancellations)

$$f(\mathbf{x}; p|s) = \binom{n}{s}^{-1}\left[\prod_{i=1}^n I_{\{0,1\}}(x_i)\right] I_{\{\sum_{i=1}^n x_i = s\}}(\mathbf{x})$$

for any choice of $s \in \{0, 1, \ldots, n\}$, which does *not* depend on the parameter $p$. Thus, $S = \sum_{i=1}^n X_i$ is a sufficient statistic for $p$.

Note that the conditional density states that, *given* $\sum_{i=1}^n x_i = s$, all outcomes of $(X_1,\ldots,X_n)$ are equally likely with probability $\binom{n}{s}^{-1}$, and thus the probability of a particular pattern of sales and no sales occurring for $(x_1,\ldots,x_n)$, given that $\sum_{i=1}^n x_i = s$, has *nothing to do with the value of p*. It follows that only the fact that $\sum_{i=1}^n x_i = s$ provides any information about $p$—the particular pattern of 0's and 1's in $(x_1,\ldots,x_n)$ is irrelevant. This is consistent with intuition in that it is the total number of sales in $n$ phone calls, and not the particular pattern of sales and no sales, that provides information in a relative-frequency sense about the probability, $p$, of obtaining a sale on a phone call solicitation. Furthermore, if $Y = g(\mathbf{X})$ is any other function of the random sample, then it can provide no *additional* information about $p$ other than that already provided by $s(\mathbf{X})$. This follows from the fact that $h(y|s(\mathbf{x}) = s)$ will not depend on $p$, because the conditional density of $Y$ will have been derived from a conditional density of $\mathbf{X}$ that is independent of $p$, i.e.,

$$h(y|s(\mathbf{x}) = s) = P(y|s(\mathbf{x}) = s) = \sum_{\{\mathbf{x}:g(\mathbf{x})=y\}} f(\mathbf{x}; p|s) = \sum_{\{\mathbf{x}:g(\mathbf{x})=y\}} h(x)$$

since $f(\mathbf{x}; p|s) = h(\mathbf{x})$ if s is a sufficient statistic. □

In any problem of estimating $q(\Theta)$, once the outcome of a set of sufficient statistics is observed, the random-sample outcome $(x_1,\ldots,x_n)$ can effectively be ignored for the remainder of the point estimation problem since $\mathbf{s}(\mathbf{x})$ captures all of the relevant information that the sample has to offer regarding $q(\Theta)$. Essentially, it is *sufficient* that the outcome of $\mathbf{s}$ be observed. For example, with reference to Ex. 7.9, if a colleague were to provide the information that 123 sales were observed in a total of 250 phone calls, i.e., $\sum_{i=1}^{250} x_i = 123$, we would have no need to examine any other characteristic of the random-sample outcome $(x_1,\ldots,x_{250})$ when estimating $p$, or $q(p)$.

A significant practical problem in the use of sufficient statistics is knowing how to identify them. A criterion that is sometimes useful for identifying sufficient statistics is given by the Neyman factorization theorem:

**Theorem 7.1** (Neyman's Factorization Theorem) *Let $f(\mathbf{x};\Theta)$ be the density function of the random sample $(X_1,\ldots,X_n)$. The statistics $S_1,\ldots,S_r$ are sufficient statistics for $f(\mathbf{x};\Theta)$ iff $f(\mathbf{x};\Theta)$ can be factored as*

$$f(\mathbf{x};\Theta) = g(s_1(\mathbf{x}),\ldots,s_r(\mathbf{x});\Theta)h(\mathbf{x}),$$

*where g is a function of only $s_1(\mathbf{x}),\ldots,s_r(\mathbf{x})$ and $\Theta$, and $h(\mathbf{x})$ does not depend on $\Theta$.*

**Proof** The proof of the theorem in the continuous case is quite difficult, and we leave it to a more advanced course of study (see E. Lehmann (1986), *Testing Statistical Hypotheses.* New York: John Wiley, pp. 54–55). We provide a proof for the discrete case.

Sufficiency: Suppose the factorization criterion is met. Let $B(\mathbf{a}) = \{(x_1,\ldots,x_n) : s_i(\mathbf{x}) = a_i,\ i = 1,\ldots,r;\ \mathbf{x} \in R(\mathbf{X})\}$ be such that $P(B(\mathbf{a})) > 0$, and note that

$$P(B(\mathbf{a})) = \sum_{(x_1,\ldots,x_n)\in B(\mathbf{a})}\cdots\sum f(\mathbf{x};\Theta) = g(a_1,\ldots,a_r;\Theta)\sum_{(x_1,\ldots,x_n)\in B(\mathbf{a})}\cdots\sum h(x_1,\ldots,x_n).$$

Therefore,

$$f(\mathbf{x};\Theta|s_i(\mathbf{x}) = a_i, i = 1,\ldots,r) =$$

$$\begin{cases} \dfrac{g(a_1,\ldots,a_r;\Theta)h(x_1,\ldots,x_n)}{g(a_1,\ldots,a_r;\Theta)\sum\cdots\sum_{(x_1,\ldots,x_n)\in B(a)} h(x_1,\ldots,x_n)} = h^*(x_1,\ldots,x_n) & \text{when } \mathbf{x} \text{ is such that } s_i(\mathbf{x}) = a_i \quad \forall i \\ 0 & \text{otherwise,} \end{cases}$$

which does not depend on $\Theta$, and hence $S_1,\ldots,S_r$ are sufficient statistics.

Necessity: Suppose $S_1,\ldots,S_r$ are sufficient statistics, and note by the definition of the conditional density that

$$f(\mathbf{x};\Theta) = f(\mathbf{x}|s_i(\mathbf{x}) = a_i, i = 1,\ldots,r)P(s_i(\mathbf{x}) = a_i, i = 1,\ldots,r),$$

where the conditional density function does *not* depend on $\Theta$ by the sufficiency of S. Then we have factored $f(\mathbf{x};\Theta)$ into the product of a function of $s_1(\mathbf{x}),\ldots,s_r(\mathbf{x})$ and $\Theta$ (i.e., $P(s_i(\mathbf{x}) = a_i, i = 1,\ldots,r)$ will depend on $\Theta$) and a function that does *not* depend on $\Theta$. ■

As we have alluded to previously, a practical advantage of sufficient statistics is that they can often greatly reduce the number of random variables required to represent the sample information relevant for estimating $q(\Theta)$, as seen in Ex. 7.9 and in the following example of the use of the Neyman factorization theorem.

**Example 7.10** Let $(X_1, \ldots, X_n)$ be a random sample from the exponential population distribution $\theta^{-1}e^{-z/\theta}I_{(0,\infty)}(z)$ representing waiting times between customer arrivals at a retail store. Note that the joint density of the random sample is given by

$$f(x_1, \ldots, x_n; \theta) = \theta^{-n}e^{-\sum_{i=1}^n x_i/\theta}\prod_{i=1}^n I_{(0,\infty)}(x_i).$$

The joint density can be factored into the form required by the Neyman factorization theorem by defining $g(\sum_{i=1}^n x_i; \theta) = \theta^{-n}e^{-\sum_{i=1}^n x_i/\theta}$ and $h(\mathbf{x}) = \prod_{i=1}^n I_{(0,\infty)}(x_i)$. Then, from the theorem, we can conclude that $S = \sum_{i=1}^n X_i$ is a sufficient statistic for $f(\mathbf{x}; \theta)$. It follows that the value of the sum of the random-sample outcomes contains all the information in the sample outcomes relevant for estimating $q(\theta)$. □

Successful use of the Neyman factorization theorem for identifying sufficient statistics requires that one be ingenious enough to define the appropriate $g(\mathbf{s}(\mathbf{x}); \Theta)$ and $h(\mathbf{x})$ functions that achieve the required joint probability density factorization. Since the appropriate function definitions will *not* always be readily apparent, an approach introduced by Lehmann and Scheffé[13] can sometimes be quite useful for providing direction to the search for sufficient statistics. We will discuss this useful result in the context of minimal sufficient statistics, described next.

### Minimal Sufficient Statistics

At the beginning of our discussion of sufficient statistics, we remarked that one objective of using sufficient statistics is to reduce the number of functions of the random sample required to represent all of the information in the random sample relevant for estimating $q(\Theta)$. A natural question to consider is what is the *smallest* number of functions of the random sample that can represent all of the relevant sample information in a given point estimation problem? This relates to the concept of a *minimal sufficient statistic*, which is essentially the sufficient statistic for a given $f(\mathbf{x}; \Theta)$ that is defined using the fewest number of (functionally independent) coordinate functions of the random sample.

The statement of subsequent definitions and theorems will be facilitated by introducing the notion of the **range of X over the parameter space Ω**, defined as

$$R_\Omega(\mathbf{X}) = \{\mathbf{x} : f(\mathbf{x}; \Theta) > 0 \text{ for some } \Theta \in \Omega\}$$

The set $R_\Omega(X)$ represents all the values of $\mathbf{x}$ that are assigned a nonzero density weighting by $f(\mathbf{x}; \Theta)$ for at least one $\Theta \in \Omega$. In other words, $R_\Omega(\mathbf{X})$ is the union of the supports of the densities $f(\mathbf{x}; \Theta)$ for $\Theta \in \Omega$ and thus corresponds to the set of relevant $\mathbf{x}$ outcomes for the statistical model $\{f(\mathbf{x}; \Theta), \Theta \in \Omega\}$. If the support of the density $f(\mathbf{x}; \Theta)$ does not change with $\Theta$ (e.g., normal, Gamma, binomial

[13] E. L. Lehmann, and H. Scheffé (1950), "Completeness, similar regions, and unbiased estimation. *Sankhyā*, 10, p. 305.

with $n$ fixed) then $R_\Omega(\mathbf{X}) = R(\mathbf{X}) = \{x : f(\mathbf{x}; \Theta) > 0\}$, where $\Theta \in \Omega$ can be chosen arbitrarily, and we henceforth treat the range of $\mathbf{X}$ as being synonymous with the support of its density.

**Definition 7.19** ***Minimal sufficient statistics***

A sufficient statistic $\mathrm{S} = \mathrm{s}(\mathbf{X})$ for $f(\mathbf{x}; \Theta)$ is said to be a **minimal sufficient statistic** if, for every other sufficient statistic $\mathbf{T} = \mathbf{t}(\mathbf{X})$, $\exists$ a function $\mathbf{h}_\mathrm{T}(\cdot)$ such that $\mathrm{s}(\mathbf{x}) = \mathbf{h}_\mathrm{T}(\mathbf{t}(\mathbf{x}))\ \forall\, \mathbf{x} \in R_\Omega(\mathbf{X})$.

In order to motivate what is "minimal" about the sufficient statistic S in Def. 7.19, first note that S will have the fewest elements in its range compared to all sufficient statistics for $f(\mathbf{x}; \Theta)$. This follows from the fact that a function can never have more elements in its range than in its domain (recall the definition of a function, which requires that there is only one range point associated with each domain element, although there can be many domain elements associated with each range element). Thus, if $\mathrm{S} = \mathbf{h}_\mathrm{T}(\mathbf{T})$ for any other sufficient statistic $\mathbf{T}$, then the number of elements in $R(\mathrm{S})$ must not exceed the number of elements in $R(\mathbf{T})$, for *any* sufficient statistic $\mathbf{T}$. So, in this sense, S utilizes the *minimal* set of points for representing the sample information relevant for estimating $q(\Theta)$.

It can also be shown that a minimal sufficient statistic can be chosen to have the fewest number of coordinate functions relative to any other sufficient statistic, i.e., the number of coordinate functions defining the minimal sufficient statistic is *minimal.* A rigorous proof of this fact is quite difficult and is deferred to a more advanced course of study.[14] In order to at least motivate the plausibility of this fact, first note that since a minimal sufficient statistic, say S, is a function of all other sufficient statistics, then if $\mathbf{T}$ is any other sufficient statistic, $\mathbf{t}(\mathbf{x}) = \mathbf{t}(\mathbf{y}) \Rightarrow \mathrm{s}(\mathbf{x}) = \mathbf{h}_\mathrm{T}(\mathbf{t}(\mathbf{x})) = \mathbf{h}_\mathrm{T}(\mathbf{t}(\mathbf{y})) = \mathrm{s}(\mathbf{y})$. It follows that

$$A_\mathrm{T} = \{(\mathbf{x}, \mathbf{y}) : \mathbf{t}(\mathbf{x}) = \mathbf{t}(\mathbf{y})\} \subset \{(\mathbf{x}, \mathbf{y}) : \mathrm{s}(\mathbf{x}) = \mathrm{s}(\mathbf{y})\} = B$$

no matter which sufficient statistic, $\mathbf{T}$, is being referred to. If $B$ is to contain the set $A_\mathrm{T}$, then the constraints on $(\mathbf{x}, \mathbf{y})$ representing the set-defining conditions of $B$ cannot be more constraining than the constraints defining $A_\mathrm{T}$. In particular, the number of nonredundant constraints[15] defining $B$ cannot be more than the number defining $A_\mathrm{T}$. Thus the number of nonredundant coordinate functions defining S must be no larger than the number of nonredundant coordinate functions defining any other sufficient statistic so that the number of coordinate functions defining S is *minimal.* Identifying minimal sufficient statistics can often be facilitated by the following approach suggested by Lehmann and Scheffé.

[14] See E. W. Barankin and M. Katz (1959), "Sufficient statistics of minimal dimension." *Sankhyā*, 21, pp. 217–246; R. Shimizu (1966), "Remarks on sufficient statistics." *Ann. Inst. Statist. Math.*, 18, pp. 49–66; D. A. S. Fraser (1963), "On sufficiency and the exponential family." *J. Roy. Statist. Soc.*, Series B, 25, p. 115–123.

[15] By "nonredundant," we mean that none of the constraints is implied by the others. Redundant constraints are constraints that are ineffective or unnecessary in defining sets.

**Theorem 7.2** (Lehmann–Scheffé's Minimal Sufficiency Theorem) *Let* $\mathbf{X} \sim f(\mathbf{x};\Theta)$*. If the statistic* $\mathrm{S} = \mathrm{s}(\mathbf{X})$ *is such that* $\forall\, \mathbf{x}$ *and* $\mathbf{y} \in R_\Omega(\mathbf{X})$

$f(\mathbf{x};\Theta) = \tau(\mathbf{x},\mathbf{y})f(\mathbf{y};\Theta)$ iff $(\mathbf{x},\mathbf{y})$ satisfies $\mathrm{s}(\mathbf{x}) = \mathrm{s}(\mathbf{y})$,

*then* $\mathrm{S} = \mathrm{s}(\mathbf{X})$ *is a minimal sufficient statistic for* $f(\mathbf{x};\Theta)$.

***Proof** Define $A(\mathrm{s}) = \{\mathbf{x} : \mathrm{s}(\mathbf{x}) = \mathrm{s}\}$, and let $\mathbf{x}_\mathrm{s} \in A(\mathrm{s}) \cap R_\Omega(\mathbf{X})$, be chosen as a representative element of $A(\mathrm{s})$, $\forall\, \mathrm{s} \in R(\mathrm{S})$. Define $\eta(\mathbf{x}) = \mathbf{x}_\mathrm{s}\ \forall\, \mathbf{x} \in A(\mathrm{s})$ and $\forall\, \mathrm{s} \in R(\mathrm{S})$. Thus, $A(\mathrm{s})$ is the set of $\mathbf{x}$ outcomes whose image $\mathrm{s}(\mathbf{x})$ is s, and $\eta(\mathbf{x})$ is the representative element of the set $A(\mathrm{s})$ to which $\mathbf{x}$ belongs.

Assume that $(\mathbf{x},\mathbf{y}) \in \{(\mathbf{x},\mathbf{y}) : \mathrm{s}(\mathbf{x}) = \mathrm{s}(\mathbf{y})\} \Rightarrow f(\mathbf{x};\Theta) = \tau(\mathbf{x},\mathbf{y})f(\mathbf{y};\Theta)\ \forall\, \mathbf{x}$ and $\mathbf{v} \in R_\Omega(\mathbf{X})$. Then for $\mathbf{x} \in A(\mathrm{s}) \cap R_\Omega(\mathbf{X})$,

$$\begin{aligned}\mathrm{s}(\mathbf{x}) = \mathrm{s}(\mathbf{x}_\mathrm{s}) \Rightarrow f(\mathbf{x};\Theta) &= \tau(\mathbf{x},\mathbf{x}_\mathrm{s})f(\mathbf{x}_\mathrm{s};\Theta)\\ &= \tau(\mathbf{x},\eta(\mathbf{x}))f(\eta(\mathbf{x});\Theta) \quad (\text{substitute } \eta(\mathbf{x}) = \mathbf{x}_\mathrm{s})\\ &= h(\mathbf{x})g(\mathrm{s}(\mathbf{x});\Theta),\end{aligned}$$

where $h(\mathbf{x}) \equiv \tau(\mathbf{x},\eta(\mathbf{x}))$, $g(\mathrm{s}(\mathbf{x});\Theta) \equiv f(\eta(\mathbf{x});\Theta)$, and the $g$ function in the latter identity can be defined from the fact that $\eta(\mathbf{x}) = \mathbf{x}_\mathrm{s}$ iff $\mathrm{s}(\mathbf{x}) = \mathrm{s}$, so that $\eta(\mathbf{x}) \Leftrightarrow \mathrm{s}(\mathbf{x})$. If $\mathbf{x} \notin R_\Omega(\mathbf{X})$, then $f(\mathbf{x};\Theta) = h(\mathbf{x})g(\mathrm{s}(\mathbf{x});\Theta)$ by defining $h(\mathbf{x}) = 0$. Since Neyman factorization holds, $s(\mathbf{X})$ is a sufficient statistic.

Now assume $f(\mathbf{x};\Theta) = \tau(\mathbf{x},\mathbf{y})f(\mathbf{y};\Theta) \Rightarrow (\mathbf{x},\mathbf{y}) \in \{(\mathbf{x},\mathbf{y}) : \mathrm{s}(\mathbf{x}) = \mathrm{s}(\mathbf{y})\}\ \forall\, \mathbf{x}$ and $\mathbf{y} \in R_\Omega(\mathbf{X})$. Let $\mathrm{s}_*(\mathbf{x})$ be any other sufficient statistic for $f(\mathbf{x};\Theta)$. Then, by Neyman factorization, for some $g_*(\cdot)$ and $h_*(\cdot)$ functions,

$f(\mathbf{x};\Theta) = g_*(\mathrm{s}_*(\mathbf{x});\Theta)h_*(\mathbf{x})$.

If $\mathrm{s}_*(\mathbf{x}) = \mathrm{s}_*(\mathbf{y})$, then since $g_*(\mathrm{s}_*(\mathbf{x});\Theta) = g_*(\mathrm{s}_*(\mathbf{y});\Theta)$, it follows that

$f(\mathbf{x};\Theta) = [h_*(\mathbf{x})/h_*(\mathbf{y})]f(\mathbf{y};\Theta) = \tau(\mathbf{x},\mathbf{y})f(\mathbf{y};\Theta)$

whenever $h_*(\mathbf{y}) \neq 0$, so that $\mathrm{s}_*(\mathbf{x}) = \mathrm{s}_*(\mathbf{y}) \Rightarrow \mathrm{s}(\mathbf{x}) = \mathrm{s}(\mathbf{y})$. Values of $\mathbf{y}$ for which $\mathbf{h}_*(\mathbf{y}) = 0$ are such that $f(\mathbf{y};\Theta) = 0\ \forall\, \Theta \in \Omega$ by Neyman factorization and are thus irrelevant to the minimal sufficiency of S (recall Def. 7.19). Then s is a function of $\mathrm{s}_*$, as $\mathrm{s}(\mathbf{x}) = \gamma(\mathrm{s}_*(\mathbf{x}))$, $\forall\, \mathbf{x} \in R_\Omega(\mathbf{X})$, since for a representative $\mathbf{y}_{\mathrm{s}^*} \in \{\mathbf{x} : \mathrm{s}_*(\mathbf{x}) = \mathrm{s}_*\}$, $\mathrm{s}_*(\mathbf{x}) = \mathrm{s}_*(\mathbf{y}_{\mathrm{s}^*}) = \mathrm{s}_* \Rightarrow \mathrm{s}(\mathbf{x}) = \mathrm{s}(\mathbf{y}_{\mathrm{s}^*}) = \mathrm{s}$, and thus $\mathrm{s}(\mathbf{x}) = \mathrm{s} = \gamma(\mathrm{s}_*) = \gamma(\mathrm{s}_*(\mathbf{x}))$. Therefore, by Def. 7.19, $\mathrm{s}(\mathbf{X})$ is a minimal sufficient statistic. ■

Before proceeding to applications of the theorem, we present two corollaries that are informative and useful in practice.

**Corollary 7.1** (Lehmann–Scheffé Sufficiency) *Let* $\mathbf{X} \sim f(\mathbf{x};\Theta)$*. If the statistic* $\mathrm{S} = \mathrm{s}(\mathbf{X})$ *is such that* $\forall\, \mathbf{x}$ *and* $\mathbf{y} \in R_\Omega(\mathbf{X})$

$(\mathbf{x},\mathbf{y}) \in \{(\mathbf{x},\mathbf{y}) : \mathrm{s}(\mathbf{x}) = \mathrm{s}(\mathbf{y})\} \Rightarrow f(\mathbf{x};\Theta) = \tau(\mathbf{x},\mathbf{y})f(\mathbf{y};\Theta)$,

*then* $\mathrm{S} = \mathrm{s}(\mathbf{X})$ *is a sufficient statistic for* $f(\mathbf{x};\Theta)$.

**Proof** The proof of this corollary was given in the first part of the proof of Theorem 7.2. ■

The corollary indicates that the "only if" part of the condition in Theorem 7.2 is not required for the *sufficiency* of $\mathbf{s}(\mathbf{X})$, but it is the addition of the "only if" part that results in the *minimality* of $\mathbf{s}(\mathbf{X})$.

**Corollary 7.2** (Minimal Sufficiency When $R(\mathbf{X})$ Is Independent of $\Theta$) *Let* $\mathbf{X} \sim f(\mathbf{x};\Theta)$, *and suppose that* $R(\mathbf{X})$ *does not depend on* $\Theta$. *If the statistic* $\mathbf{S} = \mathbf{s}(\mathbf{X})$ *is such that* $f(\mathbf{x};\Theta)/f(\mathbf{y};\Theta)$ *does not depend on* $\Theta$ *iff* $(\mathbf{x},\mathbf{y})$ *satisfies* $\mathbf{s}(\mathbf{x}) = \mathbf{s}(\mathbf{y})$, *then* $\mathbf{S} = \mathbf{s}(\mathbf{X})$ *is a minimal sufficient statistic.*

**Proof** This follows from Theorem 7.2 by dividing through by $f(\mathbf{y};\Theta)$ on the left-hand side of the iff condition, which is admissible for all $\mathbf{x}$ and $\mathbf{y}$ in the density support $R(\mathbf{X}) = \{\mathbf{x} : f(\mathbf{x};\Theta) > 0\}$. Values of $\mathbf{x}$ and $\mathbf{y} \notin R(\mathbf{X})$ are irrelevant to minimal sufficiency (recall Def. 7.19). ■

Of course, using the preceding results for defining a minimal sufficient statistic still requires that one is observant enough to recognize an appropriate (vector) function $\mathbf{S}$. However, in many cases, the Lehmann–Scheffé approach transforms the problem into one where a choice of $\mathbf{S}$ is readily apparent. The following examples illustrate the use of the procedure for discovering minimal sufficient statistics.

**Example 7.11** Let $(X_1,\ldots,X_n)$ be a random sample from a nondegenerate Bernoulli population distribution representing whether or not a customer contact results in a sale, so that

$$f(\mathbf{x};p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}\prod_{i=1}^{n} I_{\{0,1\}}(x_i) \quad \text{for } p \in (0,1).$$

In an attempt to define a sufficient statistic for $f(\mathbf{x};p)$, follow the Lehmann-Scheffé procedure by examining

$$\frac{f(\mathbf{x};p)}{f(\mathbf{y};p)} = \frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}\prod_{i=1}^n I_{\{0,1\}}(x_i)}{p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i}\prod_{i=1}^n I_{\{0,1\}}(y_i)}$$

for all values of $\mathbf{x}$ and $\mathbf{y} \in R(\mathbf{X}) = \times_{i=1}^n\{0,1\}$. The ratio will be independent of $p$ iff the constraint $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ is imposed. A minimal sufficient statistic for $f(\mathbf{x};p)$ is thus $s(\mathbf{X}) = \sum_{i=1}^n X_i$, by Corollary 7.2. □

**Example 7.12** Let $(X_1,\ldots,X_n)$ be a random sample from a Gamma population distribution representing the operating life until failure of a certain brand and type of personal computer, so that

$$f(\mathbf{x};\alpha,\beta) = \frac{1}{\beta^{n\alpha}\Gamma^n(\alpha)}\left(\prod_{i=1}^n x_i\right)^{\alpha-1} e^{-\sum_{i=1}^n x_i/\beta}\prod_{i=1}^n I_{(0,\infty)}(x_i).$$

Using the Lehmann–Scheffé procedure for defining a sufficient statistic for $f(\mathbf{x}; \alpha, \beta)$, examine

$$\frac{f(\mathbf{x}; \alpha, \beta)}{f(\mathbf{y}; \alpha, \beta)} = \frac{\left(\prod_{i=1}^n x_i\right)^{\alpha-1} e^{-\sum_{i=1}^n x_i/\beta} \prod_{i=1}^n I_{(0,\infty)}(x_i)}{\left(\prod_{i=1}^n y_i\right)^{\alpha-1} e^{-\sum_{i=1}^n y_i/\beta} \prod_{i=1}^n I_{(0,\infty)}(y_i)}$$

for all values of $\mathbf{x}$ and $\mathbf{y} \in R(\mathbf{X}) = \times_{i=1}^n (0, \infty)$. (Note that the term $(\beta^{n\alpha}\Gamma^n(\alpha))$ has been algebraically cancelled in the density ratio.) The ratio will be independent of both $\alpha$ and $\beta$ iff both the constraints $\prod_{i=1}^n x_i = \prod_{i=1}^n y_i$ *and* $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ are imposed. A minimal sufficient statistic for $f(\mathbf{x}; \alpha, \beta)$ is then bivariate and is given by $s_1(\mathbf{X}) = \prod_{i=1}^n X_i$ and $s_2(\mathbf{X}) = \sum_{i=1}^n X_i$, by Corollary 7.2. □

**Example 7.13** Let $(X_1, \ldots, X_n)$ be a random sample from a uniform population distribution representing the number of minutes that a shipment is delivered before $(x_i < 0)$ or after $(x_i > 0)$ its scheduled arrival time, so that

$$f(\mathbf{x}; a, b) = (b - a)^{-n} \prod_{i=1}^{n} I_{[a,b]}(x_i).$$

Unlike the previous examples, here the range of $\mathbf{X}$ depends on the parameters $a$ and $b$. Referring to the Lehmann–Scheffé procedure for defining a minimal sufficient statistic for $f(\mathbf{x}; a, b)$ as given by Theorem 7.2, examine

$$\tau(\mathbf{x}, \mathbf{y}, a, b) = f(\mathbf{x}; a, b)/f(\mathbf{y}; a, b) = \prod_{i=1}^{n} I_{[a,b]}(x_i)/\prod_{i=1}^{n} I_{[a,b]}(y_i)$$

for all values of $\mathbf{x} \in R_\Omega(\mathbf{X}) = \{\mathbf{x} : f(\mathbf{x}; a, b) > 0$ for some $(a, b)$ satisfying $-\infty < a < b < \infty\}$,[16] and for values of $\mathbf{y}$ for which the denominator is greater than 0. (Note we have algebraically cancelled the $(b - a)^{-n}$ term that appears in both the numerator and the denominator of the ratio.) The $\mathbf{x}$ and $\mathbf{y}$ vectors under consideration will be $n$-element vectors, with $\mathbf{x}$ being any point in $R^n$ (why?) and $\mathbf{y}$ being any point in $\times_{i=1}^n [a, b]$. The ratio will be independent of $a$ and $b$ iff $\min(x_1, \ldots, x_n) = \min(y_1, \ldots, y_n)$ and $\max(x_1, \ldots, x_n) = \max(y_1, \ldots, y_n)$, in which case the ratio will be equal to 1. The preceding conditions also ensure that $f(\mathbf{x}; a, b) = 0$ when $f(\mathbf{y}; a, b) = 0$, so that $f(\mathbf{x}; a, b) = \tau(\mathbf{x}, \mathbf{y}) f(\mathbf{y}; a, b)$ holds $\forall \mathbf{x}$ and $\mathbf{y} \in R_\Omega(\mathbf{X})$. A minimal sufficient statistic for $f(\mathbf{x}; a, b)$ is then bivariate and is given by the *order statistics* $s_1(\mathbf{X}) = \min(X_1, \ldots, X_n)$ and $s_2(\mathbf{X}) = \max(X_1, \ldots, X_n)$, by Theorem 7.2. □

### Sufficient Statistics in the Exponential Class

The exponential class of densities represents a collection of parametric families of density functions for which sufficient statistics are straightforwardly

[16] It may be more appropriate to assume finite lower and upper bounds for $a$ and $b$, respectively. Doing so will not change the final result of the example.

defined. Furthermore, the sufficient statistics are generally *minimal* sufficient statistics.

**Theorem 7.3** (Exponential Class and Sufficient Statistics) *Let $f(\mathbf{x};\Theta)$ be a member of the exponential class of density functions*

$$f(\mathbf{x};\Theta) = \exp\left[\sum_{i=1}^{k} c_i(\Theta)g_i(\mathbf{x}) + d(\Theta) + z(\mathbf{x})\right] I_A(\mathbf{x}).$$

*Then* $\mathrm{s}(\mathbf{X}) = (g_1(\mathbf{X}),\ldots,g_k(\mathbf{X}))$ *is a k-variate sufficient statistic, and if* $c_i(\Theta)$, $i = 1,\ldots,k$, *are linearly independent, the sufficient statistic is a minimal sufficient statistic.*

**Proof** That $\mathrm{s}(\mathbf{X})$ is a sufficient statistic follows immediately from the Neyman factorization theorem by defining $g(g_1(\mathbf{x}),\ldots,g_k(\mathbf{x});\Theta) = \exp[\sum_{i=1}^{k} c_i(\Theta)g_i(\mathbf{x}) + d(\Theta)]$ and $h(\mathbf{x}) = \exp[z(\mathbf{x})]I_A(\mathbf{x})$ in the theorem.

That $\mathrm{s}(\mathbf{X})$ is a *minimal* sufficient statistic follows from the fact that $\mathrm{s}(\mathbf{X})$ can be derived using the Lehmann–Scheffé approach of Corollary 7.2. To see this, note that

$$f(\mathbf{x};\Theta)/f(\mathbf{y};\Theta) = \exp\left[\sum_{i=1}^{k} c_i(\Theta)[g_i(\mathbf{x}) - g_i(\mathbf{y})] + z(\mathbf{x}) - z(\mathbf{y})\right]\frac{I_A(\mathbf{x})}{I_A(\mathbf{y})}$$

will be independent of $\Theta$ iff $\mathbf{x}$ and $\mathbf{y}$ satisfy $g_i(\mathbf{x}) = g_i(\mathbf{y})$, for $i = 1,\ldots,k$, assuming $c_i(\Theta)$, $i = 1,\ldots,k$, are linearly independent.[17] □

As an exercise, the reader should utilize Theorem 7.3 as an alternative approach for discovering minimal sufficient statistics in the problems of random sampling examined in Ex. 7.11 and Ex. 7.12.

### Sufficiency and the MSE Criterion

In addition to generating a condensed representation of the information in a sample relevant for estimating $q(\Theta)$, sufficient statistics can also facilitate the discovery of estimators of $q(\Theta)$ that are relatively efficient in terms of MSE. In particular, in the pursuit of estimators with low MSE, *only* functions of sufficient statistics need to be examined, which is the implication of the Rao–Blackwell theorem.

[17] If one (or more) $c_i(\Theta)$ were linearly dependent on the other $c_j(\Theta)$'s, then "only if" would not apply. To see this, suppose $c_k(\Theta) = \sum_{i=1}^{k-1} a_i c_i(\Theta)$. Then the exp term could be rewritten as

$$\exp\left[\sum_{i=1}^{k-1} c_i(\Theta)\,[g_i(\mathbf{x}) - g_i(\mathbf{y}) + a_i[g_k(\mathbf{x}) - g_k(\mathbf{y})]]\right],$$

and so $g_i(\mathbf{x}) = g_i(\mathbf{y})$, $i = 1,\ldots,k$, is sufficient but not *necessary* for the term to be independent of $\Theta$, and thus $\mathrm{s}(\mathbf{X})$ would not be *minimal.*

**Theorem 7.4** (Rao–Blackwell's Theorem (scalar case)) *Let* $\mathbf{S} = [S_1, \ldots, S_r]'$ *be an r-variate sufficient statistic for* $f(\mathbf{x};\Theta)$, *and let* $t^*(\mathbf{X})$ *be any estimator of the scalar* $q(\Theta)$ *having finite variance. Define* $t(\mathbf{X}) = \mathrm{E}(t^*(\mathbf{X})|S_1, \ldots, S_r) = h(S_1, \ldots, S_r)$. *Then* $t(\mathbf{X})$ *is an estimator of* $q(\Theta)$ *for which* $\mathrm{MSE}_\Theta(t(\mathbf{X})) \leq \mathrm{MSE}_\Theta(t^*(\mathbf{X}))\ \forall\ \Theta \in \Omega$, *where the equality is attained only if* $P_\Theta[t(\mathbf{x}) = t^*(\mathbf{x})] = 1$.

**Proof** First note that since $\mathbf{S} = [S_1, \ldots, S_r]'$ is an $r$-variate sufficient statistic, $f(\mathbf{x}|\mathbf{s})$ does not depend on $\Theta$, and thus neither does the function $t(\mathbf{X})$ (since it is defined as a conditional expectation using $f(\mathbf{x}|\mathbf{s})$), so $t(\mathbf{X})$ is a *statistic* that can be used as an estimator of $q(\Theta)$. Now, by the double expectation theorem, $\mathrm{E}t(\mathbf{X}) = \mathrm{E}\mathrm{E}(t^*(\mathbf{X})|S_1, \ldots, S_r) = \mathrm{E}t^*(\mathbf{X})$, so that $t(\mathbf{X})$ and $t^*(\mathbf{X})$ have precisely the same expectation.

Next, examine

$$\mathrm{MSE}(t^*(\mathbf{X})) = \mathrm{E}(t^*(\mathbf{X}) - q(\Theta))^2 = \mathrm{E}(t^*(\mathbf{X}) - t(\mathbf{X}) + t(\mathbf{X}) - q(\Theta))^2$$
$$= \mathrm{E}(t^*(\mathbf{X}) - t(\mathbf{X}))^2 + 2\mathrm{E}(t^*(\mathbf{X}) - t(\mathbf{X}))(t(\mathbf{X}) - q(\Theta)) + \mathrm{E}(t(\mathbf{X}) - q(\Theta))^2.$$

The cross-product term is zero. To see this, first note that $\mathrm{E}[(t^*(\mathbf{X}) - t(\mathbf{X}))(t(\mathbf{X}) - q(\Theta))] = \mathrm{E}[t(\mathbf{X})(t^*(\mathbf{X}) - t(\mathbf{X}))]$ since $\mathrm{E}(t^*(\mathbf{X}) - t(\mathbf{X}))q(\Theta) = 0$ because $\mathrm{E}t^*(\mathbf{X}) = \mathrm{E}t(\mathbf{X})$. Now note that by definition, $t(\mathbf{X})$ is a function of only sufficient statistics, and so $t(\mathbf{X})$ is a constant *given* $s_1, \ldots, s_r$. Therefore,

$$\mathrm{E}[t(t^*(\mathbf{X}) - t)|s_1, \ldots, s_r] = [t\mathrm{E}(t^*(\mathbf{X}) - t)|s_1, \ldots, s_r)] = 0$$

since $\mathrm{E}(t^*(\mathbf{X})|s_1, \ldots, s_r) = t$ by definition, so that $\mathrm{E}[t(\mathbf{X})(t^*(\mathbf{X}) - t(\mathbf{X}))] = 0$ by the double expectation theorem.

Then, dropping the nonnegative term $\mathrm{E}(t^*(\mathbf{X}) - t(\mathbf{X}))^2$ on the right-hand side of the expression defining $\mathrm{MSE}(t^*(\mathbf{X}))$ above yields $\mathrm{MSE}_\Theta(t^*(\mathbf{X})) \geq \mathrm{E}_\Theta(t(\mathbf{X}) - q(\Theta))^2 = \mathrm{MSE}_\Theta(t(\mathbf{X}))\ \forall\ \Theta \in \Omega$. The equality is attained iff $\mathrm{E}_\Theta(t^*(\mathbf{X}) - t(\mathbf{X}))^2 = 0$, which requires that $P_\Theta[t^*(\mathbf{x}) = t(\mathbf{x})] = 1$. ■

The point of the theorem is that for any estimator $t^*(\mathbf{X})$ of $q(\Theta)$, there always exists an alternative estimator that is at least as good as $t^*(\mathbf{X})$ in terms of MSE and that is a function of *any* set of sufficient statistics. Thus, the Rao–Blackwell theorem suggests that the search for estimators of $q(\Theta)$ with low MSEs can always be restricted to an examination of functions of sufficient statistics, where hopefully the number of sufficient statistics required to fully represent the information about $q(\Theta)$ is substantially less than the size of the random sample itself.[18] Note that if attention is restricted to the unbiased class of estimators, so that $t^*(\mathbf{X})$ is an unbiased estimator in the statement of the theorem, then the Rao–Blackwell theorem implies that the search for a minimum-variance estimator within the class of unbiased estimators can also be restricted to functions of sufficient statistics. As an illustration, in Ex. 7.10 we know that $\sum_{i=1}^n X_i$ is a

[18] The reader will recall that the random sample, $(X_1, \ldots, X_n)$, is by definition a set of sufficient statistics for $f(\mathbf{x}; \Theta)$. However, it is clear that no improvement (decrease) in the MSE of an unbiased estimator will be achieved by conditioning on $(X_1, \ldots, X_n)$, i.e., the reader should verify that this is a case where $\mathrm{E}(t^*(\mathbf{X}) - t(\mathbf{X}))^2 = 0$ and MSE equality is achieved in the Rao–Blackwell theorem.

sufficient statistic for $f(\mathbf{x};\theta)$. Thus, our search for estimators of $\theta$ with low MSE can be confined to functions of the sufficient statistic, i.e., $t(\mathbf{X}) = h(\sum_{i=1}^n X_i)$. For future reference, we note that $h(\sum_{i=1}^n X_i) = n^{-1}(\sum_{i=1}^n X_i)$ is the MVUE of $\theta$.

The Rao–Blackwell theorem can be extended to the vector case as follows.

**Theorem 7.5** (Rao–Blackwell Theorem (vector case)) *Let* $\mathbf{S} = [S_1,\ldots,S_r]'$ *be an r-variate sufficient statistic for* $f(\mathbf{x};\boldsymbol{\Theta})$*, and let* $\mathbf{t}^*(\mathbf{X})$ *be an estimator of the* $(k\times 1)$ *vector function* $\mathbf{q}(\boldsymbol{\Theta})$ *having a finite covariance matrix. Define* $\mathbf{t}(\mathbf{X}) = \mathrm{E}(\mathbf{t}^*(\mathbf{X})|S_1,\ldots,S_r) = \mathbf{h}(S_1,\ldots,S_r)$. *Then* $\mathbf{t}(\mathbf{X})$ *is an estimator of* $\mathbf{q}(\boldsymbol{\Theta})$ *for which* $\mathbf{MSE}_{\Theta}(\mathbf{t}(\mathbf{X})) \preceq \mathbf{MSE}_{\Theta}(\mathbf{t}^*(\mathbf{X}))\ \forall\ \boldsymbol{\Theta}\in\Omega$, *where the equality is attained only if* $P_{\Theta}(\mathbf{t}(\mathbf{x}) = \mathbf{t}^*(\mathbf{x})) = 1$.

**Proof** The proof is analogous to the proof in the scalar case, except for MSE matrices replacing scalar MSEs in establishing that $\mathbf{MSE}(\mathbf{t}(\mathbf{X}))$ is smaller than $\mathbf{MSE}(\mathbf{t}^*(\mathbf{X}))$. The details are left to the reader. ■

The implications of Theorem 7.5 are analogous to those for the scalar case. Namely, one need only examine *vector* functions of sufficient statistics for estimating the *vector* $\mathbf{q}(\boldsymbol{\Theta})$ if the objective is to obtain an estimator with a small MSE matrix. Furthermore, the search for an MVUE of $\mathbf{q}(\boldsymbol{\Theta})$ can also be restricted to functions of sufficient statistics. As stated previously, this can decrease substantially the dimensionality of the data used in a point estimation problem if the minimal sufficient statistics for the problem are few in number. Revisiting Ex. 7.11, we note for future reference that

$$\mathbf{T} = \mathbf{h}\left(\sum_{i=1}^{n} X_i\right) = \begin{bmatrix} \left(\sum_{i=1}^n X_i\right)/n \\ \left(n\left(\sum_{i=1}^n X_i\right) - \left(\sum_{i=1}^n X_i\right)^2\right)/n(n-1) \end{bmatrix}$$

is the MVUE for $(p, p(1-p))$, the mean and variance of the Bernoulli population distribution in the example.

### Complete Sufficient Statistics

If a sufficient statistic, S, has the property of being *complete*, then any unbiased estimator of $q(\Theta)$ that is defined as a function of S is *unique*. We state the formal definition of completeness, and then motivate the uniqueness of unbiased estimators based on complete sufficient statistics.

**Definition 7.20** ***Complete sufficient statistics***

Let $\mathbf{S} = [S_1,\ldots,S_r]'$ be a sufficient statistic for $f(\mathbf{x};\Theta)$. The sufficient statistic S is said to be **complete** iff the only real-valued function $h$ defined on the range of S which satisfies $\mathrm{E}_{\Theta}h(\mathbf{S}) = 0\ \forall\ \Theta\in\Omega$ is the function defined as $h(\mathbf{S}) = 0$ with probability 1 $\forall\ \Theta\in\Omega$.

If a sufficient statistic, S, is complete, it follows that two *different* functions of S cannot have the same expected value. To see this, suppose that $\mathrm{E}t(\mathbf{S}) = \mathrm{E}t^*(\mathbf{S}) = q(\Theta)$, and define $\tau(\mathbf{S}) = t(\mathbf{S}) - t^*(\mathbf{S})$. Then $\mathrm{E}\tau(\mathbf{S}) = 0$, and since $\tau(\mathbf{S})$

is a function of the complete sufficient statistic S, it must be the case that $\tau(s) = t(s) - t^*(s) = 0$ occurs with probability 1 for all $\Theta$. Thus, $t(s)$ and $t^*(s)$ are the *same* function with probability 1. An important implication of this result is that any unbiased estimator of $q(\Theta)$ that is a function of the complete sufficient statistic is *unique*—there cannot be more than one unbiased estimator of $q(\Theta)$ defined in terms of complete sufficient statistics. This uniqueness property leads to an important procedure for defining MVUEs that will be discussed in Section 7.5.

The following example illustrates the process of verifying the completeness property of a sufficient statistic. Such verification often requires considerable ingenuity.

**Example 7.14** Let $(X_1, \ldots, X_n)$ be a random sample from the Bernoulli population distribution $p^z(1-p)^{1-z}I_{\{0,1\}}(z)$ representing whether or not the administration of a particular drug cures the disease of a patient, and suppose we wish to estimate $q(p) = p(1-p) = \sigma^2$. Note that the joint density of the random sample in this case is given by

$$f(x_1, \ldots, x_n; p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \prod_{i=1}^n I_{\{0,1\}}(x_i).$$

Using the Neyman factorization criterion, we know that $S = \sum_{i=1}^n X_i$ is a sufficient statistic for $f(x_1, \ldots, x_n; p)$, since the joint density can be factored into the product of

$$g(s; p) = p^s(1-p)^{n-s} \qquad \text{and } h(x_1, \ldots, x_n) = \prod_{i=1}^n I_{\{0,1\}}(x_i).$$

To determine whether $S = \sum_{i=1}^n X_i$ is a *complete* sufficient statistic, we need to determine whether the only real-valued function $h$ defined on the range of $S$ which satisfies $\mathrm{E}h(S) = 0 \; \forall \, p \in [0, 1]$ is the function defined as $h(S) = 0$ with probability $1 \; \forall \, p \in [0, 1]$. Note that since $S$ has a *binomial* density, in this case we know that for a sample size of $n$, $\mathrm{E}h(S) = 0 \; \forall \, p \in [0, 1]$ implies

$$\mathrm{E}h(S) = \sum_{i=0}^n h(i)\binom{n}{i}p^i(1-p)^{n-i} = 0 \quad \forall p \in [0, 1].$$

Let $p \in (0, 1)$, and note, by dividing through by $(1-p)^n$, that the preceding summation condition can be rewritten as

$$\tau(z) = \sum_{i=0}^n h(i)\binom{n}{i}z^i = 0,$$

which is a polynomial in $z = p/(1-p)$. Differentiating $n$ times with respect to $z$ yields

$$\frac{d^n \tau(z)}{dz^n} = n!h(n) = 0,$$

which implies that $h(n) = 0$. Differentiating $\tau(z)$ only $(n-1)$ times with respect to $z$, and using the fact that $h(n) = 0$, yields

$$\frac{d^{n-1}\tau(z)}{dz^{n-1}} = n!h(n-1) = 0.$$

The process can be continued to lead ultimately to the conclusion that $h(i) = 0$ for $i = 0, \ldots, n$ is required. Thus, necessarily $h(S) = 0$ with probability 1 if $\mathrm{E}h(S) = 0$, and thus $S = \sum_{i=1}^n X_i$ is a *complete* sufficient statistic for $f(x_1, \ldots, x_n; p)$. □

In general, the method used to verify completeness must be devised on a case-by-case basis. However, the following theorem identifies a large collection of parametric families for which complete sufficient statistics are relatively straightforward to identify.

**Theorem 7.6** (Completeness in the exponential class) *Let the joint density, $f(\mathbf{x}; \Theta)$, of the random sample $(X_1, \ldots, X_n)$ be a member of a parametric family of densities belonging to the exponential class of densities. If the range of $\mathbf{c}(\Theta) = [c_1(\Theta), \ldots, c_k(\Theta)]'$, $\Theta \in \Omega$, contains an open k-dimensional rectangle, then $\mathbf{g}(\mathbf{X}) = [g_1(\mathbf{X}), \ldots, g_k(\mathbf{X})]'$ is a set of complete sufficient statistics for $f(\mathbf{x}; \Theta)$, $\Theta \in \Omega$.*

**Proof** See E. Lehmann (1986), *Testing Statistical Hypotheses*. New York: John Wiley, pp. 142–143 and P. Bickel and K. Doksum (1977), *Mathematical Statistics*. San Francisco: Holden Day, p. 123. ■

Theorem 7.6 implies that if we are dealing with a parametric family of densities from the exponential class, once we verify that the range of $\mathbf{c}(\Theta)$ contains an open $k$-dimensional rectangle, we will have immediately identified a set of complete sufficient statistics given by $\mathbf{g}(\mathbf{X})$ in the exponential class representation. Regarding the open rectangle condition, it will often be readily apparent from the definition of the range of $\mathbf{c}(\Theta)$ whether there exists an open $k$-dimensional rectangle contained in $R(\mathbf{c})$. Alternatively, it can be shown that $R(\mathbf{c})$ will contain such an open rectangle if $\mathbf{c}(\Theta)$ is continuously differentiable $\forall\Theta$ in some open $k$-dimensional rectangle, $\Gamma$, contained in the parameter space $\Omega$ and $\partial\mathbf{c}/\partial\Theta$ has full rank for at least one $\Theta \in \Gamma$ (Bartle, Real Anaylsis, p. 381).

**Example 7.15** Let $(X_1, \ldots, X_n)$ be a random sample from a normal population distribution with mean $\mu$ and variance $\sigma^2$ representing the package weights of a certain type of cereal produced by a large company. The joint density function for the random sample is then given by

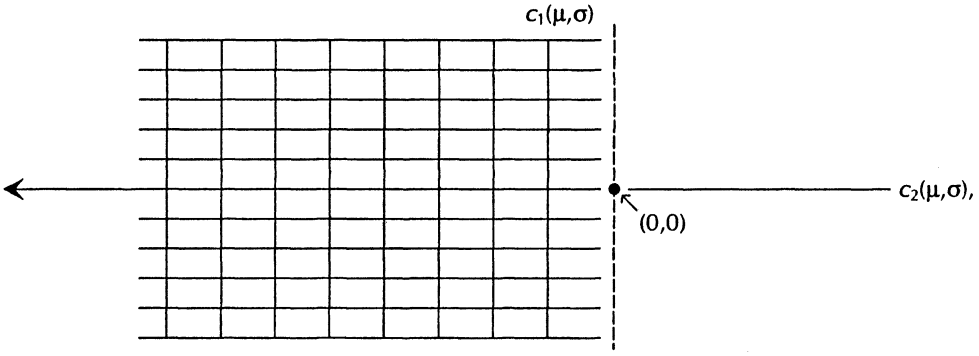
$$f(x_1, \ldots, x_n; \mu, \sigma) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-(1/2)\sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2\right).$$

The density is a member of the exponential class of density functions

$$\exp\left(\sum_{i=1}^{2} c_i(\mu,\sigma)g_i(\mathbf{x}) + d(\mu,\sigma) + z(\mathbf{x})\right) I_A(x_1,\ldots,x_n),$$

where $c_1(\mu,\sigma) = \mu/\sigma^2$, $g_1(\mathbf{x}) = \sum_{i=1}^n x_i$, $c_2(\mu,\sigma) = -(1/2\sigma^2)$, $g_2(\mathbf{x}) = \sum_{i=1}^n x_i^2$, $d(\mu,\sigma) = (-n/2)((\mu^2/\sigma^2) + \ln(2\pi\sigma^2))$, $z(\mathbf{x}) = 0$, and $A = \times_{i=1}^n(-\infty,\infty)$.

Now note that $\mathbf{c}(\mu,\sigma) = (c_1(\mu,\sigma), c_2(\mu,\sigma))'$ has the set $(-\infty,\infty) \times (-\infty,0)$ for its range, i.e., graphically



since by definition the range of $\mathbf{c}(\mu,\sigma)$ in this case is

$$R(\mathbf{c}) = \{(c_1,c_2) : c_1 = \mu/\sigma^2, c_2 = -1/2\sigma^2, \mu \in (-\infty,\infty), \sigma \in (0,\infty)\}.$$

The range of $\mathbf{c}(\mu,\sigma)$ clearly contains an open two-dimensional rectangle, i.e., there exists a set of points $\{(x_1,x_2) : a_i < x_i < b_i,\ i = 1,2\} \subset R(\mathbf{c})$, and so we know from Theorem 7.6 that $\{\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\}$ is a set of complete sufficient statistics for the multivariate normal family of densities in this example.

As an alternative verification of the open rectangle condition, note that $\mathbf{c}(\mu,\sigma)$ is continuously differentiable for all $\mu \in (-\infty,\infty)$ and $\sigma > 0$. Furthermore, letting $\Theta = [\mu,\sigma]'$,

$$\frac{\partial \mathbf{c}'}{\partial \Theta} = \left[\begin{array}{c|c} \dfrac{\partial c_1}{\partial \mu} & \dfrac{\partial c_1}{\partial \sigma} \\ \hline \dfrac{\partial c_2}{\partial \mu} & \dfrac{\partial c_2}{\partial \sigma} \end{array}\right] = \left[\begin{array}{c|c} 1/\sigma^2 & -\dfrac{2\mu}{\sigma^3} \\ \hline 0 & \dfrac{1}{\sigma^3} \end{array}\right],$$

so that $\det[\partial \mathbf{c}/\partial\Theta] = 1/\sigma^5 > 0\ \forall\ \sigma > 0$. Thus $\partial\mathbf{c}/\partial\Theta$ has full rank $\forall \mu \in (-\infty,\infty)$ and $\forall \sigma \in (0,\infty)$, and the open rectangle condition is verified. □

The reader might have noticed that in all of the preceding examples in which sampling was from a population distribution belonging to the exponential class, the joint density function for the random sample was also a member of the exponential class of densities. This was *not* just a coincidence, as the following theorem makes clear.

**Theorem 7.7** (Joint Density of Random Sample from an Exponential Class Distribution) *Let $(X_1, \ldots, X_n)$ be a random sample from a population distribution that belongs to the exponential class of density functions. Then the joint density function of the random sample also belongs to the exponential class of density functions.*

**Proof** Suppose

$$X_j \sim f(x_j; \Theta) = \exp\left[\sum_{i=1}^{k} c_i(\Theta) g_i(x_j) + d(\Theta) + z(x_j)\right] I_A(x_j) \quad \forall j.$$

Then

$$(X_1, \ldots, X_n) \sim \prod_{j=1}^{n} f(x_j; \Theta) = \exp\left[\sum_{i=1}^{k} c_i(\Theta) \sum_{j=1}^{n} g_i(x_j) + nd(\Theta) + \sum_{j=1}^{n} z(x_j)\right] \prod_{j=1}^{n} I_A(x_j)$$

$$= \exp\left[\sum_{i=1}^{k} c_i(\Theta) g_i^*(x_1, \ldots, x_n) + d^*(\Theta) + z^*(x_1, \ldots, x_n)\right] I_{A^*}(x_1, \ldots, x_n),$$

where $g_i^*(x_1, \ldots, x_n) = \sum_{j=1}^{n} g_i(x_j)$, $d^*(\Theta) = nd(\Theta)$, $z^*(x_1, \ldots, x_n) = \sum_{j=1}^{n} z(x_j)$, and $A^* = \times_{i=1}^{n} A$. Thus, the joint density function of $(X_1, \ldots, X_n)$ belongs to the exponential class. ■

The point is that if random sampling is from an exponential class population distribution, then Theorem 7.6 could potentially be useful for finding complete sufficient statistics, since we know that the joint density of the random sample is in fact in the exponential class of probability distributions.

If a sufficient statistic is complete, then it is also minimal, although the converse does *not* hold in general.

**Theorem 7.8** s(**X**) *is complete* $\Rightarrow$ s(**X**) *is minimal.*

**Proof** E. L. Lehmann and H. Scheffé, 1950, Unbiased Estimation. ■

Thus, if one finds a set of complete sufficient statistics, one also knows there is no smaller set of sufficient statistics available for the statistical model being analyzed. Furthermore, one knows that the search for complete sufficient statistics can be limited to an examination of minimal sufficient statistics.

### Sufficiency, Minimality, and Completeness of Functions of Sufficient Statistics

Sufficient statistics are not unique. In fact, any one-to-one (i.e., invertible) function of a sufficient statistic S, say $\tau$(S), is also a sufficient statistic, and if S is complete or minimal, then it is also true that $\tau$(S) is complete or minimal, respectively. We formalize this observation in the following theorem.

**Theorem 7.9** (Sufficiency of Invertible Functions of Sufficient Statistics) *Let* $\mathbf{S} = \mathbf{s}(\mathbf{X})$ *be an* $(r \times 1)$ *sufficient statistic for* $f(\mathbf{x};\Theta)$. *If* $\boldsymbol{\tau}(\mathbf{s}(\mathbf{X}))$ *is an* $(r \times 1)$ *invertible function of* $\mathbf{s}(\mathbf{X})$, *then*

**a.** $\boldsymbol{\tau}(\mathbf{s}(\mathbf{X}))$ *is an* $(r \times 1)$ *sufficient statistic for* $f(\mathbf{x};\Theta)$;

**b.** *if* $\mathbf{s}(\mathbf{X})$ *is a minimal sufficient statistic, then* $\boldsymbol{\tau}(\mathbf{s}(\mathbf{X}))$ *is a minimal sufficient statistic;*

**c.** *if* $\mathbf{s}(\mathbf{X})$ *is a complete sufficient statistic, then* $\boldsymbol{\tau}(\mathbf{s}(\mathbf{X}))$ *is a complete sufficient statistic.*

**Proof**

**a.** If $\mathbf{s}(\mathbf{X})$ is sufficient, then by Neyman factorization $f(\mathbf{x};\Theta) = g(\mathbf{s}(\mathbf{x});\Theta)h(\mathbf{x})$. Since $\boldsymbol{\tau}(\mathbf{s}(\mathbf{X}))$ is invertible, it follows that $\mathbf{s}(\mathbf{x}) = \boldsymbol{\tau}^{-1}(\boldsymbol{\tau}(\mathbf{s}(\mathbf{x})))$, so that $g(\mathbf{s}(\mathbf{x});\Theta) = g(\boldsymbol{\tau}^{-1}(\boldsymbol{\tau}(\mathbf{s}(\mathbf{x})));\Theta) = g_*(\boldsymbol{\tau}(\mathbf{s}(\mathbf{x}));\Theta)$. By Neyman factorization, since $f(\mathbf{x};\Theta) = g_*(\boldsymbol{\tau}(\mathbf{s}(\mathbf{x}));\Theta)h(\mathbf{x})$, $\boldsymbol{\tau}(\mathbf{s}(\mathbf{X}))$ is a sufficient statistic for $f(\mathbf{x};\Theta)$.

**b.** If it can be shown that $\boldsymbol{\tau}(\mathbf{s}(\mathbf{X}))$ satisfies the conditions of the Lehmann–Scheffé minimal sufficiency theorem, then we know that $\boldsymbol{\tau}$ is a minimal sufficient statistic. Note that $\boldsymbol{\tau}(\mathbf{s}(\mathbf{x})) = \boldsymbol{\tau}(\mathbf{s}(\mathbf{y})) \Leftrightarrow \mathbf{s}(\mathbf{x}) = \mathbf{s}(\mathbf{y})$ by the invertibility of $\boldsymbol{\tau}$. It follows that $\{(\mathbf{x},\mathbf{y}) : \boldsymbol{\tau}(\mathbf{s}(\mathbf{x})) = \boldsymbol{\tau}(\mathbf{s}(\mathbf{y}))\} = \{(\mathbf{x},\mathbf{y}) : \mathbf{s}(\mathbf{x}) = \mathbf{s}(\mathbf{y})\}$ so that by Theorem 7.2, if $\mathbf{S}$ is a minimal sufficient statistic, then so is $\boldsymbol{\tau}(\mathbf{S})$.

**c.** Suppose $\mathrm{E}h_*(\boldsymbol{\tau}(\mathbf{S})) = 0 \;\forall\; \boldsymbol{\Theta} \in \Omega$. Since $\boldsymbol{\tau}(\mathbf{S})$ is invertible, it follows that $\mathrm{E}h \circ \boldsymbol{\tau}^{-1}(\boldsymbol{\tau}(\mathbf{S})) = 0 \;\forall\; \Theta \in \Omega$, where $h$ is such that $h_* = h \circ \boldsymbol{\tau}^{-1}$. But since $\mathbf{S}$ is complete, $P(h(\mathbf{s}) = 0) = P(h_*(\boldsymbol{\tau}(\mathbf{s})) = 0) = 1 \;\forall\; \Theta \in \Omega$, so that $\boldsymbol{\tau}(\mathbf{S})$ is complete by Def. 7.20. ■

The implication of Theorem 7.9 is that we can transform a set of sufficient statistics, via invertible functions, in any way that is useful or convenient. For example, the minimal sufficient statistic $\sum_{i=1}^n X_i$ in Ex. 7.11 could alternatively be defined as $\bar{X}_n$; the minimal sufficient statistic $(\sum_{i=1}^n X_i, \prod_{i=1}^n X_i)$ in Ex. 7.12 could alternatively be defined as $(\bar{X}_n, (\prod_{i=1}^n X_i)^{1/n})$; and the complete sufficient statistic $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ in Ex. 7.15 could alternatively be represented as $(\bar{X}_n, S_n^2)$. All of the alternative representations are justified by Theorem 7.9 since each of the alternative representations can be defined via invertible functions of the original sufficient statistics.

## 7.5 Results on MVUE Estimation

This section provides number of results that can aid in the search for an MVUE of $\mathbf{q}(\boldsymbol{\Theta})$. However, it should be noted at the outset that MVUEs do not always exist. When they do exist, MVUEs may be difficult to determine even with the aid of the theorems presented in this section. In Chapter 8 we will present tractable analytical procedures for deriving estimators of $\mathbf{q}(\boldsymbol{\Theta})$ that often lead to estimators with good properties and that can also lead to MVUEs. Thus, when the results in this section cannot be applied successfully in a given estimation problem, there are nevertheless useful alternative approaches that can be used to define an estimator of $\mathbf{q}(\boldsymbol{\Theta})$ for a wide range of problems of practical interest.

We begin with a theorem that identifies necessary and sufficient conditions for an unbiased estimator of a scalar $q(\Theta)$ to be a MVUE. Henceforth, we tacitly restrict the class of estimators under consideration to those with *finite variances* or *finite covariance matrices*, since estimators with infinite variances will clearly not minimize variance and in fact cannot even be compared on the basis of variance.

**Theorem 7.10** (Scalar MVUE: Necessary and Sufficient Conditions) *A necessary and sufficient condition that an estimator $T = t(\mathbf{X})$, of the scalar $q(\Theta)$, be a MVUE is that $\mathrm{cov}(T,S) = 0\ \forall\ \Theta \in \Omega$ and $\forall\ S = s(\mathbf{X}) \in \upsilon_o$, where $\upsilon_o$ is the set of functions of $\mathbf{X}$ whose expectations equal 0.*

**Proof**

**a.** Necessity: If $T$ is unbiased for $q(\Theta)$, and $\mathrm{E}S = 0$, then $T + \lambda S$ is unbiased for $q(\Theta)$. Given $\Theta$ and $\lambda$, the variance of $T + \lambda S$ equals $\mathrm{var}(T + \lambda S) = \mathrm{var}(T) + [2\lambda\,\mathrm{cov}(T, S) + \lambda^2\,\mathrm{var}(S)]$. Suppose $\mathrm{cov}(T, S) < 0$. Then the bracketed term can be made negative, and thus $\mathrm{var}(T + \lambda S) < \mathrm{var}(T)$, for $\lambda \in (0, -2\,\mathrm{cov}(T, S)/\,\mathrm{var}(S))$. Alternatively, suppose $\mathrm{cov}(T, S) > 0$. Then the bracketed term can be made negative, and thus $\mathrm{var}(T + \lambda S) < \mathrm{var}(T)$, for $\lambda \in (-2\,\mathrm{cov}(T, S)/\,\mathrm{var}(S), 0)$. Thus, $t(\mathbf{X})$ cannot be a MVUE if $\mathrm{cov}(T, S) \neq 0$.

**b.** Sufficiency: Let $T^*$ be any other unbiased estimator of $q(\Theta)$. Then $(T - T^*) \in \upsilon_o$, i.e., $(T - T^*)$ has an expectation equal to zero. Suppose that $\forall\ \Theta \in \Omega$ and $T^* \in \upsilon_{q(\Theta)}$, where $\upsilon_{q(\Theta)}$ is the class of unbiased estimators of $q(\Theta)$,

$$\mathrm{cov}(T, (T - T^*)) = \mathrm{E}(T - q(\Theta))(T - T^*) = \mathrm{E}T(T - T^*) = \mathrm{E}T^2 - \mathrm{E}TT^* = 0,$$

which is implied in the statement of the theorem upon defining $s(\mathbf{X}) = t(\mathbf{X}) - t^*(\mathbf{X})$. It follows that $\mathrm{E}T^2 = \mathrm{E}TT^*$, and thus

$$\mathrm{cov}(T, T^*) = \mathrm{E}TT^* - \mathrm{E}T\mathrm{E}T^* = \mathrm{E}T^2 - (q(\Theta))^2 = \mathrm{var}(T)$$

since both $T$ and $T^*$ are unbiased for $q(\Theta)$. The preceding result can be algebraically manipulated to yield

$$[\mathrm{var}(T)]^{1/2} = \left[\frac{\mathrm{cov}(T, T^*)}{(\mathrm{var}(T))^{1/2}(\mathrm{var}(T^*))^{1/2}}\right](\mathrm{var}(T^*))^{1/2} = \rho(\mathrm{var}(T^*))^{1/2} \le (\mathrm{var}(T^*))^{1/2}$$

since the correlation coefficient $\rho \in [0, 1]$ in this case. Thus $T$ has minimum variance. ■

The result in Theorem 7.10 facilitates the proof of the following proposition that establishes a very important relationship between *scalar* and *vector* MVUEs.

**Theorem 7.11** (Relationships Between Scalar and Vector MVUEs) *Let $\mathbf{T} = \mathbf{t}(\mathbf{X})$ be a $(k \times 1)$ vector estimator of the $(k \times 1)$ vector function $\mathbf{q}(\Theta)$. Then any of the following statements implies the others:*

1. *$\mathbf{T}$ is a MVUE for $\mathbf{q}(\Theta)$.*
2. *$\mathbf{T}_i$ is a MVUE for $q_i(\Theta)$, $i = 1,\ldots,k$.*

**3.** *$\ell'\mathbf{T}$ is a MVUE for $\ell'\mathbf{q}(\Theta)$, $\forall\, \ell \neq [0]$.*

***Proof** We prove the theorem by demonstrating that (1) ⇔ (2) and (2) ⇔ (3).

*Proof that* (2) ⇔ (3):

Assume (2). By Theorem 7.10, for $Z = \sum_{i=1}^{k} \ell_i T_i$ to be a MVUE of $\sum_{i=1}^{k} \ell_i q_i(\Theta)$, it is necessary and sufficient that $\text{cov}(Z, S) = 0\ \forall\ S \in \upsilon_o$ and $\forall\ \Theta \in \Omega$. Note that $\forall\ S \in \upsilon_o$

$$\text{cov}(Z, S) = \text{E}\left[\left[\sum_{i=1}^{k} \ell_i T_i - \sum_{i=1}^{k} \ell_i q_i(\Theta)\right] S\right] = \sum_{i=1}^{k} \text{E}\ell_i\,[T_i - q_i(\Theta)]\,S$$
$$= \sum_{i=1}^{k} \ell_i \,\text{cov}(T_i, S) = 0,$$

where the last equality follows from the fact that $T_i$ is a MVUE for $q_i(\Theta)$, and thus by Theorem 7.10 $\text{cov}(T_i, S) = 0$. From Theorem 7.10, it follows that $\sum_{i=1}^{k} \ell_i T_i$ is a MVUE of $\sum_{i=1}^{k} \ell_i q_i(\Theta)$.

Assume (3). Defining $\ell$ so that $\ell'\mathbf{T} = T_i$ ($\ell$ is a vector of zeros except for a 1 in the $i$th position), $T_i$ is a MVUE for $q_i(\Theta)$, $i = 1, \ldots, k$.

*Proof that* (1) ⇔ (2):

Assume (1). Note that $T_i = \ell'\mathbf{T}$ and $\text{var}(T_i) = \ell'\mathbf{Cov}(\mathbf{T})\ell$, where $\ell$ is a zero vector except for a 1 in the $i$th position. Clearly, $T_i$ is unbiased for $q_i(\Theta)$. Suppose $\text{var}(T_i) = \ell'\mathbf{Cov}(\mathbf{T})\ell > \text{var}(T_i^*)$ for some other unbiased estimator, $T_i^*$, of $q_i(\Theta)$. But then $\mathbf{T}$ could not be a MVUE of $\mathbf{q}(\Theta)$, since defining $\mathbf{T}^*$ equal to $\mathbf{T}$, except for $T_i^*$ replacing $T_i$, we would have $\ell'\mathbf{Cov}(\mathbf{T})\ell > \ell'\mathbf{Cov}(\mathbf{T}^*)\ell$, contradicting the fact that $\mathbf{Cov}(\mathbf{T}) - \mathbf{Cov}(\mathbf{T}^*)$ would be negative semidefinite for all other unbiased estimators $\mathbf{T}^*$ if $\mathbf{T}$ were a MVUE for $\mathbf{q}(\Theta)$. Thus (1) ⇒ (2).

Assume (2). Suppose $T_i$ is a MVUE for $q_i(\Theta)$, $i = 1, \ldots, k$. Then clearly $\text{E}\mathbf{T} = \mathbf{q}(\Theta)$, so that the vector $\mathbf{T}$ is an unbiased estimator of $\mathbf{q}(\Theta)$. Now suppose there exists another unbiased estimator, $\mathbf{T}^*$, of $\mathbf{q}(\Theta)$ such that $\ell'\mathbf{Cov}(\mathbf{T}^*)\ell < \ell'\mathbf{Cov}(\mathbf{T})\ell$ for some real vector $\ell$ and for some $\Theta \in \Omega$. Then $\ell'\mathbf{T}^*$ would be an unbiased estimator of $\ell'\mathbf{q}(\Theta)$ having a smaller variance than the unbiased estimator $\ell'\mathbf{T}$. However, the latter estimator is a MVUE of $\ell'\mathbf{q}(\Theta)$ by the earlier proof of (2) ⇔ (3). This is a contradiction, and so there is no $\Theta$ for which $\exists\,\ell$ such that $\ell'\mathbf{Cov}(T^*)\ell < \ell'\mathbf{Cov}(\mathbf{T})\ell$, i.e., $\ell'\mathbf{Cov}(\mathbf{T})\ell \leq \ell'\mathbf{Cov}(\mathbf{T}^*)\ell\ \forall\ \ell$ and $\forall\ \Theta \in \Omega$. But this implies that $\mathbf{Cov}(\mathbf{T}) \leq \mathbf{Cov}(\mathbf{T}^*)\ \forall\ \Theta \in \Omega$ and for any unbiased estimator $\mathbf{T}^*$, so that $\mathbf{T}$ is the MVUE of $\mathbf{q}(\Theta)$. ■

The primary practical implication of Theorem 7.11 is that it allows a vector MVUE of $\mathbf{q}(\Theta)$ to be defined *elementwise* with the knowledge that once each MVUE, $T_i$, for $q_i(\Theta)$ has been *individually* defined for $i = 1, \ldots, k$, then $\mathbf{T} = [T_1, \ldots, T_k]'$ is a MVUE of $\mathbf{q}(\Theta)$ in the *vector* sense. In addition, the theorem indicates that a MVUE for *any* linear combination of the entries in $\mathbf{q}(\Theta)$ is then immediately known to be $\ell'\mathbf{T}$.

Up to this point, we have been referring to "a" MVUE of $\mathbf{q}(\Theta)$, rather than to "the" MVUE of $\mathbf{q}(\Theta)$. The following theorem implies that if we have discovered a MVUE of $\mathbf{q}(\Theta)$, it is unique with probability 1. We present the result for the scalar case and then use it to provide the extension to vector situations.

**Theorem 7.12** (Uniqueness of Scalar MVUEs) *If $T$ and $T^*$ are both scalar MVUEs for $q(\Theta)$, then $P[t = t^*] = 1$.*

**Proof** If $T$ and $T^*$ are both MVUEs for $q(\Theta)$, then necessarily $\text{var}(T) = \text{var}(T^*) \ \forall \ \Theta \in \Omega$. From the sufficiency proof of Theorem 7.10, we know that $(\text{var}(T))^{1/2} = \rho(\text{var}(T^*))^{1/2}$, where $\rho$ is the correlation between $T$ and $T^*$. It also follows from the sufficiency proof of Theorem 7.10 that since $\text{var}(T) = \text{var}(T^*)$, $\rho = \text{var}(T)/\,\text{var}(T^*) = 1$. From Theorem 3.35, $\rho = 1$ implies that $P[t = a_1 + bt^*] = 1$, with $a_1 = (\mu_T\sigma_{T^*} - \mu_{T^*}\sigma_T)/\sigma_{T^*} = 0$ and $b = \sigma_T/\sigma_{T^*} = 1$, since $\mu_T = \mu_{T^*} = q(\Theta)$ and $\sigma_{T^*} = \sigma_T$, so that $P(t = t^*) = 1$. ■

**Theorem 7.13** (Uniqueness of Vector MVUEs) *If estimators $\mathbf{T}$ and $\mathbf{T}^*$ are both MVUE's for the $(k \times 1)$ vector $\mathbf{q}(\Theta)$, then $\mathbf{P}[\mathbf{t} = \mathbf{t}^*] = 1$.*

**Proof** Let $A_i$ represent the event that $T_i = T_i^*$ for $i = 1, \ldots, k$. Then since $P(\cap_{i=1}^k A_i) \geq 1 - \sum_{i=1}^k P(\bar{A}_i)$ by Bonferroni's inequality,

$$P[t_i = t_i^*, i = 1, \ldots, k] = P[\mathbf{t} = \mathbf{t}^*] \geq 1 - \sum_{i=1}^{k} P[t_i \neq t_i^*] = 1 - \sum_{i=1}^{k} 0 = 1,$$

where the next-to-last equality follows from Theorem 7.12 and from the fact that both $T_i$ and $T_i^*$ are MVUEs for $q_i(\Theta)$. ■

Given Theorems 7.12 and 7.13, we will henceforth refer to *the* MVUE of $\mathbf{q}(\Theta)$, since a MVUE of $\mathbf{q}(\Theta)$ is essentially unique.[19]

### Cramér–Rao Lower Bound

We now examine a lower bound to the covariance matrix of an unbiased estimator of $\mathbf{q}(\Theta)$. If an unbiased estimator of $\mathbf{q}(\Theta)$ can be found whose covariance matrix attains this lower bound, then the estimator is the MVUE of $\mathbf{q}(\Theta)$. The bound is called the **Cramér–Rao Lower Bound** (CRLB), and its applicability relies on a number of so-called regularity conditions on the underlying joint density function, $f(\mathbf{x}; \Theta)$, of the random sample under investigation. We state the regularity conditions below. The interpretation and application of the regularity conditions can be quite challenging, and the reader may wish to skim the discussion of these conditions on first reading.

[19] Of course, by "essentially" we mean that if two or more MVUEs do exist, their outcomes differ only on a set of probability zero and, as a practical matter, the difference can be ignored.

### *CRLB Regularity Conditions

**Definition 7.21** ***CRLB regularity conditions***

1. The parameter space, $\Omega$, of the family of densities to which $f(\mathbf{x};\Theta)$ belongs is an open rectangle.[20]
2. The support of $f(\mathbf{x};\Theta)$, $A=\{\mathbf{x}: f(\mathbf{x};\Theta)>0\}$ is the same $\forall\,\Theta\in\Omega$.
3. $\partial \ln f(\mathbf{x};\Theta)/\partial\Theta_i$ exists and is finite $\forall\, i$, $\forall\,\mathbf{x}\in A$, and $\forall\,\Theta\in\Omega$.
4. $\forall\, i$ and $j$ and $\forall$ unbiased estimator $\mathbf{t}(\mathbf{X})$ of $\mathbf{q}(\Theta)$ having a finite covariance matrix, one can differentiate under the integral or summation sign as follows.

Continuous Case:

$$\text{(a)}\quad \frac{\partial}{\partial\Theta_j}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x};\Theta)dx_1\ldots dx_n=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial f(\mathbf{x};\Theta)}{\partial\Theta_j}dx_1\ldots dx_n$$

$$\text{(b)}\quad \frac{\partial}{\partial\Theta_j}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t_i(\mathbf{x})f(\mathbf{x};\Theta)dx_1\ldots dx_n$$

$$=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t_i(\mathbf{x})\frac{\partial f(\mathbf{x};\Theta)}{\partial\Theta_j}dx_1\ldots dx_n$$

Discrete Case:

$$\text{(a)}\quad \frac{\partial}{\partial\Theta_j}\sum_{\mathbf{x}\in A}\cdots\sum f(\mathbf{x};\Theta)=\sum_{\mathbf{x}\in A}\cdots\sum\frac{\partial f(\mathbf{x};\Theta)}{\partial\Theta_j}$$

$$\text{(b)}\quad \frac{\partial}{\partial\Theta_j}\sum_{\mathbf{x}\in A}\cdots\sum t_i(\mathbf{x})f(\mathbf{x};\Theta)=\sum_{\mathbf{x}\in A}\cdots\sum t_i(\mathbf{x})\frac{\partial f(\mathbf{x};\Theta)}{\partial\Theta_j}$$

In practice, conditions (1), (2), and (3) are generally not difficult to verify, but condition (4) can be problematic. To see what is involved in verifying regularity condition (4a) for the continuous case, note that since $\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(\mathbf{x};\Theta)d\mathbf{x}=1$ because $f(\mathbf{x};\Theta)$ is a density function, then the left-hand side of the equality in (4a) is equal to zero. Then a regularity condition equivalent to (4a) is

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial f(\mathbf{x};\Theta)}{\partial\Theta_j}dx_1\ldots dx_n=0,\quad \forall\, j.$$

[20] Recall that by "open rectangle" we mean that the parameter space can be represented as $\Omega=\{(\Theta_1,\ldots,\Theta_k): a_i<\Theta_i<b_i,\ i=1,\ldots,k\}$, where any of the $a_i$'s could be $-\infty$ and any of the $b_i$'s could be $\infty$. This condition can actually be weakened to requiring only that the parameter space be an open subset of $R^k$, and not necessarily an open rectangle, and the CRLB would still apply. Such added generality is unnecessary for most applications.

Since $\partial \ln f(\mathbf{x};\Theta)/\partial\Theta_j = [f(\mathbf{x};\Theta)]^{-1}(\partial f(\mathbf{x};\Theta)/\partial\Theta_j)\ \forall\ \mathbf{x} \in A$, another regularity condition equivalent to (4a) is

$$\underset{\mathbf{x}\in A}{\int \cdots \int} \frac{\partial \ln f(\mathbf{x};\Theta)}{\partial\Theta_j} f(\mathbf{x};\Theta)dx_1 \ldots dx_n = E\left[\frac{\partial \ln f(\mathbf{X};\Theta)}{\partial\Theta_j}\right] = 0 \quad \forall\, j.$$

Thus, condition (4a) can be verified by showing that certain functions of $\mathbf{X}$ have zero expectations. By replacing integration with summation, the preceding two equivalent conditions suffice to verify condition (4a) in the discrete case as well.

Verification of regularity condition (4b) is complicated by the fact that it must hold true for *all* unbiased estimators of $\mathbf{q}(\Theta)$ with finite covariance matrices. In general, demonstrating (4b) can be a challenging exercise in real analysis. We present without proof a sufficient condition for the validity of (4b) which is based on results in advanced calculus relating to the validity of *differentiating under the integral sign* (e.g., T. M. Apostol (1974), *Mathematical Analysis*, 2nd ed. Reading, MA: Addison-Wesley, p. 167).

**Theorem 7.14** *Let CRLB regularity conditions (1)–(4a) hold. Then regularity condition (4b) holds if*

$$\mathrm{E}\left[\left(\frac{\partial \ln f(\mathbf{X};\boldsymbol{\xi})}{\partial\xi_j}\right)^2\right] < \tau(\Theta) < \infty$$

$\forall\, j$, $\forall\, \boldsymbol{\xi} \in \Gamma(\Theta)$, *and* $\forall\, \Theta \in \Omega$, *where* $\Gamma(\Theta)$ *is an open rectangle containing* $\Theta$.

Note that since $\mathrm{E}(\partial \ln f(\mathbf{X};\boldsymbol{\xi})/\partial\xi_j) = 0$ by condition (4a), Theorem 7.14 can be interpreted as a boundedness condition on the *variance* of the random variable $\partial \ln f(\mathbf{X};\boldsymbol{\xi})/\partial\xi_j\ \forall\, j$ and $\forall\, \Theta \in \Omega$. The following example illustrates the CRLB regularity verification process.

**Example 7.16** Let $(X_1,\ldots,X_n)$ be a random sample from an exponential population distribution representing the waiting time between customer arrivals at a retail store, so that

$$f(\mathbf{x};\theta) = \theta^{-n} e^{-\sum_{i=1}^{n} x_i/\theta} \prod_{i=1}^{n} I_{(0,\infty)}(x_i).$$

It is clear that the parameter space, $\Omega$, is an open rectangle (interval in this case), since $\theta > 0$ for the exponential density. The set, $A$, of $\mathbf{x}$ values that satisfy $f(\mathbf{x};\theta) > 0$ is the same for all $\theta \in \Omega$, i.e., $A = \times_{i=1}^{n}(0,\infty)$, regardless of the value of $\theta > 0$. Also, $\partial \ln f(\mathbf{x};\theta)/\partial\theta = (-n/\theta) + (\sum_{i=1}^{n} x_i/\theta^2)$ exists $\forall\ \theta \in \Omega$ and $\mathbf{x} \in A$ and is finite $\forall\ \theta \in \Omega$. Thus, CRLB regularity conditions (1)–(3) are met. Regarding condition (4), first note that

$$\mathrm{E}\left(\frac{\partial \ln f(\mathbf{X};\theta)}{\partial\theta}\right) = \mathrm{E}\left(-\frac{n}{\theta} + \frac{\sum_{i=1}^{n} X_i}{\theta^2}\right) = \frac{-n}{\theta} + \frac{n}{\theta} = 0$$

since $\mathrm{E}(\sum_{i=1}^{n} X_i) = n\theta$, and thus (4a) is met. Regarding (4b), note that

$$\mathrm{E}\left[\left(\frac{\partial \ln f(\mathbf{X};\theta)}{\partial \theta}\right)^2\right] = \mathrm{E}\left[\left(-\frac{n}{\theta} + \frac{\sum_{i=1}^{n} X_i}{\theta^2}\right)^2\right] = \mathrm{E}\left(\frac{n^2}{\theta^2} - \frac{2n\sum_{i=1}^{n} X_i}{\theta^3} + \frac{(\sum_{i=1}^{n} X_i)^2}{\theta^4}\right)$$

$$= \frac{n^2}{\theta^2} - \frac{2n^2\theta}{\theta^3} + \frac{2n\theta^2 + n(n-1)\theta^2}{\theta^4} = \frac{n}{\theta^2},$$

where the next-to-last equality follows from the fact that $\mathrm{E}X_i^2 = 2\Theta^2 \; \forall \; i$ and the fact that the $X_i$'s are iid. It follows that *whatever* the value of $\theta \in \Omega$, there exists an open interval (i.e., one-dimensional open rectangle) $\Gamma(\theta) = (\theta - \varepsilon, \theta + \varepsilon)$ for some $\varepsilon > 0$, and a positive number $\tau(\theta) = n/(\theta - \varepsilon)^2$ such that

$$\mathrm{E}\left[\left(\frac{\partial \ln f(\mathbf{X};\xi)}{\partial \theta}\right)^2\right] = \frac{n}{\xi^2} < \tau(\theta) < \infty \quad \forall \; \xi \in \Gamma(\theta),$$

so that CRLB condition (4b) is met. Thus, the CRLB regularity conditions are met for this problem. □

In practice, there are cases where it can be quite difficult to verify all of the CRLB regularity conditions. It is useful to note that there is a wide class of densities, the *exponential class* discussed in Chapter 4, that satisfies the conditions quite generally, as we state below without proof.

**Theorem 7.15** *Let $f(\mathbf{x};\Theta)$ be a member of the exponential class of densities, as defined in Def. 4.2, and let $\mathbf{c}(\Theta)$ and $d(\Theta)$ be continuously differentiable with $\partial\mathbf{c}(\Theta)/\partial\Theta$ having full rank for $\Theta \in \Omega$. Then $f(\mathbf{x};\Theta)$ satisfies the CRLB regularity conditions, provided $\Omega$ is defined to be an open rectangle.*

**Proof** A proof of the theorem can be constructed using the result in E. Lehmann (1986), *Testing Statistical Hypotheses*. New York: John Wiley, pp. 59–60. ■

**Example 7.17** Let $(X_1, \ldots, X_n)$ be a random sample from a geometric population distribution representing the number of customer contacts required for a salesperson to make her first sale, so that

$$f(\mathbf{x};p) = p^n(1-p)^{\sum_{i=1}^{n} x_i - n} \prod_{i=1}^{n} I_{\{1,2,3,\ldots\}}(x_i).$$

The PDF of the random sample is a member of the exponential class of densities with $c(p) = \ln(1-p)$, $d(p) = \ln[p/(1-p)]^n$, $g(\mathbf{x}) = \sum_{i=1}^{n} x_i$, $z(\mathbf{x}) = 0$, and $A = \times_{i=1}^{n}\{1,2,3,\ldots\}$. Note that $\partial c(p)/\partial p = -1/(1-p) \neq 0$ (and thus is a $(1 \times 1)$ matrix of full rank) and is continuous for all $p \in (0,1)$, $\partial d(p)/\partial p = n/p(1-p)$ is continuous for all $p \in (0,1)$, and $\Omega = (0,1)$ is an open interval (or rectangle). Then, by Theorem 7.15, $f(\mathbf{x};p)$ satisfies the CRLB regularity conditions. □

*The CRLB* Using the CRLB regularity conditions, we now present the CRLB.

**Theorem 7.16** (Cramér–Rao Lower Bound) *Let* $\mathbf{t}(\mathbf{X})$ *be an unbiased estimator of the* $(k \times 1)$ *vector function* $\mathbf{q}(\Theta)$ *and let* $\mathbf{t}(\mathbf{X})$ *have a finite covariance matrix. Let the* $(k \times m)$ *Jacobian matrix* $\partial \mathbf{q}(\Theta)/\partial \Theta$ *exist* $\forall\ \Theta_{(m\times 1)} \in \Omega$. *Assume the CRLB regularity conditions of Def. 7.21 hold, where* $f(\mathbf{x};\Theta)$ *is the joint density function of the random sample* $\mathbf{X}$. *Then*

$$\mathbf{Cov}_\Theta(\mathbf{t}(\mathbf{X})) \succeq \left[\frac{\partial \mathbf{q}(\Theta)}{\partial \Theta}\right]' \left[\mathrm{E}_\Theta \frac{\partial \ln f(\mathbf{X};\Theta)}{\partial \Theta}\frac{\partial \ln f(\mathbf{X};\Theta)'}{\partial \Theta}\right]^{-1} \left[\frac{\partial \mathbf{q}(\Theta)}{\partial \Theta}\right] \quad \forall\ \Theta \in \Omega,$$

*provided the inverse matrix exists.*

**Proof** We prove the result for the continuous case. The discrete case is analogous and is left to the reader.

First note that CRLB regularity condition (4a) implies that (the $\Theta$ subscripts are suppressed) $\mathrm{E}[\partial \ln f(\mathbf{X};\Theta)/\partial \Theta_i] = 0,\ i = 1,\ldots,m$. Then defining the $(m \times 1)$ random vector $s(\mathbf{X}) = [\partial \ln f(\mathbf{X};\Theta)/\partial \Theta_1, \ldots, \partial \ln f(\mathbf{X};\Theta)/\partial \Theta_m)]'$, it follows that $\mathrm{E}\mathbf{s}(\mathbf{X}) = [\mathbf{0}]$, and

$$\mathbf{Cov}(\mathbf{s}(\mathbf{X})) = \mathrm{E}\mathbf{s}(\mathbf{X})\mathbf{s}(\mathbf{X})' = \mathrm{E}\left[\frac{\partial \ln f(\mathbf{X};\Theta)}{\partial \Theta}\frac{\partial \ln f(\mathbf{X};\Theta)'}{\partial \Theta}\right].$$

Now, since $\mathbf{t}(\mathbf{X})$ is an unbiased estimator of $\mathbf{q}(\Theta)$,

$$q_i(\Theta) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t_i(\mathbf{x}) f(\mathbf{x};\Theta)dx_1\cdots dx_n,$$

so that by CRLB regularity condition (4b)

$$\begin{aligned}\frac{\partial q_i(\Theta)}{\partial \Theta_j} &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t_i(\mathbf{x})\frac{\partial f(\mathbf{x};\Theta)}{\partial \Theta_j}dx_1\cdots dx_n,\\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t_i(\mathbf{x})\frac{\partial \ln f(\mathbf{x};\Theta)}{\partial \Theta_j}f(\mathbf{x};\Theta)dx_1\cdots dx_n\\ &= \mathrm{E}\left[t_i(\mathbf{X})\frac{\partial \ln f(\mathbf{X};\Theta)}{\partial \Theta_j}\right] = \mathrm{E}[t_i(\mathbf{X})s_j(\mathbf{X})] \quad \text{for } i,j = 1,\ldots,m.\end{aligned}$$

Examine the covariance matrix of the $(k+m) \times 1$ random vector $\mathbf{Z} = \left[\begin{smallmatrix}\mathbf{t}(\mathbf{X})\\ \mathbf{s}(\mathbf{X})\end{smallmatrix}\right]$. Since $\mathrm{E}\mathbf{Z} = \left[\begin{smallmatrix}\mathbf{q}(\Theta)\\ [0]\end{smallmatrix}\right]$, it follows that

$$\begin{aligned}\mathbf{Cov}(\mathbf{Z}) &= \left[\begin{array}{c|c}\mathbf{Cov}(\mathbf{t}(\mathbf{X})) & \mathrm{E}\mathbf{t}(\mathbf{X})\mathbf{s}(\mathbf{X})' \\ \hline \mathrm{E}\mathbf{s}(\mathbf{X})\mathbf{t}(\mathbf{X})' & \mathbf{Cov}(\mathbf{s}(\mathbf{X}))\end{array}\right]\\ &= \left[\begin{array}{c|c}\mathbf{Cov}(\mathbf{t}(\mathbf{X})) & \dfrac{\partial \mathbf{q}(\Theta)'}{\partial \Theta} \\ \hline \dfrac{\partial \mathbf{q}(\Theta)}{\partial \Theta} & \mathbf{Cov}(\mathbf{s}(\mathbf{X})).\end{array}\right]\end{aligned}$$

Premultiply **Cov(Z)** by

$$\mathbf{D} = \left[ \underset{(k\times k)}{\mathbf{I}} \;\middle|\; -\frac{\partial \mathbf{q}(\Theta)'}{\partial \Theta} \underset{(k\times m)}{[\mathbf{Cov}(\mathbf{s}(\mathbf{X}))]^{-1}} \right]$$

and then postmultiply by $\mathbf{D}'$ to obtain $\mathbf{DCov(Z)D}' = \mathbf{Cov}(\mathbf{t}(\mathbf{X})) - [\partial \mathbf{q}(\Theta)/\partial\Theta]' [\mathbf{Cov}(\mathbf{s}(\mathbf{X})]^{-1}[\partial\mathbf{q}(\Theta)/\partial\Theta] \succeq [\mathbf{0}]$, where the relation $\succeq$ is due to the fact that **Cov(Z)** is necessarily positive semidefinite, and then $\mathbf{DCov(Z)D}'$ is necessarily positive semidefinite for any conformable matrix **D**.[21] Then, given the definition of **Cov(s(X))** above, the result of the theorem follows. ■

An **alternative form of the CRLB** that utilizes second order partial derivatives of $f(\mathbf{x};\Theta)$ with respect to $\Theta$ is sometimes easier to use in practice when applicable. Specifically, it is often the case that

$$\mathrm{E}\left[\frac{\partial \ln f(\mathbf{X};\Theta)}{\partial\Theta}\frac{\partial \ln f(\mathbf{X};\Theta)'}{\partial\Theta}\right] = -\mathrm{E}\left[\frac{\partial^2 \ln f(\mathbf{X};\Theta)}{\partial\Theta\partial\Theta'}\right],$$

which, when substituted into the CRLB in Theorem 7.16, provides an alternative expression for the CRLB. To see what is required for the above equality to hold, note that under the assumption that the derivatives and expectations exist,

$$\begin{aligned}\mathrm{E}\left[\frac{\partial^2 \ln f(\mathbf{X};\Theta)}{\partial\Theta_i\partial\Theta_j}\right] &= \mathrm{E}\left[\frac{\partial}{\partial\Theta_j}\left[\frac{1}{f(\mathbf{X};\Theta)}\frac{\partial f(\mathbf{X};\Theta)}{\partial\Theta_i}\right]\right]\\ &= \mathrm{E}\left[\frac{1}{f(\mathbf{X};\Theta)}\frac{\partial^2 f(\mathbf{X};\Theta)}{\partial\Theta_i\partial\Theta_j} - \frac{1}{(f(\mathbf{X};\Theta))^2}\frac{\partial f(\mathbf{X};\Theta)}{\partial\Theta_i}\frac{\partial f(\mathbf{X};\Theta)}{\partial\Theta_j}\right]\\ &= \mathrm{E}\left[\frac{1}{f(\mathbf{X};\Theta)}\frac{\partial^2 f(\mathbf{X};\Theta)}{\partial\Theta_i\partial\Theta_j} - \frac{\partial \ln f(\mathbf{X};\Theta)}{\partial\Theta_i}\frac{\partial \ln f(\mathbf{X};\Theta)}{\partial\Theta_j}\right]\end{aligned}$$

$\forall\, i, j$. Thus, if the first term in brackets in the last equality is such that[22]

$$\mathrm{E}\left[\frac{1}{f(\mathbf{X};\Theta)}\frac{\partial^2 f(\mathbf{X};\Theta)}{\partial\Theta_i\partial\Theta_j}\right] = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial^2 f(\mathbf{x};\Theta)}{\partial\Theta_i\partial\Theta_j}dx_1\cdots dx_n = 0 \quad \forall\, i,j,$$

then we can replace the matrix being inverted in Theorem 7.16 by the expectation of the negative of the second-order partial derivative matrix of $\ln f(\mathbf{X};\Theta)$ with respect to $\Theta$, i.e., we would have as the alternative form of the CRLB that

$$\mathbf{Cov}(\mathbf{t}(\mathbf{X})) \succeq \left[\frac{\partial\mathbf{q}(\Theta)}{\partial\Theta}\right]'\left[-\mathrm{E}\frac{\partial^2 \ln f(\mathbf{X};\Theta)}{\partial\Theta\partial\Theta'}\right]^{-1}\left[\frac{\partial\mathbf{q}(\Theta)}{\partial\Theta}\right].$$

To allow this substitution, note that if, $\forall\, i$ and $j$,

$$\frac{\partial}{\partial\Theta_j}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial f(\mathbf{x};\Theta)}{\partial\Theta_i}dx_1\cdots dx_n = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial^2 f(\mathbf{x};\Theta)}{\partial\Theta_i\partial\Theta_j}dx_1\ldots dx_n,$$

[21] This can be easily seen, since $\boldsymbol{\ell}'\mathbf{DCov(Z)D}'\boldsymbol{\ell} = \boldsymbol{\ell}_*'\mathbf{Cov(Z)}\boldsymbol{\ell}_* \geq 0$, where $\boldsymbol{\ell}^* = \mathbf{D}'\boldsymbol{\ell}$, $\forall\, \boldsymbol{\ell}$.

[22] We proceed for the continuous case—the discrete case is analogous, with integration replaced with summation.

then the integral on the right-hand side is zero as desired, since by the CRLB regularity conditions the left-hand side of the equality must be zero, because

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial f(\mathbf{x};\Theta)}{\partial\Theta_i}dx_1\cdots dx_n=\frac{\partial}{\partial\Theta_i}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(\mathbf{x};\Theta)dx_1\cdots dx_n=\frac{\partial}{\partial\Theta_i}(1)=0$$

and then the derivative of the constant zero is also zero. Thus, the substitution follows if we may *twice* differentiate $\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(\mathbf{x};\Theta)dx_1\ldots dx_n$ under the integral sign.

We note that the alternative form of the CRLB is valid if $f(\mathbf{x};\Theta)$ is a member of the exponential class of densities satisfying Theorem 7.15 and if $d(\Theta)$ and $\mathbf{c}(\Theta)$ are *twice* continuously differentiable with respect to $\Theta\in\Omega$ (see the Lehmann reference cited in Theorem 7.15). If $f(\mathbf{x};\Theta)$ is not in the exponential class, the admissibility of differentiating under the integral sign a second time would need to be established, or else it would need to be verified by *direct evaluation* that the integrals of the second-order derivatives of $f(\mathbf{x};\Theta)$ are indeed zero. For the sake of illustration, we provide an example of how the direct evaluation process would proceed using a normal distribution with $\sigma^2=1$ as the population distribution. The reader will note that since the normal distribution is a member of the exponential class satisfying Theorem 7.15, and since $\partial^2c(\Theta)/\partial\Theta^2$ and $\partial^2d(\Theta)/\partial\Theta^2$ exist and are continuous, we already know that the outcome of our evaluation process will result in the alternative form of the CRLB being applicable.

**Example 7.18** Let $(X_1,\ldots,X_n)$ be a random sample from a normal population distribution with $\sigma^2=1$, so that

$$f(\mathbf{x};\mu)=\frac{1}{(2\pi)^{n/2}}e^{-(1/2)\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Note that

$$\frac{\partial f(\mathbf{x};\mu)}{\partial\mu}=f(\mathbf{x};\mu)\sum_{i=1}^{n}(x_i-\mu)$$

and

$$\frac{\partial^2 f(\mathbf{x};\mu)}{\partial\mu^2}=-nf(\mathbf{x};\mu)+\left[\sum_{i=1}^{n}(x_i-\mu)\right]^2 f(\mathbf{x};\mu).$$

It follows that

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial^2 f}{\partial\mu^2}dx_1\ldots dx_n=-n\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(\mathbf{x};\mu)dx_1\cdots dx_n$$

$$+\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\sum_{i=1}^{n}(x_i-\mu)^2 f(\mathbf{x},\mu)dx_1\cdots dx_n$$

$$=-n+n=0$$

since $\sigma_i^2 = 1 \;\forall\; i$, and so the alternative form of the CRLB is applicable. Thus, in this case (the reader is encouraged to evaluate the derivatives),

$$\mathrm{E}\left[\left(\frac{\partial \ln f(\mathbf{X};\mu)}{\partial \mu}\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2 \ln f(\mathbf{X};\mu)}{\partial \mu^2}\right] = \mathrm{E}[n] = n.$$

If interest centers on estimating $\mu$ (as opposed to estimating some function of $\mu$), then $q(\mu) = \mu$ in the statement of the CRLB in Theorem 7.16, so that $\partial q(\mu)/\partial\mu = 1$, and the CRLB for the variance of any unbiased estimator of $\mu$ is equal to $n^{-1}$. □

Note that when the equality

$$\mathrm{E}\left[\left(\frac{\partial \ln f(\mathbf{X};\Theta)}{\partial \Theta_i}\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2 \ln f(\mathbf{X};\Theta)}{\partial \Theta_i^2}\right]$$

is true, the latter expectation expression can be substituted into the previously mentioned sufficient conditions for checking the validity of CRLB regularity condition (4b) (Theorem 7.14). In particular, the substitution would imply that $-\mathrm{E}[\partial^2 \ln f(\mathbf{x};\boldsymbol{\xi})/\partial\Theta_i^2] < \tau(\Theta) < \infty \;\forall\; \boldsymbol{\xi}$ in an open rectangle $\Gamma(\boldsymbol{\Theta})$ containing $\boldsymbol{\Theta}$, $\forall\; \boldsymbol{\Theta} \in \Omega$, is the condition that would need to be verified.

Two special cases of the CRLB often arise in practice. The first case occurs when $q(\Theta)$ is a scalar function of a scalar parameter $\Theta$.

**Corollary 7.3** (CRLB for Scalar Function Case) *Let $k = m = 1$ in Theorem 7.16. Then the unbiased estimator $T = t(\mathbf{X})$ of the scalar function $q(\Theta)$ of the scalar parameter $\Theta$ is such that*

$$\mathrm{var}_\Theta(t(\mathbf{X})) \geq \left[\frac{dq(\Theta)}{d\Theta}\right]^2 \Big/ \mathrm{E}_\Theta\left[\left(\frac{d \ln f(\mathbf{X};\Theta)}{d\Theta}\right)^2\right] \quad \forall\; \Theta \in \Omega.$$

The second special case is when $\mathbf{q}(\boldsymbol{\Theta}) = \boldsymbol{\Theta}$, i.e., the function defined by $\mathbf{q}(\boldsymbol{\Theta})$ is the identity function. Since in this case, $\partial\mathbf{q}(\boldsymbol{\Theta})/\partial\boldsymbol{\Theta} = \mathbf{I}$, we have the following result.

**Corollary 7.4** *Let $\mathbf{q}(\boldsymbol{\Theta}) = \boldsymbol{\Theta}$ in Theorem 7.16. Then the unbiased estimator $\mathbf{T} = \mathbf{t}(\mathbf{X})$ of the vector $\boldsymbol{\Theta}$ is such that*

$$\mathbf{Cov}_{\boldsymbol{\Theta}}(\mathbf{t}(\mathbf{X})) \succeq \left[\mathrm{E}_{\boldsymbol{\Theta}} \frac{\partial \ln f(\mathbf{X};\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} \frac{\partial \ln f(\mathbf{X};\boldsymbol{\Theta})'}{\partial \boldsymbol{\Theta}}\right]^{-1} \quad \forall\; \boldsymbol{\Theta} \in \Omega.$$

We now turn to examples of the use of the CRLB to discover MVUEs.

**Example 7.19** Let $(X_1,\ldots,X_n)$ be a random sample from a Bernoulli population distribution representing whether or not a vaccine developed by a pharmaceutical company prevents the flu, so that

$$f(\mathbf{x};p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}\prod_{i=1}^n I_{\{0,1\}}(x_i).$$

Find the MVUE of the probability that the vaccine prevents the flu.
**Answer:** Note that the support of $f(\mathbf{x};p)$ does not depend on $p \in (0,1)$, and we concentrate attention on $p \in (0,1)$ to achieve the open rectangle regularity condition on the parameter space. Also, $f(\mathbf{x};p)$ is continuous in $p$, and

$$\ln f(\mathbf{x};p) = \sum_{i=1}^{n} x_i \ln p + (n - \sum_{i=1}^{n} x_i)\ln(1-p) + \ln\left[\prod_{i=1}^{n} I_{\{0,1\}}(x_i)\right],$$

so that

$$\frac{d\ln f(\mathbf{x};p)}{dp} = \frac{\sum_{i=1}^n x_i}{p} - \frac{(n-\sum_{i=1}^n x_i)}{(1-p)}$$

exists and is finite $\forall\, p \in (0,1)$. Regarding CRLB regularity condition (4), we note that since $df(\mathbf{x};p)/dp = f(\mathbf{x};p)(d\ln f(\mathbf{x};p)/dp)$ exists $\forall\, p \in (0,1)$, and since we are dealing with *finite* sums when defining expectations, condition (4) is met. Now note that

$$\frac{d^2\ln f(\mathbf{x};p)}{dp^2} = \frac{-\sum_{i=1}^n x_i}{p^2} - \frac{(n-\sum_{i=1}^n x_i)}{(1-p)^2}$$

also exists. Furthermore, since $\sum_{x_1=0}^{1} \cdots \sum_{x_n=0}^{1} df(\mathbf{x};p)/dp$ is a sum involving a *finite* number of terms, differentiation, *under the summation sign* a second time is permissible, so that the substitution of $-\mathrm{E}[d^2 \ln f(\mathbf{X};\Theta)/d\Theta^2]$ in the CRLB, as discussed earlier, is allowed. Note that

$$\mathrm{E}\left[\frac{d^2\ln f(\mathbf{X};p)}{dp^2}\right] = \frac{-np}{p^2} - \frac{(n-np)}{(1-p)^2} = \frac{-n}{p} - \frac{n}{1-p} = \frac{-n(1-p)-np}{p(1-p)} = \frac{-n}{p(1-p)},$$

and thus the CRLB for the variance of an unbiased estimator of $q(p) = p$ is given by $p(1-p)/n$. Since $\mathrm{var}(\bar{X}_n) = p(1-p)/n$, and since $\mathrm{E}\bar{X}_n = p$, we now know that $\bar{X}_n$ is the MVUE of the probability of flu prevention, $p$. (As an exercise, the reader should alternatively demonstrate CRLB regularity by showing that $f(\mathbf{x};p)$ is a member of the exponential class, satisfying Theorem 7.15.) □

**Example 7.20** Let $(X_1,\ldots,X_n)$ be a random sample from an exponential population distribution representing the waiting time between customer arrivals at a retail store, where

$$f(\mathbf{x};\theta) = \theta^{-n} e^{-\sum_{i=1}^n x_i/\theta} \prod_{i=1}^{n} I_{(0,\infty)}(x_i).$$

Find the MVUE of the mean waiting time between customers, $\theta$.
**Answer:** Since $f(\mathbf{x};\theta)$ is a member of the exponential class, i.e.,

$$f(\mathbf{x};\theta) = e^{c(\theta)g(\mathbf{x})+d(\theta)+z(\mathbf{x})} I_A(\mathbf{x}),$$

with $c(\theta) = -\theta^{-1}$, $g(\mathbf{x}) = \sum_{i=1}^n x_i$, $d(\theta) = \ln(\theta^{-n})$, $z(\mathbf{x}) = 0$, $A = \times_{i=1}^n (0,\infty)$, and since $d(\theta)$ and $c(\theta)$ are *twice* continuously differentiable and $dc(\theta)/d\theta \neq 0$ for $\theta > 0$, we know by Theorem 7.15 that the CRLB regularity conditions are met and also that the alternative form of the CRLB applies. Note that

$$\ln f(\mathbf{x};\theta) = -n\ln\theta - \sum_{i=1}^{n} x_i/\theta + \sum_{i=1}^{n} \ln I_{(0,\infty)}(x_i),$$

so that

$$\frac{d\ln f(\mathbf{x};\theta)}{d\theta} = \frac{-n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} \quad \text{and} \quad \frac{d^2\ln f(\mathbf{x};\theta)}{d\theta^2} = \frac{n}{\theta^2} - \frac{2\sum_{i=1}^{n} x_i}{\theta^3},$$

and thus

$$\mathrm{E}\left[\frac{d^2\ln f(\mathbf{x};\theta)}{d\theta^2}\right] = \frac{n}{\theta^2} - \frac{2n\theta}{\theta^3} = \frac{-n}{\theta^2}.$$

The CRLB for the variance of an unbiased estimator of $\theta$ is then $\theta^2/n$. Since $\mathrm{E}\bar{X}_n = \theta$ and $\mathrm{var}(\bar{X}_n) = \theta^2/n$, it follows that $\bar{X}_n$ is the MVUE of the mean waiting time, $\theta$. □

*CRLB Attainment* The use of the CRLB for identifying the MVUE of $\mathbf{q}(\Theta)$ will only be useful if there actually exists an unbiased estimator whose covariance matrix equals the CRLB. A natural question to ask is under what conditions will an unbiased estimator, $\mathbf{T}$, of $\mathbf{q}(\Theta)$ have a covariance matrix that actually *achieves* the CRLB? It turns out that the CRLB is achieved only when the definition of the estimator $\mathbf{T}$ has the special form given in the following theorem.

**Theorem 7.17** (Attainment of the CRLB) *Assume the CRLB regularity conditions hold, and let* $\mathbf{T} = \mathbf{t}(\mathbf{X})$ *be an unbiased estimator of* $\mathbf{q}(\Theta)$. *Then* $\mathbf{Cov}(\mathbf{t}(\mathbf{X}))$ *equals the CRLB iff*

$$\mathbf{t}(\mathbf{X}) = \mathbf{q}(\Theta) + \frac{\partial \mathbf{q}(\Theta)'}{\partial\Theta}\left[\mathrm{E}\frac{\partial\ln f(\mathbf{X};\Theta)}{\partial\Theta}\frac{\partial\ln f(\mathbf{X};\Theta)'}{\partial\Theta}\right]^{-1}\frac{\partial\ln f(\mathbf{X};\Theta)}{\partial\Theta}$$

*with probability 1.*

**Proof** Recalling the proof of the CRLB, it is recognized that the covariance matrix of $\mathbf{t}(\mathbf{X})$ will equal the CRLB iff $\mathbf{Cov}(\mathbf{Y}) = [0]$, where the random variable $\mathbf{Y}$ is defined as

$$\mathbf{Y} = \left[\mathbf{I} \mid \frac{-\partial\mathbf{q}(\Theta)'}{\partial\Theta}[\mathbf{Cov}(\mathbf{s}(\mathbf{X}))]^{-1}\right]\begin{bmatrix}\mathbf{t}(\mathbf{X})\\ \mathbf{s}(\mathbf{X})\end{bmatrix} \quad \text{and } \mathbf{s}(\mathbf{X}) = \frac{\partial\ln f(\mathbf{X};\Theta)}{\partial\Theta}.$$

($\mathbf{Y}$ is the random variable that has the covariance matrix $\mathbf{DCov}(\mathbf{Z})\mathbf{D}'$ in the proof of Theorem 7.16, and the CRLB is attained when $\mathbf{DCov}(\mathbf{Z})\mathbf{D}' = [0]$.)

Since $\mathrm{E}\mathbf{Y} = \mathrm{E}\mathbf{t}(\mathbf{X}) = \mathbf{q}(\Theta)$, because $\mathrm{E}\mathbf{s}(\mathbf{X}) = [0]$, then setting $\mathbf{Cov}(\mathbf{Y}) = [0]$ for the attainment of the CRLB implies that $P[\mathbf{y} = \mathbf{q}(\Theta)] = 1$, which in turn implies $P[\mathbf{t}(\mathbf{x}) - (\partial\mathbf{q}(\Theta)/\partial\Theta)'[\mathbf{Cov}(\mathbf{s}(\mathbf{X}))]^{-1}\mathbf{s}(\mathbf{x}) = \mathbf{q}(\Theta)] = 1$. Upon substitution for $[\mathbf{Cov}(\mathbf{s}(\mathbf{X}))]^{-1}$ and $\mathbf{s}(\mathbf{x})$, the result is proved. ■

The CRLB attainment theorem suggests an explicit procedure for deriving the MVUE of $\mathbf{q}(\Theta)$ when it and the CRLB exist. Namely, given a statistical model $\{f(\mathbf{x}; \Theta), \Theta \in \Omega\}$, define

$$\mathbf{t}(\mathbf{X}) = \mathbf{q}(\Theta) + \frac{\partial \mathbf{q}(\Theta)'}{\partial \Theta}\left[\mathrm{E}\frac{\partial \ln f(\mathbf{X};\Theta)}{\partial \Theta}\frac{\partial \ln f(\mathbf{X};\Theta)'}{\partial \Theta}\right]^{-1}\frac{\partial \ln f(\mathbf{X};\Theta)}{\partial \Theta}$$

as a tentative estimator for $\mathbf{q}(\Theta)$. If the right-hand side does not depend on $\Theta$, i.e., a statistic is defined, then $\mathbf{t}(\mathbf{X})$ is the MVUE for $\mathbf{q}(\Theta)$ since then $\mathbf{t}(\mathbf{X})$ is an estimator that achieves the CRLB. (Note that, unbiasedness of $\mathbf{t}(\mathbf{X})$ follows immediately from the fact that $\mathrm{E}[\partial \ln f(\mathbf{X};\Theta)/\partial\Theta] = [\mathbf{0}]$ by CRLB regularity condition (4a)).

**Example 7.21** Reexamine Ex. 7.20. Because we are estimating $\theta$, we know that $dq(\theta)/d\theta = 1$. We also know from the example that $d \ln f(\mathbf{X};\theta)/\partial\theta = (-n/\theta) + (\sum_{i=1}^{n} X_i/\theta^2)$, and that

$$\mathrm{E}\left[\frac{d \ln f(\mathbf{X};\theta)}{d\theta}\frac{d \ln f(\mathbf{X};\theta)'}{d\theta}\right] = -\mathrm{E}\left[\frac{d^2 \ln f(\mathbf{X};\theta)}{d\theta^2}\right] = \frac{n}{\theta^2}.$$

Then, using Theorem 7.17, we define $\mathbf{t}(\mathbf{X})$ as

$$t(\mathbf{X}) = \theta + \frac{\theta^2}{n}\left[\frac{-n}{\theta} + \frac{\sum_{i=1}^{n} X_i}{\theta^2}\right] = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X},$$

and since $t(\mathbf{X})$ is a statistic, $t(\mathbf{X})$ is the MVUE of $\theta$. □

It turns out that we can be quite specific about the class of parametric families of density functions that will both satisfy the CRLB regularity conditions and allow an unbiased estimator of $\mathbf{q}(\Theta)$ actually to achieve the CRLB. Specifically, the exponential class of densities contains the parametric families of densities for which the CRLB is attained for an unbiased estimator of some $\mathbf{q}(\Theta)$.

**Theorem 7.18** (Exponential Class and CRLB) *Let $f(\mathbf{x};\Theta)$ satisfy the CRLB regularity conditions, and suppose there exists an unbiased estimator of $\mathbf{q}(\Theta)$ that achieves the CRLB. Then $f(\mathbf{x};\Theta)$ belongs to the exponential class of density functions.*

**Proof** The proof for the scalar parameter case can be found in P. Bickel and K. Doksum (1977), *Mathematical Statistics*. San Francisco: Holden-Day, pp. 130–131; a discussion of the multivariate case can be found in S. Zacks (1971), *The Theory of Statistical Inference*. New York: John Wiley, pp. 194–201, and in N. N. Cencov (1982), *Statistical Decision Rules and Optimal Inference*. Providence: American Mathematical Society, pp. 219–225. ■

Theorem 7.18 suggests that the procedure for defining the MVUE for $\mathbf{q}(\Theta)$ implied by Theorem 7.17 has potential for success only for statistical models where the joint density of the random sample belongs to the exponential class.

**Asymptotic Efficiency and the CRLB* To this point, we have been using the CRLB concept in the context of examining finite sample properties of estimators. The CRLB plays a prominent role in defining the concept of an asymptotically efficient estimator of $\mathbf{q}(\Theta)$, which relies on the properties of consistency and asymptotic normality as the following definition indicates.

**Definition 7.22** ***Asymptotic efficiency or best asymptotically normal (BAN)***

$\mathbf{T}_n = \mathbf{t}_n(\mathbf{X})$ is an asymptotically efficient estimator of $\mathbf{q}(\Theta)$ iff $\mathbf{t}_n(\mathbf{X})$ is a consistent estimator of $\mathbf{q}(\Theta)$ and

$$\left(\left[\frac{\partial \mathbf{q}(\Theta)}{\partial \Theta}\right]'\left[\mathrm{E}\left(\frac{\partial \ln f(\mathbf{X};\Theta)}{\partial \Theta}\frac{\partial \ln f(\mathbf{X};\Theta)'}{\partial \Theta}\right)\right]^{-1}\left[\frac{\partial \mathbf{q}(\Theta)}{\partial \Theta}\right]\right)^{-1/2}(\mathbf{T}_n - \mathbf{q}(\Theta)) \overset{d}{\to} N([0], \mathbf{I}),$$

so that

$$\mathbf{T}_n \overset{a}{\sim} N\left(\mathbf{q}(\Theta), \left[\frac{\partial \mathbf{q}(\Theta)}{\partial \Theta}\right]'\left[\mathrm{E}\left(\frac{\partial \ln f(\mathbf{X};\Theta)}{\partial \Theta}\frac{\partial \ln f(\mathbf{X};\Theta)'}{\partial \Theta}\right)\right]^{-1}\left[\frac{\partial \mathbf{q}(\Theta)}{\partial \Theta}\right]\right).$$

Thus, if $\mathbf{T}_n$ is asymptotically efficient, then it has an asymptotic normal distribution with mean $\mathbf{q}(\Theta)$ and covariance matrix equal to the CRLB, and $\mathbf{T}_n$ is then *approximately* MVUE for $\mathbf{q}(\Theta)$, based on the characteristics of its asymptotic (or approximate) probability distribution. Even more important than the preceding observation is the result by L. LeCam[23] stating that an asymptotically efficient estimator in the CAN class has an asymptotic covariance matrix that is smaller than the asymptotic covariance matrix of any other estimator of $\mathbf{q}(\Theta)$ in the CAN class except, perhaps, on a set of $\Theta$ values having Lebesque measure zero. (Recall the previous discussion of asymptotic efficiency, Def. 7.17.)

In order to illustrate the asymptotic efficiency concept, we demonstrate that in the case of sampling from an exponential density function (recall Ex. 7.20), $\bar{X}_n$ satisfies the conditions of Def. 7.22 and is thus an *asymptotically efficient* estimator of $\theta$.

**Example 7.22** Recall Ex. 7.20. Since $\mathrm{E}\bar{X}_n = \theta$, so that $\bar{X}_n$ is unbiased, and since $\mathrm{var}(\bar{X}_n) = \theta^2/n \to 0$ as $n \to \infty$, then $\bar{X}_n \overset{m}{\to} \theta$, which is sufficient for $\mathrm{plim}\,\bar{X}_n = \theta$ and so $\bar{X}_n$ is a consistent estimator of $\theta$. The CRLB for unbiasedly estimating $\theta$ in this case is given by $\theta^2/n$ (again, recall Ex. 7.20). By the Lindberg–Levy CLT, $Z_n = (\bar{X}_n - \theta)/(n^{-1/2}\theta) \overset{d}{\to} N(0, 1)$ and thus $\bar{X}_n$ is in the CAN class of estimators. Then, by Def. 7.22, $\bar{X}_n$ is an asymptotically efficient estimator of $\theta$. The asymptotic distribution of $\bar{X}_n$ can be written as $\bar{X}_n \overset{a}{\sim} N(\theta, \theta^2/n)$. □

There is a generic way of defining asymptotic probability distributions for asymptotically efficient estimators which is inherent in the definition of

[23] L. LeCam (1953), "On some asymptotic properties of maximum likelihood estimates and related Bayes estimates." *Univ. of California Publications in Statistics*, 1, pp. 277–330.

asymptotic efficiency. In particular, the asymptotic distribution of an asymptotically efficient estimator can always be specified as normal, with a mean equal to whatever is being estimated and a covariance matrix equal to the CRLB.

### Complete Sufficient Statistics and MVUEs

If complete sufficient statistics exist for the statistical model $\{f(\mathbf{x}; \Theta), \Theta \in \Omega\}$, then an alternative to the CRLB approach is available to aid in the search for the MVUE of $\mathbf{q}(\Theta)$. The approach is based on the Lehmann–Scheffé completeness theorem.

**Theorem 7.19** (Lehmann–Scheffé's Completeness Theorem) *Let $S_1,\ldots,S_r$ be a set of complete sufficient statistics for $f(\mathbf{x};\Theta)$. Let $\mathbf{T} = \mathbf{t}(S_1,\ldots,S_r)$ be an unbiased estimator for the $(k \times 1)$ vector function $\mathbf{q}(\Theta)$. Then $\mathbf{T} = \mathbf{t}(S_1,\ldots,S_r)$ is the MVUE of $\mathbf{q}(\Theta)$.*

**Proof** Let $\mathbf{S}$ be an $r$-variate complete sufficient statistic, and let $\mathbf{t}_1(\mathbf{S}) = \mathrm{E}(\mathbf{t}_1{}^*(\mathbf{X})| S_1,\ldots,S_r)$ and $\mathbf{t}_2(\mathbf{S}) = \mathrm{E}(\mathbf{t}_2{}^*(\mathbf{X})|S_1,\ldots,S_r)$ be two unbiased estimators of $\mathbf{q}(\Theta)$ that have been defined by application of the Rao–Blackwell procedure to any two unbiased estimators of $\mathbf{q}(\Theta)$, say, $\mathbf{t}_1{}^*(\mathbf{X})$ and $\mathbf{t}_2{}^*(\mathbf{X})$. Then, letting $h(\mathbf{S}) = \boldsymbol{\ell}'[\mathbf{t}_1(\mathbf{S}) - \mathbf{t}_2(\mathbf{S})]$ for any conformable $\boldsymbol{\ell}$-vector, it follows from unbiasedness that

$$\mathrm{E}h(\mathbf{S}) = \boldsymbol{\ell}'\mathrm{E}[\mathbf{t}_1(\mathbf{S}) - \mathbf{t}_2(\mathbf{S})] = \boldsymbol{\ell}'[\mathrm{E}\mathbf{t}_1(\mathbf{S}) - \mathrm{E}\mathbf{t}_2(\mathbf{S})] = 0 \quad \forall\, \Theta \in \Omega.$$

But since $\mathbf{S}$ is a set of *complete* sufficient statistics, necessarily $h(\mathbf{S}) = \boldsymbol{\ell}'[\mathbf{t}_1(\mathbf{S}) - \mathbf{t}_2(\mathbf{S})] = 0$ with probability 1 since $\boldsymbol{\ell}'[\mathbf{t}_1(\mathbf{S}) - \mathbf{t}_2(\mathbf{S})]$ is a function of the set of complete sufficient statistics that has an expectation of zero $\forall\, \Theta \in \Omega$. Since $\boldsymbol{\ell}$ can be chosen to have all zero entries except for a 1 in any position, it follows that the $(k \times 1)$ vectors $\mathbf{t}_1(\mathbf{S})$ and $\mathbf{t}_2(\mathbf{S})$ have the same $k$ elements and thus are the same unbiased estimator, say $\mathbf{t}(\mathbf{S})$, of $\mathbf{q}(\Theta)$ with probability 1, regardless of the choice of unbiased estimators $\mathbf{t}_1{}^*(\mathbf{X})$ and $\mathbf{t}_2{}^*(\mathbf{X})$ used in the definitions of $\mathbf{t}_1(\mathbf{S})$ and $\mathbf{t}_2(\mathbf{S})$, respectively. By the Rao–Blackwell theorem, $\mathbf{t}(\mathbf{S}) = \mathrm{E}(\mathbf{t}^*(\mathbf{X})|S_1,\ldots,S_r)$ must be the MVUE of $\mathbf{q}(\Theta)$, regardless of the choice of unbiased estimator $\mathbf{t}^*(\mathbf{X})$ since $\mathbf{Cov}(\mathbf{t}(\mathbf{S})) \preceq \mathbf{Cov}(\mathbf{t}^*(\mathbf{X}))$ for any choice of unbiased estimator $\mathbf{t}^*(\mathbf{X})$.

We now show that *regardless* of the procedure used to define $\mathbf{t}(\mathbf{S})$, if $\mathrm{E}\mathbf{t}(\mathbf{S}) = \mathbf{q}(\Theta)$, then $\mathbf{t}(\mathbf{S})$ is the MVUE of $\mathbf{q}(\Theta)$. Let $\mathbf{t}(\mathbf{S})$ be an unbiased estimator for $\mathbf{q}(\Theta)$. Note that, by definition, $\mathrm{E}(\mathbf{t}(\mathbf{S})|\mathbf{S}) = \mathbf{t}(\mathbf{S})$. Then, by the argument presented above, $\mathbf{t}(\mathbf{S})$ is the MVUE of $\mathbf{q}(\Theta)$ (regardless of how we were led to the definition of $\mathbf{t}(\mathbf{S})$). ■

The point of the Lehmann–Scheffé completeness theorem is that if a set of complete sufficient statistics exists for $f(\mathbf{x}; \Theta)$, then our search for the MVUE of $\mathbf{q}(\Theta)$ is complete when we have found a function of the set of complete sufficient statistics whose expectation is $\mathbf{q}(\Theta)$. Viewed another way, Theorem 7.19 implies that if $\mathbf{S}$ is complete, then for *any* function of $\mathbf{S}$, say the function defined by $\tau(\mathbf{S})$, we have that $\tau(\mathbf{S})$ is the MVUE for its expectation, i.e., $\tau(\mathbf{S})$ is the MVUE for $\mathrm{E}\tau(\mathbf{S})$.

The Lehmann–Scheffé completeness theorem then suggests two possible procedures for defining the MVUE of $\mathbf{q}(\Theta)$ if a set of complete sufficient statistics, $\mathbf{S}$, exists, which we will refer to as the **Lehmann–Scheffé MVUE approach**:

1. Find a statistic of the form $\mathbf{t}(\mathbf{S})$ such that $\mathrm{E}\mathbf{t}(\mathbf{S}) = \mathbf{q}(\Theta)$. Then $\mathbf{t}(\mathbf{S})$ is necessarily the MVUE of $\mathbf{q}(\Theta)$.
2. Find *any* unbiased estimator of $\mathbf{q}(\Theta)$, say $\mathbf{t}^*(\mathbf{X})$. Then $\mathbf{t}(\mathbf{S}) = \mathrm{E}(\mathbf{t}^*(\mathbf{X})|\mathbf{S})$ is the MVUE of $\mathbf{q}(\Theta)$.

**Example 7.23** Recall Ex. 7.14 and Ex. 7.15. Since $\sum_{i=1}^n X_i$ is a complete sufficient statistic for the statistical model of Ex. 7.14, and since $T = t(\sum_{i=1}^n X_i) = n^{-1}(\sum_{i=1}^n X_i) = \bar{X}$ is such that $\mathrm{E}T = p$, we know by the Lehmann–Scheffé completeness theorem that $\bar{X}$ is the MVUE for $p$. In Ex. 7.15, $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n X_i^2$, or alternatively by Theorem 7.9, $\bar{X}$ and $nS^2/(n-1)$ are complete sufficient statistics for the statistical model. Since $\mathrm{E}\bar{X} = \mu$ and $\mathrm{E}nS^2(n-1) = \sigma^2$, it follows by the Lehmann–Scheffé completeness theorem that $(\bar{X}, nS^2/(n-1))$ is the MVUE for $(\mu, \sigma^2)$. □

## Key Words, Phrases, and Symbols

nonparametric estimation
true value of $\Theta$
true value of $q(\Theta)$
true PDF of $X$
parametric estimation
distribution-specific case
distribution-free case
statistical model
distinct PDFs
parameter identifiability
estimating $\Theta$ or $q(\Theta)$
an estimate of $\Theta$ or $\mathbf{q}(\Theta)$
point estimator
estimand
point estimate
unbiased estimator
biased estimator
$T_i$, $T_{(n)}$
mean square error, MSE
bias
bias vector
mean square error matrix
bias matrix
relative efficiency, relatively more efficient
estimator admissibility
strong mean-square error (SMSE) criterion
$\mathbf{A} \preceq \mathbf{B}$, matrix $\mathbf{A}$ is smaller than matrix $\mathbf{B}$
Minimum Variance Unbiased Estimator, MVUE
best linear unbiased estimator, BLUE
minimum variance linear unbiased estimator, MVLUE
efficient
asymptotic MSE
asymptotic relative efficiency
asymptotically relatively more efficient
consistent estimator
consistency
asymptotically unbiased
asymptotic efficiency
sufficient statistics
Neyman factorization theorem
minimal sufficient statistics
range of $X$ over the parameter space $\Omega$, $R_\Omega(X)$
Lehmann–Scheffé minimal sufficiency theorem
exponential class and sufficient statistics
Rao–Blackwell theorem
complete sufficient statistics
completeness in the exponential class
Cramér–Rao lower bound, CRLB
CRLB regularity conditions
alternative form of the CRLB
Lehmann–Scheffé completeness theorem
Lehmann–Scheffé MVUE approach

## Problems

**1.** The operating life of a certain brand of math coprocessor installed in a personal computer can be represented as the outcome of a random variable having an exponential density function, as

$$Z \sim f(z; \theta) = \frac{1}{\theta} e^{-z/\theta} I_{(0,\infty)}(z),$$

where $z$ = the number of hours the math coprocessor functions until failure, measured in thousands of hours.

A random sample, $(X_1, \ldots, X_n)$, of the operating lives of 200 coprocessors is taken, where the objective is to estimate a number of characteristics of the operating-life distribution of the coprocessors. The outcome of the sample mean was $\bar{x} = 28.7$.

a. Define a minimal sufficient statistic for $f(\mathbf{x}; \theta)$, the joint density of the random sample.

b. Define a complete sufficient statistic for $f(\mathbf{x}; \theta)$.

c. Define the MVUE for $\mathrm{E}Z = \theta$ if it exists. Estimate $\theta$.

d. Define the MVUE for $\mathrm{Var}(Z) = \theta^2$ if it exists. Estimate $\theta^2$.

e. Define the MVUE for $\mathrm{E}Z^2 = 2\theta^2$ if it exists. Estimate $2\theta^2$.

f. Define the MVUE for $q(\theta)_{(3\times 1)} = \begin{bmatrix} \theta \\ \theta^2 \\ 2\theta^2 \end{bmatrix}$ if it exists. Estimate $\mathbf{q}(\theta)$.

g. Is the second sample moment about the origin, i.e., $M_2' = \sum_{i=1}^n X_i^2/n$, the MVUE for $\mathrm{E}Z^2$?

h. Is the sample variance, $S^2$, the MVUE for $\mathrm{var}(Z)$?

i. Suppose we want the MVUE for $F(b) = P(z \le b) = 1 - e^{-b/\theta}$, where $F(b)$ is the probability that the coprocessor fails before 1000 $b$ hours of use. It can be shown that

$$t(\mathbf{X}) = 1 - \left(1 - \frac{b}{(\sum_{i=1}^n X_i)}\right)^{n-1} I_{[b,\infty)}\left(\sum_{i=1}^n X_i\right)$$

is such that $\mathrm{E}t(\mathbf{X}) = 1 - e^{-b/\theta}$. Is $t(\mathbf{X})$ the MVUE for $P(z \le b)$? Why or why not? Estimate $P(z \le 20)$.

j. Is $t_*(\mathbf{X}) = 1 - e^{-b/\bar{X}}$ a MVUE for $F(b)$? Is $t_*(\mathbf{X})$ a consistent estimator of $F(b)$?

**2.** Use the Lehmann–Scheffé minimal sufficiency theorem, or some other argument, to find a set of minimal sufficient statistics for each case below.

a. You are random sampling from a log-normal population distribution given by

$$f(z; \mu, \sigma^2) = \frac{1}{(2\pi)^{1/2}\sigma z} \exp\left(-\frac{1}{2\sigma^2}(\ln z - \mu)^2\right) I_{(0,\infty)}(z),$$

where $\mu \in (-\infty, \infty)$ and $\sigma^2 > 0$.

b. You are random sampling from a "power function" population distribution given by

$$f(z; \lambda) = \lambda z^{\lambda-1} I_{(0,1)}(z), \quad \text{where } \lambda > 0.$$

c. You are random sampling from a Poisson population distribution

$$f(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!} I_{\{0,1,2,\ldots\}}(x).$$

d. You are random sampling from a negative binomial density

$$f(x; r_0, p) = \frac{(x-1)!}{(r_0-1)!(x-r_0)!} p^{r_0}(1-p)^{x-r_0} I_{\{r_0, r_0+1,\ldots\}}(x),$$

where $r_0$ is a *known* positive integer.

e. You are random sampling from an $N(\mu, \sigma^2)$ population distribution.

f. You are random sampling from a continuous uniform density

$$f(x; \Theta) = \frac{1}{\Theta} I_{(0,\Theta)}(x).$$

g. You are sampling from a Beta distribution

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{(0,1)}(x).$$

**3.** Identify which of the minimal sufficient statistics in (2) are *complete* sufficient statistics.

**4.** The operating life of a small electric motor manufactured by the AJAX Electric Co. can be represented as a random variable having a probability density given as

$$Z \sim f(z; \Theta) = \frac{1}{6\Theta^4} z^3 e^{-z/\Theta} I_{(0,\infty)}(z),$$

where $\Theta \in \Omega = (0, \infty)$, $\mathrm{E}Z = 4\Theta$, $\mathrm{var}(Z) = 4\Theta^2$, and $z$ is measured in thousand of hours. A random sample $(X_1, \ldots, X_{100})$ of the operating lives of 100 electric motors has an outcome that is summarized as $\bar{x} = 7.65$ and $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2/100 = 1.73$.

a. Define a minimal, complete sufficient statistic for estimating the expected operating life of the electric motors produced by the AJAX Co.

b. Define the MVUE for estimating $EZ = 4\Theta$. Justify the MVUE property of your estimator. Generate an estimate of $EZ$ using the MVUE.

c. Is the MVUE estimator a consistent estimator of $EZ = 4\Theta$? Why or why not?

d. Does the variance of the estimator you defined in (b) attain the Cramér–Rao lower bound? (The CRLB regularity conditions hold for the joint density of the random sample. Furthermore, the alternative form of the CRLB, expressed in terms of second-order derivatives, applies in this case if you want to use it.)

**5.** The number of customers that enter the corner grocery store during the noon hour has a Poisson distribution, i.e.,

$$f(z;\lambda) = \frac{e^{-\lambda}\lambda^z}{z!} I_{\{0,1,2,3,\ldots\}}(z).$$

Assume that $(X_1, X_2, \ldots, X_n)$ is a random sample from this Poisson population distribution.

a. Show that the Cramér–Rao lower bound regularity conditions hold for the joint density of the random sample.

b. Derive the CRLB for unbiased estimation of the parameter $\lambda$. Is $\bar{X}$ the MVUE for estimating $\lambda$? Why or why not?

c. Use the CRLB attainment theorem to derive the MVUE for estimating $\lambda$. Suppose $n = 100$ and $\sum_{i=1}^{100} x_i = 283$. Estimate $\lambda$ using the MVUE.

d. Is $\bar{X}$ a member of the CAN class of estimators? Is $\bar{X}$ asymptotically efficient?

e. Define the CRLB for estimating $P(z = 0) = e^{-\lambda}$. Does there exist an unbiased estimator of $e^{-\lambda}$ that achieves the CRLB? Why or why not?

**6.** Polly Pollster wants to estimate the proportion of voters in the state of Washington that are in favor of an antitax initiative. She will be using a random sample (with replacement) of 1,000 voters, and she will record their preference regarding the antitax initiative. She needs some statistical advice from you.

a. Define a statistical model for the problem of estimating the proportion of voters in favor of the initiative.

b. Define the MVLUE for the proportion of voters in favor of the initiative. Justify that your estimator really is a MVLUE.

c. Is the estimator that you defined in (b) a consistent estimator of the proportion of voters in favor of the initiative? Is it a CAN estimator? Is it asymptotically efficient?

d. Assume there are two million voters in the state. What is the probability that the estimator you defined in (b) generates an estimate that is within ±.03 of the true proportion of voters favoring the initiative? (Note: This may be a function of unknown parameters!)

e. Polly summarized the outcome of the random sample as $\sum_{i=1}^{1000} x_i = 670$, where $x_i = 1$ indicates that the $i$th sample voter was in favor of the initiative, and $x_i = 0$ otherwise. Estimate the proportion of voters in favor of the initiative. Given your result in (d), if the election were held today, would you predict that the initiative would pass? Why or why not?

**7.** Two economics professors are arguing about the appropriate estimator to use in estimating the mean of the population distribution of incoming freshmen's I.Q.'s. The estimators will be based on a random sample from the population distribution. One professor suggests that they simply calculate the sample mean I.Q., $\bar{x}_n$, and use it as an estimate of the mean of the population distribution. The other prefers to use an estimator of the form $t(\mathbf{X}) = \sum_{i=1}^{n} X_i/(n+k)$, where $n$ is the random sample size and $k$ is some positive integer, and she argues that her estimator has less variance than the sample mean and that for an appropriate choice of $k$, her estimator would be superior on the basis of MSE.

a. We know that $\bar{X}$ is unbiased, asymptotically unbiased, BLUE, and consistent for estimating the mean of the population distribution. Which of these properties apply to the alternative estimator?

b. Define asymptotic distributions for both estimators. On the basis of their asymptotic distributions, do you favor one estimator over the other?

c. Define the MSEs of the estimators. Is there any validity to the statement that "for an appropriate choice of $k$" $t(\mathbf{X})$ will be superior to $\bar{X}$ in terms of MSE? Explain.

d. Can you foresee any practical problems in using $t(\mathbf{X})$ to generate estimates of the population mean?

**8.** The diameters of blank compact disks manufactured by the Dandy Disk Co. can be represented as outcomes of a random variable

$$Z \sim f(z;\Theta) = \frac{1}{\Theta} I_{[4,4+\Theta]}(z), \quad \text{for some } \Theta > 0,$$

where $z$ is measured in inches. You will be using a random sample $(X_1, X_2, \ldots, X_n)$ from the population distribution $f(z;\Theta)$ to answer the questions below.

a. Based on the random sample, define an unbiased estimator of the parameter $\Theta$.

b. Is the estimator you defined in (a) a BLUE for $\Theta$? If not, find a BLUE for $\Theta$, if it exists.

c. Is your estimator a consistent estimator for $\Theta$? Why or why not?

d. Define an asymptotic distribution for your estimator.

e. A random sample of size $n = 1000$ from $f(z;\Theta)$ results in $\sum_{i=1}^{1000} x_i = 4{,}100$. Use your estimator to estimate the value of $\Theta$. Using your estimate of $\Theta$, what is the estimated probability that $z \in (4.05, 4.15)$?

**9.** Your company sells trigger mechanisms for air bags that are used in many modern domestic and foreign-built passenger cars. The reliability of such trigger mechanisms is obviously critical in the event that an air bag-equipped vehicle is involved in an accident. One large Detroit automobile manufacturer said that it would be willing to purchase trigger mechanisms from your company if you could provide convincing support for the statement that, in repeated simulations of an impact of 15 mph, the expected number of impacts needed to obtain the first failure of your trigger (i.e., the trigger does not signal an air bag deployment) was greater than or equal to 1,000.

Management has randomly chosen 10,000 of the trigger mechanisms for a testing program in which the number of simulated impacts needed to obtain the first failure will be observed for each mechanism. You need to estimate the expected number of impacts needed to obtain the first failure of the trigger mechanisms you manufacture. You intend to use the outcome of the sample mean as your estimate of this expected number of impacts.

a. Define an appropriate statistical model for the sampling experiment, and justify your choice.

b. Is the sample mean an unbiased estimator in this case? Why?

c. Is the sample mean an asymptotically unbiased estimator? Why?

d. Is the sample mean a consistent estimator? Why?

e. Is the sample mean a BLUE (or equivalently, a MVLUE)? Why?

f. Derive the Cramér–Rao lower bound for the variance of unbiased estimators of the expected number of impacts to obtain the first failure. (Hint: You may use the alternative (second-derivative) form of the bound—it might be a little easier to work with in this case.)

g. Is the sample mean a MVUE? Why?

h. Use Theorem 7.17 on the attainment of the Cramér–Rao lower bound to derive the MVUE of the expected number of impacts needed to obtain the first failure.

i. Define an appropriate asymptotic distribution for the sample mean in this case. Is the sample mean asymptotically efficient?

j. The 10,000 observations resulted in $\sum_{i=1}^{10{,}000} x_i = 1.5 \times 10^7$. What is your estimate of the expected number of impacts needed to obtain the first failure?

k. Is $(\bar{X}_n)^{-1}$ a consistent estimator of $p$, the probability that a trigger successfully signals deployment of the air bag on any given trial?

l. Define an asymptotic distribution for the estimator $(\bar{X}_n)^{-1}$ of $p$. Is the estimator asymptotically efficient?

m. Use the estimator $(\bar{X}_n)^{-1}$ to estimate $p$, and use this estimated value in the asymptotic distribution you defined for $\bar{X}_n$ to estimate the probability that an estimate generated by $\bar{X}_n$ would be within ±50 units of the population mean.

n. What, if anything, can you say to the Detroit manufacturer to convince the company to buy your trigger mechanisms?

**10.** Suppose random sampling is from an exponential population distribution representing the waiting time between customer arrivals at a bank, so that the statistical model is given by

$$f(\mathbf{x};\theta) = \theta^{-n} e^{-\sum_{i=1}^{n} x_i/\theta} \prod_{i=1}^{n} I_{(0,\infty)}(x_i)$$

for $\theta > 0$.

Your objective is to use an outcome of a random sample of size $n$ ($n$ fixed and known) to estimate $q(\theta) = \theta^2$, the variance of waiting times.

a. Define a complete sufficient statistic for $f(\mathbf{x}; \theta)$.

b. Define the CRLB for unbiasedly estimating $\theta^2$.

c. Does there exist an unbiased estimation of $\theta^2$ whose variance is equal to the CRLB?

d. Define the MVUE of $\theta^2$ by finding an appropriate function of the complete sufficient statistic—if you can.

e. Is the sample variance, $S_n^2$, the MVUE for the population variance $\theta^2$?

f. If $\sum_{i=1}^{100} x_i = 257$, generate a MVUE estimate of the variance of waiting times.

**11.** One hundred one-acre test plots are being used to assess the yield potential of a new variety of wheat genetically engineered by Washington State University. The actual yield-per-acre observations on the test plots can be viewed as iid observations from some log-normal population, so that the statistical model for the experiment is given by

$$f(\mathbf{y}; \mu, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n \prod_{i=1}^n y_i} \times \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (\ln y_i - \mu)^2\right] \prod_{i=1}^n I_{(0,\infty)}(y_i)$$

for $\mu \in (-\infty, \infty)$ and $\sigma > 0$.

a. Define minimal sufficient statistics for $f(\mathbf{y}; \mu, \sigma^2)$.

b. Are the sufficient statistics you defined in (a) *complete* sufficient statistics?

c. Is $t_1(\mathbf{Y}) = n^{-1}\sum_{i=1}^n \ln Y_i$ the MVUE of the parameter $\mu$? Is it consistent? Why?

d. Is $t_2(\mathbf{Y}) = (n-1)^{-1}\sum_{i=1}^n (\ln Y_i - n^{-1}\sum_{i=1}^n \ln Y_i)^2$ the MVUE of the parameter $\sigma^2$? Is it consistent?

e. Define a consistent estimator of $q(\mu, \sigma) = e^{\mu+\sigma^2/2}$, which is the mean of the log-normal population. Justify your answer. (The MVUE of the mean exists, but it is quite complicated to define and calculate. See D. J. Finney (1941), "On the distribution of a variate whose logarithm is normally distributed." *Roy. Statistical Society*, Series B, 7, pp. 155–161.)

f. An overworked, underpaid, gaunt-looking research assistant hands you an envelope that contains only summary information on the results of the experiment. In particular, the information is that

$$\sum_{i=1}^{100} \ln Y_i = 375.00,$$

$$\sum_{i=1}^{100} (\ln Y_i)^2 = 1455.75.$$

Generate an estimate of $q(\mu, \sigma^2)_{(2\times 1)} = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$ using the MVUE of $q(\mu, \sigma^2)$. Generate an estimate of the mean of the lognormal distribution using a consistent estimator.

**12.** In each case below, determine whether the estimator under consideration is unbiased, asymptotically unbiased, and/or consistent.

a. The random sample $(X_1, \ldots, X_n)$ is generated from a Gamma population distribution. The estimator $t(\mathbf{X}) = \sum_{i=1}^n X_i/n$ will be used to estimate $\mathrm{E}X_i = \alpha\beta$.

b. The random sample $(X_1, \ldots, X_n)$ is generated from an exponential population distribution. The estimator $t(\mathbf{X}) = (1/2)\sum_{i=1}^n X_i^2/n$ will be used to estimate $\mathrm{var}(X_i) = \theta^2$.

c. The random sample $(X_1, \ldots, X_n)$ is generated from a geometric population distribution. The estimator $t(\mathbf{X}) = (S^2 + \bar{X})$ will be used to estimate $\mathrm{E}X_i^2 = p^{-2}$.

d. The random sample $(X_1, \ldots, X_n)$ is generated from a Bernoulli population distribution. The estimator $t(\mathbf{X}) = \bar{X}(1 - \bar{X})$ will be used to estimate $\mathrm{var}(X_i) = p(1-p)$. (Hint: $2\sum_{i=1}^n \sum_{j>i}^n a = n(n-1)a$).

**13.** In problem 12(b), above, consider the alternative estimator $t^*(\mathbf{X}) = S^2$ for estimating $\theta^2$. In a mean square error sense, which estimator would you prefer for estimating $\theta^2$, $t(\mathbf{X})$, or $t^*(\mathbf{X})$? Note that $\mu_r' = r!\theta^r$ for an exponential PDF.

**14.** An incoming shipment of 1000 toys from a toy manufacturer is received by a large department store for a pre-Christmas sale. The store randomly samples 50 toys from the shipment, *without replacement*, and records whether or not the sample item is defective. The store wants to generate a MVUE estimate of the proportion of defectives in the shipment of toys. The statistical model it uses for the sampling experiment is given by the hypergeometric density with the following parameterization:

$$f(x; \Theta) = \frac{\binom{1000\Theta}{x}\binom{1000(1-\Theta)}{50-x}}{\binom{1000}{50}} I_{\{0,1,2,\ldots,50\}}(x),$$

where $\Theta \in \Omega = \{0, .001, .002, \ldots, 1\}$ represents the proportion of defectives in the shipment.

*a. Show that $X \sim f(x; \Theta)$ is a minimal, complete sufficient statistic for $f(x; \Theta)$.

b. Define the MVUE for the proportion of defectives in the shipment.

c. Suppose that the outcome of $X$ was 3. Define a MVUE estimate of the proportion of defectives in the shipment.

d. Define an MVUE for the *number* of defective toys in the shipment, and provide an MVUE estimate of this number.

**15.** The rates of return per dollar invested in two common stocks over a given investment period can be viewed as the outcome of a bivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The rates are independent between investment periods. An investment firm intends to use a random sample from the $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population distribution of rates of return to generate estimates of the expected rates of return, $\boldsymbol{\mu}$, as well as the variances in the rates of return, given by the diagonal of $\boldsymbol{\Sigma}$.

a. Find a minimal, complete (vector) sufficient statistic for $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

b. Define the MVUE for $\boldsymbol{\mu}$.

c. Define the MVUE for $\boldsymbol{\Sigma}$ and for diag($\boldsymbol{\Sigma}$).

d. Define the MVUE for the vector $[\mu_1, \mu_2, \sigma_1^2, \sigma_2^2]'$.

e. A random sample of size 50 has an outcome that is summarized by $\bar{x} = [.048 \;\; .077]'$, $s_1^2 = .5 \times 10^{-3}$, $s_2^2 = .3 \times 10^{-4}$, and $s_{12} = .2 \times 10^{-4}$. Calculate the MVUE outcome for $[\mu_1, \mu_2, \sigma_1^2, \sigma_2^2]'$.

f. Is the MVUE of $[\mu_1, \mu_2, \sigma_1^2, \sigma_2^2]'$ consistent?

g. If an investor invests \$500 in each of the two investments, what is the MVUE of her expected dollar return on the investment during the investment period under consideration?

# 8 Point Estimation Methods

## 8.1 Introduction

In our discussion of point estimation, we have yet to present any generally applicable procedures that lead to specific functional forms for estimators of $\mathbf{q}(\Theta)$ and that can be relied upon to define estimators that generally have good estimator properties. In fact, the only result in the previous chapter that could be used directly to define the functional form of an estimator is the theorem on the attainment of the CRLB, which is useful only if the statistical model $\{f(\mathbf{x}; \Theta), \Theta \in \Omega\}$ and the estimand $\mathbf{q}(\Theta)$ are such that the CRLB is actually attainable. Of course, we did examine a number of important results that could be used either to narrow the search for a good estimator of $\mathbf{q}(\Theta)$, to improve upon an unbiased estimator that we already discovered, or to recognize when an unbiased estimator was actually the best in the sense of minimizing variance (or in the sense of having the smallest covariance matrix). However, since the functional form of an estimator of $\mathbf{q}(\Theta)$ having good estimator properties is often not apparent even with the aid of the results assembled in Chapter 7, we now examine procedures that suggest functional forms of estimators.

Currently, there is no single procedure for generating estimators of $\mathbf{q}(\Theta)$ that will always lead to the best estimator or even to an estimator that always has "good" properties. In this chapter, we will examine three specific estimation procedures that, under appropriate circumstances, will often lead to good estimators of $\mathbf{q}(\Theta)$.

## 8.2 Least Squares and the General Linear Model

The **general linear model** (GLM) is a specific form of model used to characterize how the outcomes of a random sample are generated. In particular, the GLM decomposes the $i$th random variable, $Y_i$, in the random sample $(Y_1, \ldots, Y_n)$ into the sum of its expectation and the deviation from its expectation, as

$$Y_i = \mu_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\mathrm{E}Y_i = \mu_i$, so that $\mathrm{E}\varepsilon_i = 0$. At this level of generality, the decomposition is *always* valid as long as $\mathrm{E}Y_i$ exists. The GLM specializes this representation by further assuming that $\mu_i$ is defined via a function of the $m$ explanatory variables contained in the $(m \times 1)$ vector $\mathbf{z}_i$ and that this function is known except for $k$ unknown parameters that enter the function *linearly*, leading to the representation

$$\mu_i = \sum_{j=1}^{k} h_j(\mathbf{z}_i)\beta_j, \quad i = 1, \ldots, n.$$

We emphasize that *linearity* in the general *linear* model refers to the assumption that the function of the explanatory variables defining $\mu_i$ is **linear in parameters.** The representation need not be linear in the explanatory variables. However, it is customary to define the variables $x_{ij} = h_j(\mathbf{z}_i)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, k$, so that the representation of the mean $\mu_i$ is then also linear in the $x_{ij}$'s, as

$$\mu_i = \sum_{j=1}^{k} x_{ij}\beta_j, \quad i = 1, \ldots, n.$$

A compact matrix representation of the GLM in terms of the $x_{ij}$ variables can then be defined as

$$\underset{(n\times 1)}{\mathbf{Y}} = \underset{(n\times k)}{\mathbf{x}}\ \underset{(k\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

The GLM can be used to represent a random sample $\mathbf{Y}$ whether the random sample is from a population distribution or from a composite experiment. If the elements of $\mathbf{Y}$ are iid random variables from a population distribution, then $\mathbf{x}$ is a column vector of 1's, $\beta$ is a scalar representing the common mean of the $Y_i$'s, and the $\varepsilon_i$'s are iid with mean zero. More generally, if the elements of $\mathbf{Y}$ are generated by a composite experiment, so that the $Y_i$'s need not be identically distributed nor independent, then linear combinations of the elements in the rows of $\mathbf{x}$ are used to represent the various mean values of the $Y_i$'s, and the $\varepsilon_i$'s have a joint probability distribution with a zero mean vector and appropriate variances and covariances.

In order to illustrate the generality of the GLM, note that

$$Y_i = \beta_1 + \beta_2 z_{i2}^2 + \beta_3 \sin(z_{i3}) + \beta_4 \left(\frac{z_{i4}}{z_{i5}}\right) + \varepsilon_i, \quad i = 1, \ldots, n,$$

is consistent with the GLM, and a representation that is linear in explanatory variables can be obtained upon defining $x_{i1} = 1$, $x_{i2} = z_{i2}^2$, $x_{i3} = \sin(z_{i3})$, and $x_{i4} = (z_{i4}/z_{i5})$. More generally, the GLM can serve as an approximation in situations where $\mathrm{E}Y_i = \mu(z_{i1}, \ldots, z_{im})$ is a *continuous* function of $(z_{i1}, \ldots, z_{im})$, at least *in principle*, since by Weierstrass's approximation theorem any continuous function can be approximated arbitrarily closely by a polynomial of sufficiently high degree (see Bartle, Real Analysis, pp. 185–186). That is, if $\mu(z_{i1}, \ldots, z_{im})$ represents a continuous function, then

$$\mu(z_{i1}, \ldots, z_{im}) \approx \sum \cdots \sum_{(j_1, \ldots, j_m) \in A} c_{j_1 \cdots j_m} z_{i1}^{j_1} z_{i2}^{j_2} \cdots z_{im}^{j_m},$$

where $A = \{(j_1, \ldots, j_m) \colon 0 \leq j_1 + \cdots + j_m \leq d,\ j_i\text{'s are nonnegative integers}\}$, and $d$ represents the degree of the polynomial. Letting the $c_{j_1 \ldots j_m}$'s be the entries in the $\boldsymbol{\beta}$ vector, and the product terms $(z_{i1}^{j_1} z_{i2}^{j_2} \ldots z_{im}^{j_m})$ be the entries in the $\mathbf{x}$ matrix, it is clear that the polynomials in the Weierstrass approximation theorem can be incorporated into the GLM framework.[1]

Finally, a relationship between $\mathbf{Y}$, $\mathbf{x}$, and $\boldsymbol{\varepsilon}$ that is initially *nonlinear* in the parameters might be *transformable* into the GLM form. For example, suppose

$$Y_i = \beta_1 k_i^{\beta_2} \ell_i^{\beta_3} \varepsilon_i$$

which might represent a Cobb–Douglas production function with $y_i$ being output, and $k_i$ and $\ell_i$ being capital and labor inputs, respectively. Applying a logarithmic transformation obtains

$$\ln Y_i = \ln \beta_1 + \beta_2 \ln k_i + \beta_3 \ln \ell_i + \ln \varepsilon_i,$$

or

$$Y_i^* = \beta_1^* + \beta_2 k_i^* + \beta_3 \ell_i^* + \varepsilon_i^*,$$

which is in the GLM form, assuming $Y_i^* = \ln Y_i$, $\beta_1^* = \ln \beta_1$, $k_i^* = \ln k_i$, $\ell_i^* = \ln \ell_i$, and $\mathrm{E}\varepsilon_i^* = \mathrm{E} \ln \varepsilon_i = 0 \ \forall\, i$.

### The Classical GLM Assumptions

In the GLM, $\mathbf{Y}$ is referred to as the **dependent variable** vector, $\mathbf{x}$ is called the matrix of **independent or explanatory variables**, and $\boldsymbol{\varepsilon}$ is called the **disturbance**,

[1] We note, however, that for highly nonlinear functions, the degree of polynomial required to provide an adequate approximation to $\mu(z_{i1}, \ldots, z_{im})$ may be so high that there will not be enough sample observations to estimate the unknown $\beta_j$'s. This is related to a requirement that $\mathbf{x}$ have full column rank, which will be discussed shortly.

**error, or residual vector.** The main objective in analyzing the GLM will be to estimate (and later, test hypotheses about) the entries and/or functions of the entries in the parameter vector $\boldsymbol{\beta}$. In particular, we will utilize an outcome, $(y_1, \ldots, y_n)$, of the random sample, $(Y_1, \ldots, Y_n)$, together with knowledge of the values of the explanatory variables, $\mathbf{x}$, to estimate the unknown entries in $\boldsymbol{\beta}$. Note that the entries in $\boldsymbol{\varepsilon}$ are *unobservable* random variables, since they represent deviations of $(y_1, \ldots, y_n)$ from the *unknown* mean vector $\mathbf{x}\boldsymbol{\beta}$, and thus outcomes of $\boldsymbol{\varepsilon}$ will not be useful in estimating $\boldsymbol{\beta}$, per se.

At this point, we have not yet defined the characteristics of the GLM sufficiently for point estimation purposes. The following set of assumptions, which will be referred to as the **classical assumptions of the GLM**, are made.

**Definition 8.1**
***Classical assumptions of the GLM***

1. $\mathrm{E}\mathbf{Y} = \mathbf{x}\boldsymbol{\beta}$ and $\mathrm{E}\boldsymbol{\varepsilon} = [\mathbf{0}]$.
2. $\mathbf{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I} = \mathbf{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'$.
3. $\mathbf{x}$ is nonrandom with rank $\mathbf{x} = k$ (full column rank).

The first assumption simply restates what we have already asserted concerning the mean of $\mathbf{Y}$, namely, $\mathbf{x}\boldsymbol{\beta}$ is representing or explaining $\mathrm{E}\mathbf{Y}$, which then necessarily implies that $\mathrm{E}\boldsymbol{\varepsilon} = [\mathbf{0}]$. In practice, $\mathrm{E}Y_i = \mathbf{x}_{i.}\boldsymbol{\beta}$ (recall that $\mathbf{x}_{i.}$ refers to the $i$th row of the $\mathbf{x}$ matrix) means that no matter what real-world experiment is under investigation, the expected value of the distribution associated with outcomes of $Y_i$ is a linear function of $\mathbf{x}_{i.}$. Graphically, for $k = 2$ with $\mathbf{x}_{.1} = [1\ 1\ \ldots\ 1]'$, we have the representation in Figure 8.1.

Note that under the classical assumptions, $\mathbf{x}$ is a fixed matrix of numbers. Regarding the experiments under consideration, this implies that an observa-

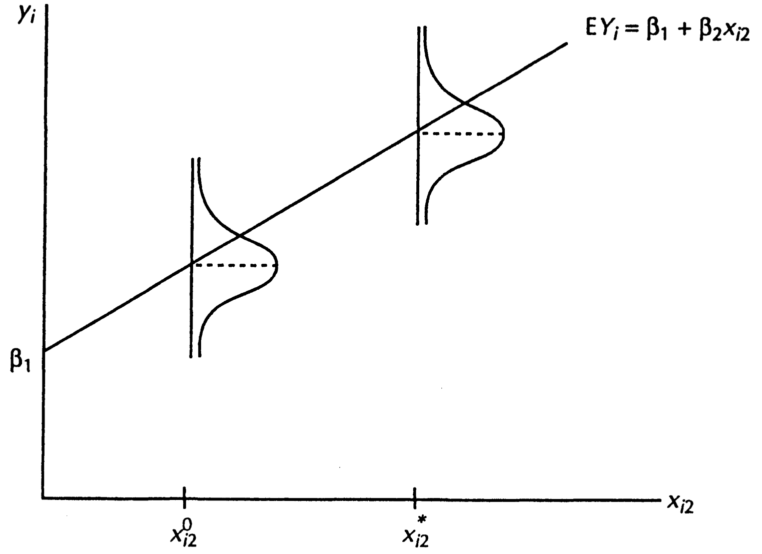


**Figure 8-1**
GLM representation of $\mathrm{E}Y_i$ in bivariate case.

tion $(y_i, x_{i1}, \ldots, x_{ik})$ can be interpreted as having been generated from an outcome of the vector $(Y_i, x_{i1}, \ldots, x_{ik})$, where this latter vector can be considered a random vector, pairing the (generally) *nondegenerate* random variable $Y_i$ with the *degenerate* random vector $(x_{il}, \ldots, x_{ik})$. When observing repeated outcomes of $(Y_i, x_{i1}, \ldots, x_{ik})$, it is implied that $(x_{i1}, \ldots, x_{ik})$ remains constant, while the outcome $y_i$ that is paired with the vector $(x_{i1}, \ldots, x_{ik})$ generally varies. The distribution of $y_i$ outcomes for given $(x_{i1}, \ldots, x_{ik})$ values, when $k = 2$, is depicted in Figure 8.1 by a distribution centered at $EY_i$.

It is useful to distinguish two contexts in which the $\mathbf{x}$ matrix can be considered fixed. One context is the situation where elements of the $\mathbf{x}$ matrix are controlled or fixed by the individual performing the experiment. In this case, the $\mathbf{x}$ matrix is sometimes referred to as the **design matrix**, meaning that the researcher can essentially design the vectors $(x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, in a way that is of particular interest to the investigation at hand, and then she can observe the $y_i$ values associated with the chosen design matrix. As an example of this type of interpretation for $\mathbf{x}$, suppose various levels of inputs to a production process were *chosen* and an observation was made on the output corresponding to each fixed level of the inputs. We might expect that, "on the average," a certain level of output, say $EY_i$, would be produced given the levels of inputs specified by $\mathbf{x}_{i.}$. However, on any given observation with input level $\mathbf{x}_{i.}$, deviations from the average level of output might occur due to a myriad of noninput-type factors that are not controlled in the production process (e.g., machine function variation, labor efficiency, weather, temperature, etc.). Then our observations on output levels, given the chosen (or "designed") levels of inputs represented by the elements in $\mathbf{x}$, can be conceptualized as observations on the vector $(Y_i, x_{i1}, \ldots, x_{ik})$, as the GLM classical assumptions imply. In general, the underlying rationale for interpreting the expectation of the random variable $Y_i$ as a function of the explanatory variables is based on the existence of some underlying systematic economic, sociological, biological, or physical linear relationship relating the mean of $Y_i$ to the value of $\mathbf{x}_{i.}$. As we have alluded to previously, sometimes a relationship that is a polynomial function of explanation variables, and thus a linear function of parameters, is assumed as an approximation.

Not all experimental situations allow the $\mathbf{x}$ matrix to be controlled or designed. For many experimental situations in economics, business, or the social sciences in general, the researcher is a passive observer of values of $(y_i, x_{i1}, \ldots, x_{ik})$ that have been generated by consumers, entrepreneurs, the economy, markets, or the actions of society. In such cases, it is more natural to consider some or all of the elements in the $\mathbf{x}$ matrix as having been generated by an outcome of a random matrix $\mathbf{X}$ while the vector $\mathbf{y}$ is interpreted to be the outcome of the random vector $\mathbf{Y}$. Then the assumption $E\mathbf{Y} = \mathbf{x}\boldsymbol{\beta}$ of the GLM is interpreted in a *conditional* sense, i.e., the expected value of $\mathbf{Y}$ is conditional on the outcome $\mathbf{x}$ for $\mathbf{X}$. A more revealing notation for this interpretation would be given by $E(\mathbf{Y}|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$. The reader will recall from Chapter 3 that this conditional expectation is then simply referring to the *regression function* of $\mathbf{Y}$

on $\mathbf{x}$, and thus the GLM maintains that the regression function is actually a regression *hyperplane*.[2]

Note that since $\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is assumed by the GLM, it follows that in the context of operating conditionally on an outcome of $\mathbf{X}$, we must then be referring to a *conditional distribution* of the vector $\mathbf{Y}$ when we say that outcomes of $\mathbf{Y}$ can be decomposed into the mean vector, $\mathbf{x}\boldsymbol{\beta}$, and deviations, $\boldsymbol{\varepsilon}$, from the mean vector. We might use the notation

$$\mathbf{Y} \mid \mathbf{x} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \mid \mathbf{x}$$

to emphasize this fact, implying that $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{x}) = [\mathbf{0}]$. In this situation, the entire analysis is then cast in the framework of being conditional on the $\mathbf{x}$-matrix observed. Once this interpretation is affixed to the GLM together with the implicit assumption that *all* expectations are conditional on $\mathbf{x}$, the situation (and method of analysis) is analogous to the previous, inherently fixed, $\mathbf{x}$ matrix case. We will generally suppress the conditional-on-$\mathbf{x}$ notation, leaving it to the reader to provide the proper context for interpreting the GLM in a given problem situation.

The assumption $\mathbf{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I} = \mathbf{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'$ implies that the covariance matrix of $\mathbf{Y}$, and of $\boldsymbol{\varepsilon}$, is represented by

$$\mathbf{Cov}(\mathbf{Y}) = \mathbf{Cov}(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ & \sigma^2 & \dots & 0 \\ & & \ddots & \vdots \\ & & & 0 \\ \text{(symmetric)} & & & \sigma^2 \end{bmatrix}$$

i.e., the covariance matrix is a diagonal matrix with $\sigma^2$'s along the diagonal and zeroes on the off-diagonal. The fact that all of the variances of the elements in $\mathbf{Y}$ (or $\boldsymbol{\varepsilon}$) have the same value, $\sigma^2$, is referred to as the property of **homoskedasticity** (from *homoskedastic,* meaning "same scatter"). The fact that off-diagonal entries are zero implies that the covariance, or equivalently, the correlation between any two elements of $\mathbf{Y}$ (or $\boldsymbol{\varepsilon}$), is zero, which is called the property of **zero autocovariance** or **zero autocorrelation**. This assumption on the covariance matrix of $\mathbf{Y}$ (or $\boldsymbol{\varepsilon}$) then implies that the random sample $(Y_1, \dots, Y_n)$ is a collection of uncorrelated random variables all having the same variance or measure of "spread." The reader is reminded that this assumption may be referring to the conditional distribution of $\mathbf{Y}$ (and $\boldsymbol{\varepsilon}$), if we are conditioning on an outcome of the random $\mathbf{X}$-matrix.

The assumption that rank $\mathbf{x} = k$, (i.e., the $\mathbf{x}$ matrix has full column rank) simply implies that there are no linear dependencies among the columns of the $\mathbf{x}$ matrix. That is, no column of the $\mathbf{x}$ matrix is representable as some linear combination of the remaining columns in the $\mathbf{x}$ matrix. This necessarily implies

[2] Note that by tradition, the previous case where $\mathbf{x}$ is "designed" and not random is also loosely referred to as a regression of $\mathbf{Y}$ on $\mathbf{x}$, even though a conditional expectation in the usual sense is not implied.

that $n \geq k$, i.e., we must have at least as many sample observations as there are unknown parameters in the $\beta$-vector and columns in the **x** matrix.

To this point we still have not fully specified the *statistical model* since we have not yet specified a parametric family of densities for the random vector **Y** (or $\varepsilon$). In fact, all we have essentially assumed, so far, is that $(Y_1, \ldots, Y_n)$ is a random sample from some (generally) composite experiment such that the joint density of the random sample has a mean vector $\mathbf{x}\beta$, a covariance matrix $\sigma^2\mathbf{I}$, with **x** having full column rank. Nevertheless, it is possible to suggest an estimator for the parameter vector $\beta$ that has a number of useful properties. We will proceed to define such an estimator of $\beta$ in the absence of a specific parametric family specification, reserving such a specification until we have exhausted the progress that can be made in its absence. We will also define a useful estimator of $\sigma^2$.

### Estimator for $\beta$ Under Classical GLM Assumptions

How should the estimator of $\beta$ be defined? We will examine two approaches to the problem that lead to the same estimator for $\beta$ but that offer different insights into the definition of the estimator of $\beta$.

*Least-Squares Estimator of $\beta$* One approach, called the **method of least squares,** defines the estimator of $\beta$ by associating with each observation on $(\mathbf{y}, \mathbf{x})$ the $(k \times 1)$ vector **b** that solves the minimization problem

$$\min_{\mathbf{b}} (\mathbf{y} - \mathbf{x}\mathbf{b})'(\mathbf{y} - \mathbf{x}\mathbf{b}) = \min_{\mathbf{b}} \mathbf{e}'\mathbf{e},$$

where $\mathbf{e} = \mathbf{y} - \mathbf{x}\mathbf{b}$, i.e., $\mathbf{b} = \arg\min_{\mathbf{b}}(\mathbf{y} - \mathbf{x}\mathbf{b})'(\mathbf{y} - \mathbf{x}\mathbf{b})$.[3] Note that this is equivalent to finding the vector $\hat{\mathbf{y}} = \mathbf{x}\mathbf{b}$ that is the minimum distance from **y**, since the minimum of the distance $d(\mathbf{y}, \hat{\mathbf{y}}) = [(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})]^{1/2}$ and the minimum of $(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$ occur at precisely the same value of **b**.[4] From the point of view of attempting to explain **y** as best one can in terms of a linear function of the explanatory variable values, **x**, the least squares approach has intuitive appeal. To solve the minimization problem, note that $(\mathbf{y} - \mathbf{x}\mathbf{b})'(\mathbf{y} - \mathbf{x}\mathbf{b}) = \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{x}'\mathbf{y} + \mathbf{b}'\mathbf{x}'\mathbf{x}\mathbf{b}$, and the $k$ first-order conditions for a minimum can be represented in matrix form as

$$\frac{\partial(\mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{x}'\mathbf{y} + \mathbf{b}'\mathbf{x}'\mathbf{x}\mathbf{b})}{\partial\mathbf{b}} = -2\mathbf{x}'\mathbf{y} + 2\mathbf{x}'\mathbf{x}\mathbf{b} = \underset{(k \times 1)}{[\mathbf{0}]},$$

where we have used a result contained in the following lemma concerning matrix calculus.

[3] In general, $\arg\min_w f(w)$ denotes the *argument* of $f(w)$ that *minimizes* $f(w)$, where argument is taken to mean *the value of w*.

[4] This follows because $z^{1/2}$ is a monotonic transformation of $z$ for $z \geq 0$, so that the minimum and maximum of $z^{1/2}$ and $z$ occur at the same values of $z$, $z \in D$, $D$ being some set of nonnegative numbers.

**Lemma 8.1**

Let $\mathbf{z}$ be a $(k \times 1)$ vector, let $\mathbf{A}$ be a $(k \times j)$ matrix, and let $\mathbf{w}$ be a $(j \times 1)$ vector. Then

**a.** $\partial \mathbf{z}'\mathbf{A}\mathbf{w}/\partial \mathbf{z} = \mathbf{A}\mathbf{w}$,

**b.** $\partial \mathbf{z}'\mathbf{A}\mathbf{w}/\partial \mathbf{w} = \mathbf{A}'\mathbf{z}$,

**c.** $\partial \mathbf{z}'\mathbf{A}\mathbf{z}/\partial \mathbf{z} = 2\mathbf{A}\mathbf{z}$ $(k = j$ and $\mathbf{A}$ is symmetric),

**d.** $\partial \mathbf{z}'\mathbf{w}/\partial \mathbf{z} = \mathbf{w}$ $(k = j)$,

**e.** $\partial \mathbf{z}'\mathbf{w}/\partial \mathbf{w} = \mathbf{z}$ $(k = j)$.

Since $\mathbf{x}$ has full column rank, the $(k \times k)$ matrix $(\mathbf{x}'\mathbf{x})$ necessarily has full rank and is thus invertible.[5] The first-order conditions can then be solved for $\mathbf{b}$ as

$$2\mathbf{x}'\mathbf{x}\mathbf{b} = 2\mathbf{x}'\mathbf{y} \Leftrightarrow \mathbf{b} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y},$$

which defines the estimate for $\boldsymbol{\beta}$ implied by the least-squares method. The estimator of $\boldsymbol{\beta}$ is then defined by the $(k \times 1)$ random vector[6]

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}.$$

Note that the second-order conditions for the minimization problem are $\partial^2(\mathbf{e}'\mathbf{e})/\partial \mathbf{b}\partial \mathbf{b}' = 2\mathbf{x}'\mathbf{x}$, and since $\mathbf{x}'\mathbf{x}$ is positive definite, the second order conditions are satisfied. Thus, $b$ minimizes $\mathbf{e}'\mathbf{e}$.

*Coefficient of Determination, $R^2$* A function of the minimized value of the distance between $\mathbf{y}$ and $\mathbf{xb}$ is often used as a measure of how well the $\mathbf{Y}$ outcomes have been "explained" by the $\mathbf{x}\boldsymbol{\beta}$ outcomes. In particular, letting $\iota$ be an $(n \times 1)$ vector of 1's and $\hat{\mathbf{y}} = \mathbf{x}b$, the measure is given by

$$R^2 = 1 - \frac{d^2(\mathbf{y}, \hat{\mathbf{y}})}{d^2(\mathbf{y}, \iota\bar{y})} = 1 - \frac{(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})}{(\mathbf{y} - \iota\bar{y})'(\mathbf{y} - \iota\bar{y})} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

and is called the **coefficient of determination** or "$R$-squared." It is clear that the closer $\hat{\mathbf{y}}$ is to $\mathbf{y}$ in terms of distance, the higher $R^2$ is, and its *maximum* value of 1 is achieved iff $\mathbf{y} = \hat{\mathbf{y}}$.

*If* $\mathbf{x}$ contains a column of 1's, so that $\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ contains an intercept term, then $R^2$ is lower bounded by zero, and in fact $R^2$ is identically the square of the sample correlation between $\mathbf{y}$ and $\hat{\mathbf{y}} = \mathbf{x}b$. To see this, first define $\hat{\mathbf{e}}$ so that $\mathbf{y} = \mathbf{x}b + \hat{\mathbf{e}}$ and note that $\mathbf{y}'\mathbf{y}$ can be decomposed as

$$\mathbf{y}'\mathbf{y} = \mathbf{b}'\mathbf{x}'\mathbf{x}\mathbf{b} + \hat{\mathbf{e}}'\hat{\mathbf{e}}$$

[5] A matrix $\mathbf{x}$ has full column rank iff $\mathbf{x}'\mathbf{x}$ has full rank. See C. Rao, Statistical Inference, p. 30.

[6] Henceforth, we will use $\hat{\boldsymbol{\beta}}$ to denote an *estimator* and $\mathbf{b}$ to denote an *outcome* of $\hat{\boldsymbol{\beta}}$.

since $\hat{\mathbf{y}}'\hat{\mathbf{e}} = b'\mathbf{x}'\hat{\mathbf{e}} = b'(\mathbf{x}'\mathbf{y} - \mathbf{x}'\mathbf{x}b) = 0$ by the first-order conditions of the least-squares minimization problem. Also note that $\iota'\hat{\mathbf{e}} = \iota'\mathbf{y} - \iota'\hat{\mathbf{y}} = 0$ since $\mathbf{x}'(\mathbf{y} - \mathbf{x}b) = \mathbf{x}'\hat{\mathbf{e}} = [\mathbf{0}]$ from the first-order conditions and since $\mathbf{x}'$ contains a row of 1's. It follows that $\bar{y} = \bar{\hat{y}}$. Subtracting $n\bar{y}^2$ from both sides of the preceding decomposition of $\mathbf{y}'\mathbf{y}$ and then dividing by $\mathbf{y}'\mathbf{y} - n\bar{y}^2$, imply

$$R^2 = \frac{b'\mathbf{x}'\mathbf{x}b - n\bar{y}^2}{(\mathbf{y} - \iota\bar{y})'(\mathbf{y} - \iota\bar{y})} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Finally, since $\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 = \sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})$,

$$R^2 = \frac{\left[\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right]^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2} = r^2_{\mathbf{Y}\hat{\mathbf{Y}}},$$

i.e., $R^2$ is the square of the sample correlation between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$.

It follows from our discussion of the *sample correlation* in Section 6.4 that $R^2$ can be interpreted as the proportion of the sample variance in the $y_i$'s that is explained by the corresponding values of the $\hat{y}_i$'s. This follows from the fact that the vector of values $a + b\hat{y}_i$, $i = 1, \ldots, n$, has the smallest expected squared distance from $y_1, \ldots, y_n$ when

$$b = s_{Y\hat{Y}}/s^2_{\hat{Y}} = s^2_{\hat{Y}}/s^2_{\hat{Y}} = 1 \quad \text{and} \quad a = \bar{y} - b\bar{\hat{y}} = \bar{y} - 1 \cdot \bar{y} = 0.$$

*BLUE of $\boldsymbol{\beta}$* A second approach for defining an estimator of $\boldsymbol{\beta}$ begins with the objective of defining the BLUE of $\boldsymbol{\beta}$. In order to be in the *linear* class, the estimator must be of the form $\hat{\boldsymbol{\beta}} = \mathbf{AY} + \mathbf{d}$ for some nonrandom $(k \times n)$ matrix $\mathbf{A}$ and some nonrandom $(k \times 1)$ vector $\mathbf{d}$. If the estimator is to be in the *unbiased* class, then $\mathrm{E}\hat{\boldsymbol{\beta}} = \mathbf{Ax}\boldsymbol{\beta} + \mathbf{d} = \boldsymbol{\beta}, \forall\, \boldsymbol{\beta}$, which requires that $\mathbf{Ax} = \mathbf{I}$ and $\mathbf{d} = [\mathbf{0}]$. For the estimator to be the *best* in the class of linear unbiased estimators, its covariance matrix must be as small or smaller than the covariance matrix of any other estimator in the linear unbiased class. Since $\hat{\boldsymbol{\beta}} = \mathbf{AY}$ and $\mathbf{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$, it follows that $\mathbf{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{AA}'$ is the covariance matrix that must be minimized through choice of the matrix $\mathbf{A}$ subject to the unbiasedness constraint that $\mathbf{Ax} = \mathbf{I}$. Note that $\mathbf{Cov}(\hat{\boldsymbol{\beta}})$ will be minimized iff $\boldsymbol{\ell}'\mathbf{AA}'\boldsymbol{\ell}$ is minimized for every choice of $\boldsymbol{\ell} \neq [\mathbf{0}]$. This follows from the fact that if $\hat{\boldsymbol{\beta}}^* = \mathbf{DY}$, with $\mathbf{Cov}(\hat{\boldsymbol{\beta}}^*) = \sigma^2\mathbf{DD}'$, is any other linear unbiased estimator of $\boldsymbol{\beta}$, then $\sigma^2\mathbf{AA}' - \sigma^2\mathbf{DD}' = \mathbf{Cov}(\hat{\boldsymbol{\beta}}) - \mathbf{Cov}(\hat{\boldsymbol{\beta}}^*)$ will be a negative semidefinite matrix, and thus $\hat{\boldsymbol{\beta}}$ would have the smaller covariance matrix, iff $\boldsymbol{\ell}'\mathbf{DD}'\boldsymbol{\ell} \geq \boldsymbol{\ell}'\mathbf{AA}'\boldsymbol{\ell}\ \forall\ \boldsymbol{\ell} \neq [\mathbf{0}]$.

The problem at this point is then one of finding the solution, if it exists, to the following problem:

$$\min_{\mathbf{A}} \boldsymbol{\ell}'\mathbf{AA}'\boldsymbol{\ell} \quad \text{such that } \mathbf{Ax} = \mathbf{I}, \quad \forall\ \boldsymbol{\ell} \neq [\mathbf{0}].$$

The feasible choices of $\mathbf{A}$ can be represented as $\mathbf{A} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' + \mathbf{D}$, where $\mathbf{D}$ is any matrix such that $\mathbf{Dx} = [\mathbf{0}]$. To see this, first note that *any* matrix $\mathbf{A}$ can be equivalently represented as $(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' + \mathbf{D}$ by simply choosing $\mathbf{D} = \mathbf{A} - (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$. Now for $\mathbf{A}$ to satisfy $\mathbf{Ax} = \mathbf{I}$, it must be the case that $\mathbf{Ax} = [(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' + \mathbf{D}]\mathbf{x} =$

$\mathbf{I} + \mathbf{Dx} = \mathbf{I}$, so that $\mathbf{Dx} = [\mathbf{0}]$ is implied. Substituting $(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' + \mathbf{D}$ for $\mathbf{A}$ in the previous minimization problem then results in the equivalent minimization problem

$$\min_{\mathbf{D}} \boldsymbol{\ell}'(\mathbf{x}'\mathbf{x})^{-1}\boldsymbol{\ell} + \boldsymbol{\ell}'\mathbf{DD}'\boldsymbol{\ell} \quad \text{such that } \mathbf{Dx} = [\mathbf{0}], \quad \forall\, \boldsymbol{\ell} \neq [\mathbf{0}].$$

The value $\mathbf{D} = [\mathbf{0}]$ is the solution to the minimization problem, yielding $\boldsymbol{\ell}'(\mathbf{x}'\mathbf{x})^{-1}\boldsymbol{\ell}$ as the minimum value of the objective function $\forall\ \boldsymbol{\ell} \neq [\mathbf{0}]$, since any other choice of $\mathbf{D} \neq [\mathbf{0}]$ satisfying $\mathbf{Dx} = [\mathbf{0}]$ will result in $\boldsymbol{\ell}'\mathbf{DD}'\boldsymbol{\ell} > 0$ for *some* $\boldsymbol{\ell} \neq [\mathbf{0}]$, which is a nonoptimal objective function value $> \boldsymbol{\ell}'(\mathbf{x}'\mathbf{x})^{-1}\boldsymbol{\ell}$ for some $\boldsymbol{\ell} \neq [\mathbf{0}]$. Then $\mathbf{A} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$ is the optimal choice of $\mathbf{A}$ in the original minimization problem.

The preceding argument establishes that $\hat{\boldsymbol{\beta}} = \mathbf{AY} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ is the BLUE of $\boldsymbol{\beta}$, and the covariance matrix of $\boldsymbol{\beta}$ is $\mathbf{Cov}(\boldsymbol{\beta}) = \sigma^2\mathbf{AA}' = \sigma^2(\mathbf{x}'\mathbf{x})^{-1}$. We have thus proved the celebrated Gauss–Markov theorem.

**Theorem 8.1** (Gauss–Markov Theorem) *Under the classical assumptions of the GLM,* $\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ *is the best linear unbiased estimator of* $\boldsymbol{\beta}$.

**Proof** This is proven by the derivation of the BLUE of $\boldsymbol{\beta}$ preceding the theorem. ■

It is also true that $\boldsymbol{\ell}'\hat{\boldsymbol{\beta}} = \boldsymbol{\ell}'(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ is the BLUE of $\boldsymbol{\ell}'\boldsymbol{\beta}$ for any choice of $\boldsymbol{\ell} \neq 0$. Thus, once $\hat{\boldsymbol{\beta}}$ is known, one can calculate BLUE estimates of any linear combination of the entries in $\boldsymbol{\beta}$ by simply calculating the corresponding linear combination of the entries in $\hat{\boldsymbol{\beta}}$. We formalize this result in the following theorem.

**Theorem 8.2** (BLUE for $\boldsymbol{\ell}'\boldsymbol{\beta}$) *Under the classical assumptions of the GLM,* $\boldsymbol{\ell}'\hat{\boldsymbol{\beta}} = \boldsymbol{\ell}'(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ *is the best linear unbiased estimator of* $\boldsymbol{\ell}'\boldsymbol{\beta}\ \forall\ \boldsymbol{\ell} \neq [\mathbf{0}]$.

**Proof** Let $T = t(\mathbf{Y}) = \mathbf{c}'\mathbf{Y} + d$ be any linear estimator of the scalar $\boldsymbol{\ell}'\boldsymbol{\beta}$, where $\mathbf{c}$ is an $(n \times 1)$ vector and $d$ is a scalar. For $T$ to be unbiased, it is required that

$$\mathrm{E}T = \mathrm{E}(\mathbf{c}'\mathbf{Y} + d) = \mathbf{c}'\mathbf{x}\boldsymbol{\beta} + d = \boldsymbol{\ell}'\boldsymbol{\beta} \quad \forall\, \boldsymbol{\beta},$$

which in turn requires that $\mathbf{x}'\mathbf{c} = \boldsymbol{\ell}$ and $d = 0$. Note the following lemma.

**Lemma 8.2**

Let $\mathbf{x}$ be an $(n \times k)$ matrix of rank $k$, and let $\boldsymbol{\ell}$ be any $(k \times 1)$ vector. Then $\mathbf{x}'\mathbf{c} = \boldsymbol{\ell}$ iff

$$\mathbf{c} = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\boldsymbol{\ell} + [\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}']\mathbf{h},$$

where the $(n \times 1)$ vector $\mathbf{h}$ can be chosen arbitrarily.

**Proof**: See F. A. Graybill (1983), *Matrices with applications in statistics*, 2nd ed. Belmont, CA: Wadsworth, p. 153.

Using the solution for $\mathbf{c}$ given by the lemma, note that

$$\mathrm{var}(\mathbf{c}'\mathbf{Y}) = \sigma^2[\boldsymbol{\ell}'(\mathbf{x}'\mathbf{x})^{-1}\boldsymbol{\ell} + \mathbf{h}'[\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}']\mathbf{h}].$$

The variance is minimized by setting $\mathbf{h} = [0]$, and thus $T = \mathbf{c}'\mathbf{Y} = \boldsymbol{\ell}'(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y} = \boldsymbol{\ell}'\hat{\boldsymbol{\beta}}$ is the BLUE for $\boldsymbol{\ell}'\boldsymbol{\beta}$. ■

The estimator $\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ is referred to in the literature as the **least-squares estimator** of $\boldsymbol{\beta}$ due to its definition via the least-squares approach described previously. Without additional assumptions on the statistical model, the BLUE property exhausts the properties that we can attribute to $\hat{\boldsymbol{\beta}}$. In particular, we cannot demonstrate consistency of $\hat{\boldsymbol{\beta}}$, nor can we make progress toward defining a MVUE of $\boldsymbol{\beta}$.

### Estimator for $\sigma^2$ Under Classical GLM Assumptions

The classical assumptions are sufficient to allow the definition of an unbiased estimator of the common variance, $\sigma^2$, of the $Y_i$'s (or equivalently, the common variance of the $\varepsilon_i$'s). Since $\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y} = \boldsymbol{\beta} + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\varepsilon}$,

$$\begin{aligned}
\mathrm{E}(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}}) &= \mathrm{E}\boldsymbol{\varepsilon}'(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\boldsymbol{\varepsilon} \\
&= \mathrm{tr}(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\mathrm{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \quad (\text{since } (\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}') \text{ is symmetric and idempotent,} \\
&\quad \mathrm{tr}\,\mathbf{ABC} = \mathrm{tr}\,\mathbf{BCA}, \text{ and } \mathrm{E}\,\mathrm{tr}(\mathbf{W}) = \mathrm{tr}(\mathbf{EW})) \\
&= \sigma^2[\mathrm{tr}\,\mathbf{I} - \mathrm{tr}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{x}] \quad (\text{since } \mathrm{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \sigma^2\mathbf{I}) \\
&= \sigma^2(n - k).
\end{aligned}$$

It follows that if we define

$$\hat{S}^2 = (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})/(n - k),$$

then $\mathrm{E}\hat{S}^2 = \sigma^2$, and so $\hat{S}^2$ is an unbiased estimator of $\sigma^2$ (we use the hat, ^, to distinguish this unbiased estimator from the sample variance, $S^2$, introduced in Chapter 6). Furthermore, since $\mathbf{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{x}'\mathbf{x})^{-1}$ (recall the derivation of the BLUE of $\boldsymbol{\beta}$), it follows that $\hat{S}^2(\mathbf{x}'\mathbf{x})^{-1}$ is an unbiased estimator of $\mathbf{Cov}(\hat{\boldsymbol{\beta}})$. Without additional assumptions on the statistical model, unbiasedness exhausts the properties that we can attribute to the estimator $\hat{S}^2$.

The following is an example of the use of both the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{S}^2$.

**Example 8.1** Let the production of a commodity in a given period of time be represented by the Cobb–Douglas production function

$$Y_t = \beta_1 x_{t1}^{\beta_2} x_{t2}^{\beta_3} \mathrm{e}^{\varepsilon_t}$$

where $y_t$ = thousands of units produced in time period $t$, $(x_{t1}, x_{t2})$ = units of labor and capital applied to the production process in year $t$, and $\varepsilon_t$ = disturbance term value in year $t$. Assume that $\mathrm{E}\varepsilon_t = 0$ and $\mathrm{var}(\varepsilon_t) = \sigma^2$, ∀ $t$, and let $\varepsilon_i$ and $\varepsilon_j$ be uncorrelated ∀ $i \neq j$.

Ten time periods' worth of observations on $y_t$, $x_{t1}$, and $x_{t2}$ were obtained, yielding

$$\mathbf{y} = \begin{bmatrix} 47.183 \\ 53.005 \\ 43.996 \\ 38.462 \\ 54.035 \\ 59.132 \\ 79.763 \\ 67.252 \\ 55.267 \\ 38.972 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} 7 \\ 8 \\ 6 \\ 4 \\ 5 \\ 9 \\ 11 \\ 9 \\ 8 \\ 4 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 2 \\ 3 \\ 2 \\ 5 \\ 6 \\ 4 \\ 8 \\ 5 \\ 5 \\ 4 \end{bmatrix}.$$

Calculate a BLUE estimate of $\ln(\beta_1)$, $\beta_2$, $\beta_3$, and of $\beta_2+\beta_3$. Calculate an unbiased estimate of $\sigma^2$.

**Answer**: Before we can apply the least-squares estimator, we must transform the model to be in GLM form. By taking the natural logarithm of both sides of the production function relationship, we obtain

$$\ln Y_t = \ln(\beta_1) + \beta_2 \ln(x_{t1}) + \beta_3 \ln(x_{t2}) + \varepsilon_t,$$

or

$$Y_t^* = \beta_1^* + \beta_2 x_{t1}^* + \beta_3 x_{t2}^* + \varepsilon_t.$$

The reader can check to see that the transformed model satisfies the classical assumptions of the GLM. Defining $\underset{(10\times 1)}{\mathbf{y}_*} = \ln(\mathbf{y})$ and $\underset{(10\times 3)}{\mathbf{x}_*} = [\iota \mid \ln(\mathbf{x}_1) \mid \ln(\mathbf{x}_2)]$, where $\iota$ is a $(10 \times 1)$ vector with each element equal to 1, we can represent the observations on the transformed production function relationship as outcomes of $\mathbf{Y}_* = \mathbf{x}_*\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}_*' = (\ln(\beta_1), \beta_2, \beta_3)'$. The least-squares (BLUE) estimate of $\boldsymbol{\beta}_*$ is then given by

$$\mathbf{b} = (\mathbf{x}_*'\mathbf{x}_*)^{-1}\mathbf{x}_*'\mathbf{y}_* = \begin{bmatrix} 2.638 \\ .542 \\ .206 \end{bmatrix}.$$

The BLUE estimate of $\beta_2 + \beta_3$, which in this case is an estimate of the degree of homogeneity of the production function, is given by $b_2 + b_3 = .748$ by Theorem 8.2 with $\boldsymbol{\ell}' = [0 \;\; 1 \;\; 1]$. An unbiased estimate of $\sigma^2$ is given by $\hat{s}^2 = (\mathbf{y}_* - \mathbf{x}_*\mathbf{b})'(\mathbf{y}_* - \mathbf{x}_*\mathbf{b})/7 = .0043$. The reader is encouraged to provide an unbiased estimate of $\mathbf{Cov}(\hat{\boldsymbol{\beta}})$.

We emphasize that in the process of transforming the production relationship to GLM form, we also transformed the parameter $\beta_1$ to $\ln(\beta_1)$, which then resulted in a situation where the least-squares estimator was estimating *not* $\beta_1$, but $\ln(\beta_1)$. The phenomenon of transforming parameters often occurs when a relationship must be transformed in order to convert it into GLM form. □

### Consistency of $\hat{\boldsymbol{\beta}}$

If, *in addition to* the classical assumptions of the GLM, it can be assumed that $(\mathbf{x}'\mathbf{x})^{-1} \to [0]$ as $n \to \infty$, then $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$.

**Theorem 8.3** (Consistency of $\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$) *Under the classical assumptions of the GLM, if* $(\mathbf{x}'\mathbf{x})^{-1} \to [0]$ *as* $n \to \infty$, *then* $\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y} \xrightarrow{p} \boldsymbol{\beta}$, *so that* $\hat{\boldsymbol{\beta}}$ *is a consistent estimator of* $\boldsymbol{\beta}$.

**Proof** Given the classical assumptions of the GLM, we know that $\mathrm{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, and the covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by $\mathbf{Cov}(\hat{\boldsymbol{\beta}}) = \mathbf{Cov}((\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}) = \sigma^2(\mathbf{x}'\mathbf{x})^{-1}$. It follows from Corollary 5.2 that if $\mathbf{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{x}'\mathbf{x})^{-1} \to [0]$, then $\hat{\boldsymbol{\beta}}$ converges in mean square to $\boldsymbol{\beta}$, i.e., $\hat{\boldsymbol{\beta}} \xrightarrow{m} \boldsymbol{\beta}$. But since convergence in mean square implies convergence in probability (Theorem 5.13), it follows that plim $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, so that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$. ■

*Conditions for Which* $(\mathbf{x}'\mathbf{x})^{-1} \to [0]$ As a practical matter, what is required of the $\mathbf{x}$ matrix for $\lim_{n\to\infty}(\mathbf{x}'\mathbf{x})^{-1} = [0]$? The convergence to the zero matrix will occur iff each *diagonal* entry of $(\mathbf{x}'\mathbf{x})^{-1}$ converges to zero, the necessity of this condition being obvious. The sufficiency follows from the fact that $(\mathbf{x}'\mathbf{x})^{-1}$ is positive definite since $\mathbf{x}'\mathbf{x}$ is positive definite,[7] and then the $(i, j)$th entry of $(\mathbf{x}'\mathbf{x})^{-1}$ is upper bounded in absolute value by the product of the square roots of the $(i, i)$th and $(j, j)$th (diagonal) entries of $(\mathbf{x}'\mathbf{x})^{-1}$. To see the boundedness of the $(i, j)$th entry of $(\mathbf{x}'\mathbf{x})^{-1}$, note that for any symmetric, positive definite matrix, all principal submatrices formed by deleting one or more diagonal entries together with the rows and columns in which they appear must also be positive definite.[8] In particular, retaining only the $i$th and $j$th diagonal entries results in a $(2 \times 2)$ principal submatrix of the form

$$\begin{bmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{bmatrix},$$

where $a_{ij}$ is used to denote the $(i, j)$th entry in $(\mathbf{x}'\mathbf{x})^{-1}$. The conditions for the submatrix to be positive definite are the standard ones, i.e., $a_{ii} > 0$ and $a_{ii}a_{jj} > a_{ij}a_{ji}$, the latter condition implying $a_{ii}a_{jj} > (a_{ij})^2$ by the symmetry of $(\mathbf{x}'\mathbf{x})^{-1}$, and thus $|a_{ij}| < (a_{ii})^{1/2}(a_{jj})^{1/2}$, which is the boundedness result mentioned above. Therefore, if $a_{ii} \to 0 \ \forall\, i$, then $a_{ij} \to 0 \ \forall\, i \neq j$.

Regarding the convergence of the diagonal entries of $(\mathbf{x}'\mathbf{x})^{-1}$ to zero, examine the (1,1) entry of $(\mathbf{x}'\mathbf{x})^{-1}$. We will again have use for the concept of partitioned inversion, which we restate here for convenience, as Lemma 8.3.

[7] The inverse of a symmetric positive definite matrix is necessarily a symmetric positive definite matrix. This can be shown by noting that a symmetric matrix is positive definite iff all of its characteristic roots are positive, and the characteristic roots of $\mathbf{A}^{-1}$ are the reciprocals of the characteristic roots of $\mathbf{A}$.

[8] Suppose $\mathbf{B}$ is a $(k \times k)$ symmetric, positive definite matrix. Let $\mathbf{C}$ be an $(r \times k)$ matrix such that each row consists of all zeroes except for a 1 in one position, and let the $r$ rows of $\mathbf{C}$ be linearly independent. Then $\mathbf{CBC}'$ defines an $(r \times r)$ principal submatrix of $\mathbf{B}$. Now note that $\ell'\mathbf{CBC}'\ell = \ell_*'\mathbf{B}\ell_* > 0 \ \forall\, \ell \neq [0]$ since $\ell_* = \mathbf{C}'\ell \neq [0] \ \forall\, \ell \neq [0]$ by the definition of $\mathbf{C}$, and $\mathbf{B}$ is positive definite. Therefore, the principal submatrix $\mathbf{CBC}'$ is positive definite.

**Lemma 8.3**
***Partitioned Inversion***

Let $\mathbf{B}$ be any $(n \times n)$ nonsingular matrix that is partitioned as $\mathbf{B} = \left[\begin{smallmatrix}\mathbf{B}_{11} & \mathbf{B}_{12}\\ \mathbf{B}_{21} & \mathbf{B}_{22}\end{smallmatrix}\right]$ where $\mathbf{B}_{11}$ and $\mathbf{B}_{22}$ are nonsingular matrices. Then

$$\mathbf{B}^{-1} = \begin{bmatrix} (\mathbf{B}_{11} - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21})^{-1} & -\mathbf{B}_{11}^{-1}\mathbf{B}_{12}(\mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12})^{-1} \\ -\mathbf{B}_{22}^{-1}\mathbf{B}_{21}(\mathbf{B}_{11} - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21})^{-1} & (\mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12})^{-1} \end{bmatrix}$$

Now partition the $\mathbf{x}$ matrix as $\mathbf{x} = [\mathbf{x}_{.1} \;\; \mathbf{x}_*]$, where $\mathbf{x}_{.1}$ denotes the first column of $\mathbf{x}$, and $\mathbf{x}_*$ denotes the remaining $(k - 1)$ columns of the $\mathbf{x}$ matrix. Then

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} \mathbf{x}'_{.1}\mathbf{x}_{.1} & \mathbf{x}'_{.1}\mathbf{x}_* \\ \mathbf{x}'_*\mathbf{x}_{.1} & \mathbf{x}'_*\mathbf{x}_* \end{bmatrix},$$

and using the partitioned inversion lemma, the (1,1) entry in $(\mathbf{x}'\mathbf{x})^{-1}$ is represented by

$$(\mathbf{x}'\mathbf{x})^{-1}_{1,1} = (\mathbf{x}'_{.1}\mathbf{x}_{.1} - \mathbf{x}'_{.1}\mathbf{x}_*(\mathbf{x}'_*\mathbf{x}_*)^{-1}\mathbf{x}'_*\mathbf{x}_{.1})^{-1}.$$

Thus, for the (1,1) entry in $(\mathbf{x}'\mathbf{x})^{-1}$ to have a limit of zero as $n \to \infty$, it is clear that the expression in parentheses on the right-hand side of the above equality must $\to \infty$ as $n \to \infty$. Since $(\mathbf{x}'_*\mathbf{x}_*)$ is positive definite (recall footnote 8), this necessarily requires that $\mathbf{x}'_{.1}\mathbf{x}_{.1} \to \infty$ as $n \to \infty$ since a nonnegative quantity, $\mathbf{x}'_{.1}\mathbf{x}_*(\mathbf{x}'_*\mathbf{x}_*)^{-1}\mathbf{x}'_*\mathbf{x}_{.1}$, is being subtracted from $\mathbf{x}'_{.1}\mathbf{x}_{.1}$.

Now note further that $(\mathbf{x}'_{.1}\mathbf{x}_{.1} - \mathbf{x}'_{.1}\mathbf{x}_*(\mathbf{x}'_*\mathbf{x}_*)^{-1}\mathbf{x}'_*\mathbf{x}_{.1})$ is in fact the sum of the squared deviations of $\mathbf{x}_{.1}$ from a vector of predictions of $\mathbf{x}_{.1}$ generated by a least-squares–estimated linear explanation of $\mathbf{x}_{.1}$ (the dependent variable $\mathbf{x}_{.1}$ is being predicted by a linear function of $\mathbf{x}_*$, which are the explanatory variables in this context). To see this, note that the vector $\mathbf{x}_*\hat{\mathbf{b}}$ is the least distance from $\mathbf{x}_{.1}$ when we choose $\hat{\mathbf{b}}$ to be $\hat{\mathbf{b}} = (\mathbf{x}'_*\mathbf{x}_*)^{-1}\mathbf{x}'_*\mathbf{x}_{.1}$, as we have argued previously with regard to the least-squares procedure. Then the deviations between entries in $\mathbf{x}_{.1}$ and $\mathbf{x}_*\hat{\mathbf{b}}$ are represented by the vector

$$\mathbf{e}_1 = \mathbf{x}_{.1} - \mathbf{x}_*(\mathbf{x}'_*\mathbf{x}_*)^{-1}\mathbf{x}'_*\mathbf{x}_{.1},$$

and straightforward matrix multiplication then demonstrates that

$$\mathbf{e}'_1\mathbf{e}_1 = \mathbf{x}'_{.1}\mathbf{x}_{.1} - \mathbf{x}'_{.1}\mathbf{x}_*(\mathbf{x}'_*\mathbf{x}_*)^{-1}\mathbf{x}'_*\mathbf{x}_{.1},$$

which is the *reciprocal* of the (1,1) entry of $(\mathbf{x}'\mathbf{x})^{-1}$. Thus, for the (1,1) entry of $(\mathbf{x}'\mathbf{x})^{-1}$ to converge to zero, we require that $\mathbf{e}'_1\mathbf{e}_1 \to \infty$ as $n \to \infty$. We thus specifically rule out the possibility that $\mathbf{e}'_1\mathbf{e}_1 < \delta < \infty$, i.e., that the sum of squared deviations is bounded as $n \to \infty$. As a practical matter, it is *sufficient* that the average squared error, $n^{-1}\mathbf{e}'_1\mathbf{e}_1$, in predicting entries in $\mathbf{x}_{.1}$ via linear combinations of corresponding entries in the remaining columns of the $\mathbf{x}$ matrix exhibits some positive lower bound and thus does not converge to zero as $n \to \infty$. Roughly speaking, it is sufficient to assume that a linear dependence between $\mathbf{x}_{.1}$ and the column vectors in $\mathbf{x}_*$ never develops regardless of sample size.

Note that the column of $\mathbf{x}$ that we utilize in applying the partitioned inversion result is arbitrary—we can always rearrange the columns of the $\mathbf{x}$ matrix to place whichever explanatory variable we choose in the "first position." Thus, our above discussion applies to each diagonal entry in the $(\mathbf{x}'\mathbf{x})^{-1}$ matrix. Then in summary, our sufficient condition for the consistency of $\hat{\boldsymbol{\beta}}$ is that $\lim_{n\to\infty}(\mathbf{x}'\mathbf{x})^{-1} = [\mathbf{0}]$,[9] which holds if $(\mathbf{x}'_{.i}\mathbf{x}_{.i}) \to \infty \;\forall\, i$, and if $\exists$ some positive number $\delta > 0$ such that no column of $\mathbf{x}$ can be represented via a linear combination of the remaining columns of $\mathbf{x}$ with an average squared prediction error less than $\delta$ as $n \to \infty$.

### Consistency of $\hat{S}^2$

If, *in addition to* the classical assumptions of the GLM, it can be assumed that the $\varepsilon_i$'s are iid, then $\hat{S}^2$ is a consistent estimator of $\sigma^2$.

**Theorem 8.4** (Consistency of $\hat{S}^2$-iid Residuals) *Under the classical assumptions of the GLM, if the elements of the residual vector $\boldsymbol{\varepsilon}$ are iid, then $\hat{S}^2 \xrightarrow{p} \sigma^2$, so that $\hat{S}^2$ is a consistent estimator of $\sigma^2$.*

**Proof** Recall from our discussion of the unbiasedness of $\hat{S}^2$ that $\hat{S}^2 = (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})/(n-k) = \boldsymbol{\varepsilon}'(I - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\boldsymbol{\varepsilon}/(n-k)$. Examine

$$\operatorname{plim}\hat{S}^2 = \operatorname{plim}\left(\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n-k}\right) - \operatorname{plim}\left(\frac{\boldsymbol{\varepsilon}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\varepsilon}}{n-k}\right),$$

and focus on the last term first. Note that $(n-k)^{-1}(\boldsymbol{\varepsilon}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}\boldsymbol{\varepsilon})$ is a nonnegative-valued random variable since $(\mathbf{x}'\mathbf{x})^{-1}$ is positive definite. Also note that under the classical assumptions, $\mathrm{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \sigma^2\mathbf{I}$, so that

$$\begin{aligned}\mathrm{E}(n-k)^{-1}\boldsymbol{\varepsilon}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}\boldsymbol{\varepsilon} &= (n-k)^{-1}\operatorname{tr}[\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathrm{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] \\ &= \sigma^2(n-k)^{-1}\operatorname{tr}[(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{x}] = \sigma^2\left(\frac{k}{n-k}\right).\end{aligned}$$

Then, by Markov's inequality, $\forall\, c > 0$,

$$P[(n-k)^{-1}\boldsymbol{\varepsilon}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}\boldsymbol{\varepsilon} \geq c] \leq \frac{\sigma^2(k/(n-k))}{c},$$

and since $\lim_{n\to\infty}\sigma^2(k/(n-k)) = 0$, it follows that $\operatorname{plim}(n-k)^{-1}\boldsymbol{\varepsilon}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}\boldsymbol{\varepsilon} = 0$.

Given the preceding result, it is clear that for plim $\hat{S}^2 = \sigma^2$, it is necessary and sufficient that plim $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/(n-k) = \sigma^2$. Recalling that $\operatorname{plim}(WZ) =$

[9] Consistency of $\hat{\boldsymbol{\beta}}$ can be proven under alternative conditions on $\mathbf{x}$. Judge, et al. (1982), *Introduction to the Theory and Practice of Econometrics*. New York: John Wiley, pp. 268-269, prove the result using the stronger condition that $\lim_{n\to\infty} n^{-1}\mathbf{x}'\mathbf{x} = \mathbf{Q}$, where $\mathbf{Q}$ is a finite, positive definitive matrix. Halbert White (1984), *Asymptotic Theory for Econometricians*, Cambridge, MA: Academic Press, p. 20, assumes that $n^{-1}\mathbf{x}'\mathbf{x}$ is bounded and uniformly positive definite, which is also a stronger condition than the one we use here.

$\text{plim}(W)\,\text{plim}(Z))$ by Theorem 5.6, and since $(n-k)/n \to 1$ as $n \to \infty$, an equivalent necessary and sufficient condition is that $\text{plim}\, n^{-1}\varepsilon'\varepsilon = \sigma^2$. If the $\varepsilon_i$'s are not only uncorrelated, but also iid, then Khinchin's WLLN (Theorem 5.19) allows us to conclude the preceding plim result, since $\text{E}\varepsilon_i^2 = \sigma^2 \; \forall \; i$, the $\varepsilon_i^2$'s are iid, and $n^{-1}\varepsilon'\varepsilon = n^{-1}\sum_{i=1}^n \varepsilon_i^2$, which is precisely in the form of Khinchin's theorem applied to the sequence of iid random variables $\{\varepsilon_1^2, \varepsilon_2^2, \varepsilon_3^2, \ldots\}$. Thus, $\text{plim}\; \hat{S}^2 = \sigma^2$, and $\hat{S}^2$ is a consistent estimator of $\sigma^2$. ■

Consistency of $\hat{S}^2$ for estimating $\sigma^2$ can be demonstrated without requiring that the $\varepsilon_i$'s be iid. In the following result, we replace the iid assumption with alternative assumptions that still result in $\hat{S}^2 \xrightarrow{p} \sigma^2$.

***Theorem 8.5** (Consistency of $\hat{S}^2$—Non-iid Residuals) *Assume the classical assumptions of the GLM. Then $\hat{S}^2$ is a consistent estimator of $\sigma^2$ if $\text{E}\varepsilon_i^4 \le \tau < \infty \; \forall \; i$ and either of the following conditions holds:*

**a.** $\sum_{\substack{j=1 \\ j \neq i}}^n \text{Cov}(\varepsilon_i^2, \varepsilon_j^2) = o(n^1) \; \forall \; i$;

**b.** $\{\varepsilon_n\}$ *is an m-dependent sequence.*

**Proof** **a.** Under the stated conditions, the sequence of squared residuals $\{\varepsilon_n^2\} = \{\varepsilon_1^2, \varepsilon_2^2, \ldots\}$ satisfies the assumptions of the WLLN stated in Theorem 5.22. To see this, note that $\text{var}(\varepsilon_i^2) = \text{E}\varepsilon_i^4 - (\text{E}\varepsilon_i^2)^2 = \text{E}\varepsilon_i^4 - \sigma^4 \le \tau < \infty \; \forall \; i$, so that

$$\text{var}(n^{-1}\varepsilon'\varepsilon) = n^{-2}\left[\sum_{i=1}^n \text{var}(\varepsilon_i^2) + \sum_{i \neq j}\sum \text{cov}(\varepsilon_i^2, \varepsilon_j^2)\right] \le \left[\tau + o(n^1)\right]/n \to 0 \quad \text{as } n \to \infty.$$

It follows immediately that $\text{plim}\;(n^{-1}\varepsilon'\varepsilon) = \sigma^2$. Also, given the classical assumptions of the GLM, it can be shown that $\text{plim}\;[(n-k)^{-1}\varepsilon'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}\varepsilon] = 0$ using Markov's inequality in exactly the same way as in the proof of Theorem 8.4. It follows that

$$\text{plim}\, S^2 = \underbrace{\text{plim}(n^{-1}\varepsilon'\varepsilon)}_{\sigma^2} - \underbrace{\text{plim}[(n-k)^{-1}\varepsilon'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon]}_{0} = \sigma^2.$$

**b.** The conditions in part (b) *imply* the conditions in part (a), since $\text{cov}(\varepsilon_i^2, \varepsilon_j^2) = 0$ when $|i-j| > m$ by $m$-dependence. Thus, $\hat{S}^2 \xrightarrow{p} \sigma^2$. ■

Regarding the conditions in Theorem 8.5, first note that the boundedness of the moments of the residuals, i.e., $\text{E}\varepsilon_i^4 \le \tau < \infty$ for some $\tau > 0$, is a relatively weak assumption that can often be comfortably adopted in practice. Note, for example, that these moment-boundedness assumptions are satisfied for *every* parametric family of density functions that we examined in Chapter 4. The remaining conditions on the residuals relate to the degree of association between outcomes of $\varepsilon_i$ and $\varepsilon_j$ as $i$ and $j$ become further and further apart. Essentially, (a) implies that the average covariance between $\varepsilon_i^2$ and all of the remaining squared residuals, $\varepsilon_j^2$ for $j \neq i$, converges to zero $\forall \; i$; and (b) states that $\varepsilon_i$ and

$\varepsilon_j$ become independent of one another when $|i-j|$ exceeds a fixed finite value. If the random sample refers to a time series of observations, then $|i-j|$ refers to separation of $\varepsilon_i$ and $\varepsilon_j$ in the time dimension. If the random sample refers to a cross section of observations, then one must search for an ordering of the elements in the random sample to which the conditions of Theorem 8.5 can be applied.

### Asymptotic Normality of $\hat{\beta}$

We establish the asymptotic normality of $\hat{\beta}$ by initially stipulating that the $\varepsilon_i$'s are iid, that $|x_{ij}| < \xi < \infty \ \forall \ i, j$, i.e., the explanatory variable values are bounded in absolute value, and that $P(|\varepsilon_i| < m) = 1 \ \forall \ i$, for $m < \infty$, i.e., the disturbance terms are bounded in absolute value with probability 1. As a practical matter, it is often reasonable to assume that real-world explanatory variables in a linear model do not have values that increase without bound; e.g., all published data on the economy, census data, samples of sociodemographic information, and the like contain finite numbers, no matter how vast the collection. Indeed, it is safe to say that all of the numerical data ever measured and recorded can be bounded in absolute value by some large enough finite number, $\xi$. The boundedness assumption on the explanatory variables thus has wide applicability.[10] Similar reasoning suggests that, in practice, boundedness of the disturbance terms in a linear model is often a reasonable assumption. That is, if $y_i$ is an observation on real-world economic or social data and thus bounded in magnitude, it follows that $\varepsilon_i = Y_i - \mathrm{E}Y_i$ would also be bounded, so that $P(|\varepsilon_i| < m) = 1$.

We finally assume at the outset that $\lim_{n\to\infty} n^{-1}\mathbf{x}'\mathbf{x} = \mathbf{Q}$, where $\mathbf{Q}$ is a finite, positive definite matrix. This of course implies that $n^{-1}\mathbf{x}'\mathbf{x}$ has full rank, not just for a finite number of observations, which would be ensured by rank $\mathbf{x} = k$, but also in the limit as $n \to \infty$. It also requires that the entries in $n^{-1}\mathbf{x}'\mathbf{x}$ be bounded (or else the limit would not exist). Note that this assumption poses no conceptual difficulty if $\mathbf{x}$ is fixed and designed by the researcher (or someone else), since the $\mathbf{x}$ matrix can then always be designed to behave in the required manner. If $\mathbf{x}$ is not designed, it must be assumed that the $\mathbf{x}_{ij}$'s are generated in a way that allows the limit of $n^{-1}\mathbf{x}'\mathbf{x}$ to be finite and positive definite.

**Theorem 8.6** (Asymptotic Normality of $\hat{\beta}$—iid Residuals) *Assume the classical assumptions of the GLM. In addition, assume that $\{\varepsilon_n\}$ is a collection of iid random variables and that $P(|\varepsilon_i| < m) = 1$ for $m < \infty$ and $\forall \ i$. Finally, assume that the explanatory variables are such that $|x_{ij}| < \xi < \infty \ \forall \ i$ and $j$, and that*

[10] We should note, however, that in the specification of some linear models, certain proxy variables might be used to explain y that literally violate the boundedness assumption. For example, a linear "time trend" $t = 1, 2, 3, 4, \ldots$ is sometimes used to explain an apparent upward or downward trend in $\mathrm{E}Y_t$, and the trend clearly violates the boundedness constraint. In such cases, one may wonder whether it is really to be believed that $t \to \infty$ is relevant in explaining y, or whether the time trend is just an artifice relevant for a certain range of observations, but for which extrapolation ad infinitum is inadvisable.

$\lim_{n\to\infty} n^{-1}\mathbf{x}'\mathbf{x} = \mathbf{Q}$, *where* $\mathbf{Q}$ *is a finite, positive definite matrix. Then*

$$n^{1/2}(\hat{\boldsymbol\beta} - \boldsymbol\beta) \xrightarrow{d} N([0], \sigma^2\mathbf{Q}^{-1}) \quad \textit{and} \quad \hat{\boldsymbol\beta} \overset{a}{\sim} N(\boldsymbol\beta, n^{-1}\sigma^2\mathbf{Q}^{-1}).$$

**Proof** Recall that the least-squares estimator can be expressed as $\hat{\boldsymbol\beta} = \boldsymbol\beta + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol\varepsilon$. Then

$$n^{1/2}(\hat{\boldsymbol\beta} - \boldsymbol\beta) = n^{1/2}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol\varepsilon = (n^{-1}\mathbf{x}'\mathbf{x})^{-1}n^{-1/2}\mathbf{x}'\boldsymbol\varepsilon.$$

Note that $n^{-1/2}\mathbf{x}'\boldsymbol\varepsilon$ can be written as the sum of $n$ $k$-vectors as

$$n^{-1/2}\mathbf{x}'\boldsymbol\varepsilon = n^{-1/2}\sum_{i=1}^{n} \underset{(k\times 1)}{\mathbf{x}'_{i.}\varepsilon_i}$$

(where $\mathbf{x}_{i.}$ is the $(1 \times k)$ vector representing the $i$th row of $\mathbf{x}$) since

$$\underset{(k\times n)(n\times 1)}{\mathbf{x}'\boldsymbol\varepsilon} = [\mathbf{x}'_{1.}\mathbf{x}'_{2.}\ldots\mathbf{x}'_{n.}]\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{bmatrix}.$$

The expectation of the $(k\times 1)$ random vector $\mathbf{x}'_{i.}\varepsilon_i$ is given by $\mathrm{E}\mathbf{x}'_{i.}\varepsilon_i = [\mathbf{0}]_{(k\times 1)}$ since $\mathrm{E}\varepsilon_i = 0$. The covariance matrix of the random vector is given by $\mathbf{Cov}\,(\mathbf{x}'_{i.}\varepsilon_i)_{(k\times k)} = \mathrm{E}(\mathbf{x}'_{i.}\varepsilon_i)(\mathbf{x}'_{i.}\varepsilon_i)' = \mathrm{E}\varepsilon_i^2\mathbf{x}'_{i.}\mathbf{x}_{i.} = \sigma^2\mathbf{x}'_{i.}\mathbf{x}_{i.}$.

The random vectors $\mathbf{x}'_{i.}\varepsilon_i$, $i = 1, \ldots, n$, are independent because the $\varepsilon_i$'s are independent, and the entries in the random vector $\mathbf{x}'_{i.}\varepsilon_i$ are bounded in absolute value with probability 1, $\forall\, i$, since $P(|\varepsilon_i| < m) = 1$ and $|x_{ij}| < \xi$, so that $P(|x_{ij}\varepsilon_i| < m\xi) = 1\ \forall\, i, j$. Given that[11]

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\mathbf{Cov}(\mathbf{x}'_{i.}\varepsilon_i) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\sigma^2\mathbf{x}'_{i.}\mathbf{x}_{i.} = \sigma^2\lim_{n\to\infty} n^{-1}(\mathbf{x}'\mathbf{x}) = \sigma^2\mathbf{Q},$$

and since $\mathbf{Q}$ is a finite, positive definite matrix, it follows directly from Theorem 5.38, concerning a multivariate CLT for independent bounded random vectors, that

$$n^{-1/2}\sum_{i=1}^{n}\mathbf{x}'_{i.}\varepsilon_i = n^{-1/2}\mathbf{x}'\boldsymbol\varepsilon \xrightarrow{d} N\left(\underset{(k\times 1)}{[0]}, \underset{(k\times k)}{\sigma^2\mathbf{Q}}\right).$$

Then, by Slutsky's theorem, since $\lim_{n\to\infty} n^{-1}\mathbf{x}'\mathbf{x} = \operatorname{plim} n^{-1}\mathbf{x}'\mathbf{x} = \mathbf{Q}$, and thus $\lim_{n\to\infty}(n^{-1}\mathbf{x}'\mathbf{x})^{-1} = \operatorname{plim}(n^{-1}\mathbf{x}'\mathbf{x})^{-1} = (\operatorname{plim} n^{-1}\mathbf{x}'\mathbf{x})^{-1} = \mathbf{Q}^{-1}$, we have that

$$n^{1/2}(\hat{\boldsymbol\beta} - \boldsymbol\beta) = n^{1/2}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol\varepsilon = (n^{-1}\mathbf{x}'\mathbf{x})^{-1}n^{-1/2}\mathbf{x}'\boldsymbol\varepsilon \xrightarrow{d} N([0], \sigma^2\mathbf{Q}^{-1}),$$

which proves the limiting distribution part of the theorem.

[11] Note that $\mathbf{x}'\mathbf{x} = [\mathbf{x}'_{1.}\ \mathbf{x}'_{2.}\ \ldots\ \mathbf{x}'_{n}]\begin{bmatrix}\mathbf{x}_{1.}\\ \mathbf{x}_{2.}\\ \vdots\\ \mathbf{x}_{n.}\end{bmatrix}$ so that $\sum_{i=1}^{n}\mathbf{x}'_{i.}\mathbf{x}_{i.} = \mathbf{x}'\mathbf{x}$.

The asymptotic distribution then follows directly from Def. 5.10. ■

While the previous theorem establishes the asymptotic normality of the least-squares or BLUE estimator $\hat{\beta}$, the asymptotic distribution is expressed in terms of the generally unobservable limit matrix $\mathbf{Q}$. We now present a corollary to Theorem 8.6 that replaces $n^{-1}\mathbf{Q}^{-1}$ with the observable matrix $(\mathbf{x}'\mathbf{x})^{-1}$, leading to a covariance matrix for the asymptotic distribution that is identical to the finite sample covariance matrix of $\hat{\beta}$.

**Corollary 8.1** (Alternative Normal Asymptotic Distribution for $\hat{\beta}$—iid Residuals) *Under the conditions of Theorem 8.6,*

$$(\mathbf{x}'\mathbf{x})^{1/2}(\hat{\beta}-\beta) \xrightarrow{d} N([0],\sigma^2\mathbf{I}) \quad \textit{and} \quad \hat{\beta} \overset{a}{\sim} N(\beta,\sigma^2(\mathbf{x}'\mathbf{x})^{-1}).$$

**Proof** By Theorem 8.6 and Slutsky's theorem, it follows that

$$\mathbf{Q}^{1/2}n^{1/2}(\hat{\beta}-\beta) \xrightarrow{d} N([0],\sigma^2\mathbf{I}).$$

Since $(n^{-1}\mathbf{x}'\mathbf{x})^{1/2}\mathbf{Q}^{-1/2} \rightarrow \mathbf{Q}^{1/2}\mathbf{Q}^{-1/2} = \mathbf{I}$ because $n^{-1}\mathbf{x}'\mathbf{x} \rightarrow \mathbf{Q}$,[12] Slutsky's theorem also implies that

$$(n^{-1}\mathbf{x}'\mathbf{x})^{1/2}\mathbf{Q}^{-1/2}\mathbf{Q}^{1/2}n^{1/2}(\hat{\beta}-\beta) = (\mathbf{x}'\mathbf{x})^{1/2}(\hat{\beta}-\beta) \xrightarrow{d} N([0],\sigma^2\mathbf{I}),$$

so that $\hat{\beta} \overset{a}{\sim} N(\beta,\sigma^2(\mathbf{x}'\mathbf{x})^{-1})$. ■

The asymptotic normality of the least-squares estimator can be demonstrated more generally without reliance on the boundedness of the disturbance terms, without the iid assumption on the error terms, and without assuming the existence of a finite, positive definite limit of $n^{-1}\mathbf{x}'\mathbf{x}$ as $n \rightarrow \infty$. The demonstration relies on somewhat more complicated central limit arguments. We state one such result next.

**Theorem 8.7** (Asymptotic Normality of $\hat{\beta}$—Non-iid Residuals) *Assume the classical assumptions of the GLM. Then* $(\mathbf{x}'\mathbf{x})^{1/2}(\hat{\beta}-\beta) \xrightarrow{d} N([0],\sigma^2\mathbf{I})$ *and* $\hat{\beta}\overset{a}{\sim}N(\beta,\sigma^2(\mathbf{x}'\mathbf{x})^{-1})$ *if the elements of the sequence* $\{\varepsilon_n\}$ *are independent random variables such that* $\mathrm{E}\varepsilon_i^4 \leq \tau < \infty\, \forall\, i$, *and the explanatory variables are such that* $|x_{ij}| < \xi < \infty\, \forall\, i$ *and* $j$ *with* $\det(n^{-1}\mathbf{x}'\mathbf{x}) > \eta > 0$ *and with* $\{n^{-1}\mathbf{x}'\mathbf{x}\}$ *being* $O(1)$.

***Proof** The proof is based on the Liapounov CLT for triangular arrays (Theorem 5.34) and the Cramer-Wold device (Corollary 5.4). Define $\mathbf{V}_n = n^{-1}\mathbf{x}'\mathbf{x}$, and note that

$$(\mathbf{x}'\mathbf{x})^{1/2}(\hat{\beta}-\beta) = n^{-1/2}\mathbf{V}_n^{-1/2}\mathbf{x}'\varepsilon = n^{-1/2}\sum_{t=1}^{n}\mathbf{V}_n^{-1/2}\mathbf{x}'_{t.}\varepsilon_t.$$

[12] Regarding the symmetric square root matrix $\mathbf{A}^{1/2}$, note that $\mathbf{P}'\mathbf{A}\mathbf{P} = \Lambda$, where $\mathbf{P}_{(k\times k)}$ is the orthogonal matrix of characteristic vectors of the symmetric, positive semidefinite matrix $\mathbf{A}$, and $\Lambda_{(k\times k)}$ is the diagonal matrix of characteristic roots. Then $\mathbf{A}^{1/2} = \mathbf{P}\Lambda^{1/2}\mathbf{P}'$, where $\Lambda^{1/2}$ is the diagonal matrix formed from $\Lambda$ by taking the square root of each diagonal element. Note that since $\mathbf{P}'\mathbf{P} = \mathbf{I}$, $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{P}\Lambda^{1/2}\mathbf{P}'\mathbf{P}\Lambda^{1/2}\mathbf{P}' = \mathbf{P}\Lambda^{1/2}\Lambda^{1/2}\mathbf{P}' = \mathbf{P}\Lambda\mathbf{P}' = \mathbf{A}$. The matrix square root is a continuous function of the elements of $\mathbf{A}$. Therefore, $\lim_{n\rightarrow\infty}(\mathbf{A}_n)^{1/2} = (\lim_{n\rightarrow\infty}\mathbf{A}_n)^{1/2}$.

Examine

$$\ell'(\mathbf{x}'\mathbf{x})^{1/2}(\hat{\beta}-\beta) = n^{-1/2}\sum_{t=1}^{n}\ell'\mathbf{V}_n^{-1/2}\mathbf{x}'_{t.}\varepsilon_t = n^{-1/2}\sum_{t=1}^{n}W_{nt},$$

where $\ell$ is any conformable vector such that $\ell'\ell = 1$. Note that $\mathrm{E}W_{nt} = 0$ and $\mathrm{var}(W_{nt}) = \sigma^2\ell'\mathbf{V}_n^{-1/2}\mathbf{x}'_{t.}\mathbf{x}_{t.}\mathbf{V}_n^{-1/2}\ell < \infty \;\forall\; n$ and $t$, where the finiteness of $\mathrm{var}(W_{nt})$ follows from the boundedness of $\mathrm{E}W_{nt}^4$, which is shown below.

Since $\mathrm{E}W_{nt} = 0$, $\mathrm{E}|W_{nt} - \mathrm{E}W_{nt}|^4 = \mathrm{E}W_{nt}^4$, and then

$$\mathrm{E}W_{nt}^4 = (\mathrm{E}\varepsilon_t^4)[\ell'\mathbf{V}_n^{-1/2}\mathbf{x}'_{t.}\mathbf{x}_{t.}\mathbf{V}_n^{-1/2}\ell]^2 \leq \tau[\ell'\mathbf{V}_n^{-1/2}\mathbf{x}'_{t.}\mathbf{x}_{t.}\mathbf{V}_n^{-1/2}\ell]^2.$$

Now note that $\ell'\mathbf{V}_n^{-1/2}\mathbf{x}'_{t.}\mathbf{x}_{t.}\mathbf{V}_n^{-1/2}\ell < c < \infty \;\forall\; n$ and $t$. This follows from the fact that $|\mathbf{V}_n| > \eta > 0 \;\forall\; n$ with $\{\mathbf{V}_n\}$ being $0(1) \Rightarrow \{\mathbf{V}_n^{-1/2}\}$ is $0(1)$ (H. White (1982), "Instrumental variables regression with independent observations." *Econometrica*, p. 484–485), and since $\mathbf{x}'_{t.}\mathbf{x}_t$ is a positive semidefinite matrix whose elements are all bounded in absolute value by $\xi^2$, it follows that $\mathbf{V}_n^{-1/2}\mathbf{x}'_{t.}\mathbf{x}_{t.}\mathbf{V}_n^{-1/2}$ is an $0(1)$ positive semidefinite matrix, so that $\ell'\mathbf{V}_n^{-1/2}\mathbf{x}'_{t.}\mathbf{x}_{t.}\mathbf{V}_n^{-1/2}\ell$ is bounded as claimed. Then, $\mathrm{E}W_{nt}^4 < \tau c^2 = \gamma < \infty \;\forall\; n$ and $t$.

Given the preceding results, Theorem 5.34 is applicable. Note that the $W_{nt}$'s can be represented in the form of a triangular array (Def. 5.17) with typical row $(W_{n1}, \ldots, W_{nn})$, $\forall\; n$, and the $W_{nt}$'s are independent within rows since the $\varepsilon_t$'s are independent. The limit condition of Theorem 5.34 is met, since

$$0 \leq \lim_{n\to\infty}\left[\frac{\sum_{t=1}^{n}\mathrm{E}W_{nt}^4}{\left[\sum_{t=1}^{n}\mathrm{var}(W_{nt})\right]^2}\right] \leq \gamma\lim_{n\to\infty}\left[\frac{n}{n^2\sigma^4}\right] = 0.$$

Then, because $n^{-1}\sum_{t=1}^{n}\mathrm{var}(W_{nt}) = \sigma^2$, it follows from Theorem 5.34 that

$$\ell'(\mathbf{x}'\mathbf{x})^{1/2}(\hat{\beta}-\beta) = n^{1/2}\left[n^{-1}\sum_{t=1}^{n}W_{nt}\right] \xrightarrow{\mathrm{d}} N(0,\sigma^2),$$

and by the Cramer–Wold device $(\mathbf{x}'\mathbf{x})^{1/2}(\hat{\beta}-\beta) \xrightarrow{\mathrm{d}} N([\mathbf{0}],\sigma^2\mathbf{I})$, so that $\hat{\beta} \overset{\mathrm{a}}{\sim} N(\beta,\sigma^2(\mathbf{x}'\mathbf{x})^{-1})$. ■

The difference in assumptions between our previous result on the asymptotic normality of $\hat{\beta}$ and the result in Theorem 8.7 is that the existence of moments of the $\varepsilon_i$'s of order four replaces the assumption that the errors are bounded with probability 1, the $\varepsilon_i$'s are no longer required to be identically distributed (although first- and second-order moments are assumed to be the same by the classical GLM assumptions), and bounds on the determinant and elements of $n^{-1}\mathbf{x}'\mathbf{x}$ replace the assumption that $\lim_{n\to\infty} n^{-1}\mathbf{x}'\mathbf{x}$ exists and is positive definite. The latter assumptions are less restrictive than the assumption that $\lim_{n\to\infty} n^{-1}\mathbf{x}'\mathbf{x}$ exists and is a nonsingular positive definite matrix, especially in cases where there is no control over how the $\mathbf{x}$ matrix was generated and no knowledge that the process generating the $\mathbf{x}$ matrix would inherently lead to the existence of a limit for $n^{-1}\mathbf{x}'\mathbf{x}$.

Regarding the moment condition on the $\varepsilon_i$'s, note that our previous assumption $P(|\varepsilon_i| < m) = 1 \; \forall \, i$ implies boundedness of the moments and thus implies the moment condition in Theorem 8.7. However, the moment conditions in Theorem 8.7 also allow situations in which there is no absolute upper bound to the value of the error term that holds with probability 1, and thus, for example, allows the $\varepsilon_i$'s to have a normal distribution or a mean-shifted (to zero) Gamma distribution, whereas the former assumption would not.

In summary, the least-squares estimator $\hat{\beta}$ can be approximately normally distributed, even if $\varepsilon$ is not multivariate normally distributed, given certain conditions on the values of the explanatory variables and certain assumptions regarding distributional characteristics of the disturbance vector. Note that asymptotic normality of $\hat{\beta}$ can be demonstrated using still weaker assumptions—the interested reader is directed to Chapter 5 of H. White (1984), *Asymptotic Theory for Econometricians*, Cambridge, MA: Academic Press, for further details.

### Asymptotic Normality of $\hat{S}^2$

If, in addition to the classical GLM assumptions, it is assumed that the $\varepsilon_i$'s are iid and have bounded fourth-order moments about the origin, then $\hat{S}^2$ will be asymptotically normally distributed.

**Theorem 8.8** (Asymptotic Normality of $\hat{S}^2$—iid Residuals) *Under the classical assumptions of the GLM, if the elements of the residual vector, $\varepsilon$, are iid, and if* $\mathrm{E}\varepsilon_i^4 \le \tau < \infty$, *then*

$$n^{1/2}(\hat{S}^2 - \sigma^2) \xrightarrow{d} N(0, \mu_4' - \sigma^4) \quad \textit{and} \quad \hat{S}^2 \overset{a}{\sim} N(\sigma^2, n^{-1}(\mu_4' - \sigma^4)).$$

**Proof** Note that

$$n^{1/2}(\hat{S}^2 - \sigma^2) = \frac{n^{1/2}}{(n-k)}\varepsilon'(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\varepsilon - n^{1/2}\sigma^2 = \left[\frac{n^{1/2}\varepsilon'\varepsilon}{(n-k)} - n^{1/2}\sigma^2\right] - \frac{n^{1/2}}{n-k}\varepsilon'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon.$$

Based on Markov's inequality,

$$P\left(\left(\frac{n^{1/2}}{n-k}\right)\varepsilon'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon \ge c\right) \le \frac{\sigma^2 n^{1/2} k}{c(n-k)}$$

(recall the proof of Theorem 8.4), and since $\sigma^2 n^{1/2} k / c(n-k) \to 0$ as $n \to \infty$, $\text{plim}(n^{1/2}/(n-k))\varepsilon'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon = 0$. Then it follows from Slutsky's theorem that the limiting density of $n^{1/2}(\hat{S}^2 - \sigma^2)$ will depend only on the bracketed term in its definition above. Also by Slutsky's theorem, the limiting density is unaffected if we multiply the bracketed term by $(n-k)/n$, since $(n-k)/n \to 1$ as $n \to \infty$, or if we then also add the term $-k\sigma^2/n^{1/2}$ since $-k\sigma^2/n^{1/2} \to 0$ as $n \to \infty$. Thus, to establish the limiting density of $n^{1/2}(\hat{S}^2 - \sigma^2)$, it suffices to examine the limiting density of

$$\frac{\varepsilon'\varepsilon}{n^{1/2}} - \frac{(n-k)\sigma^2}{n^{1/2}} - \frac{k\sigma^2}{n^{1/2}} = \frac{\varepsilon'\varepsilon - n\sigma^2}{n^{1/2}}.$$

Since $\mathrm{E}\varepsilon_i^4 \leq \tau < \infty \Leftrightarrow \mathrm{var}(\varepsilon_i^2) \leq \tau - \sigma^4 < \infty$, a direct application of the Lindberg-Levy CLT to the sequence of iid random variables $\{\varepsilon_1^2, \varepsilon_2^2, \varepsilon_3^2, \ldots\}$ yields

$$\frac{\varepsilon'\varepsilon - n\sigma^2}{n^{1/2}(\mu_4' - \sigma^4)^{1/2}} \xrightarrow{d} N(0,1).$$

Consequently, by Slutsky's theorem,

$$\frac{\varepsilon'\varepsilon - n\sigma^2}{n^{1/2}} \xrightarrow{d} N(0, \mu_4' - \sigma^4).$$

Therefore,

$$n^{1/2}(\hat{S}^2 - \sigma^2) \xrightarrow{d} N(0, \mu_4' - \sigma^4) \quad \text{and} \quad \hat{S}^2 \overset{a}{\sim} N\left(\sigma^2, \frac{\mu_4' - \sigma^4}{n}\right). \qquad \blacksquare$$

The asymptotic normality of $\hat{S}^2$ can be established without assuming that the disturbance terms are iid. We present one such result in the following theorem.

**Theorem 8.9** (Asymptotic Normality of $\hat{S}^2$—Non-iid Residuals) *Assume the classical assumptions of the GLM. Also assume that the $\varepsilon_i$'s are independent and* $\mathrm{E}|\varepsilon_i^2 - \sigma^2|^{2+\delta} \leq \tau < \infty \ \forall\, i$ *for some* $\delta > 0$. *Then, letting* $\xi_n = n^{-1}\sum_{i=1}^n \mathrm{var}(\varepsilon_i^2) \geq \eta > 0$ $\forall\, n$, $n^{1/2}(\hat{S}^2 - \sigma^2)/\xi_n^{1/2} \xrightarrow{d} N(0,1)$ *and* $\hat{S}^2 \overset{a}{\sim} N(\sigma^2, n^{-1}\xi_n)$.

***Proof** Similar to the proof of Theorem 8.8, we can show that the limiting distributions of $n^{1/2}(\hat{S}^2 - \sigma^2)/\xi_n^{1/2}$ and of $(\sum_{i=1}^n \varepsilon_i^2 - n\sigma^2)/n^{1/2}\xi_n^{1/2}$ coincide. To see this, note that by Markov's inequality,

$$P\left(\frac{n^{1/2}}{(n-k)\xi_n^{1/2}}\varepsilon'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}\varepsilon \geq c\right) \leq \frac{\sigma^2 n^{1/2} k}{c(n-k)\xi_n^{1/2}},$$

and if $\xi_n \geq \eta > 0$, $\forall\, n$, then the RHS of the inequality $\rightarrow 0$ as $n \rightarrow \infty$, so that $\mathrm{plim}(n^{1/2}/(n-k)\xi_n^{1/2})\varepsilon'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon = 0$. The remaining steps for showing the equivalence of the limiting distributions are analogous to those in the proof of Theorem 8.8.

Liapounov's CLT (Theorem 5.33) can now be applied to

$$\left(\sum_{i=1}^n \varepsilon_i^2 - n\sigma^2\right) \bigg/ n^{1/2}\xi_n^{1/2}.$$

Note that the $\varepsilon_i^2$'s are independent, with $\mathrm{E}\varepsilon_i^2 = \sigma^2$ and $\mathrm{var}(\varepsilon_i^2) = \mathrm{E}(\varepsilon_i^2 - \sigma^2)^2 < \infty$ $\forall\, i$ (the existence of $\mathrm{var}(\varepsilon_i^2)$ follows from the existence of $\mathrm{E}|\varepsilon_i^2 - \sigma^2|^{2+\delta}$ for $\delta > 0$;

recall Theorem 3.22). Furthermore, the limit condition of the Liapounov CLT is met, since

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} \mathrm{E}|\varepsilon_i^2 - \sigma^2|^{2+\delta}}{\left[\sum_{i=1}^{n} \mathrm{var}\,(\varepsilon_i^2)\right]^{1+\delta/2}} \le \lim_{n\to\infty} \frac{n\tau}{(n\eta)^{1+\delta/2}} = \lim_{n\to\infty} (\tau/\eta^{1+\delta/2}) n^{-\delta/2} = 0.$$

Therefore, by the Liapounov CLT,

$$\left(\sum_{i=1}^{n} \varepsilon_i^2 - n\sigma^2\right) \Big/ (n\xi_n)^{1/2} \xrightarrow{\mathrm{d}} N(0,1),$$

which in turn implies that

$$\frac{n^{1/2}(\hat{S}^2 - \sigma^2)}{\xi_n^{1/2}} \xrightarrow{\mathrm{d}} N(0,1) \quad \text{and} \quad \hat{S}^2 \overset{\mathrm{a}}{\sim} N(\sigma^2, n^{-1}\xi_n). \qquad \blacksquare$$

### Summary of Estimator Properties

Table 8.1 summarizes assumptions and the resultant properties of $\hat{\beta}$ and $\hat{S}^2$ as estimators of $\beta$ and $\sigma^2$ in the GLM context.

**Example 8.2** GLM When Random Sampling from a Population Distribution

Let $(Y_1, \ldots, Y_n)$ be a random sample from a population distribution having finite mean $\beta$ and finite variance $\sigma^2$ (*both* scalars). The GLM representation of the vector **Y** is given by

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \mathbf{x}\beta + \boldsymbol{\varepsilon} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where $\mathrm{E}\boldsymbol{\varepsilon} = [\mathbf{0}]$, and since the $\varepsilon_i$'s are iid, $\mathrm{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \sigma^2\mathbf{I}$. Then note that $\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y} = n^{-1}\sum_{i=1}^{n} Y_i = \bar{Y}$, so that the least-squares estimator of $\beta$ is the sample mean. Furthermore, $\hat{S}^2 = (\mathbf{Y} - \mathbf{x}\hat{\beta})'(\mathbf{Y} - \mathbf{x}\hat{\beta})/(n-k) = \sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)$ is used as an estimator of $\sigma^2$. The reader can verify that cases 2 and 3 in Table 8.1 are met, and thus $\hat{\beta}$ is BLUE and consistent for $\beta$, while $\hat{S}^2$ is an unbiased and consistent estimator for $\sigma^2$. Under additional assumptions such as cases 5–8 in Table 8.1, or recall those discussed in Chapter 6 concerning the asymptotic normality of the sample mean and sample variance, $\hat{\beta}$ and $\hat{S}^2$ are both asymptotically normally distributed. □

**Example 8.3** A realtor is analyzing the relationship between the number of new one-family houses sold in a given year in the United States and various explanatory variables that she feels had important impacts on housing sales during the 1980s. She has collected the following information:

**Table 8.1** Assumptions and Properties of $\hat{\beta}$ and $\hat{S}^2$ in the GLM Point Estimation Problem

| | | | | Properties | |
|---|---|---|---|---|---|
| **Case** | **Classical GLM Assumptions** | **Additional x Assumptions** | **Additional $\varepsilon$ Assumptions** | $\hat{\beta}$ | $\hat{S}^2$ |
| 1 | Hold | — | — | BLUE | Unbiased |
| 2 | Hold | $(\mathbf{x}'\mathbf{x})^{-1} \rightarrow [0]$ | — | Consistent | — |
| 3 | Hold | — | $\varepsilon_i$'s are iid | — | Consistent |
| 4 | Hold | — | $\mathrm{E}\varepsilon_i^4 \le \tau < \infty$<br>$\left\{ \begin{array}{c} \{\varepsilon_n\} \text{ is } m\text{-dependent} \\ or \\ \operatorname{cov}(\varepsilon_i^2, \varepsilon_j^2) \rightarrow 0, \lvert i-j \rvert \rightarrow \infty \end{array} \right\}$ | —<br>— | Consistent |
| 5 | Hold | $\lvert x_{ij} \rvert < \xi < \infty$<br>$n^{-1}\mathbf{x}'\mathbf{x} \rightarrow \mathbf{Q}$ positive definite | $\varepsilon_i$'s are iid<br>$P(\lvert \varepsilon_i \rvert < m) = 1$ | $\overset{a}{\sim} N(\beta, \sigma^2(\mathbf{x}'\mathbf{x})^{-1})$ | — |
| | | *or* | | | |
| 6 | Hold | $\lvert x_{ij} \rvert < \xi < \infty$<br>$\det(n^{-1}\mathbf{x}'\mathbf{x}) > \eta > 0$<br>$\{n^{-1}\mathbf{x}'\mathbf{x}\}$ is $O(1)$ | $\varepsilon_i$'s are independent<br>$\mathrm{E}\varepsilon_i^4 \le \tau < \infty$ | | |
| 7 | Hold | — | $\varepsilon_i$'s are iid, $\mathrm{E}\varepsilon_i^4 \le \tau < \infty$<br>*or* | — | $\overset{a}{\sim} N(\sigma^2, n^{-1}(\mu_4' - \sigma^4))$ |
| 8 | Hold | — | $\varepsilon_i$'s are independent<br>$\mathrm{E}\lvert \varepsilon_i^2 - \sigma^2 \rvert^{2+\delta} \le \tau < \infty, \delta > 0$<br>$\xi_n = n^{-1} \sum_{i=1}^{n} \operatorname{var}(\varepsilon_i^2) \ge \eta > 0$ | — | $\overset{a}{\sim} N(\sigma^2, n^{-1}\xi_n)$ |

| | #New Homes | Conventional | Medium | New Home | CPI |
|---|---|---|---|---|---|
| | Sold | Mortgage | Family | Purchase Price | (1982–1984 |
| | (1000s) | Interest Rate | Income | (1000's) | = 100) |
| 1980 | 545 | 13.95 | 21,023 | 83.2 | 82.4 |
| 1981 | 436 | 16.52 | 22,388 | 90.3 | 90.9 |
| 1982 | 412 | 15.79 | 23,433 | 94.1 | 96.5 |
| 1983 | 623 | 13.43 | 24,580 | 93.9 | 99.6 |
| 1984 | 639 | 13.80 | 26,433 | 96.8 | 103.9 |
| 1985 | 688 | 12.28 | 27,735 | 105.0 | 107.6 |
| 1986 | 750 | 10.07 | 29,458 | 119.8 | 109.6 |
| 1987 | 671 | 10.17 | 30,970 | 137.2 | 113.6 |
| 1988 | 676 | 10.30 | 32,191 | 150.5 | 118.3 |
| 1989 | 650 | 10.21 | 34,213 | 160.1 | 124.0 |
| $\bar{z}$ | 609 | 12.65 | 27,242 | 113.1 | 104.6 |

*Source*: *Statistical Abstracts of the United States, 1991,* U.S. Dept. of Commerce, Washington D.C., Tables 1272, 837, 730, 825, and 771, respectively.

She specifies the dependent variable vector, **Y**, as the numbers of new homes sold, and the explanatory variable matrix, **x**, contains a column of 1's (for the intercept), followed by the remaining four variables in the table above. That is,

$\mathbf{x}[.,1] = 1\text{'s},$

$\mathbf{x}[.,2] =$ Conventional Mortgage Interest Rates,

$\mathbf{x}[.,3] =$ Median Family Incomes,

$\mathbf{x}[.,4] =$ New Home Purchase Prices,

$\mathbf{x}[.,5] =$ CPIs,

and the GLM is given by $\mathbf{Y} = \sum_{i=1}^{5} \beta_i \mathbf{x}[.,i] + \varepsilon$. She calculates the following values on her personal computer:

$$\mathbf{b} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} = \begin{bmatrix} 1074.82 \\ -46.01 \\ .0376 \\ -6.16 \\ -2.03 \end{bmatrix}, \quad \hat{s}^2 = (\mathbf{y} - \mathbf{x}\mathbf{b})'(\mathbf{y} - \mathbf{x}\mathbf{b})/5 = 905.32, \quad R^2 = .96.$$

The linear model thus explains 96 percent of the observed sample variation in new homes sold in terms of the values of mortgage interest rates, median family incomes, home prices, and the general cost of living. An estimate of the covariance matrix of the entries in $\hat{\beta}$ is calculated as

$$\hat{s}^2(\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} 79586.35 & -3389.91 & -5.9323 & 273.0789 & 896.8833 \\ & 191.4445 & .4094 & -14.5792 & -81.6034 \\ & & .0013 & -.0538 & -.2665 \\ & \text{(symmetric)} & & 3.3401 & 9.5606 \\ & & & & 60.3537 \end{bmatrix}.$$

The entries in **b** divided by their respective estimated standard deviations are given as (3.80, −3.33, 1.05, −3.37, −.26). Thus, the parameters associated with the intercept, interest rates, and new home prices are each more than three standard deviations away from zero (further insight into the significance of this observation will be provided in Chapter 10).

The realtor calculates estimates of elasticities of home sales with respect to each of the explanatory variables, other than the intercept, evaluated at the means of the observations, as

$$e_{\text{sales, rate}} = b_2 \frac{\bar{x}_2}{\bar{y}} = -46.01 \left(\frac{12.65}{609}\right) = -.96,$$

$$e_{\text{sales, income}} = b_3 \frac{\bar{x}_3}{\bar{y}} = .0376 \left(\frac{27,242}{609}\right) = 1.68,$$

$$e_{\text{sales, price}} = b_4 \frac{\bar{x}_4}{\bar{y}} = -6.16 \left(\frac{113.1}{609}\right) = -1.14,$$

$$e_{\text{sales, CPI}} = b_5 \frac{\bar{x}_5}{\bar{y}} = -2.03 \left(\frac{104.6}{609}\right) = -.35.$$

Based on the estimated elasticities, home sales appear to be quite responsive to changes in interest rates, income levels, and home prices. Response to changes in the general cost of living is notably inelastic. □

### Violations of Classic GLM Assumptions

The classical assumptions of the GLM given in Def. 8.1 form essentially the base-level set of assumptions on which useful properties of the least-squares estimator depend. It is instructive to examine the effect that a violation in each of the classical assumptions has on the basic estimator properties of unbiasedness, BLUE, and consistency.

*Assumption Violation:* $\mathrm{E}\boldsymbol{\varepsilon} \neq [\mathbf{0}]$ In this section we assume that $\mathrm{E}\boldsymbol{\varepsilon} \neq [\mathbf{0}]$, but we retain the other classical GLM assumptions. Note that the covariance assumption must then be restated to accommodate the nonzero mean of $\boldsymbol{\varepsilon}$, as $\mathbf{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I} = \mathbf{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon} - \mathrm{E}\boldsymbol{\varepsilon})(\boldsymbol{\varepsilon} - E\boldsymbol{\varepsilon})'$.

*$\hat{\boldsymbol{\beta}}$ Properties* If $\mathrm{E}\boldsymbol{\varepsilon} = \boldsymbol{\phi} \neq [\mathbf{0}]$, then since $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\varepsilon}$, $\mathrm{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\phi}$. Except for the special case where $\mathbf{x}'\boldsymbol{\phi} = [\mathbf{0}]$, the estimator $\hat{\boldsymbol{\beta}}$ is a *biased* estimator of $\boldsymbol{\beta}$ since $(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\phi} = [\mathbf{0}]$ iff $\mathbf{x}'\boldsymbol{\phi} = (\mathbf{x}'\mathbf{x})[\mathbf{0}] = [\mathbf{0}]$. If $\mathbf{x}'\boldsymbol{\phi} \neq [\mathbf{0}]$, then $\hat{\boldsymbol{\beta}}$ is clearly *not* BLUE since $\hat{\boldsymbol{\beta}}$ is not even unbiased. In the special case where $\mathbf{x}'\boldsymbol{\phi} = [\mathbf{0}]$, $\hat{\boldsymbol{\beta}}$ retains the BLUE property, as can be verified by repeating the previous derivation of the BLUE for $\boldsymbol{\beta}$ using the conditions $E\boldsymbol{\varepsilon} = \boldsymbol{\phi}$ and $\mathbf{x}'\boldsymbol{\phi} = [\mathbf{0}]$.

Regarding consistency, it follows from $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\varepsilon}$, that $\operatorname{plim}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ iff $\operatorname{plim}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\varepsilon} = [\mathbf{0}]$. If we assume as before that $(\mathbf{x}'\mathbf{x})^{-1} \to [\mathbf{0}]$ when $n \to \infty$ (Table 8.1, condition 2), then since $\mathrm{E}[(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\varepsilon} - (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\phi}] = [\mathbf{0}]$ and

$\mathbf{Cov}((\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon) = \sigma^2(\mathbf{x}'\mathbf{x})^{-1} \rightarrow [\mathbf{0}]$ as $n \rightarrow \infty$, it follows from mean-square convergence that $\text{plim}((\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon - (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\phi) = [\mathbf{0}]$. Thus, if $\lim_{n\rightarrow\infty}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\phi = \boldsymbol{\xi} \neq [\mathbf{0}]$, or if $(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\phi$ does not converge to a limit, then $\hat{\boldsymbol{\beta}}$ would *not* be consistent for $\boldsymbol{\beta}$ since $\text{plim}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon \neq [\mathbf{0}]$ and $\text{plim}\,\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$.

*$\hat{S}^2$ Properties* Since $\hat{S}^2 = (n-k)^{-1}$ $\text{tr}[(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\varepsilon\varepsilon']$, $\mathbf{Cov}(\varepsilon) = \sigma^2\mathbf{I}$, and $\text{E}\varepsilon\varepsilon' = \sigma^2\mathbf{I} + \phi\phi'$, the expectation of $\hat{S}^2$ is given by

$$\text{E}\hat{S}^2 = \sigma^2 + \frac{\phi'(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\phi}{n-k}.$$

Unless $\phi'(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\phi = 0$, the estimator $\hat{S}^2$ is a *biased* estimator of $\sigma^2$ and the bias is nonnegative, since the matrix $\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$ is symmetric and idempotent, and hence positive semidefinite.[13] Note that if $\text{E}\varepsilon = \phi \neq [\mathbf{0}]$, it is *impossible* for both $\hat{\boldsymbol{\beta}}$ and $\hat{S}^2$ to remain unbiased. To see this, first note that if $\mathbf{x}'\phi = [\mathbf{0}]$, so that $\hat{\boldsymbol{\beta}}$ is unbiased, then the bias in $\hat{S}^2$ is $(E\hat{S}^2 - \sigma^2) = \phi'\phi/(n-k) > 0$. Alternatively, if $\hat{S}^2$ is unbiased, then it must be the case that $\phi'\phi = \phi'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\phi > 0$, and since $\phi'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\phi$ is a positive *definite* quadratic form in the vector $\mathbf{x}'\phi$ (i.e., $(\mathbf{x}'\mathbf{x})^{-1}$ is positive definite), it necessarily follows that $\mathbf{x}'\phi \neq [\mathbf{0}]$ if $\phi'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\phi$ is positive, and thus $\hat{\boldsymbol{\beta}}$ is biased. Thus at least one, and generally both, of the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{S}^2$ will be biased if $\text{E}\varepsilon \neq [\mathbf{0}]$.

Regarding consistency, define $\mathbf{V} = \varepsilon - \phi$, and note that $\hat{S}^2 = (n-k)^{-1}(\phi + \mathbf{V})'(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')(\phi + \mathbf{V})$, where $\text{E}\mathbf{V} = [\mathbf{0}]$ and $\mathbf{Cov}(\mathbf{V}) = \sigma^2\mathbf{I}$. Letting $\mathbf{M} = \mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$, it follows that $\hat{S}^2 = (n-k)^{-1}[\phi'\mathbf{M}\phi - 2\phi'\mathbf{M}\mathbf{V} + \mathbf{V}'\mathbf{M}\mathbf{V}]$, so that $\hat{S}^2$ is consistent iff the expression on the right-hand side of the equality has a probability limit of $\sigma^2$. To illustrate that this is generally not true, assume that the $\mathbf{V}_i$'s are iid, which is tantamount to assuming that the $\varepsilon_i$'s are independent and identically distributed once they are transformed to have mean zero, as $\varepsilon_i - \phi_i$. Then $\text{plim}(n-k)^{-1}\mathbf{V}'\mathbf{M}\mathbf{V} = \sigma^2$, which follows from an argument identical to that used in the proof of Theorem 8.4. It follows that $\hat{S}^2$ will be consistent for $\sigma^2$ iff $(n-k)^{-1}[\phi'\mathbf{M}\phi - 2\phi'\mathbf{M}\mathbf{V}] \xrightarrow{p} 0$.

Now assume $|\phi_i| \leq c < \infty$, so that the means of the residuals are finite and $\phi'\phi = O(n^1)$, or at least assume that $\phi'\phi$ is $o(n^2)$, implying that the sum of the squared means approaches infinity at a rate less than $n^2$. It follows that $\phi'\mathbf{M}\phi$ is $o(n^2)$, since $0 \leq \phi'\mathbf{M}\phi = \phi'\phi - \phi'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\phi \leq \phi'\phi$. Then since $\text{E}(2(n-k)^{-1}\phi'\mathbf{M}\mathbf{V}) = 0$ and $\text{var}(2(n-k)^{-1}\phi'\mathbf{M}\mathbf{V}) = 4\sigma^2\phi'\mathbf{M}\phi/(n-k)^2 \leq 4\sigma^2 o(n^2)/(n-k)^2 \rightarrow 0$ as $n \rightarrow \infty$, $\text{plim}(2(n-k)^{-1}\phi'\mathbf{M}\mathbf{V}) = 0$ follows from mean-square convergence. Thus, except for the special case $\lim_{n\rightarrow\infty}(n-k)^{-1}\phi'\mathbf{M}\phi = 0$, $\hat{S}^2$ will *not* be consistent for $\sigma^2$.

### *Assumption Violation:* $\text{E}\varepsilon\varepsilon' \neq \sigma^2\mathbf{I}$

*$\hat{\boldsymbol{\beta}}$ Properties* If $\text{E}\varepsilon\varepsilon' = \boldsymbol{\Phi} \neq \sigma^2\mathbf{I}$, so that the error terms are **heteroskedastic** and/or **autocorrelated**, then the estimator $\hat{\boldsymbol{\beta}}$ is nonetheless still *unbiased* for $\boldsymbol{\beta}$.

[13] Positive semidefiniteness can be deduced from the fact that the characteristic roots of a symmetric idempotent matrix are all nonnegative, being a collection of 0's and 1's.

This follows because the representation $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\varepsilon}$ is unaffected, and taking expectations still yields $\mathrm{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathrm{E}\boldsymbol{\varepsilon} = \boldsymbol{\beta}$.[14]

To examine consistency, note that now $\mathbf{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\Phi}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}$. If it were true that $\mathrm{var}(\hat{\beta}_i) \to 0$ as $n \to \infty \; \forall \, i$, then since $\mathrm{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, we could conclude by convergence in mean square that $\mathrm{plim}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. Under mild conditions on the $\boldsymbol{\Phi}$ matrix, and assuming as before that $(\mathbf{x}'\mathbf{x})^{-1} \to [\mathbf{0}]$, convergence of $\mathrm{var}(\hat{\beta}_i)$ to zero $\forall \, i$ will be achieved. The argument is facilitated by the following matrix theory lemma.

**Lemma 8.4**

Let $A$ and $B$ be symmetric, positive semidefinite matrices of order $(n \times n)$. Then $\mathrm{tr}(\mathbf{AB}) \le \lambda_L(\mathbf{A})\,\mathrm{tr}\,\mathbf{B}$, where $\lambda_L(\mathbf{A})$ represents the value of the largest characteristic root of the matrix $\mathbf{A}$.

**Proof**: Let $\mathbf{P}$ be the $(n \times n)$ characteristic vector matrix of the symmetric matrix $\mathbf{A}$, so that $\mathbf{P}'\mathbf{AP} = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is the diagonal matrix of characteristic roots of $\mathbf{A}$. Since $\mathbf{PP}' = \mathbf{I}$ by the orthogonality of $\mathbf{P}$, it follows by a property of the trace operator $(\mathrm{tr}(\mathbf{DFG}) = \mathrm{tr}(\mathbf{GDF}))$ that $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{APP}'\mathbf{BPP}') = \mathrm{tr}(\mathbf{P}'\mathbf{APP}'\mathbf{BP}) = \mathrm{tr}(\boldsymbol{\Lambda}\mathbf{C})$, where $\mathbf{C} = \mathbf{P}'\mathbf{BP}$. But then $\mathrm{tr}\,(\boldsymbol{\Lambda}\mathbf{C}) = \sum_{i=1}^{n} \lambda_i C_{ii} \le \lambda_L(\mathbf{A}) \sum_{i=1}^{n} C_{ii}$ where $\lambda_i$ is the $i$th diagonal entry in the diagonal matrix of characteristic roots $\boldsymbol{\Lambda}$, $C_{ii}$ represents the $(i,i)$th entry in $\mathbf{C}$ and $\lambda_i \ge 0$ and $C_{ii} \ge 0 \; \forall \, i$ by the positive semidefiniteness of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{P}'\mathbf{BP}$. But since $\sum_{i=1}^{n} C_{ii} = \mathrm{tr}\,\mathbf{C} = \mathrm{tr}\,\mathbf{P}'\mathbf{BP} = \mathrm{tr}\,\mathbf{BPP}' = \mathrm{tr}\,\mathbf{B}$, we have that $\mathrm{tr}(\mathbf{AB}) \le \lambda_L(\mathbf{A})\,\mathrm{tr}\,\mathbf{B}$.

Using the lemma and properties of the trace operator, note that

$$\mathrm{tr}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\Phi}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} = \mathrm{tr}\,\boldsymbol{\Phi}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-2}\mathbf{x}' \le \lambda_L(\boldsymbol{\Phi})\,\mathrm{tr}\,\mathbf{x}(\mathbf{x}'\mathbf{x})^{-2}\mathbf{x}' = \lambda_L(\boldsymbol{\Phi})\,\mathrm{tr}(\mathbf{x}'\mathbf{x})^{-1}.$$

Thus, if $\lambda_L(\boldsymbol{\Phi}) < \tau < \infty$, i.e., if the largest characteristic root of $\boldsymbol{\Phi}$ is bounded, then since $\mathrm{tr}(\mathbf{x}'\mathbf{x})^{-1} \to 0$ as $n \to \infty$, $\mathrm{tr}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\Phi}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} \to 0$ as $n \to \infty$. It follows that all of the diagonal entries in $(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\Phi}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}$ must converge to zero, since the $(k \times k)$ matrix is (at least) positive semidefinite and thus has nonnegative diagonal entries, and the sum of these $k$ nonnegative numbers converging to zero requires that each diagonal entry converge to zero. Thus, by convergence in mean square, $\mathrm{plim}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ even if $\mathrm{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \boldsymbol{\Phi} \ne \sigma^2\mathbf{I}$.

As a practical matter, the assumption that $\lambda_L(\boldsymbol{\Phi})$ is bounded is not very restrictive. The following matrix theory lemma will be useful for establishing sufficient conditions for the boundedness of $\lambda_L(\boldsymbol{\Phi})$.

**Lemma 8.5**

The absolute value of any characteristic root of a matrix, $\mathbf{A}$, is less than or equal to the sum of the absolute values of the elements in the row of $\mathbf{A}$ for which the sum is largest.

[14] We are making the tacit assumption that x contains no lagged values of the dependent variable, which is as it must be if it is presumed that x can be held fixed. In the event that x contains lagged values of the dependent variable and the error terms are autocorrelated, then in general $\mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \ne [\mathbf{0}]$, and $\hat{\boldsymbol{\beta}}$ is biased. Issues related to this case are discussed in a subsequent subsection.

**Proof**: See S. J. Hammarling (1970), *Latent Roots and Latent Vectors*, Univ. of Toronto Press, p. 9.

Using Lemma 8.5, we now examine three different sufficient conditions on the covariance matrix $\mathbf{\Phi}$ that ensure the boundedness of $\lambda_L(\mathbf{\Phi})$ and thus the consistency of $\hat{\boldsymbol{\beta}}$. Assume that $\text{var}(\varepsilon_i) = \sigma_{ii} < \xi < \infty \ \forall\, i$, so that the variances of the disturbance terms exhibit some (perhaps very large) upper bound. As a first sufficient condition, if the disturbance vector is *heteroskedastic* but there is *zero autocorrelation*, then it is immediate that $\lambda_L(\mathbf{\Phi}) < \xi$, since $\mathbf{\Phi}$ is a diagonal matrix of variances, and the characteristic roots of a diagonal matrix are directly the diagonal elements. Thus, $\lambda_L(\mathbf{\Phi})$ is bounded and plim $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

If the disturbance vector is either *heteroskedastic* or *homoskedastic* and exhibits *nonzero autocorrelation*, then $\lambda_L(\mathbf{\Phi})$ will be bounded if the covariance terms in $\mathbf{\Phi}$ decline sufficiently fast as $|i-j|$ increases. The most straightforward sufficient condition to examine in this case is $\sigma_{ij} = 0$ for $|i-j| > m$, where $m$ is some (perhaps very large) positive integer $m$, which is true if (but *not* only if) the $\varepsilon_i$'s are $m$-dependent (recall Def. 5.18). Then let row $i$ of $\mathbf{\Phi}$ be the row for which the sum of the absolute values of the elements in the row is largest in comparison with the other rows of $\mathbf{\Phi}$. Because $\sigma_{ij} = 0$ when $|i-j| > m$, the *maximum* number of nonzero entries in row $i$ is $2m+1$. Also, by the bound on covariances, $|\sigma_{ij}| \leq |\sigma_{ii}|^{1/2}|\sigma_{jj}|^{1/2} \ \forall\, i, j$. It follows that the sum of the absolute values of the entries in the $i$th row of the covariance matrix $\mathbf{\Phi}$ is upper bounded by $(2m+1)\boldsymbol{\xi}$, and thus by Lemma 8.5, $\lambda_L(\mathbf{\Phi}) < (2m+1)\boldsymbol{\xi}$. Then since $\lambda_L(\mathbf{\Phi})$ is bounded, plim $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

Now suppose instead that $\sigma_{ij} \neq 0$ and there is no value of $|i-j|$ beyond which the covariance is assumed to be zero. Then again using Lemma 8.5, a third sufficient condition for $\lambda_L(\mathbf{\Phi}) < \tau < \infty$ is that $\sum_{j=1}^{n} |\sigma_{ij}| < \tau < \infty \ \forall\, i$ and $\forall\, n$. Since we are assuming $\sigma_{ii} < \boldsymbol{\xi} \ \forall\, i$, the sufficient condition can alternatively be stated in terms of the boundedness of $\sum_{j \neq i} |\sigma_{ij}| \ \forall\, i$. Thus, if covariances decline sufficiently fast so that $\sum_{j \neq i} |\sigma_{ij}| < \eta < \infty \ \forall\, i$, then plim $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

We should note that the consistency of $\hat{\boldsymbol{\beta}}$ does *not* require that $\lambda_L(\mathbf{\Phi})$ be bounded. It is sufficient that $\lambda_L(\mathbf{\Phi})$ increase at a rate slower than the rate at which $\text{tr}(\mathbf{x}'\mathbf{x})^{-1} \rightarrow 0$, for then it would still be true that $\lambda_L(\mathbf{\Phi})\,\text{tr}(\mathbf{x}'\mathbf{x})^{-1} \rightarrow 0$. For additional results on the consistency of $\hat{\boldsymbol{\beta}}$, the reader is referred to the book by H. White, Asymptotic Theory.

Having seen that $\hat{\boldsymbol{\beta}}$ is unbiased and consistent for $\boldsymbol{\beta}$ under general conditions on $\mathrm{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \mathbf{\Phi}$, one may wonder whether $\hat{\boldsymbol{\beta}}$ retains the BLUE property. In the general (and usual) case where $\mathbf{\Phi}$ is unknown, *no* linear estimator exists that has the BLUE property. To see this, note that if $\mathrm{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \mathbf{\Phi}$, then

$$\mathbf{\Phi}^{-1/2}\mathbf{Y} = \mathbf{\Phi}^{-1/2}\mathbf{x}\boldsymbol{\beta} + \mathbf{\Phi}^{-1/2}\boldsymbol{\varepsilon} \quad \text{or} \quad \mathbf{Y}_* = \mathbf{x}_*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_*$$

satisfies the classical assumptions of the GLM (assuming, of course, that all of the classical assumptions *except* $\mathbf{Cov}(\mathbf{Y}) = \mathbf{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ apply to $\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$). In particular, note that $\mathrm{E}\boldsymbol{\varepsilon}_*\boldsymbol{\varepsilon}_*' = \mathbf{\Phi}^{-1/2}\mathrm{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{\Phi}^{-1/2} = \mathbf{\Phi}^{-1/2}\mathbf{\Phi}\mathbf{\Phi}^{-1/2} = \mathbf{I}$ and $\mathrm{E}\boldsymbol{\varepsilon}_* =$

$\mathrm{E}\Phi^{-1/2}\varepsilon = \Phi^{-1/2}\mathrm{E}\varepsilon = [\mathbf{0}]$. Then the Gauss–Markov theorem applied to the transformed linear model implies that

$$\hat{\beta} = (\mathbf{x}_*'\mathbf{x}_*)^{-1}\mathbf{x}_*'\mathbf{Y}_* = (\mathbf{x}'\Phi^{-1}\mathbf{x})^{-1}\mathbf{x}'\Phi^{-1}\mathbf{Y}$$

would be the BLUE for $\beta$. However, the definition of the BLUE estimator depends on the unknown value of $\Phi$, and since then there does not exist a fixed choice of $\mathbf{A}$ and $\mathbf{b}$ such that $\mathbf{AY} + \mathbf{d}$ is BLUE for all potential values of $\Phi$, no BLUE estimator of $\beta$ exists when $\Phi$ is unknown.

In the *very* special case where $\Phi$ is known up to a scalar multiple, i.e., $\Phi = \sigma^2\Omega$ with $\Omega$ *known*, then the linear estimator $\hat{\beta}^* = (\mathbf{x}'\Phi^{-1}\mathbf{x})^{-1}\mathbf{x}'\Phi^{-1}\mathbf{Y} = (\mathbf{x}'\Omega^{-1}\mathbf{x})^{-1}\mathbf{x}'\Omega^{-1}\mathbf{Y}$ is the BLUE estimator. This special estimator is referred to in the literature as the **generalized least-squares estimator of $\beta$.**

$\hat{S}^2$ *Properties* The estimator $\hat{S}^2$ under the condition $\mathrm{E}\varepsilon\varepsilon' = \Phi \neq \sigma^2\mathbf{I}$ is useful only if there exists a counterpart to the parameter $\sigma^2$ which can be estimated. Two such situations arise when $\Phi = \sigma^2\Omega$, and either $\Omega$ is a known positive definite matrix, or else $\varepsilon$ is homoskedastic with common variance $\sigma^2$. In either case,

$$\mathrm{E}\hat{S}^2 = (n-k)^{-1}\mathrm{E}\varepsilon'(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\varepsilon = (n-k)^{-1}\,\mathrm{tr}(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\mathrm{E}\varepsilon\varepsilon'$$
$$= \sigma^2\,\mathrm{tr}\,\frac{[\Omega(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')]}{(n-k)}.$$

Thus $\hat{S}^2$ is generally *biased* as an estimator of $\sigma^2$, since the trace of the matrix divided by $(n-k)$ in the preceding expression will generally not be equal to 1. Of course, in the *very* special case where $\Omega$ were known, $\hat{S}^2$ could be scaled by the known value of the trace to define an unbiased estimator of $\sigma^2$.

In certain cases, the bias in $\hat{S}^2$ converges to 0 as $n \to \infty$. Note that Lemma 8.4 implies that

$$\sigma^2\,\mathrm{tr}\,\frac{(\Omega\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')}{(n-k)} \leq \sigma^2\lambda_L(\Omega)\left(\frac{k}{(n-k)}\right),$$

and if $\lambda_L(\Phi)$ is bounded or at least $o(n^1)$ as we have argued above, then $\sigma^2\lambda_L(\Omega) = \lambda_L(\Phi)$[15] is bounded and $\sigma^2\lambda_L(\Omega)(k/(n-k)) \to 0$ as $n \to \infty$. Then since $\mathrm{tr}(\Omega\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}') \geq 0$,[16] it must be the case that $\sigma^2\,\mathrm{tr}\,\Omega\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')/(n-k) \to 0$ as $n \to \infty$. It follows that $\mathrm{E}\hat{S}^2 \to \sigma^2$ as $n \to \infty$ iff $\mathrm{tr}\,\Omega/(n-k) \to 1$ as $n \to \infty$, which *does* occur if $\varepsilon$ is homoskedastic ($\varepsilon$ may still exhibit nonzero autocorrelation) with $\sigma^2$ representing the common variance of the $\varepsilon_i$'s, for then $\mathrm{tr}(\Omega) = n$, since the diagonal entries of $\Omega$ would all be equal to the number 1, and then $n/(n-k) \to 1$ as $n \to \infty$. In the heteroskedastic case, there is no reason to expect that the preceding condition will hold.

[15] The characteristic roots of $\tau\mathbf{A}$ are equal to the characteristic roots of $\mathbf{A}$ times the scalar $\tau$.

[16] Note that $\mathrm{tr}\;\Omega\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' = \mathrm{tr}\,\Omega^{1/2}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\Omega^{1/2}$, and since $(\mathbf{x}'\mathbf{x})^{-1}$ is positive definite, $\Omega^{1/2}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\Omega^{1/2}$ is at least positive semidefinite, and its trace must be nonnegative.

Regarding consistency of $\hat{S}^2$ for $\sigma^2$, Markov's inequality can be used as in the proof of Theorem 8.4 to show that plim $(n-k)^{-1}\varepsilon'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon = 0$, assuming $\lambda_L(\mathbf{\Phi})$ is bounded. If $\varepsilon$ is *homoskedastic* with variance $\sigma^2$, and if the conditions of Theorem 8.5 other than $\mathrm{E}\varepsilon\varepsilon' = \sigma^2\mathbf{I}$ can be assumed to hold, then the weak law of large numbers given by Theorem 5.22 implies that plim $(n-k)^{-1}\varepsilon'\varepsilon = \mathrm{plim}\, n^{-1}\varepsilon'\varepsilon = \sigma^2$, and thus $\hat{S}^2$ would be consistent for $\sigma^2$ (consistency under alternative conditions on $\varepsilon$ can be demonstrated as well—recall Table 8.1). If $\varepsilon$ is heteroskedastic with $\mathbf{Cov}(\varepsilon) = \sigma^2\mathbf{\Omega}$ while maintaining the remaining preceding assumptions, then by Theorem 5.22 plim $(n-k)^{-1}[\varepsilon'\varepsilon - \sum_{i=1}^{n}\sigma^2\omega_{ii}] = \mathrm{plim}\, n^{-1}[\varepsilon'\varepsilon - \sigma^2\sum_{i=1}^{n}\omega_{ii}] = 0$, and except for the very special case $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\omega_{ii} = 1$, $\hat{S}^2$ will not be consistent for $\sigma^2$.

*Assumption Violation: Rank* $\mathbf{x} < k$, *or* $|\mathbf{x}'\mathbf{x}| \approx 0$ In this section we assume that either $\mathbf{x}$ is less than full column rank so that rank $\mathbf{x} < \mathbf{k}$, or else there is "nearly" a linear dependency among the columns of $\mathbf{x}$, so that $|\mathbf{x}'\mathbf{x}| \approx 0$. Note that this latter assumption is *not* a violation of the classical GLM assumption, but rather "nearly" a violation. We retain the other classical GLM assumptions.

$\hat{\beta}$ *Properties* If the classical GLM assumption concerning rank $\mathbf{x} = k$ is violated, then the least-squares estimator $\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ does not exist, since $(\mathbf{x}'\mathbf{x})^{-1}$ does not exist. In this case, there is an *infinite* number of solutions to the problem of minimizing $(\mathbf{y}-\mathbf{xb})'(\mathbf{y}-\mathbf{xb})$ through choice of the vector $\mathbf{b}$. To see this, recall that the first-order conditions for the minimization problem are given by $(\mathbf{x}'\mathbf{x})\mathbf{b} = \mathbf{x}'\mathbf{y}$, which is a system of $k$ equations in the $k$ unknowns represented by the vector $\mathbf{b}$. If $\mathbf{x}$ is less than full column rank, so that $(\mathbf{x}'\mathbf{x})$ is less than full rank, then this system of linear equations effectively contains one or more redundant equations, so that there are essentially more unknowns than equations. Then there is an infinite number of solution values for $\mathbf{b}$. In this case, the parameter vector $\beta$ cannot be estimated uniquely. This problem is referred to in the literature as **perfect multicollinearity**.

More prevalent in applications is the case where rank $\mathbf{x} = k$ but $\mathbf{x}'\mathbf{x}$ is *nearly singular*, i.e., its determinant is near zero. In such cases, $(\mathbf{x}'\mathbf{x})^{-1}$ tends to have large diagonal entries, implying that the variances of elements of $\hat{\beta}$, given by the diagonal elements of $\sigma^2(\mathbf{x}'\mathbf{x})^{-1}$, are very large. This follows from our previous application of Lemma 8.3 on partitioned inversion to the matrix $(\mathbf{x}'\mathbf{x})^{-1}$. In particular, the lemma implies that the variance of $\hat{\beta}_i$ (i.e., the $(i,i)$th entry in $\sigma^2(\mathbf{x}'\mathbf{x})^{-1}$) is given by

$$\sigma^2(\mathbf{x}'_{.i}\mathbf{x}_{.i} - \mathbf{x}'_{.i}\mathbf{x}_*(\mathbf{x}'_*\mathbf{x}_*)^{-1}\mathbf{x}'_*\mathbf{x}_{.i})^{-1} = \sigma^2(\mathbf{e}'_i\mathbf{e}_i)^{-1},$$

where $\mathbf{e}_i$ represents a vector of deviations of $\mathbf{x}_{.i}$ from a least-squares prediction of $\mathbf{x}_{.i}$ based on a linear combination of the columns of $\mathbf{x}$ other than $\mathbf{x}_{.i}$ (these remaining $(k-1)$ columns are represented by $\mathbf{x}_*$). The more closely $\mathbf{x}_{.i}$ can be approximated by $\mathbf{x}_*\hat{\mathbf{b}}$ (where $\hat{\mathbf{b}} = (\mathbf{x}'_*\mathbf{x}_*)^{-1}\mathbf{x}'_*\mathbf{x}_{.i}$), and thus the more closely that $\mathbf{x}_{.i}$ is linearly related to the remaining column vectors in the $\mathbf{x}$ matrix, the smaller $\mathbf{e}_i$ and $\mathbf{e}'_i\mathbf{e}_i$ are and thus the larger $\mathrm{var}(\hat{\beta}_i) = \sigma^2(\mathbf{e}'_i\mathbf{e}_i)^{-1}$ is. Note that as long as rank $\mathbf{x} = k$, and the other appropriate assumptions hold, $\hat{\beta}$ remains BLUE and consistent as an estimator of $\beta$. However, the large variances associated with the $\hat{\beta}_i$'s imply that in small samples, outcomes of $\hat{\beta}$ can be quite distant from $\beta$ with high probability.

*$\hat{S}^2$ Properties* In the case of perfect multicollinearity, $\hat{S}^2$ can nonetheless be used to generate estimates of $\sigma^2$, and with minor modifications to the derivations used previously to establish properties of $\hat{S}^2$, it can be shown that $\hat{S}^2$ is unbiased and consistent for $\sigma^2$. Of course, the definition of $\hat{S}^2$ cannot be expressed in terms of $(\mathbf{Y}-\mathbf{x}\hat{\beta})'(\mathbf{Y}-\mathbf{x}\hat{\beta})/(n-k)$, since $\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ does not exist in this case. However, $\hat{s}^2 = \mathbf{e}'\mathbf{e}/(n-k)$ can nevertheless be calculated, where $\mathbf{e}'\mathbf{e}$ is the minimum sum of squared errors calculated from any choice of $\hat{\beta}_*$ which solves the first-order conditions $(\mathbf{x}'\mathbf{x})\hat{\beta}_* = \mathbf{x}'\mathbf{y}$. A rigorous demonstration of the properties of $\hat{S}^2$ under perfect multicollinearity relies on the notion of generalized inverses of matrices (e.g., see F. Graybill (1976), *Theory and Application of the Linear Model*. North Sciuate, MA: Duxbury Press, pp. 23–39), and given that perfect multicollinearity is by far the exception rather than the rule, we will not pursue the details here.

In the case where rank $\mathbf{x} = k$ but $(\mathbf{x}'\mathbf{x})$ is *nearly singular*, the proofs of unbiasedness and consistency of $\hat{S}^2$ nonetheless apply *exactly* as stated previously, and thus these properties are attained by $\hat{S}^2$ under the assumptions introduced heretofore. It is also useful to note that unlike the variances of the $\hat{\beta}_i$'s, which can increase without bound as the multicollinearity becomes increasingly severe, the variance of $\hat{S}^2$ exhibits a finite bound, which is *unaffected* by the degree of multicollinearity. To derive such a bound, assume the classical GLM assumptions hold, assume the $\varepsilon_i$'s are iid, and let $\mathrm{E}\varepsilon_i^4 = \mu_4' < \infty$ exist. Recall that $\mathrm{var}(\hat{S}^2) = \mathrm{E}\hat{S}^4 - (\mathrm{E}\hat{S}^2)^2 = \mathrm{E}\hat{S}^4 - \sigma^4$, where we have used the fact that $\mathrm{E}\hat{S}^2 = \sigma^2$. Now note that

$$\hat{S}^2 = (n-k)^{-1}(\varepsilon'\varepsilon - \varepsilon'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon) \le (n-k)^{-1}\varepsilon'\varepsilon,$$

since $\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$ is positive semidefinite, which implies

$$\begin{aligned}\mathrm{E}\hat{S}^4 &\le (n-k)^{-2}\mathrm{E}(\varepsilon'\varepsilon)^2 \le (n-k)^{-2}\mathrm{E}\left[\sum_{i=1}^{n}\varepsilon_i^4 + 2\sum\sum_{i<j}\varepsilon_i^2\varepsilon_j^2\right]\\ &\le (n-k)^{-2}\left[n\mu_4' + n(n-1)\sigma^4\right] = \tau(n),\end{aligned}$$

where $\tau(n) \ge \sigma^4$, $\tau(n)$ is monotonically decreasing in $n \,\forall\, n > k$, and $\lim_{n\to\infty}\tau(n) = \sigma^4$.

Then the variance of $\hat{S}^2$ is bounded as

$$\mathrm{var}(\hat{S}^2) \le \tau(n) - \sigma^4$$

regardless of the severity of the multicollinearity, as long as rank $\mathbf{x} = k$. Other bounds can be derived in cases where the $\varepsilon_i$'s are not iid.

*Assumption Violation: Stochastic* $\mathbf{X}$ *with* $\mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \ne [\mathbf{0}]$ *and* $\mathrm{plim}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \ne [\mathbf{0}]$ In this section we focus explicitly on the case where the explanatory variable matrix $\mathbf{X}$ is a nondegenerate random matrix. Note the condition

$$\mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \neq [\mathbf{0}]$$

necessarily implies (by the double expectation theorem) that $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{x}) \neq [\mathbf{0}]$ with positive probability, and so the conditional mean of $\boldsymbol{\varepsilon}$ is *not* independent of $\mathbf{X}$. This in turn implies that $\mathrm{E}(\mathbf{Y} \mid \mathbf{x}) \neq \mathbf{x}\boldsymbol{\beta}$ with positive probability, and thus assumption 1 of the classical assumptions of the GLM, applied explicitly to the stochastic $\mathbf{X}$ case, is violated. Furthermore, the dependence between $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ persists in the limit, as indicated by $\mathrm{plim}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \neq [\mathbf{0}]$. The interdependence of $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ is typically caused by either *measurement error* in the $\mathbf{x}$ outcomes or by *simultaneous determination* of $\mathbf{Y}$ and $\mathbf{X}$ outcomes and is discussed in the econometric literature under the rubric of *errors in variables* and *simultaneous equations*, respectively.

*$\hat{\boldsymbol{\beta}}$ Properties* Since $\mathrm{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \neq \boldsymbol{\beta}$, because $\mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \neq [\mathbf{0}]$, it follows that the estimator is a *biased* estimator of $\boldsymbol{\beta}$, so that $\hat{\boldsymbol{\beta}}$ is also *not* BLUE. Regarding consistency, $\mathrm{plim}\ \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathrm{plim}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \neq \boldsymbol{\beta}$ because $\mathrm{plim}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \neq [\mathbf{0}]$ and so $\hat{\boldsymbol{\beta}}$ is also *not* consistent.

*$\hat{S}^2$ Properties* Letting $\boldsymbol{\delta}(\mathbf{x}) = E(\boldsymbol{\varepsilon} \mid \mathbf{x})$, the expectation of $(n-k)\hat{S}^2$ can be written via the double expectation theorem as

$$\begin{aligned}
\mathrm{E}(n-k)\hat{S}^2 &= \mathrm{E}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - \mathrm{E}\left[\mathrm{E}(\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \mid \mathbf{X})\right] \\
&= n\sigma^2 - \mathrm{E}\left[\mathrm{tr}\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X})\right] \\
&= n\sigma^2 - \mathrm{E}\left[\mathrm{tr}\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Cov}(\boldsymbol{\varepsilon} \mid \mathbf{X}) + \boldsymbol{\delta}(\mathbf{X})\boldsymbol{\delta}(\mathbf{X})')\right] \\
&= (n-k)\sigma^2 - \mathrm{E}\boldsymbol{\delta}(\mathbf{X})'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\delta}(\mathbf{X})
\end{aligned}$$

assuming $\mathbf{Cov}(\boldsymbol{\varepsilon} \mid \mathbf{x}) = \sigma^2\mathbf{I}$, $\forall\ \mathbf{x}$.[17] Thus $\mathrm{E}\hat{S}^2 < \sigma^2$ in general, so that $\hat{S}^2$ is *biased.*

Regarding the consistency of $\hat{S}^2$, first note that $\hat{S}^2 = (n-k)^{-1}\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} = n^{-1}\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} + o_p(1) = n^{-1}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - n^{-1}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} + o_p(1)$, and since $(n^{-1}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - \sigma^2) \xrightarrow{p} 0$,

$$(\hat{S}^2 - \sigma^2) = -\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(n^{-1}\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} + o_p(1).$$

If $n^{-1}\mathbf{X}'\mathbf{X} \xrightarrow{p} \mathbf{Q}$, a symmetric, positive definite matrix, and if $\mathrm{plim}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} = \boldsymbol{\xi} \neq [\mathbf{0}]$, then $(\hat{S}^2 - \sigma^2) \xrightarrow{p} -\boldsymbol{\xi}'\mathbf{Q}\boldsymbol{\xi} < 0$, and $\hat{S}^2$ is not consistent. Even if neither $n^{-1}\mathbf{X}'\mathbf{X}$ nor $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$ converges at all, so long as $\det(n^{-1}\mathbf{X}'\mathbf{X}) > \eta > 0\ \forall\ n$, as in Table 8.1, and $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \overset{p}{\nrightarrow} [\mathbf{0}]$, it follows that $(\hat{S}^2 - \sigma^2) \overset{p}{\nrightarrow} 0$ and $\hat{S}^2$ is not consistent.

### GLM Assumption Violations: Property Summary and Epilogue

The basic properties of the OLS-based estimators of $\boldsymbol{\beta}$ and $\sigma^2$ under the various violations of the classical GLM assumptions are summarized in Table 8.2. The

[17] Recall that $\mathrm{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \mathbf{Cov}(\boldsymbol{\varepsilon}) + \mathrm{E}\boldsymbol{\varepsilon}\mathrm{E}\boldsymbol{\varepsilon}'$.

row of the table identified as "terminology" indicates the label under which the problem is generally discussed in the econometric literature.

A major focus of research based on the GLM concerns the detection and remedies for violations of the assumptions needed to achieve the desirable properties of the estimators of $\beta$ and $\sigma^2$. Indeed, fields such as **econometrics** and **sociometrics** were necessitated by the fact that, unlike disciplines in which the experiments under investigation can be designed and controlled to achieve the conditions needed for the least-squares estimator to have optimal properties, experiments in economics, business, sociology, and other social sciences are often not under the control of the researcher. Variations on the least-squares estimator, such as *restricted least squares*, the *feasible generalized least-squares estimator, instrumental variables estimators, two- and three-stage least-squares estimators, general method of moments estimators*, and *limited and full-information maximum likelihood estimators*, represent procedures for remedying violations of the GLM classical assumptions that the reader will encounter in subsequent studies of statistics and econometrics, e.g., see the books by G. Judge, et al., (1985), *The Theory and Practice of Econometrics*, 2nd ed., New York: John Wiley, and T. Amemiya (1985), *Advanced Econometrics* Cambridge, Harvard University Press. We will examine some hypothesis tests for detecting GLM assumption violations in Chapter 10.

### Least Squares Under Normality

Before examining the implications of making a specific parametric family assumption for the PDF of $\mathbf{Y}$, we emphasize that the type of random sample envisioned for $(Y_1, \ldots, Y_n)$ in the GLM context is generally of the composite experiment type, i.e., note that $\mathrm{E}Y_i$ does not necessarily equal $\mathrm{E}Y_j$ for $i \neq j$, nor is it necessarily assumed that $(Y_1, \ldots, Y_n)$ or $(\varepsilon_1, \ldots, \varepsilon_n)$ are iid. We now examine the assumption of normality for $(Y_1, \ldots, Y_n)$ and, hence, for $(\varepsilon_1, \ldots, \varepsilon_n)$.

In the GLM, under the classical GLM assumptions, let $Y \sim N(\mathbf{x}\beta, \sigma^2\mathbf{I})$. Under the normality assumption, the classical GLM assumptions *necessarily* imply the $\varepsilon_i$'s are iid normal, with mean zero and variance $\sigma^2$. Furthermore, *all* of the assumptions in Table 8.1 corresponding to the disturbance vector hold, *except* for $P(|\varepsilon_i| < m) = 1 \;\forall\, i$, which does not hold. Of course, our entire preceding discussion concerning the properties of $\hat{\beta}$ and $\hat{S}^2$ then applies equally well to the case where $\mathbf{Y}$ is multivariate normally distributed. The question this section addresses is what additional properties can be attributed to the estimators when $\mathbf{Y}$ is multivariate normally distributed?

One immediate property is that $\hat{\beta}$ is multivariate normally distributed for every $n \geq k$, and not just asymptotically normally distributed. This follows straightforwardly from the fact that $\hat{\beta}$ is a linear function of the entries in $\mathbf{Y}$, so that (by Theorem 4.9) $\hat{\beta}$ is normally distributed with mean $\mathrm{E}\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'(\mathbf{x}\beta) = \beta$, and covariance matrix $\mathbf{Cov}(\hat{\beta}) = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'[\sigma^2\mathbf{I}]\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} = \sigma^2(\mathbf{x}'\mathbf{x})^{-1}$, i.e., $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{x}'\mathbf{x})^{-1})$.

**Table 8.2** General Least-Squares Estimator Properties Under GLM Assumption Violations
Appropriate conditions in Table 9.1 hold, with exceptions such that:

| Violation | $E\varepsilon = \phi \neq [\mathbf{0}]$ | $E\varepsilon\varepsilon' = \Phi \neq \sigma^2\mathbf{I}$ | $\det(\mathbf{x}'\mathbf{x}) \approx 0$ | Random $\mathbf{X}$, $E(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \neq [\mathbf{0}]$, $\operatorname{plim}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon) \neq [\mathbf{0}]$ |
|---|---|---|---|---|
| terminology | specification error | heteroskedasticity and/or autocorrelation | multicollinearity | errors in variables, simultaneity |
| $\hat{\beta}$ | | | | |
| Unbiased | **generally no**<br>yes if $\mathbf{x}'\phi = [\mathbf{0}]$ | **yes** | **yes** | **no** |
| BLUE | **generally no**<br>yes if $\mathbf{x}'\phi = [\mathbf{0}]$ | **no** | **yes** | **no** |
| Consistent | **generally no**<br>yes if $(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\phi \to [\mathbf{0}]$ | **generally yes**<br>if $\lambda_L(\Phi)\,\mathrm{tr}(\mathbf{x}'\mathbf{x})^{-1} \to 0$ | **yes** | **no** |
| $S^2$ | | | | |
| Unbiased | **generally no**<br>yes if $\phi'(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\phi \to 0$ | **generally no** | **yes** | **no** |
| Consistent | **generally no**<br>yes if $\phi'(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\phi \to 0$ | **generally no**<br>yes if $\lambda_L(\Phi)$ is $o(n^1)$ and $\varepsilon_i$'s are homoskedastic | **yes** | **no** |

Another property is that $(n-k)\hat{S}^2/\sigma^2 \sim \chi^2_{n-k}$. To see this, first note that

$$\frac{(n-k)\hat{S}^2}{\sigma^2} = \frac{\boldsymbol{\varepsilon}'[\mathbf{I}-\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}']\boldsymbol{\varepsilon}}{\sigma^2} = \sigma^{-1}\boldsymbol{\varepsilon}'\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'\boldsymbol{\varepsilon}\sigma^{-1},$$

where $\mathbf{I}-\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$, and $\boldsymbol{\Lambda}$ and $\mathbf{P}$ are, respectively, the diagonal matrix of characteristic roots and the matrix of characteristic vectors (stored columnwise) associated with the symmetric and idempotent matrix $(\mathbf{I}-\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')$ (recall the proof of Theorem 6.12.b). Examine the probability distribution of $\mathbf{Z} = \mathbf{P}'\boldsymbol{\varepsilon}\sigma^{-1}$. Since $\boldsymbol{\varepsilon} \sim N([0],\sigma^2\mathbf{I})$, then $\mathbf{Z} \sim N(\sigma^{-1}\mathbf{P}'[0],\sigma^{-1}\mathbf{P}'\sigma^2\mathbf{I}\mathbf{P}\sigma^{-1}) = N([0],\mathbf{I})$, because $\mathbf{P}'\mathbf{P} = \mathbf{I}$ by the orthogonality of $\mathbf{P}$. Then

$$\frac{(n-k)\hat{S}^2}{\sigma^2} = \mathbf{Z}'\boldsymbol{\Lambda}\mathbf{Z} = \sum_{i=1}^{n-k} Z_i^2 \sim \chi^2_{n-k},$$

i.e., we have the sum of the squares of $(n-k)$ iid standard normal random variables, which has a chi-square distribution with $n-k$ degrees of freedom.

Since by the preceding result $\hat{S}^2$ can be characterized as a $\chi^2$ random variable that has been multiplied by the constant $\sigma^2/(n-k)$, it follows that $\hat{S}^2 \sim \text{Gamma}((n-k)/2,\ 2\sigma^2/(n-k))$, as can be shown by deriving the MGF of $\hat{S}^2$. Using properties of the Gamma density, this in turn implies that $\mathrm{E}\hat{S}^2 = ((n-k)/2)(2\sigma^2/(n-k)) = \sigma^2$ and $\mathrm{var}(\hat{S}^2) = ((n-k)/2)(2\sigma^2/(n-k))^2 = 2\sigma^4/(n-k)$.

Still another property is that $\hat{\boldsymbol{\beta}}$ and $\hat{S}^2$ are *independent* random variables. This follows from an application of Theorem 6.11 along the lines of the proof of Theorem 6.12a and is left as an exercise for the reader.

*MVUE Property of $\hat{\boldsymbol{\beta}}$ and $\hat{S}^2$* Perhaps the most important additional property that results when $\mathbf{Y} \sim N(\mathbf{x}\boldsymbol{\beta},\sigma^2\mathbf{I})$ is that $\left[\begin{smallmatrix}\hat{\boldsymbol{\beta}}\\ \hat{S}^2\end{smallmatrix}\right]$ is then the MVUE for $\left[\begin{smallmatrix}\boldsymbol{\beta}\\ \sigma^2\end{smallmatrix}\right]$.

**Theorem 8.10** (MVUE Property of $(\hat{\boldsymbol{\beta}},\hat{S}^2)$ Under Normality) *Assume the classical assumptions of the GLM, and assume that $\mathbf{Y} \sim N(\mathbf{x}\boldsymbol{\beta},\sigma^2\mathbf{I})$. Then $(\hat{\boldsymbol{\beta}},\hat{S}^2)$ is the MVUE for $(\boldsymbol{\beta},\sigma^2)$.*

**Proof** Note that the multivariate normal density belongs to the exponential class of densities

$$\exp\left(\sum_{i=1}^{k+1} c_i(\boldsymbol{\beta},\sigma^2)g_i(\mathbf{y}) + d(\boldsymbol{\beta},\sigma^2) + z(\mathbf{y})\right) I_A(\mathbf{y}) = \exp(\mathbf{c}(\boldsymbol{\beta},\sigma^2)'\mathbf{g}(\mathbf{y}) + d(\boldsymbol{\beta},\sigma^2) + z(\mathbf{y}))I_A(\mathbf{y}),$$

where

$$\mathbf{c}(\boldsymbol{\beta},\sigma^2) = \begin{bmatrix} c_1(\boldsymbol{\beta},\sigma^2) \\ \vdots \\ c_k(\boldsymbol{\beta},\sigma^2) \\ c_{k+1}(\boldsymbol{\beta},\sigma^2) \end{bmatrix} = \begin{bmatrix} \dfrac{\beta_1}{\sigma^2} \\ \vdots \\ \dfrac{\beta_k}{\sigma^2} \\ \dfrac{-1}{2\sigma^2} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\sigma^2}\boldsymbol{\beta} \\ \dfrac{-1}{2\sigma^2} \end{bmatrix}, \qquad \mathbf{g}(\mathbf{y}) = \begin{bmatrix} g_1(\mathbf{y}) \\ \vdots \\ g_{k+1}(\mathbf{y}) \end{bmatrix} = \begin{bmatrix} \mathbf{x}'\mathbf{y} \\ \mathbf{y}'\mathbf{y} \end{bmatrix},$$

$d(\boldsymbol{\beta}, \sigma^2) = -((\boldsymbol{\beta}'\mathbf{x}'\mathbf{x}\boldsymbol{\beta})/2\sigma^2) - \ln(2\pi\sigma^2)^{n/2}, z(\mathbf{y}) = 0$, and $A = \times_{i=1}^{n}(-\infty, \infty) = R^n$. Then by Theorem 7.3, $\left[\begin{smallmatrix}\mathbf{x}'\mathbf{Y}\\ \mathbf{Y}'\mathbf{Y}\end{smallmatrix}\right]$ is a vector of *minimal* sufficient statistics for estimating $\boldsymbol{\beta}$ and $\sigma^2$. These sufficient statistics are also *complete* sufficient statistics, since the range of the vector function represented by $\mathbf{c}(\boldsymbol{\beta}, \sigma^2)$ contains an open $(k+1)$-dimensional rectangle (recall Theorem 7.6). In particular, note that the range is given by $R(\mathbf{c}) = \{(c_1, \ldots, c_{k+1}) : -\infty < c_i < \infty, i = 1, \ldots, k, c_{k+1} < 0\}$ since $\beta_i \in (-\infty, \infty)$ for $i = 1, \ldots, k$, and $\sigma^2 > 0$, so that any open rectangle of the form $A = \{(c_1, \ldots, c_{k+1}) \colon a_i < c_i < b_i, i = 1, \ldots, k+1\}$, for $a_i < b_i$, and $b_{k+1} < 0$ is a $(k+1)$-dimensional open rectangle subset of $R(\mathbf{c})$.[18] Then since $\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ and $\hat{S}^2 = (n-k)^{-1}(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}}) = (\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y})/(n-k)$ are functions of the complete sufficient statistics $\left[\begin{smallmatrix}\mathbf{x}'\mathbf{Y}\\ \mathbf{Y}'\mathbf{Y}\end{smallmatrix}\right]$, and since we have shown previously that $\hat{\boldsymbol{\beta}}$ and $\hat{S}^2$ are unbiased for $\boldsymbol{\beta}$ and $\sigma^2$ under the classical assumptions of the GLM, it follows from the Lehmann–Scheffé completeness theorem (Theorem 7.19) that $\left[\begin{smallmatrix}\hat{\boldsymbol{\beta}}\\ \hat{S}^2\end{smallmatrix}\right]$ is the MVUE for $\left[\begin{smallmatrix}\boldsymbol{\beta}\\ \sigma^2\end{smallmatrix}\right]$. ■

The reader should note that had he or she attempted to establish that $\left[\begin{smallmatrix}\hat{\boldsymbol{\beta}}\\ \hat{S}^2\end{smallmatrix}\right]$ was MVUE by utilizing the Cramer–Rao lower bound approach, the CRLB would have been found to be

$$\mathrm{E}\left[\frac{\partial \ln f(\mathbf{Y};\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} \frac{\partial \ln f(\mathbf{Y};\boldsymbol{\Theta})'}{\partial \boldsymbol{\Theta}}\right]^{-1} = \left[\begin{array}{c|c} \sigma^2(\mathbf{x}'\mathbf{x})^{-1} & [0] \\ \hline [0] & \dfrac{2\sigma^4}{n} \end{array}\right],$$

where $\boldsymbol{\Theta} = \binom{\boldsymbol{\beta}}{\sigma^2}$. Then since $\mathbf{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{x}'\mathbf{x})^{-1}$, $\hat{\boldsymbol{\beta}}$ achieves the CRLB for unbiased estimators of $\boldsymbol{\beta}$ proving that $\hat{\boldsymbol{\beta}}$ is the MVUE for $\boldsymbol{\beta}$. However, since $\mathrm{var}(\hat{S}^2) = 2\sigma^4/(n-k) > 2\sigma^4/n$, $\hat{S}^2$ does *not* achieve the CRLB for unbiased estimators of $\sigma^2$, and the CRLB approach would have left the question unanswered regarding whether $\hat{S}^2$ was the MVUE for $\sigma^2$. The approach using complete sufficient statistics demonstrates that no unbiased estimator of $\sigma^2$ can achieve the CRLB in this case, for $\hat{S}^2$ *is* indeed the MVUE.

*On the Assumption of Normality* The reader may wonder what considerations would lead to a specification of the normal family of densities for the probability distribution of $\mathbf{Y}$. Of course, if there is an underlying theoretical or physical rationale for why $\mathbf{Y}$ is multivariate normally distributed, the assumption is obviously supported. However, the underlying rationale for normality is often not clear. In these cases, normality is sometimes rationalized by an appeal to a central limit theorem. In particular, it is often argued that the elements in the disturbance vector are themselves defined as the summation of a large number

[18] Note that $\beta_i \in (-\infty, \infty)$, $\forall\, i$, and $\sigma^2 > 0$ are the *admissible* parameter values for $\mathbf{Y} \sim N(\mathbf{x}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. It may be the case that only a subset of these values for the parameters are deemed to be relevant in a given estimation problem (e.g., a price effect may be restricted to be of one sign, or the realistic magnitude of the effect of an explanatory variable may be bounded). Restricting the parameter space comes under the realm of *prior information* models, which we do not pursue here. So long as the admissible values of $\boldsymbol{\beta}$ and $\sigma^2$ form an open rectangle themselves, the range of $\mathbf{c}(\boldsymbol{\beta}, \sigma^2)$ will contain an open rectangle.

of random variables, i.e., $\varepsilon_i = \sum_{j=1}^{m} V_{ij}$ for large $m$. The $V_{ij}$'s may represent a myriad of neglected explanatory factors that affect $\mathrm{E}Y_i$, $i = 1, \ldots, n$, which, because of the need for tractability, parsimony, or because of a lack of data, are not explicitly represented in the specification of $\mathrm{E}Y_i = \mathbf{x}_{i.}\boldsymbol{\beta}$. Alternatively, the $V_{ij}$'s may be intrinsic to the error-term specification if the GLM is being used to represent a summation of individual microrelations. For example, an aggregate short-run supply function in an aggregate economic analysis might be represented as

$$\sum_{j=1}^{m} Q_{ij} = \sum_{j=1}^{m} (\beta_{1j} + \beta_{2j} p_i + V_{ij}) \Rightarrow Q_i^* = \beta_1^* + \beta_2^* p_i + \varepsilon_i,$$

where $Q_i^* = \sum_{j=1}^{m} Q_{ij}$ is the $i$th observation on aggregate supply, $\beta_1^* = \sum_{j=1}^{m} \beta_{1j}$ is the intercept term of the aggregate supply relationship, $\beta_2^* = \sum_{j=1}^{m} \beta_{2j}$ is the aggregate price effect, $p_i$ is the $i$th observation on supply price, and $\varepsilon_i = \sum_{j=1}^{m} V_{ij}$ is the disturbance term for the $i$th observation, defined as the sum of the individual disturbance terms of the $m$ micro supply functions. Our investigation of CLTs in Chapter 5 suggested that under a variety of conditions, sums of large numbers of random variables are asymptotically normally distributed, i.e., $\boldsymbol{\varepsilon} = \sum_{j=1}^{m} \mathbf{V}_{.j} \overset{a}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

A prudent approach to the assumption of normality is to view arguments such as those presented above as *suggestive* of a normal approximation to the true distribution of $\boldsymbol{\varepsilon}$, but the assumption should be tested for acceptability whenever possible. We will investigate testing the hypothesis of normality (and indeed, hypotheses of other parametric families as well) in Chapter 10.

## 8.3 The Method of Maximum Likelihood

The method of maximum likelihood (ML) can be used to estimate the unknown parameters, or functions of unknown parameters, corresponding to the joint density function of a random sample. The procedure leads to an estimate of $\boldsymbol{\Theta}$ or $\mathbf{q}(\boldsymbol{\Theta})$ by maximizing the *likelihood function* of the parameters, given the observed outcome of the random sample. At the outset we will focus on the problem of estimating $\boldsymbol{\Theta}$ itself. It will be seen later that maximum likelihood estimation of $\mathbf{q}(\boldsymbol{\Theta})$ is easily implemented through an *invariance principle* once the problem of estimating $\Theta$ has been solved.

The **likelihood function** is *identical* in functional form to the joint density function of the random sample. However, there is a subtle interpretational difference between the two functions. In particular, while the joint density function is interpreted as a function of random-variable outcomes $(x_1, \ldots, x_n)$, *given* values of the parameters $(\Theta_1, \ldots, \Theta_k)$, the likelihood function is interpreted as a function of the parameters $(\Theta_1, \ldots, \Theta_k)$, *given* values of the random-variable outcomes $(x_1, \ldots, x_n)$. Thus,

$$L(\boldsymbol{\Theta}; \mathbf{x}) \equiv f(\mathbf{x}; \boldsymbol{\Theta})$$

defines the functional form of the likelihood function $L(\Theta; \mathbf{x})$, where $\Theta$ now *precedes* the semicolon and $\mathbf{x}$ *follows* the semicolon to denote that the likelihood function is a function of $\Theta$, for given values of $\mathbf{x}$. A **maximum likelihood** (ML) **estimate** of $\Theta$ is obtained as a solution, $\hat{\boldsymbol{\theta}} \in \Omega$, to the maximization problem

$$L(\hat{\boldsymbol{\theta}}; \mathbf{x}) = \max_{\Theta \in \Omega} L(\Theta; \mathbf{x}).$$

where $\Omega$ is the appropriate parameter space. A maximum likelihood estimate can be defined equivalently as $\hat{\boldsymbol{\theta}} = \hat{\Theta}(\mathbf{x}) = \arg\max_{\Theta \in \Omega} L(\Theta; \mathbf{x})$.[19]

Since, by definition it is also true that $\hat{\boldsymbol{\theta}} = \arg\max_{\Theta \in \Omega} f(\mathbf{x}; \Theta)$ (assuming $\mathbf{x}$ is given), the maximum likelihood procedure can be interpreted as choosing, from among all candidates, the value of the parameter vector $\Theta$ identifying the joint density function $f(\mathbf{x}; \Theta)$ that assigns the highest probability (discrete case) or highest density weighting (continuous case) to the random-sample outcome, $\mathbf{x}$, actually observed. Put another way, the maximum likelihood procedure chooses a parameter vector value so as to identify a particular member of a parametric family of densities, $f(\mathbf{x}; \Theta)$, $\Theta \in \Omega$, that assigns the highest "likelihood" to generating the random-sample outcome actually observed. Of course, whether the estimator of $\Theta$ implied by this procedure is useful depends on the properties of the **maximum likelihood estimator (MLE)** $\hat{\Theta} = \hat{\Theta}(\mathbf{X}) = \arg\max_{\Theta \in \Omega} L(\Theta; \mathbf{X})$.

### MLE Mechanics

In order to implement the ML procedure, the functional form of $f(\mathbf{x}; \Theta)$ and hence $L(\Theta; \mathbf{x})$ must be fully specified along with the feasible choices of $\Theta$, represented by the parameter space, $\Omega$. Thus the researcher essentially must specify a parametric family of density functions in order to provide a complete representation of the statistical model for the random sample under investigation. Once the parametric family is identified, the estimation problem focuses on the maximization of the likelihood function. In many cases, the likelihood function will be differentiable with respect to the parameter vector and will possess a maximum that will be interior to the parameter space (as opposed to being on the boundary of the parameter space). In these cases, the classical calculus-based approach to maximization, in which first-order conditions are used to solve for the maximizing value of the $\Theta$ vector, can be used to solve the maximization problem. That is, the ML estimate, $\hat{\boldsymbol{\theta}}$, can be found as the

[19] In general, $\arg\max f(w)$ denotes the *argument* of $f(w)$ that *maximizes* $f(w)$, where *argument* means *the value of* $w$. Also, $\arg\max_{w \in \Omega} f(w)$ denotes the value of $w \in \Omega$ that maximizes $f(w)$.

solution to the vector equation

$$\frac{\partial L(\Theta;\mathbf{x})}{\partial \Theta_{(k\times 1)}} = \begin{bmatrix} \dfrac{\partial L(\Theta;\mathbf{x})}{\partial \Theta_1} \\ \vdots \\ \dfrac{\partial L(\Theta;\mathbf{x})}{\partial \Theta_k} \end{bmatrix} = [0]_{(k\times 1)},$$

whose solution is a function of the random-sample outcome $\mathbf{x}$, as $\hat{\boldsymbol{\theta}} = \hat{\Theta}(\mathbf{x}) = \arg_{\Theta\in\Omega}[\partial L(\Theta;\mathbf{x})/\partial\Theta = [0]]$.[20]

Note that it may not be possible to *explicitly* solve the first-order conditions for $\hat{\boldsymbol{\theta}}$ in terms of a function of $\mathbf{x}$. If not, numerical methods would be required to find the value of $\hat{\boldsymbol{\theta}}$ that satisfied the first-order conditions. More generally, even if the classical maximization approach is not appropriate (e.g., there is no interior maximum, or $L(\Theta;\mathbf{x})$ is not differentiable with respect to $\Theta$), a value of $\hat{\boldsymbol{\theta}}$ that solves $\max_{\Theta\in\Omega} L(\Theta;\mathbf{x})$ is a ML estimate of $\Theta$, *no matter how it is derived*. The estimator function $\hat{\boldsymbol{\theta}} = \hat{\Theta}(\mathbf{x})$ will be either explicitly derivable, or else an *implicit* function implied by the functional dependence of $\hat{\boldsymbol{\theta}}$ on $\mathbf{x}$, and in either case $\hat{\Theta} = \hat{\Theta}(\mathbf{X})$ is referred to as a maximum likelihood estimator of $\Theta$.[21]

The following examples illustrate the derivation of MLEs. Note that in some problem situations the calculations are considerably simplified by maximizing $\ln L(\Theta;\mathbf{x})$ as opposed to $L(\Theta;\mathbf{x})$. Since the logarithmic transformation is strictly monotonically increasing, if $\hat{\boldsymbol{\theta}}$ maximizes $L(\Theta;\mathbf{x})$, it also maximizes $\ln L(\Theta;\mathbf{x})$, and vice versa, i.e.,

$$L(\hat{\boldsymbol{\theta}};\mathbf{x}) = \max_{\Theta\in\Omega} L(\Theta;\mathbf{x}) \Leftrightarrow \ln L(\hat{\boldsymbol{\theta}};\mathbf{x}) = \max_{\Theta\in\Omega} \ln L(\Theta;\mathbf{x})$$

and

$$\hat{\boldsymbol{\theta}} = \arg\max_{\Theta\in\Omega} L(\Theta;\mathbf{x}) = \arg\max_{\Theta\in\Omega} \ln L(\Theta;\mathbf{x}).$$

Thus the objective function of the ML estimation problem can be chosen to be $L(\Theta;\mathbf{x})$ or $\ln L(\Theta;\mathbf{x})$, whichever is more convenient.[22] In applying the logarith-

[20] $\arg_{\Theta\in\Omega}[g(\Theta) = c]$ represents the value of $\Theta \in \Omega$ that satisfies $g(\Theta) = c$, i.e., the *argument* of $g$ that solves $g(\Theta) = c$.

[21] We are suppressing the fact that a maximum of $L(\Theta;\mathbf{x})$ may not be attainable. For example, if the parameter space is an open interval and if the likelihood function is strictly monotonically increasing, then no maximum could be stated. If a maximum of $L(\Theta;\mathbf{x})$ for $\Theta \in \Omega$ does not exist, then the MLE of $\Theta$ does not exist.

[22] In the case where the classical first-order conditions are applicable, note that if $L(\Theta;\mathbf{x}) > 0$ (which will necessarily be true at the maximum value), then

$$\frac{\partial \ln L(\Theta;\mathbf{x})}{\partial\Theta} = \frac{1}{L(\Theta;\mathbf{x})}\frac{\partial L(\Theta;\mathbf{x})}{\partial\Theta},$$

and thus any $\Theta$ for which $\partial L(\Theta;\mathbf{x})/\partial\Theta = [0]$ also satisfies $\partial \ln L(\Theta;\mathbf{x})/\partial\Theta = [0]$. Regarding second-order conditions, note that if $\Theta$ satisfies the first-order conditions, then

$$\frac{\partial^2 \ln L(\Theta;\mathbf{x})}{\partial\Theta\partial\Theta'} = \frac{1}{L(\Theta;\mathbf{x})}\frac{\partial^2 L(\Theta;\mathbf{x})}{\partial\Theta\partial\Theta'} - \frac{\partial L(\Theta;\mathbf{x})}{\partial\Theta}\frac{\partial L(\Theta;\mathbf{x})}{\partial\Theta'} = \frac{1}{L(\Theta;\mathbf{x})}\frac{\partial^2 L(\Theta;\mathbf{x})}{\partial\Theta\partial\Theta'}$$

mic transformation, we *define* $\ln(0) \equiv -\infty$ to accommodate points where the likelihood function is zero valued.

**Example 8.4** MLE for Exponential Distribution

Let $(X_1, \ldots, X_n)$ be a random sample from an exponential population distribution representing the operating time until a work stoppage occurs on an assembly line, so that $X_i \sim \theta^{-1} \exp(-x_i/\theta) I_{(0,\infty)}(x_i)\ \forall\ i$. Define the MLE for $\theta$, the mean operating time until a work stoppage.

**Answer**: The functional form of the joint density function for the random sample, and hence the functional form of the likelihood function, is given by

$$L(\theta; x_1, \ldots, x_n) \equiv f(x_1, \ldots, x_n; \theta) = \theta^{-n} \exp\left(-\sum_{i=1}^{n} x_i/\theta\right) \prod_{i=1}^{n} I_{(0,\infty)}(x_i),$$

where $\theta \in \Omega = (0, \infty)$. Then,

$$\ln L(\theta; x_1, \ldots, x_n) = -n \ln \theta - \frac{\sum_{i=1}^{n} x_i}{\theta} + \ln \prod_{i=1}^{n} I_{(0,\infty)}(x_i).$$

The first-order condition for maximizing $\ln L(\theta; \mathbf{x})$ with respect to $\theta$ is given by

$$\frac{d \ln L}{d\theta} = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} = 0.$$

Thus the solution $\hat{\theta} = \hat{\Theta}(\mathbf{x}) = \sum_{i=1}^{n} x_i/n = \bar{x}$ is the ML estimate of $\theta$, and $\hat{\Theta} = \hat{\Theta}(\mathbf{X}) = \sum_{i=1}^{n} X_i/n = \bar{X}_n$ is the MLE of $\theta$. (Check that the second-order conditions for a maximum are met.) □

**Example 8.5** MLE for Normal Distribution

Let $(X_1, \ldots, X_n)$ be a random sample from a normal population distribution representing the actual fill volumes of 1-liter bottles of liquid laundry detergent so that

$$X_i \sim \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right], \quad \forall\ i.$$

Define the MLE for the mean and variance of fill volumes.

**Answer**: The functional form of the joint density for the random sample, and hence the functional form of the likelihood function, is given by

$$L(\mu, \sigma^2; \mathbf{x}) \equiv f(\mathbf{x}; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right],$$

since $\partial L(\Theta; \mathbf{x})/\partial \Theta = [0]$. Then since $L(\Theta; \mathbf{x}) > 0$ at the maximum, $\partial^2 \ln L(\Theta; \mathbf{x})/\partial\Theta\partial\Theta'$ is negative definite iff $\partial^2 L(\Theta; \mathbf{x})/\partial\Theta\partial\Theta'$ is negative definite.

where $\mu \geq 0$ and $\sigma^2 > 0$. Then

$$\ln L(\mu, \sigma^2; \mathbf{x}) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

The first-order conditions for the maximum of $\ln L(\mu, \sigma^2; \mathbf{x})$ are

$$\frac{\partial \ln L}{\partial \mu} = \sigma^{-2} \sum_{i=1}^{n} (x_i - \mu) = 0,$$

$$\frac{\partial \ln L}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0.$$

The solution $\hat{\mu}(\mathbf{x}) = \sum_{i=1}^{n} x_i/n$ and $\hat{\sigma}^2(\mathbf{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2/n$ defines the ML estimate of $\mu$ and $\sigma^2$, and the MLE is given by $\hat{\mu}(\mathbf{X})$ and $\hat{\sigma}^2(\mathbf{X})$. (Check that the second-order conditions for a maximum are met.) □

**Example 8.6** Maximum Likelihood Estimation in the GLM

Let $\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\mathbf{Y} \sim N(\mathbf{x}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ represent the relationship between a random sample of family expenditures on consumer durables, $\mathbf{Y}$, and the respective levels of disposable income and other sociodemographic factors for the families, $\mathbf{x}$. Define an MLE both for the marginal effects of sociodemographic variables on consumption, i.e., for $\partial \mathrm{E}Y_i/\partial \mathbf{x}_{i.}' = \boldsymbol{\beta}$, and for the variance of consumption, $\sigma^2$.

**Answer:** The likelihood function in this case is

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = N(\mathbf{x}\boldsymbol{\beta}, \sigma^2\mathbf{I}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})\right],$$

where $\boldsymbol{\beta} \in R^k$ and $\sigma^2 > 0$.[23] Then

$$\ln L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = -(n/2)\ln 2\pi - (n/2)\ln \sigma^2 - (1/2\sigma^2)(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}).$$

The first-order conditions for the maximum of $\ln L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$ are given by (recall Lemma 8.1)

$$\frac{\partial \ln L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})}{\partial \boldsymbol{\beta}} = \frac{\partial\left[-\frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{x}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{x}'\mathbf{x}\boldsymbol{\beta})\right]}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}(-2\mathbf{x}'\mathbf{y} + 2\mathbf{x}'\mathbf{x}\boldsymbol{\beta}) = [0],$$

$$\frac{\partial \ln L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) = 0.$$

The solution to the first-order conditions implies that the ML estimate is

$$\hat{\boldsymbol{\beta}}(\mathbf{y}) = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \quad \text{and} \quad \hat{\sigma}^2(\mathbf{y}) = \frac{(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}})}{n},$$

with the MLE given by $\hat{\boldsymbol{\beta}}(\mathbf{Y})$ and $\hat{\sigma}^2(\mathbf{Y})$. (Check that the second-order conditions for a maximum are met.) □

[23] Economic theory may suggest constraints on the signs of some of the entries in $\boldsymbol{\beta}$ (e.g., the effect of income on durables consumption will be positive), in which case $\boldsymbol{\beta} \in \Omega_{\boldsymbol{\beta}} \subset R^k$ may be more appropriate.

The following examples illustrate cases where the standard calculus approach cannot be utilized.

**Example 8.7** MLE for Uniform Distribution

Let $(X_1, \ldots, X_n)$ be a random sample from a uniform population distribution representing measurements of the hardness of steel, based on the Rockwell scale, produced by a certain foreign manufacturer, so that $X_i \sim (1/(b-a))I_{[a,b]}(x_i)\ \forall\ i$. Define an MLE for the lower and upper bounds to the hardness measurements, $a$ and $b$, respectively.
**Answer**: The likelihood function is given by

$$L(a, b; \mathbf{x}) \equiv f(\mathbf{x}; a, b) = (b-a)^{-n} \prod_{i=1}^{n} I_{[a,b]}(x_i),$$

where $a < b$. It is clear that for $L(a, b; \mathbf{x})$ to be maximized, $a$ and $b$ must be chosen to make $(b-a)$ as small as possible while still maintaining $\prod_{i=1}^{n} I_{[a,b]}(x_i) = 1$. Then the smallest choice for $b$ is given by $\max(x_1, \ldots, x_n)$, while the largest choice for $a$ is given by $\min(x_1, \ldots, x_n)$, yielding the smallest $(b-a) = \max(x_1, \ldots, x_n) - \min(x_1, \ldots, x_n)$. Thus, the MLE estimates are given by outcomes of the smallest and largest *order statistics*, or *extreme values*, as

$$\hat{a}(\mathbf{x}) = \min(x_1, \ldots, x_n) \quad \text{and} \quad \hat{b}(\mathbf{x}) = \max(x_1, \ldots, x_n),$$

and the MLEs are then $\hat{a}(\mathbf{X})$ and $\hat{b}(\mathbf{X})$. □

**Example 8.8** MLE for Hypergeometric Distribution

Let $X \sim$ Hypergeometric $(x; M, k, n)$ represent the number of defective parts found in a random sample without replacement of $n$ parts taken from a shipment of $M$ parts, where $M$ and $n$ are known. Define an MLE of the number of defective parts, $k$, in the shipment of $M$ parts.
**Answer**: The likelihood function is given by

$$L(k; x) \equiv f(x; k) = \frac{\binom{k}{x}\binom{M-k}{n-x}}{\binom{M}{n}}$$

where $k \in \{0, 1, 2, \ldots, M\}$. Finding the solution for $k$ that maximizes $L(k; x)$ is essentially an *integer programming problem*. Specifically, note that

$$\frac{L(k; x)}{L(k-1; x)} = \frac{\frac{k!}{x!(k-x)!} \; \frac{(M-k)!}{(n-x)!(M-k-(n-x))!}}{\frac{(k-1)!}{x!(k-x-1)!} \; \frac{(M-k+1)!}{(n-x)!(M-k-(n-x)+1)!}} = \frac{k(M-k-(n-x)+1)}{(k-x)(M-k+1)}.$$

Then $L(k; x)/L(k-1; x) \geq 1$ iff $k(M-k-(n-x)+1) \geq (k-x)(M-k+1)$, which after algebraic simplification is equivalent to $k \leq n^{-1}x(M+1)$. The implication of the preceding result is that the likelihood function increases as the integer value of $k$ increases so long as $k \leq n^{-1}x(M+1)$. Therefore, the ML estimate of $k$ equals the largest integer not exceeding $n^{-1}x(M+1)$, which we can represent as $\hat{k}(x) = \text{trunc}(n^{-1}x(M+1))$, where recall that $\text{trunc}(w)$ is the truncation function that truncates the decimal part of $w$, e.g., trunc(2.76) = 2. The MLE of $k$ would then be $\hat{k}(X) = \text{trunc}(n^{-1}X(M+1))$. □

### MLE Properties: Finite Sample

As we stated at the beginning of this section, whether the estimates produced by the ML procedure are useful depends on the properties that an MLE possesses. It turns out that there are several reasons why we might suspect that the ML procedure would lead to good estimates of $\Theta$. First of all, if an unbiased estimator of $\Theta$ exists that achieves the CRLB, then the MLE will be this estimator if the MLE can be defined by solving first-order conditions for maximizing the likelihood function.

**Theorem 8.11** (MLE Attainment of the CRLB) *If there exists an unbiased estimator,* $\mathbf{T} = \mathbf{t}(\mathbf{X})$, *of* $\Theta$ *that has a covariance matrix equal to the CRLB, and if the MLE can be defined by solving first-order conditions for maximizing the likelihood function, then the MLE is equal to* $\mathbf{T} = \mathbf{t}(\mathbf{X})$ *with probability 1.*

**Proof** If there exists an unbiased estimator, $\mathbf{t}(\mathbf{X})$, whose covariance matrix equals the CRLB, then *regardless* of the value of $\Theta \in \Omega$, the theorem on the attainment of the CRLB (Theorem 7.17) implies that the estimator has outcomes defined by

$$\mathbf{t}(\mathbf{x}) = \Theta + \left[\mathrm{E}\frac{\partial \ln L(\Theta;\mathbf{X})}{\partial \Theta}\frac{\partial \ln L(\Theta;\mathbf{X})'}{\partial \Theta}\right]^{-1}\frac{\partial \ln L(\Theta;\mathbf{x})}{\partial \Theta}$$

with probability 1, where we have expressed the result of Theorem 7.17 using likelihood function notation. Substituting the ML estimate, $\hat{\theta} = \hat{\Theta}(\mathbf{x}) \in \Omega$, for $\Theta$ in the preceding equality implies that $\mathbf{t}(\mathbf{x}) = \hat{\theta}$, since $\hat{\theta}$ would satisfy the first-order conditions $\partial \ln L(\hat{\theta};\mathbf{x})/\partial\Theta = [\mathbf{0}]$. Then outcomes of the MLE and $\mathbf{t}(\mathbf{X})$ coincide with probability 1. ■

Thus, under the conditions of Theorem 8.11, the MLE will be the MVUE for $\Theta$.

We can also show that if the MLE is uniquely defined, then the MLE can be equivalently represented as a function of *any* sufficient statistics for $f(\mathbf{x};\Theta)$ and, in particular, a function of *complete* sufficient statistics when complete sufficient statistics exist.

**Theorem 8.12** (Unique MLEs are Functions of Any Sufficient Statistics for $f(\mathbf{x};\Theta)$) *Assume that the MLE of* $\Theta$, *say* $\hat{\Theta}$, *is uniquely defined in terms of* $\mathbf{X}$. *If* $\mathbf{S} = [S_1,\ldots,S_r]'$ *is any vector of sufficient statistics for* $f(\mathbf{x};\Theta) \equiv L(\Theta;\mathbf{x})$, *then there exists a function of* $\mathbf{S}$, *say* $\tau(\mathbf{S})$, *such that* $\hat{\theta} = \tau(\mathbf{s})$.

**Proof** The Neyman factorization theorem states that

$$L(\Theta;\mathbf{x}) \equiv f(\mathbf{x};\Theta) = g(s_1,\ldots,s_r;\Theta)h(\mathbf{x}),$$

where $(s_1, \ldots, s_r)$ are sufficient statistics that can be *complete* sufficient statistics if they exist. Now since $L(\Theta; \mathbf{x}) \geq 0$, $g$ and $h$ can always be defined as nonnegative-valued functions, in which case for a *given* value of $\mathbf{x}$,

$$L(\Theta; \mathbf{x}) \propto g(s_1, \ldots, s_r; \Theta)$$

where "$\propto$" means "proportional to" and the proportionality constant is $h(\mathbf{x})$. It follows that for a *given* value of $\mathbf{x}$, if the MLE is unique, then

$$\hat{\boldsymbol{\theta}} = \arg\max_{\Theta \in \Omega} L(\Theta; \mathbf{x}) = \arg\max_{\Theta \in \Omega} g(s_1, \ldots, s_r; \Theta),$$

i.e., $\hat{\boldsymbol{\theta}}$ maximizes $L(\Theta; \mathbf{x})$ iff $\hat{\boldsymbol{\theta}}$ maximizes $g(s_1, \ldots, s_r; \Theta)$. But the latter maximization problem implies that the unique maximizing choice of $\Theta$ is then a function of the values $(s_1, \ldots, s_r)$, i.e., $\hat{\boldsymbol{\theta}} = \tau(s_1, \ldots, s_r)$.[24] ■

If the sufficient statistics $(S_1, \ldots, S_r)$ used in the Neyman factorization theorem are complete, then the unique MLE $\hat{\Theta} = \tau(S_1, \ldots, S_r)$ is a function of the complete sufficient statistics, by Theorem 8.12. It follows from the Lehmann–Scheffé completeness theorem that if $\hat{\Theta}$ is unbiased (or if $\hat{\Theta}$ can be transformed to be unbiased) then the MLE (or the **bias-adjusted MLE**) is the MVUE for $\Theta$. We formalize this observation in the following theorem.

**Theorem 8.13** (MVUE Property of Unique MLEs) *Assume that the MLE of $\Theta$, $\hat{\Theta}$, is uniquely defined in terms of $\mathbf{X}$, and that a vector of complete sufficient statistics, $\underset{\sim}{\mathbf{S}}$, exists for $f(\mathbf{x};\Theta) \equiv L(\Theta;\mathbf{x})$. If $\hat{\Theta}$ or $\eta(\hat{\Theta})$ is an unbiased estimator of $\Theta$, then $\hat{\Theta}$ or $\eta(\hat{\Theta})$ is the MVUE of $\Theta$.*

**Proof** From Theorem 8.12 it follows that the MLE is a function of the complete sufficient statistics as $\hat{\Theta} = \tau(\mathbf{S})$. It follows immediately from the Lehmann–Scheffé completeness theorem (Theorem 7.19) that if $\hat{\Theta}$ is unbiased, then $\hat{\Theta}$ is the MVUE of $\Theta$. Alternatively, if $\hat{\Theta}$ is biased, but the function of the MLE given by $\eta(\hat{\Theta})$ is an unbiased estimator of $\Theta$, then since $\eta(\hat{\Theta}) = \eta(\tau(\mathbf{S}))$ is a (composite) function of the complete sufficient statistics, $\eta(\hat{\Theta})$ is the MVUE of $\Theta$ by the Lehmann–Scheffé completeness theorem. ■

In Ex. 8.4, $\sum_{i=1}^{n} X_i$ is a complete sufficient statistic for $f(\mathbf{x}; \theta) \equiv L(\theta; \mathbf{x})$, and the MLE is unique and unbiased, and thus the MLE is MVUE. In Ex. 8.5, $\sum_{i=1}^{n} X_i^2$ and $\sum_{i=1}^{n} X_i$ are complete sufficient statistics, and the MLEs $\hat{\mu}$ and $\hat{\sigma}^2$ are unique, so that they are functions of the complete sufficient statistics. However, $\hat{\sigma}^2$ is *not* unbiased. The bias can be removed via multiplying $\hat{\sigma}^2$ by $n/(n-1)$. Then the bias-adjusted MLE of $\mu$ and $\sigma^2$, namely $(\hat{\mu}, n\hat{\sigma}^2/(n-1))$, is the MVUE by the Lehmann–Scheffé completeness theorem. The case of Ex.

[24] In the event that a MLE is *not* unique, then the *set* of MLEs is a function of any set of sufficient statistics. However, a particular MLE within the set of MLEs need not necessarily be a function of $(s_1, \ldots, s_r)$, although it is always possible to choose an MLE that *is* a function of $(s_1, \ldots, s_r)$. See D. S. Moore (1971), "Maximum likelihood and sufficient statistics," *American Mathematical Monthly*, January, pp. 50–52.

8.6 was discussed in the previous section on the least-squares estimator, where $(\hat{\beta}, n\hat{\sigma}^2/(n-k))'$, which is the bias-adjusted MLE, was found to be the MVUE of $(\beta, \sigma^2)'$. In Ex. 8.7, it can be shown that the MLE $\hat{b}(\mathbf{X}) = \max(X_1, \ldots, X_n)$ and $\hat{a}(\mathbf{X}) = \min(X_1, \ldots, X_n)$ is unique and is itself a complete sufficient statistic and that $\hat{\alpha}(X) = ((n+1)/n)\hat{a}(\mathbf{X})$, and $\hat{\beta}(\mathbf{X}) = ((n+1)/n)\hat{b}(\mathbf{X})$ are unbiased estimators of $a$ and $b$, respectively. Thus, the bias-adjusted MLE $(\hat{\alpha}(\mathbf{X}), \hat{\beta}(\mathbf{X}))$ is the MVUE for $(a, b)$. (See Bickel and Doksum, *Mathematical Statistic*, pp. 125–126, and Johnson and Kotz (1969), *Discrete Distributions*, New York: John Wiley, pp. 146–148.) Finally, in Ex. 8.8, it can be shown that $X$ is a complete sufficient statistic, and the unique MLE can be transformed as $t(X) = (M/n)X$ to define an unbiased estimator of $k$. Then the bias-adjusted MLE, $t(X) = (M/n)X$, is the MVUE for $k$ (see Bickel and Doksum, *Mathematical Statistics*, pp. 122–123).

As we have seen here, the maximum likelihood procedure is a rather straightforward approach to defining estimators of unknown parameters, and in many cases the estimator, or a simple transformation of it, will be unbiased and the MVUE. Also, since a unique MLE will always be a function of complete sufficient statistics whenever the latter exist, the unique MLE would at least appear to be a reasonable starting point in the search for the MVUE of the parameters of interest. On the other hand, the MLE need not be unbiased nor be the MVUE, and there may be no apparent transformation of the MLE that achieves the MVUE property. It is useful to examine asymptotic properties of MLEs since, even if a MLE does not possess the finite sample properties that one might desire, MLEs possess desirable, large-sample properties under general conditions, and these latter properties might still rationalize the use of an MLE for estimating the parameters of a particular statistical model.

### MLE Properties: Large Sample

There are two general approaches that one might follow in establishing asymptotic properties of a MLE. First of all, if a MLE can be explicitly solved for, so that one can analyze an explicit, real-valued function of the random sample, $\hat{\Theta} = \hat{\Theta}(\mathbf{X})$, then it might be possible to apply laws of large numbers and/or central limit theorems to $\hat{\Theta}(\mathbf{X})$ *directly*, to investigate the asymptotic properties of the MLE. Note the following example.

**Example 8.9** Recall Ex. 8.4, where it was found that the MLE of $\theta$, when random sampling from an exponential population distribution, is given by $\hat{\Theta} = n^{-1}\sum_{i=1}^{n} X_i = \bar{X}_n$. Through direct evaluation of the MLE function definition, we can establish that $\hat{\Theta}$ is a consistent, asymptotically normal, and asymptotically efficient estimator of $\theta$. In fact, the procedure for establishing these asymptotic properties has already been carried out in Ex. 7.22, where it was demonstrated that $\hat{\Theta} = \bar{X}_n$ was a consistent and asymptotically efficient estimator of $\theta$ having an asymptotically normal distribution given by $\hat{\Theta} \overset{a}{\sim} N(\theta, \theta^2/n)$. □

At times, the function defining a MLE cannot be defined explicitly, even though the MLE estimates can be calculated, or else the explicit definition of a MLE may be so complicated that it is unclear how laws of large numbers or central limit theorems could be applied. For these cases, regularity conditions on the likelihood functions have been presented in the literature that ensure the MLE is consistent, asymptotically normal, and asymptotically efficient. As an illustration of a situation in which direct evaluation of the asymptotic properties (and finite sample properties) of an MLE is not possible, consider the following example in which maximum likelihood estimation of the parameters of a Gamma density is being pursued.

**Example 8.10** MLE for Parameters of a Gamma Distribution

Let $(X_1, \ldots, X_n)$ be a random sample from a Gamma population distribution representing the time between breakdowns of a certain type of refrigeration equipment used in the frozen foods section of a major supermarket chain, so that

$$X_i \sim \frac{1}{\beta^\alpha \Gamma(\alpha)} x_i^{\alpha-1} e^{-x_i/\beta} I_{(0,\infty)}(x_i) \quad \forall\, i.$$

The likelihood function is given by

$$L(\alpha, \beta; \mathbf{x}) = \frac{1}{\beta^{n\alpha}[\Gamma(\alpha)]^n} \prod_{i=1}^n x_i^{\alpha-1} \exp\left(-\sum_{i=1}^n x_i/\beta\right) \prod_{i=1}^n I_{(0,\infty)}(x_i),$$

where $\alpha > 0$ and $\beta > 0$. The log-likelihood function is given by

$$\ln L(\alpha, \beta; \mathbf{x}) = -n\alpha(\ln \beta) - n(\ln \Gamma(\alpha)) + (\alpha - 1)\sum_{i=1}^n \ln x_i - \sum_{i=1}^n x_i/\beta + \ln \prod_{i=1}^n I_{(0,\infty)}(x_i).$$

The first-order conditions characterizing the maximum of the log-likelihood function are given by

$$\frac{\partial \ln L}{\partial \alpha} = -n(\ln \beta) - \frac{n}{\Gamma(\alpha)}\frac{d\Gamma(\alpha)}{d\alpha} + \sum_{i=1}^n \ln x_i = 0,$$

$$\frac{\partial \ln L}{\partial \beta} = -n(\alpha/\beta) + \left(\sum_{i=1}^n x_i\right)\Big/\beta^2 = 0.$$

In attempting to solve the first-order conditions for $\alpha$ and $\beta$, note that the second condition implies that $\alpha\beta = \bar{X}_n$, or $\beta = \bar{X}_n/\alpha$. Substituting this result for $\beta$ in the first condition implies

$$\ln \alpha - \frac{d\Gamma(\alpha)}{d\alpha}\Big/\Gamma(\alpha) = \ln \bar{x}_n - \left(\sum_{i=1}^n \ln x_i\right)\Big/n,$$

and there is *no* explicit solution for $\alpha$ in terms of $(x_1, \ldots, x_n)$, although $\alpha$ is *implicitly* a function of $(x_1, \ldots, x_n)$. A unique value of $\alpha$ satisfying the above equality can be solved for numerically on a computer; that value can then be used to solve for $\beta$ using the equation $\beta = \bar{X}_n/\alpha$.[25] Thus, the ML estimates for $(\alpha, \beta)$ can be calculated. However, since an explicit functional form for the MLE is not identifiable, an analysis of the estimator's finite sample and asymptotic properties is quite difficult. Regarding finite sample properties of the MLE, the reader can verify by an appeal to Theorem 7.6 that $(\sum_{i=1}^n X_i, \sum_{i=1}^n \ln X_i)$ is a set of complete sufficient statistics for this problem, and so the MLE, $(\hat{\alpha}(\mathbf{X}), \hat{\beta}(\mathbf{X}))$, is a (implicit) function of complete sufficient statistics. However, the MLE is biased, and no MVUE estimator for $(\alpha, \beta)$ has been presented in the literature. Bowman and Shenton[26] have obtained expressions for low-order moments of $(\hat{\alpha}(\mathbf{X}), \hat{\beta}(\mathbf{X}))$ that are accurate to terms having order of magnitude $n^{-6}$. For example, when $\alpha \geq 1$ and $n \geq 4$, they found that

$$E\hat{\alpha}(\mathbf{X}) \approx \alpha + \left[3\alpha - \frac{2}{3} + \frac{1}{9}\alpha^{-1} + \frac{13}{405}\alpha^{-2}\right] \Big/ (n-3),$$

and they suggest, as an "approximately unbiased" estimator of $\alpha$, the following function of the MLE for $\alpha$:

$$\hat{\alpha}^*(\mathbf{X}) = \left[(n-3)\hat{\alpha}(X) + \frac{2}{3}\right] / n.$$

Since $\hat{\alpha}^*(\mathbf{X})$ is a function of complete sufficient statistics (because the MLE $\hat{\alpha}(\mathbf{X})$ is), and since $\hat{\alpha}^*(\mathbf{X})$ is "approximately unbiased," $\hat{\alpha}^*(\mathbf{X})$ might be interpreted as being "approximately MVUE" for $\alpha$. For further details on finite sample properties of the MLE $(\hat{\alpha}(\mathbf{X}), \hat{\beta}(\mathbf{X}))$, see Bowman and Shenton. An analysis of the asymptotic properties of $(\hat{\alpha}(X), \hat{\beta}(X))$ will be developed in a subsequent example. □

A varied collection of regularity conditions on likelihood functions that represent sufficient conditions for MLEs to possess desirable asymptotic properties has been presented in the literature. Most of these conditions apply specifically to the case of random sampling from a population distribution, so that the random sample $(X_1, \ldots, X_n)$ must be a collection of iid random variables. For a survey of alternative types of regularity conditions, the reader can refer to the article by Norden.[27] We will concentrate on regularity conditions that do not cover all cases but that are relatively simple to comprehend and apply and that focus attention on key assumptions that lead to good asymptotic properties of MLEs. The conditions we present do *not* require that the random sample be a collection of iid random variables, and so the conditions can also be applied in cases other than random sampling from a population distribution.

[25] Alternatively, the solution for $\alpha$ can be determined by consulting tables generated by Chapman which were constructed specifically for this purpose (D. G. Chapman (1956), "Estimating parameters of a truncated gamma distribution." *Ann. Math. Stat.*, 27, pp. 498–506).

[26] K. O. Bowman, and L. R. Shenton (1968), *Properties of Estimators for the Gamma Distribution.* Report CTC-1, Union Carbide Corp., Oak Ridge, Tenn.

[27] R. H. Norden (1972), "A survey of maximum likelihood estimation." *International Statistical Revue*, 40, pp. 329–354 and (1973) 41, pp. 39–58.

Henceforth, we focus on the case where an MLE is the unique global maximum of the likelihood function, which covers the majority of applications. An examination of the multiple local optimum case is more complicated and is best left to a more advanced course of study.[28]

*Consistency* We first examine conditions that ensure the consistency of the MLE of $\Theta$ in the scalar case.

**Theorem 8.14** (MLE Consistency—Sufficient Conditions for Scalar $\Theta$) *Let* $\{f(\mathbf{x};\Theta), \Theta \in \Omega\}$ *be the statistical model for the random sample* $\mathbf{X}$, *where* $\Theta$ *is a scalar.*[29] *Assume that*

1. *the PDFs* $f(\mathbf{x};\Theta), \Theta \in \Omega$, *have common support,* $\Xi$,
2. *the parameter space,* $\Omega$, *is an open interval,*
3. $\ln L(\Theta;\mathbf{x})$ *is continuously differentiable with respect to* $\Theta \in \Omega \; \forall \, \mathbf{x} \in \Xi$,
4. $\partial \ln L(\Theta;\mathbf{x})/\partial\Theta = 0$ *has a unique solution for* $\Theta \in \Omega$, *and the solution defines the unique maximum likelihood estimate,* $\hat{\Theta}(\mathbf{x}), \forall \, \mathbf{x} \in \Xi$,
5. $\lim_{n\to\infty} P(\ln L(\Theta_o;\mathbf{x}) > \ln L(\Theta;\mathbf{x})) = 1$ *for* $\Theta \neq \Theta_o$, *where* $\Theta_o$ *is the true value of* $\Theta \in \Omega$.[30]

*Then* $\hat{\Theta} \xrightarrow{p} \Theta_o$, *and the MLE is thus consistent for* $\Theta$.

**Proof** Let $h > 0$ be such that $\Theta_o - h \in \Omega$ and $\Theta_o + h \in \Omega$, where such an $h$ exists by condition (2), and define the events

$$A_n = \{\mathbf{x}:\ \ln L(\Theta_o;\mathbf{x}) > \ln L(\Theta_o - h;\mathbf{x})\},$$

$$B_n = \{\mathbf{x}:\ \ln L(\Theta_o;\mathbf{x}) > \ln L(\Theta_o + h;\mathbf{x})\},$$

$$H_n = A_n \cap B_n.$$

As $n \to \infty$, $P(H_n) \to 1$, since $P(H_n) = P(A_n) + P(B_n) - P(A_n \cup B_n)$, $P(A_n)$ and $P(B_n)$ both converge to 1 as $n \to \infty$ by assumption (5), and $P(A_n \cup B_n)$ converges to 1 since $P(A_n \cup B_n) \geq P(A_n)$ and $P(A_n \cup B_n) \geq P(B_n)$.

Now note that $\mathbf{x} \in H_n \Rightarrow L(\Theta;\mathbf{x})$ exhibits its unique maximum for some value $\Theta_*$ such that $\Theta_o - h < \Theta_* < \Theta_o + h$ since, by the differentiability of $L(\Theta;\mathbf{x})$, $\Theta_*$ solves $\partial \ln L(\Theta;\mathbf{x})/\partial\Theta) = 0$ and is thus the ML estimate, $\hat{\Theta}(\mathbf{x})$, of $\Theta$ given $\mathbf{x}$. Note further that

$$H_n \subset \{\mathbf{x}:\ \Theta_o - h < \hat{\Theta}(\mathbf{x}) < \Theta_o + h\}$$

since $\mathbf{x} \in H_n$ *implies* $\Theta_o - h < \hat{\Theta}(\mathbf{x}) < \Theta_o + h$. Then $P(H_n) \to 1$ as $n \to \infty$ implies $P(H_n) \leq P(\Theta_o - h < \hat{\Theta}(\mathbf{x}) < \Theta_o + h) \to 1$ as $n \to \infty$. Since the preceding result

[28] See E. Lehmann (1983), *Theory of Point Estimation*, New York: John Wiley and Sons, pp. 420–427.

[29] We remind the reader of our tacit assumption that $\Theta$ is identified (Def. 7.2).

[30] By "true value" of $\Theta$, we again mean that $\Theta_o$ is the value of $\Theta \in \Omega$ for which $f(\mathbf{x};\Theta_o) \equiv L(\Theta_o;\mathbf{x})$ is the actual joint density function of the random sample $\mathbf{X}$. The value of $\Theta_o$ is generally unknown and in the current context is the objective of point estimation.

remains true if we decrease the values of $h$ to be arbitrarily close to zero (but still positive valued), then $\hat{\Theta} \xrightarrow{p} \Theta_o$. ■

In some cases the verification of assumption (5) in Theorem 8.14 can be challenging. If random sampling is from a population distribution, so that the random sample $X_1, \ldots, X_n$ is a collection of iid random variables, then assumption (5) is *not* needed.

**Theorem 8.15** (MLE Consistency—iid and Scalar $\Theta$) *Assume conditions (1)–(4) of Theorem 8.14, and assume further that the random sample $X_1, \ldots, X_n$ is a collection of iid random variables. Then $\hat{\Theta} \xrightarrow{p} \Theta_o$.*

**Proof** Let $\Theta_\ell = \Theta_o - \varepsilon$ and $\Theta_h = \Theta_o + \varepsilon$ for any $\varepsilon > 0$ such that $\Theta_\ell$ and $\Theta_h \in \Omega$ (such $\varepsilon$'s exist by assumption (2)). Define $H(\varepsilon) = \{\mathbf{x}: \ln L(\Theta_o; \mathbf{x}) > \ln L(\Theta_\ell; \mathbf{x})$ and $\ln L(\Theta_o; \mathbf{x}) > \ln L(\Theta_h; \mathbf{x})\}$, and note that $\mathbf{x} \in H(\varepsilon) \Rightarrow \hat{\Theta} \in (\Theta_o - \varepsilon, \Theta_o + \varepsilon)$ since the MLE is unique and is defined via $\partial \ln L(\Theta; \mathbf{x})/\partial\Theta = 0$.

Now define $A(\Theta) = \{\mathbf{x}: \ln L(\Theta_o; \mathbf{x}) > \ln L(\Theta; \mathbf{x})\}$ for $\Theta \neq \Theta_o$, and note that the event $A(\Theta)$ can be equivalently represented as

$$\tau_n(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} \ln[f(x_i; \Theta)/f(x_i; \Theta_o)] < 0.$$

Because $\tau_n(\mathbf{x})$ can be interpreted as the sample mean of $n$ iid random variables of the form $\ln[f(X_i; \Theta)/f(X_i; \Theta_o)]$, Khinchin's WLLN implies that

$$\tau_n(\mathbf{X}) \xrightarrow{p} \mathrm{E}\ln[f(X_i; \Theta)/f(X_i; \Theta_o)].$$

Also, $\ln(z)$ is *strictly* concave over its domain, and so, Jensen's inequality implies that

$$\mathrm{E}\ln[f(X_i; \Theta)/f(X_i; \Theta_o)] < \ln \mathrm{E}[f(X_i; \Theta)/f(X_i; \Theta_o)].$$

Then, since $\mathrm{E}[f(X_i; \Theta)/f(X_i; \Theta_o)] = 1$,[31] the right-hand side of the preceding inequality is zero. Thus, $\tau_n(\mathbf{X})$ converges in probability to a negative number, which implies that $\lim_{n\to\infty} P(A(\Theta)) = 1$ when $\Theta \neq \Theta_o$. This in turn implies that $\lim_{n\to\infty} P(H(\varepsilon)) = 1 \ \forall\ \varepsilon > 0$ since $H(\varepsilon) = A(\Theta_\ell) \cap A(\Theta_h)$, and thus $\lim_{n\to\infty} P(\hat{\Theta} \in (\Theta_o - \varepsilon, \Theta_o + \varepsilon)) = 1 \ \forall\ \varepsilon > 0$ and $\hat{\Theta} \xrightarrow{p} \Theta_o$. ■

**Example 8.11** Consistency of MLE for Exponential Distribution

Reexamine the case of estimating the value of $\theta$ using the MLE when random sampling is from an exponential density function (recall Ex. 8.4 and Ex. 8.9). In

[31] If $X_i$ is a continuous random variable, then since $\Theta_o$ is the true value of $\Theta$,

$$\mathrm{E}[f(X_i; \Theta)/f(X_i; \Theta_o)] = \int_{-\infty}^{\infty} \frac{f(x_i; \Theta)}{f(x_i; \Theta_o)} f(x_i; \Theta_o) dx_i = \int_{-\infty}^{\infty} f(x_i; \Theta) dx_i = 1$$

since $f(x_i; \Theta)$ is a probability density function. The discrete case is analogous.

this case, the joint density function of the random sample is given by

$$f(\mathbf{x};\theta) = \theta^{-n} e^{-\sum_{i=1}^{n} x_i/\theta} \prod_{i=1}^{n} I_{(0,\infty)}(x_i).$$

The parameter space is an open interval since $\Omega = (0,\infty)$. The log of the likelihood function is continuously differentiable for $\theta \in \Omega$, and $\partial \ln L(\theta;\mathbf{x})/\partial\theta = (-n/\theta) + (\sum_{i=1}^{n} x_i)/\theta^2 = 0$ has the unique solution $\Theta = \sum_{i=1}^{n} x_i/n$, which is the unique maximum likelihood estimate $\forall\, \mathbf{x} \in R_{+}^{n}$.[32] Therefore, it follows from Theorem 8.15 that $\hat{\Theta} \xrightarrow{p} \Theta_0$, so the MLE is a consistent estimator. □

We now examine sufficient conditions for MLE consistency when $\mathbf{\Theta}$ is $k$-dimensional. We present two sets of sufficient conditions. The set that consists of (1)–(4a) allows unbounded parameter spaces but is generally more difficult to apply. The other set, which contains (1)–(3) and (4b), is generally more tractable but requires the parameter space to be a bounded and closed rectangle (a closed and bounded *set* can also be used). As a practical matter, one can often state absolute bounds (perhaps very large) for the parameters of a statistical model based on real-world or theoretical considerations relating to an experiment. Thus, boundedness of the parameter space may not represent a serious restriction in practice.

**Theorem 8.16** (MLE Consistency—Sufficient Conditions) *Let $\{f(\mathbf{x};\mathbf{\Theta}),\mathbf{\Theta} \in \Omega\}$ be the statistical model for the random sample* $\mathbf{X}$. *Let $N(\varepsilon) = \{\mathbf{\Theta} : d(\mathbf{\Theta},\mathbf{\Theta_0}) < \varepsilon\}$ be an open $\varepsilon$-neighborhood of $\mathbf{\Theta_0}$, where $\mathbf{\Theta_0}$ is the true value of $\mathbf{\Theta}$.*[33] *Assume*

1. *the PDFs $f(\mathbf{x};\mathbf{\Theta}),\mathbf{\Theta} \in \Omega$, have common support, $\Xi$;*
2. $\ln L(\mathbf{\Theta};\mathbf{x})$ *has continuous first-order partial derivatives with respect to $\mathbf{\Theta} \in \Omega \;\forall\, \mathbf{x} \in \Xi$;*[34]
3. $\partial \ln L(\mathbf{\Theta};\mathbf{x})/\partial\mathbf{\Theta} = [0]$ *has a unique solution that defines the unique maximum likelihood estimate $\hat{\mathbf{\Theta}}(\mathbf{x}) = \arg\max_{\mathbf{\Theta}\in\Omega} L(\mathbf{\Theta};\mathbf{x}) \;\forall\, \mathbf{x} \in \Xi$;*

4a. $\lim_{n\to\infty} P(\ln L(\mathbf{\Theta_0};\mathbf{x}) > \max_{\mathbf{\Theta}\in\overline{N(\varepsilon)}} \ln L(\mathbf{\Theta};\mathbf{x})) = 1 \;\forall\, \varepsilon > 0$ *with $\Omega$ being an open rectangle containing $\mathbf{\Theta}_O$;*[35]

4b. $\lim_{n\to\infty} P(\max_{\mathbf{\Theta}\in\Omega} |n^{-1} \ln L(\mathbf{\Theta};\mathbf{x}) - G(\mathbf{\Theta})| < \varepsilon) = 1 \;\forall\, \varepsilon > 0$, *where $G(\mathbf{\Theta})$ is a continuous function that is uniquely globally maximized at $\mathbf{\Theta} = \mathbf{\Theta_0}$ and $\Omega$ is a bounded closed rectangle containing $\mathbf{\Theta_0}$.*

Then $\hat{\mathbf{\Theta}} \xrightarrow{p} \mathbf{\Theta_0}$.

[32] That this unique solution is a maximum can be demonstrated by noting that $\partial^2 \ln L(\Theta;\mathbf{x})/\partial\Theta^2 = (n/\Theta^2) - (2\sum_{i=1}^{n} x_i/\Theta^3)$, which, when evaluated at the maximum likelihood estimate $\hat{\Theta} = \sum_{i=1}^{n} x_i/n$, yields $(\partial^2 \ln L(\hat{\theta};\mathbf{x}))/\partial\Theta^2) = -n^3/(\sum_{i=1}^{n} x_i)^2 < 0$.

[33] $N(\varepsilon)$ is an open interval, the interior of a circle, the interior of a sphere, and the interior of a hypersphere in 1, 2, 3, and $\geq 4$ dimensions, respectively.

[34] It is allowable that conditions (2) and (3) be violated on a set of $\mathbf{x}$ values having probability zero.

[35] Change "max" to "sup" if max does not exist.

**Proof** Sufficiency of (1)–(4a): Define $H(\varepsilon) = \{\mathbf{x}: \ \ln L(\Theta_o; \mathbf{x}) > \max_{\Theta \in \overline{N(\varepsilon)}} L(\Theta; \mathbf{x})\}$, and note that $\mathbf{x} \in H(\varepsilon) \Rightarrow \hat{\Theta} = \arg\max_{\Theta \in \Omega} L(\Theta; \mathbf{x}) \in N(\varepsilon)$, where $\hat{\Theta}$ is unique by (3). Assumption (4a) implies that $P(\mathbf{x} \in H(\varepsilon)) \to 1$ as $n \to \infty \ \forall \ \varepsilon > 0$, which in turn implies that $P(\hat{\Theta} \in N(\varepsilon)) \to 1$ as $n \to \infty \ \forall \ \varepsilon > 0$. It follows from the definition of $N(\varepsilon)$ that $\hat{\Theta} \xrightarrow{p} \Theta_o$.

*Sufficiency of (1)–3, (4b):[36] Define $\xi(\varepsilon) = G(\Theta_o) - \max_{\Theta \in \Omega \cap \overline{N(\varepsilon)}} G(\Theta)$ (because $\Omega$ is closed and bounded, the feasible space to the maximization problem is closed and bounded by the definition of $N(\varepsilon)$, and thus the maximum exists by Weierstrass's theorem). Let the event $A_n(\varepsilon)$ be defined as

$$A_n(\varepsilon) = \{\mathbf{x}: \ \max_{\Theta \in \Omega} |n^{-1} \ln L(\Theta; \mathbf{x}) - G(\Theta)| < \xi(\varepsilon)/2\}.$$

Letting $\hat{\Theta}$ represent the unique MLE, it follows that

1. $\mathbf{x} \in A_n(\varepsilon) \Rightarrow G(\hat{\Theta}) > n^{-1} \ln L(\hat{\Theta}; \mathbf{x}) - \xi(\varepsilon)/2$,
2. $\mathbf{x} \in A_n(\varepsilon) \Rightarrow n^{-1} \ln L(\Theta_o; \mathbf{x}) > G(\Theta_o) - \xi(\varepsilon)/2$.
   By definition of the MLE, $L(\hat{\Theta}; \mathbf{x}) \geq L(\Theta_o; \mathbf{x})$, so it follows from (1) that
3. $\mathbf{x} \in A_n(\varepsilon) \Rightarrow G(\hat{\Theta}) > n^{-1} \ln L(\Theta_o; \mathbf{x}) - \xi(\varepsilon)/2$.
   Then, substituting the inequality in (2) into the right side of the inequality in (3) yields
4. $\mathbf{x} \in A_n(\varepsilon) \Rightarrow G(\hat{\Theta}) > G(\Theta_o) - \xi(\varepsilon)$,
   which upon substituting the definition of $\xi(\varepsilon)$ yields
5. $\mathbf{x} \in A_n(\varepsilon) \Rightarrow G(\hat{\Theta}) > \max_{\Theta \in \Omega \cap \overline{N(\varepsilon)}} G(\Theta) \Rightarrow \hat{\Theta} \in N(\varepsilon)$.
   Then, since $\lim_{n \to \infty} P(A_n(\varepsilon)) = 1 \ \forall \ \varepsilon > 0$, $\lim_{n \to \infty} P(\hat{\Theta} \in N(\varepsilon)) = 1 \ \forall \ \varepsilon > 0$, so that $\hat{\Theta} \xrightarrow{p} \Theta_o$. ■

We now return to the case of random sampling from the Gamma density where the MLE cannot be explicitly solved for, although $\hat{\Theta}$ is implicitly defined by first-order conditions.

**Example 8.12** Consistency of MLE for Gamma Distribution

Reexamine the case of estimating the value of $\alpha$ and $\beta$ using the MLE when random sampling is from a Gamma population distribution (recall Ex. 8.10). In an attempt to utilize Theorem 8.16 for demonstrating the consistency of the MLE, first note that $f(\mathbf{x}; \Theta) > 0$ for $\mathbf{x} \in R_+^n$, and so condition (1) is satisfied. Recalling Ex. 8.10, it is evident that $\ln L(\Theta; \mathbf{x})$ is continuously differentiable with respect to $\alpha$ and $\beta \ \forall \ \mathbf{x} \in R_+^n$ and for all $\alpha > 0$ and $\beta > 0$, validating condition (2). Also, the first-order conditions have a unique solution for $(\alpha, \beta) \ \forall \ \mathbf{x} \in R_+^n$, which defines the unique maximum likelihood estimate $(\hat{\alpha}, \hat{\beta})$ (again recall Ex. 8.10), satisfying condition (3).

[36] This proof is related to a proof by T. Amemiya (1985), *Advanced Econometrics*, Cambridge, MA: Harvard Univ. Press, p. 107, dealing with the consistency of extremum estimators.

We assume that bounds can be placed on the parameter values of $\alpha$ and $\beta$, and we examine the validity of condition (4b) of Theorem 8.16. Note that (suppressing the indicator function) $n^{-1}\ln L(\alpha,\beta;\mathbf{x}) = -\alpha\ln\beta - \ln\Gamma(\alpha) + (\alpha - 1)n^{-1}\sum_{i=1}^{n}\ln x_i - \bar{x}/\beta$. Then, since plim $\bar{X} = \alpha_o\beta_o$ and plim $n^{-1}\sum_{i=1}^{n}\ln X_i = [((d\Gamma(\alpha_o)/d\alpha)/\Gamma(\alpha_o)) + \ln\beta_o]$,[37] it follows that

$$\text{plim}\, n^{-1}\ln L(\alpha,\beta;\mathbf{x}) = G(\alpha,\beta) = \alpha\ln\left(\frac{\beta_o}{\beta}\right) + (\alpha - 1)\left[\frac{d\Gamma(\alpha_o)/d\alpha}{\Gamma(\alpha_o)}\right] - \ln\Gamma(\alpha) - \frac{\alpha_o\beta_o}{\beta} - \ln\beta_o$$

uniformly in $(\alpha,\beta)$. Furthermore, the maximum of $G(\alpha,\beta)$ occurs at $\alpha = \alpha_o$ and $\beta = \beta_o$. To see this, note that $\alpha_o$ and $\beta_o$ solve the first-order conditions for the maximization problem given by

$$\frac{\partial G(\alpha,\beta)}{\partial\alpha} = \ln\left(\frac{\beta_o}{\beta}\right) + \frac{d\Gamma(\alpha)/d\alpha}{\Gamma(\alpha_o)} - \frac{d\Gamma(\alpha)/d\alpha}{\Gamma(\alpha)} = 0$$

and

$$\frac{\partial G(\alpha,\beta)}{\partial\beta} = -\frac{\alpha}{\beta} + \frac{\alpha_o\beta_o}{\beta^2} = 0.$$

The hessian matrix is negative definite, so that the second-order conditions for a maximum are met (see Ex. 8.15 for the explicit representation of the hessian matrix), and thus condition (4b) holds. Therefore, the estimator $\hat{\Theta}$, defined *implicitly* by the first-order conditions for maximizing the likelihood function, is consistent, that is, $\hat{\Theta} \xrightarrow{p} \Theta_o$, so that both $\hat{\alpha} \xrightarrow{p} \alpha_o$ and $\hat{\beta} \xrightarrow{p} \beta_o$. □

In the next example, we revisit Ex. 8.6, in which the MLE of $(\beta,\sigma^2)$ for the GLM was defined for the case where the random sample from a composite experiment had a joint density of the multivariate normal form, $N(\mathbf{x}\beta,\sigma^2\mathbf{I})$. In this case, the MLE can be explicitly solved for, and its consistency property evaluated directly, as we have done in Section 8.2, but for the sake of illustration we reexamine consistency using Theorem 8.16.

**Example 8.13** Consistency of MLE in GLM with $\mathbf{Y} \sim N(\mathbf{x}\beta,\sigma^2\mathbf{I})$

Reexamine Ex. 8.6. To demonstrate consistency of the MLE using Theorem 8.16, first note $f(\mathbf{y};\beta,\sigma^2) > 0$ for $\mathbf{y} \in R^n$, and so condition (1) is satisfied.

It is evident from Ex. 8.6 that $L(\beta,\sigma^2;\mathbf{y})$ is continuously differentiable with respect to $\Theta \in \Omega$ for all $\mathbf{y} \in R^n$, so that condition (2) is satisfied. Also, $\partial\ln L(\beta,\sigma^2;\mathbf{y})/\partial\Theta = [\mathbf{0}]$ has a unique solution (assuming $\mathbf{x}'\mathbf{x}$ is of full rank), and this solution defines the unique MLE of $(\beta,\sigma^2)$. We assume that $\sigma^2 \in [\sigma_\ell^2,\sigma_h^2]$ and $\beta \in \times_{i=1}^{k}[\beta_{i\ell},\beta_{ih}]$ and attempt to verify condition (4b). Note that

$$\mathbf{Z}_n = n^{-1}\ln L(\beta,\sigma^2;\mathbf{y}) = -(1/2)\ln(2\pi\sigma^2) - (1/2)n^{-1}[(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta)/\sigma^2].$$

Since $\mathbf{Y} = \mathbf{x}\beta_o + \varepsilon$, $\mathbf{Z}_n$ can be expressed equivalently as

$$\mathbf{Z}_n = -(1/2)\left[\ln(2\pi\sigma^2) + (\beta_o - \beta)'(n^{-1}\mathbf{x}'\mathbf{x})(\beta_o - \beta)/\sigma^2 + n^{-1}\varepsilon'\varepsilon/\sigma^2 + 2(\beta_o - \beta)'(n^{-1}\mathbf{x}'\varepsilon)/\sigma^2\right].$$

[37] This follows from Khinchin's WLLN upon recognizing that the bracketed expression represents E $\ln X_i$.

Since the $\varepsilon_i$'s are iid because $\mathbf{Y} \sim N(\mathbf{x}\beta_o, \sigma^2\mathbf{I})$, it follows from Khinchin's WLLN that plim $n^{-1}\varepsilon'\varepsilon = \sigma_0^2$. Also note that $\mathrm{E}n^{-1}\mathbf{x}'\varepsilon = [\mathbf{0}]$ and $\mathbf{Cov}(n^{-1}\mathbf{x}'\varepsilon) = \sigma^2 n^{-2}\mathbf{x}'\mathbf{x}$, so that if we *assume* that $\mathbf{x}'\mathbf{x}$ is $o(n^2)$, so that $n^{-2}\mathbf{x}'\mathbf{x} \to [\mathbf{0}]$ as $n \to \infty$, then $n^{-1}\mathbf{x}'\varepsilon \xrightarrow{m} [\mathbf{0}]$, and thus $n^{-1}\mathbf{x}'\varepsilon \xrightarrow{p} [\mathbf{0}]$ and $(\beta_o - \beta)'(n^{-1}\mathbf{x}'\varepsilon)/\sigma^2 \xrightarrow{p} 0$. Assuming further that $n^{-1}\mathbf{x}'\mathbf{x} \to \mathbf{Q}$, a positive definite symmetric, matrix, then

$$Z_n \xrightarrow{p} G(\beta, \sigma^2) = -\frac{1}{2}\left[\ln(2\pi\sigma^2) + (\beta_o - \beta)'\mathbf{Q}(\beta_o - \beta)/\sigma^2 + \sigma_o^2/\sigma^2\right]$$

uniformly in $\beta$ and $\sigma^2$.

The function $G(\beta, \sigma^2)$ is maximized when $\beta = \beta_o$ and $\sigma^2 = \sigma_o^2$. To see this, note that $\beta_o$ and $\sigma_o^2$ satisfy the first-order conditions for maximizing $G(\beta, \sigma^2)$ given by

$$\frac{\partial G(\beta, \sigma^2)}{\partial \beta} = \mathbf{Q}(\beta_o - \beta) = [\mathbf{0}],$$

$$\frac{\partial G(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{(\beta_o - \beta)'\mathbf{Q}(\beta_o - \beta)}{2\sigma^2} + \frac{\sigma_o^2}{2\sigma^2} = 0.$$

The hessian matrix for checking second-order conditions, evaluated at $\beta = \beta_o$ and $\sigma^2 = \sigma_o^2$, is given by

$$\left[\begin{array}{c|c} \dfrac{\partial^2 G(\beta_o, \sigma_o^2)}{\partial\beta\partial\beta'} & \dfrac{\partial^2 G(\beta_o, \sigma_o^2)}{\partial\beta\partial\sigma^2} \\ \hline \dfrac{\partial^2 G(\beta_o, \sigma_o^2)}{\partial\sigma^2\partial\beta'} & \dfrac{\partial^2 G(\beta_o, \sigma_o^2)}{\partial(\sigma^2)^2} \end{array}\right] = \left[\begin{array}{c|c} -\mathbf{Q} & [\mathbf{0}] \\ \hline [\mathbf{0}] & -\dfrac{1}{2\sigma_o^4} \end{array}\right],$$

which is negative definite, and thus $G(\beta, \sigma^2)$ is indeed maximized at $\beta_o, \sigma_o^2$. Therefore, by Theorem 8.16, and under the preceding assumptions on $\mathbf{x}$, $\hat{\Theta} \xrightarrow{p} \Theta_o$ so that plim $\hat{\beta} = \beta_o$ and plim $\hat{\sigma}^2 = \sigma_o^2$. □

**Asymptotic Normality and Asymptotic Efficiency* In order for the MLE to be asymptotically normally distributed, additional conditions on the maximum likelihood estimation problem are needed. We present a collection of sufficient conditions below and remind the reader a variety of alternative sufficient conditions exists in the literature (see R. H. Norden, *Likelihood Estimation*; also see T. Amemiya, *Advanced Econometrics*, pp. 111–112, for conditions related to those presented here).

**Theorem 8.17** (MLE Asymptotic Normality–Sufficient Conditions) *In addition to conditions (1)–(4) of Theorem 8.16, assume that*

1. $\partial^2 \ln L(\Theta;\mathbf{x})/\partial\Theta\partial\Theta'$ *exists and is continuous in* $\Theta$ $\forall$ $\Theta \in \Omega$ *and* $\forall$ $\mathbf{x} \in \Xi$,
2. $\text{plim}\,[n^{-1}(\partial^2 \ln L(\Theta^*;\mathbf{X})/\partial\Theta\partial\Theta')] = \mathbf{H}(\Theta_o)$ *is a nonsingular matrix for any sequence of random variables* $\{\Theta_n^*\}$ *such that* plim $\Theta_n^* = \Theta_o$,
3. $n^{-1/2}(\partial \ln L(\Theta_o;\mathbf{X}))/\partial\Theta \xrightarrow{d} N([\mathbf{0}],\mathbf{M}(\Theta_o))$ *where* $\mathbf{M}(\Theta_o)$ *is a symmetric, positive definite matrix.*

*Then the MLE, $\hat{\Theta}$, is such that*

$$n^{1/2}(\hat{\Theta} - \Theta_0) \xrightarrow{d} N([0], \mathbf{H}(\Theta_0)^{-1}\mathbf{M}(\Theta_0)\mathbf{H}(\Theta_0)^{-1}),$$
$$\hat{\Theta} \overset{a}{\sim} N(\Theta_0, n^{-1}\mathbf{H}(\Theta_0)^{-1}\mathbf{M}(\Theta_0)\mathbf{H}(\Theta_0)^{-1}).$$

***Proof** The first-order derivative function of ln $L(\hat{\Theta}; \mathbf{x})$ can be represented by a Taylor series expansion around $\Theta_0$ as[38]

$$\frac{\partial \ln L(\hat{\Theta}; \mathbf{X})}{\partial \Theta} = \frac{\partial \ln L(\Theta_0; \mathbf{X})}{\partial \Theta} + \frac{\partial^2 \ln L(\Theta^*; \mathbf{X})}{\partial \Theta \partial \Theta'}(\hat{\Theta} - \Theta),$$

where $\Theta^* = \lambda(\mathbf{X})\hat{\Theta} + (1 - \lambda(\mathbf{X}))\Theta_0$ for $\lambda(\mathbf{X}) \in [0, 1]$. Since $\hat{\Theta}$ is the ML estimator, the value of the first-derivative vector, $\partial \ln L(\hat{\Theta}; \mathbf{X})/\partial\Theta$, is $[0]$ by the first-order conditions for the maximum likelihood problem. Then, premultiplying the Taylor series expansion by $n^{-1/2}$ obtains

$$-n^{-1}\frac{\partial^2 \ln L(\hat{\Theta}^*; \mathbf{X})}{\partial \Theta \partial \Theta'} n^{1/2}(\hat{\Theta} - \Theta_0) = n^{-1/2}\frac{\partial \ln L(\Theta_0; \mathbf{X})}{\partial \Theta} = \mathbf{Z} \xrightarrow{d} N([0], \mathbf{M}(\Theta_0)),$$

where convergence to the $N([0], \mathbf{M}(\Theta_0))$ limiting density follows from assumption (3) of the theorem.

Now note that plim $\Theta^* = \Theta_0$ since, $\forall\, \varepsilon > 0$,

$$\begin{aligned}\lim_{n\to\infty} P(\|\Theta^* - \Theta_0\| < \varepsilon) &= \lim_{n\to\infty} P(\|\lambda(\mathbf{x})\hat{\Theta}(\mathbf{x}) + (1 - \lambda(\mathbf{x}))\Theta_0 - \Theta_0\| < \varepsilon)\\ &= \lim_{n\to\infty} P(\|\lambda(\mathbf{x})(\hat{\Theta}(\mathbf{x}) - \Theta_0)\| < \varepsilon)\\ &= \lim_{n\to\infty} P(\lambda(\mathbf{x})\|(\hat{\Theta}(\mathbf{x}) - \Theta_0)\| < \varepsilon),\end{aligned}$$

where the last equality follows from the fact that $\lambda(\mathbf{x}) \in [0, 1]$, and therefore $\lambda(\mathbf{x})$ is nonnegative, and where $\|\mathbf{z}\| = (\mathbf{z}'\mathbf{z})^{1/2}$ represents the distance of $\mathbf{z}$ from $[0]$. Noting that $\{\mathbf{x}: \|\hat{\Theta}(\mathbf{x}) - \Theta_0\| < \varepsilon\} \subset \{\mathbf{x}: \lambda(\mathbf{x})\|\hat{\Theta}(\mathbf{x}) - \Theta_0\| < \varepsilon\}$ (again since $\lambda(\mathbf{x}) \in [0, 1]$), it follows that $P(\|\hat{\Theta} - \Theta_0\| < \varepsilon) \leq P(\lambda(\mathbf{x})\|\hat{\Theta} - \Theta_0\| < \varepsilon)$, and since $\hat{\Theta} \xrightarrow{p} \Theta_0$, the left-hand side of the above inequality converges to 1 as $n \to \infty$, implying that the right-hand side converges to 1 as $n \to \infty$, $\forall\, \varepsilon > 0$. Therefore, $\lim_{n\to\infty} P(\|\Theta^* - \Theta_0\| < \varepsilon) = 1$, or plim $\Theta^* = \Theta_0$.

By assumption (2) of the theorem and by Slutsky's theorem, it follows from the preceding two results that

$$\left[\mathbf{H}(\Theta_0)^{-1}n^{-1}\left[\frac{\partial^2 \ln L(\Theta^*; \mathbf{X})}{\partial \Theta \partial \Theta'}\right]\right] n^{1/2}(\hat{\Theta} - \Theta_0) \xrightarrow{d} n^{1/2}(\hat{\Theta} - \Theta_0)$$
$$\xrightarrow{d} -\mathbf{H}(\Theta_0)^{-1}\mathbf{Z} \sim N([0], \mathbf{H}(\Theta_0)^{-1}\mathbf{M}(\Theta_0)\mathbf{H}(\Theta_0)^{-1})$$

[38] See R. G. Bartle (1976), *The Elements of Real Analysis*, 2nd Ed., New York: John Wiley, p. 371.

since the probability limit of the bracketed expression is equal to the identity matrix. The statement in the theorem concerning the asymptotic distribution of $\hat{\Theta}$ then follows immediately from Def. 5.10. ■

We present one further condition that leads to the asymptotic efficiency of the MLE.

**Theorem 8.18** (MLE Asymptotic Efficiency—Sufficient Conditions) *In addition to the assumptions in Theorems 8.16 and 8.17, assume that*

$$\mathbf{M}(\Theta_o) = \lim_{n\to\infty} n^{-1}\mathrm{E}\left[\frac{\partial \ln L(\Theta_o; \mathbf{X})}{\partial \Theta}\frac{\partial \ln L(\Theta_o; \mathbf{X})}{\partial \Theta'}\right] = -\mathbf{H}(\Theta_o).$$

*Then $\hat{\Theta}$ is asymptotically efficient.*

***Proof** From Theorem 8.17, if $\mathbf{M}(\Theta_o)$ and $\mathbf{H}(\Theta_o)$ are as defined above, then $n^{1/2}(\hat{\Theta} - \Theta_o) \xrightarrow{d} \mathbf{Z} \sim N([0], \mathbf{M}(\Theta_o)^{-1})$. Referring to the definition of asymptotic efficiency (Def. 7.22), it follows that[39]

$$n^{-1/2}\left[\mathrm{E}\left[\frac{\partial \ln L(\Theta_o; \mathbf{X})}{\partial \Theta}\frac{\partial \ln L(\Theta_o; \mathbf{X})}{\partial \Theta'}\right]\right]^{1/2} n^{1/2}(\hat{\Theta} - \Theta) \xrightarrow{d} \mathbf{M}(\Theta_o)^{1/2}\mathbf{Z} \sim N([0], \mathbf{I}),$$

so that $\hat{\Theta} \overset{a}{\sim} N(\Theta_o, n^{-1}\mathbf{M}(\Theta_o)^{-1})$ is asymptotically efficient. ■

In applications, since the value of $\Theta_o$ is unknown, the asymptotic covariance of $\hat{\Theta}$ is also unknown. However, under the assumptions of Theorems 8.17 and 8.18, a consistent estimate of $\mathbf{M}(\Theta_o)^{-1}$ is obtained by using an outcome of

$$-[n^{-1}(\partial^2 \ln L(\hat{\Theta}; \mathbf{X})/\partial\Theta\partial\Theta')]^{-1}$$

(recall assumption (2) of Theorem 8.17 and the statement of Theorem 8.18).

We now revisit Ex. 8.11 and Ex. 8.12 and illustrate the use of Theorem 8.17 and Theorem 8.18 for establishing the asymptotic normality and asymptotic efficiency of the MLE.

**Example 8.14** Reexamine Ex. 8.11. Note that $\partial^2 \ln L(\theta; \mathbf{x})/\partial\theta^2 = (n/\theta^2) - (2\sum_{i=1}^n x_i/\theta^3)$ exists and is continuous for $\forall\, \theta > 0$ and $\forall\, \mathbf{x} \in \Xi$, so assumption (1) of Theorem 8.17 is satisfied. Furthermore, assuming plim $\Theta_* = \theta_o$, it follows by Slutsky's theorem

[39] Recall that the matrix square root is a *continuous* function of its arguments, so that

$$\operatorname{plim} \mathbf{A}_n^{1/2} = (\operatorname{plim} \mathbf{A}_n)^{1/2}.$$

Letting

$$\mathbf{A}_n = n^{-1}\mathrm{E}\left[\frac{\partial \ln L(\Theta_o; \mathbf{X})}{\partial \Theta}\frac{\partial \ln L(\Theta_o; \mathbf{X})}{\partial \Theta'}\right]$$

leads to $\operatorname{plim} \mathbf{A}_n^{1/2} = \mathbf{M}(\Theta_o)^{1/2}$.

that

$$\operatorname{plim} n^{-1}\frac{\partial^2 \ln L(\Theta_*;\mathbf{X})}{\partial\theta^2} = \operatorname{plim}(\Theta_*^{-2} - 2\bar{X}_n/\Theta_*^3) = (\theta_o^{-2} - 2\theta_o/\theta_o^3) = -\theta_o^{-2} = \mathbf{H}(\theta_o)$$

so that assumption (2) of Theorem 8.17 applies. Also note that

$$\begin{aligned} n^{-1/2}\frac{\partial \ln L(\theta_o;\mathbf{X})}{\partial\theta} &= -n^{1/2}\theta_o^{-1} + n^{-1/2}\left(\sum_{i=1}^{n}\frac{\mathbf{X}_i}{\theta_o^2}\right) \\ &= \frac{1}{\theta_o}\left[\frac{n^{1/2}(\bar{X}_n - \theta_o)}{\theta_o}\right] \xrightarrow{\mathrm{d}} N\left(0,\theta_o^{-2}\right) = N(0,\mathbf{M}(\theta_o)), \end{aligned}$$

where $\mathbf{M}(\theta_o) = \theta_o^{-2}$, which follows by a direct application of the Lindberg-Levy CLT to the bracketed expression and by Slutsky's theorem, since $\mathrm{E}X_i = \theta_o$ and $\operatorname{var}(X_i) = \theta_o^2\ \forall\ i$. Then, from Theorem 8.17, it follows that $n^{1/2}(\hat{\Theta} - \theta_o) \xrightarrow{\mathrm{d}} N(0,\theta_o^2)$, and $\hat{\Theta} \overset{\mathrm{a}}{\sim} N(\theta_o, n^{-1}\theta_o^2)$.

Regarding asymptotic efficiency, recall from Ex. 7.20 that

$$\mathrm{E}\left[\frac{\partial \ln L(\theta_o;\mathbf{X})}{\partial\theta}\frac{\partial \ln L(\theta_o;\mathbf{X})}{\partial\theta'}\right] = n/\theta_o^2.$$

Asymptotic efficiency of $\hat{\Theta}$ follows immediately from Theorem 8.18 since the above results demonstrate the equality

$$\mathbf{M}(\theta_o) = -\mathbf{H}(\theta_o) = \lim_{n\to\infty}\mathrm{E}n^{-1}\left[\frac{\partial \ln L(\theta_o;\mathbf{X})}{\partial\theta}\frac{\partial \ln L(\theta_o;\mathbf{X})}{\partial\theta'}\right] = \theta_o^{-2}. \qquad \square$$

**Example 8.15** Asymptotic Normality and Efficiency of MLE—Gamma Case

Reexamine Ex. 8.12. The second-order derivatives of $\ln L(\alpha,\beta;\mathbf{x})$, divided by $n$, are given by

$$\frac{1}{n}\frac{\partial^2 \ln L(\alpha,\beta;\mathbf{x})}{\partial\alpha^2} = -\left[\Gamma(\alpha)\frac{d^2\Gamma(\alpha)}{d\alpha^2} - \left(\frac{d\Gamma(\alpha)}{d\alpha}\right)^2\right]/\Gamma(\alpha)^2,$$

$$\frac{1}{n}\frac{\partial^2 \ln L(\alpha,\beta;\mathbf{x})}{\partial\alpha\partial\beta} = \frac{1}{n}\frac{\partial^2 \ln L(\alpha,\beta;\mathbf{x})}{\partial\beta\partial\alpha} = -\beta^{-1},$$

$$\frac{1}{n}\frac{\partial^2 \ln L(\alpha,\beta;\mathbf{x})}{\partial\beta^2} = \frac{\alpha}{\beta^2} - 2\frac{\bar{x}}{\beta^3},$$

and the derivatives themselves are continuous functions of $(\alpha,\beta)$ for $\alpha > 0$ and $\beta > 0$ and $\forall\,\mathbf{x} \in \Xi$, so that assumption (1) of Theorem 8.17 is satisfied.[40] Letting $\Theta = \left(\begin{smallmatrix}\alpha\\ \beta\end{smallmatrix}\right)$, $\Theta_o = \left(\begin{smallmatrix}\alpha_o\\ \beta_o\end{smallmatrix}\right)$, and $\Theta_*$ be such that $\Theta_* \xrightarrow{\mathrm{p}} \Theta_o$, it follows from continuity

[40] The Gamma function, $\Gamma(\alpha)$, is continuous in $\alpha$, and its first two derivatives are continuous in $\alpha$, for $\alpha > 0$. $\Gamma(\alpha)$ is in fact strictly convex, with its second-order derivative strictly positive for $\alpha > 0$. See Bartle, *Real Analysis*, p. 282.

that

$$\operatorname{plim} n^{-1}\frac{\partial^2 \ln L(\Theta_*;\mathbf{X})}{\partial\Theta\partial\Theta'} = -\left[\begin{array}{c|c} \left[\Gamma(\alpha_o)\dfrac{d^2\Gamma(\alpha_o)}{d\alpha^2} - \left(\dfrac{d\Gamma(\alpha_o)}{d\alpha}\right)^2\right]/\Gamma(\alpha_o)^2 & \beta_o^{-1} \\ \hline \beta_o^{-1} & \dfrac{\alpha_o}{\beta_o^2} \end{array}\right] = \mathbf{H}(\Theta_o)$$

(recall that $\operatorname{plim}\bar{X} = \alpha\beta$ in this case), so that the convergence condition in (2) of Theorem 8.17 applies.

Regarding condition (3), note that

$$n^{-1/2}\frac{\partial \ln L(\Theta_o;\mathbf{X})}{\partial\Theta} = \begin{bmatrix} -n^{1/2}\left[\ln\beta_o + \left[\dfrac{d\Gamma(\alpha_o)}{d\alpha}\right]/\Gamma(\alpha_o)\right] + n^{-1/2}\displaystyle\sum_{i=1}^{n}\ln X_i \\ -n^{1/2}\left[\dfrac{\alpha_o}{\beta_o}\right] + n^{1/2}\dfrac{\bar{X}_n}{\beta_o^2} \end{bmatrix}.$$

To establish the bivariate normal limiting density of this (2 × 1) random vector, first note that $(\mathbf{Z}_1,\ldots,\mathbf{Z}_n)$, where $\mathbf{Z}_i = \left[\begin{smallmatrix}\ln X_i \\ X_i/\beta_o^2\end{smallmatrix}\right]$, is a collection of iid random variables with[41]

$$\mathrm{E}\mathbf{Z}_i = \boldsymbol{\mu} = \begin{bmatrix} \dfrac{d\Gamma(\alpha_o)}{d\alpha}/\Gamma(\alpha_o) + \ln\beta_o \\ \alpha_o/\beta_o \end{bmatrix},$$

$$\mathbf{Cov}(\mathbf{Z}_i) = \boldsymbol{\Phi} = \left[\begin{array}{c|c} \left[\Gamma(\alpha_o)\dfrac{d^2\Gamma(\alpha_o)}{d\alpha^2} - \left(\dfrac{d\Gamma(\alpha_o)}{d\alpha}\right)^2\right]/\Gamma(\alpha_o)^2 & \beta_o^{-1} \\ \hline \beta_o^{-1} & \dfrac{\alpha_o}{\beta_o^2} \end{array}\right].$$

The multivariate Lindberg-Levy CLT then implies that

$$n^{1/2}(\bar{\mathbf{Z}} - \boldsymbol{\mu}) \xrightarrow{d} N([\mathbf{0}], \boldsymbol{\Phi}), \quad \text{where } \bar{\mathbf{Z}} = \begin{bmatrix} n^{-1}\displaystyle\sum_{i=1}^{n}\ln X_i \\ \bar{X}/\beta_o^2 \end{bmatrix}.$$

Noting that $n^{-1/2}(\partial\ln L(\Theta_o;\mathbf{X})/\partial\Theta) = n^{1/2}(\bar{\mathbf{Z}} - \boldsymbol{\mu}) \xrightarrow{d} N([\mathbf{0}], \mathbf{M}(\Theta_o))$, where $\mathbf{M}(\Theta_o) = \boldsymbol{\Phi}$, it follows that convergence to a bivariate normal density is es-

[41] A way of deriving the expectations involving $\ln X_i$ that is conceptually straightforward, albeit somewhat tedious algebraically, is first to derive the MGF of $\ln X_i$, which is given by $\beta^t\Gamma(\alpha+t)/\Gamma(\alpha)$. Then, using the MGF in the usual way establishes the mean and variance of $\ln X_i$. The covariance between $(X_i/\beta_o^2)$ and $\ln X_i$ can be established by noting that

$$\begin{aligned}\beta_o^{-2}\mathrm{E}X_i\ln X_i &= \beta_o^{-2}\left[\frac{1}{\beta_o^{\alpha_o}\Gamma(\alpha_o)}\right]\int_o^\infty (\ln x_i)x_i^{\alpha_o}e^{-x_i/\beta_o}\,dx_i \\ &= \alpha_o\beta_o^{-1}\mathrm{E}_*\ln X_i,\end{aligned}$$

where $E_*$ denotes an expectation of $\ln X_i$ using a Gamma density having parameter values $\alpha_o + 1$ and $\beta_o$. Then $\operatorname{cov}((X_i/\beta_o^2), \ln X_i)$ is equal to $\alpha_o\beta_o^{-1}E_*\ln X_i - (\mathrm{E}\ln X_i\mathrm{E}(X_i/\beta_o^2)) = \alpha_o\beta_o^{-1}(\mathrm{E}_*\ln X_i - \mathrm{E}\ln X_i) = \beta_o^{-1}$.

tablished, and thus condition (3) of Theorem 8.17 is met. Finally, note that

$$\mathbf{M}(\Theta_o) = \lim_{n\to\infty} n^{-1}\mathrm{E}\left[\frac{\partial \ln L(\Theta_o; \mathbf{X})}{\partial\Theta}\frac{\partial \ln L(\Theta_o; \mathbf{X})}{\partial\Theta'}\right] = -\mathbf{H}(\Theta_o),$$

which follows immediately upon taking the expectation and recalling that $\mathrm{E}\bar{X} = \alpha\beta$. Thus, the condition of Theorem 8.18 is met, and we can conclude that

$$n^{1/2}(\hat{\Theta} - \Theta_o) \xrightarrow{d} N([0], \mathbf{M}(\Theta_o)^{-1}), \hat{\Theta} \overset{a}{\sim} N(\Theta_o, n^{-1}\mathbf{M}(\Theta_o)^{-1}),$$

and $\hat{\Theta}$ is asymptotically efficient. □

We pause to emphasize that although we were unable to explicitly solve for the function defining the MLE in the case of random sampling from a general Gamma density, our theorems on the asymptotic properties of the MLE still allowed us to establish the asymptotic properties of the MLE which were *implicitly* defined by first-order conditions. In particular, we now know that $\hat{\Theta}$ is consistent, asymptotically normal, and asymptotically efficient as an estimator of $(\alpha, \beta)$.

The verification of the MLE's asymptotic properties can be somewhat complicated at times. Regarding the density functions commonly used in practice, it can be said that the MLE is quite generally consistent, asymptotically normal, and asymptotically efficient when the random sample is iid from a population distribution. In particular, the MLE is consistent, asymptotically normal, and asymptotically efficient when random sampling from *any* of the exponential class densities we have presented in Chapter 4, and these asymptotic properties hold quite generally for exponential class densities (see E. Lehmann, *Point Estimation*, pp. 417–418 and 438–439). In more general situations involving composite experiments, one must generally verify the asymptotic properties of the MLE on a case-by-case basis.

### MLE Invariance Principle

Our discussion of MLEs to this point has concentrated on the estimation of the parameter vector $\Theta$ itself. Fortunately, the maximum likelihood procedure has an *invariance property* that renders our preceding discussion entirely and directly relevant for estimating *functions* of the parameter vector. Before formally stating the property, we provide some additional background discussion concerning the *parameterization* of density functions.

The reader will recall that the members of parametric families of densities are indexed, or identified, by values of a parameter vector whose admissible values are identified by a set of values represented by the parameter space, $\Omega$. After a moment's reflection, it is evident that nothing is sacrosanct about a particular indexing of the members of a parametric family of densities. That is, a given collection of density functions $f(\mathbf{x}; \Theta)$, for $\Theta \in \Omega$, could be equally well represented by $f(\mathbf{x}; \mathbf{q}^{-1}(\boldsymbol{\xi}))$, where $\boldsymbol{\xi} = \mathbf{q}(\Theta)$ is any *invertible* function of $\Theta \in \Omega$. Such a parameter transformation is referred to as a **reparameterization of the**

**density family**. As a specific example, consider the exponential family

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta} I_{(0,\infty)}(x) \quad \text{for } \theta \in \Omega,$$

where $\Omega = (0, \infty)$. This same family of densities could be represented equally well by

$$h(x; \xi) = f(x; q_1^{-1}(\xi)) = \xi e^{-\xi x} I_{(0,\infty)}(x) \quad \text{for } \xi \in \Omega_\xi,$$

where $\xi = q_1(\theta) = \theta^{-1}$ for $\theta \in \Omega$, so that $\Omega_\xi = (0, \infty)$. Another parameterization of this same family of densities could be given as

$$m(x; \tau) = f(x; q_2^{-1}(\tau)) = \tau^{-1/2} \exp(-x\tau^{-1/2}) I_{(0,\infty)}(x) \quad \text{for } \tau \in \Omega_\tau,$$

where now $\tau = q_2(\theta) = \theta^2$ for $\theta \in \Omega$, so that $\Omega_\tau = (0, \infty)$. In all cases, the exact same collection of densities is identified. Of course, the interpretation of the parameter differs in each case, where $\theta$ is the mean of $X$, $\xi$ is the reciprocal of the mean of $X$, and $\tau$ is the variance of $X$ (recall $\sigma^2 = \theta^2$ in the exponential family). The fact that a density family can be **reparameterized** provides the researcher with flexibility for redefining the maximum likelihood problem in terms of a parameter or parameter vector having an interpretation that may be of more fundamental interest to research objectives. The fact that the actual reparameterization process is *unnecessary* for obtaining the maximum likelihood estimate of $\mathbf{q}(\Theta)$ is presented in the following theorem.

**Theorem 8.19** (MLE Invariance Principle—Invertible Case) *Let $\hat{\Theta}$ be a MLE of $\Theta$, and suppose that the function $\mathbf{q}(\Theta)$, for $\Theta \in \Omega$, is invertible. Then $\mathbf{q}(\hat{\Theta})$ is a MLE of $\mathbf{q}(\Theta)$.*

**Proof** The maximum likelihood estimate of $\Theta$ is given by $\hat{\theta} = \arg\max_{\Theta \in \Omega} L(\Theta; \mathbf{x})$. Suppose $\boldsymbol{\xi} = \mathbf{q}(\Theta)$ is invertible, so the likelihood function can be reexpressed as a function of $\boldsymbol{\xi}$ by substituting $\Theta = \mathbf{q}^{-1}(\boldsymbol{\xi}), \boldsymbol{\xi} \in \Omega_{\boldsymbol{\xi}} = \{\boldsymbol{\xi} : \boldsymbol{\xi} = \mathbf{q}(\Theta), \Theta \in \Omega\}$, in $L(\Theta; \mathbf{x})$. Then the maximum likelihood estimate for $\boldsymbol{\xi}$ would be $\hat{\boldsymbol{\xi}} = \arg\max_{\boldsymbol{\xi} \in \Omega_{\boldsymbol{\xi}}} L(\mathbf{q}^{-1}(\boldsymbol{\xi}); \mathbf{x})$. But since $\hat{\theta}$ maximizes $L(\Theta; \mathbf{x})$, a value of $\hat{\boldsymbol{\xi}}$ that solves the latter likelihood maximization problem satisfies $\mathbf{q}^{-1}(\hat{\boldsymbol{\xi}}) = \hat{\theta}$, which implies by the invertibility of $\mathbf{q}$ that $\hat{\boldsymbol{\xi}} = \mathbf{q}(\hat{\theta})$. ■

The invariance principle in Theorem 8.19 implies that once a MLE, $\hat{\Theta}$, of $\Theta$ has been found, one can immediately define the MLE of *any* invertible function of $\Theta$, say $\mathbf{q}(\Theta)$, as $\mathbf{q}(\hat{\Theta})$. This is obviously a very convenient property of the maximum likelihood estimation approach. In fact, the invariance principle of maximum likelihood estimation extends to more general functions of $\Theta$, invertible or not, although the intuitive interpretation of invariance in terms of reparameterizing a family of density functions no longer follows. The more general invariance principle utilizes the notion of maximizing an *induced likelihood function*, as defined in the proof of the following generalization of Theorem 8.19.

**Theorem 8.20** (MLE Invariance Principle—General Case) *Let $\hat{\Theta}$ be a MLE of the $(k \times 1)$ parameter vector $\Theta$, and let $\mathbf{q}(\Theta)$ be an $(r \times 1)$ real-valued vector function of $\Theta \in \Omega$, where $r \leq k$. Then $\mathbf{q}(\hat{\Theta})$ is a MLE of $\mathbf{q}(\Theta)$.*

**Proof** Let $\Upsilon(\tau) = \{\Theta : \mathbf{q}(\Theta) = \tau, \Theta \in \Omega\}\ \forall\, \tau \in R(\mathbf{q})$, i.e., $\Upsilon(\tau)$ is the set of $\Theta$ values having the image value $\tau$ based on the function $\mathbf{q}$. The collection of $\tau$ values represented by the range of the function, $R(\mathbf{q})$, will be called the **parameter space induced by $\mathbf{q}(\Theta)$**, and note that since there can be more than one $\Theta \in \Omega$ that satisfies $\mathbf{q}(\Theta) = \tau$, then there can be more than one density and likelihood function, $f(\mathbf{x}; \Theta) \equiv L(\Theta; \mathbf{x})$, identified by the $\Theta$ values that satisfy $\mathbf{q}(\Theta) = \tau$. Define the **likelihood function induced by $\mathbf{q}(\Theta)$** as

$$L^*(\tau; \mathbf{x}) = \underset{\Theta \in \Upsilon(\tau)}{\text{Max}}\ L(\Theta; \mathbf{x}) = \underset{\Theta \in \Upsilon(\tau)}{\text{Max}}\ f(\mathbf{x}; \Theta),$$

i.e., $L^*(\tau; \mathbf{x})$ is the largest likelihood consistent with $\mathbf{q}(\Theta) = \tau$, or equivalently, $L^*(\tau; \mathbf{x})$ is the highest density weighting assigned to the random-sample outcome, $\mathbf{x}$, by the family of densities $f(\mathbf{x}; \Theta)$, $\Theta \in \Omega$, given that $\mathbf{q}(\Theta) = \tau$. A maximum likelihood estimate of $\tau = \mathbf{q}(\Theta)$ is then given by a value of $\tau \in R(\mathbf{q})$ that maximizes the induced likelihood function, as

$$\hat{\tau} = \underset{\tau \in R(\mathbf{q})}{\arg\max}\, L^*(\tau; \mathbf{x}) = \underset{\tau \in R(\mathbf{q})}{\arg\max} \left[ \max_{\Theta \in \Upsilon(\tau)} L(\Theta; \mathbf{x}) \right] = \underset{\tau \in R(\mathbf{q})}{\arg\max} \left[ \max_{\Theta \in \Upsilon(\tau)} f(\mathbf{x}; \Theta) \right].$$

(Note that a maximizing value of $\tau$ then maintains the intuitively appealing property of associating with the sample outcome, $\mathbf{x}$, the highest probability (discrete case) or highest density weighting (continuous case) that is possible within the parametric family $f(\mathbf{x}; \Theta)$, $\Theta \in \Omega$).

Since $\hat{\theta}$ maximizes $L(\Theta; \mathbf{x}) \equiv f(\mathbf{x}; \Theta)$, it follows that $\hat{\tau} = \mathbf{q}(\hat{\theta})$ maximizes the induced likelihood function $L^*(\tau; \mathbf{x})$, so that $\mathbf{q}(\hat{\Theta})$ is the MLE of $\mathbf{q}(\Theta)$. (See P. W. Zehna (1966), "Invariance of maximum likelihood estimation." *Ann. Math. Stat.*, 37, p. 755, for further discussion of the induced likelihood function concept.) ■

The following examples illustrate both the versatility and the simplicity of Theorem's 8.19 and 8.20.

**Example 8.16** Let $(X_1, \ldots, X_n)$ be a random sample from the Bernoulli population distribution $f(z; p) = p^z(1-p)^{1-z} I_{\{0,1\}}(z)$ representing whether or not a customer contact results in a sale. The MLE of $p$ is given by $\hat{P} = \bar{X}$. Then by the invariance principle, the MLE of the variance of the Bernoulli population distribution, $\sigma^2 = q(p) = p(1-p)$, is given by $\hat{\sigma}^2 = \bar{X}(1 - \bar{X})$. □

**Example 8.17** Reexamine the case of maximum likelihood estimation in the GLM (Ex. 8.6), where the MLE of $\beta$ and $\sigma^2$ is given by $\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ and $\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{x}\hat{\beta})'(\mathbf{Y} - \mathbf{x}\hat{\beta})/n$. Then by the invariance principle, the MLE of $\sum_{i \in I} \beta_i$ is given by $\sum_{i \in I} \hat{\beta}_i$, the MLE of $\beta_1/\beta_2$ is given by $\hat{\beta}_1/\hat{\beta}_2$, and the MLE of $\sigma$ is given by $(\hat{\sigma}^2)^{1/2}$. □

While the invariance principle allows convenient and straightforward definitions of MLEs for functions of $\Theta$ once an MLE for $\Theta$ has been defined, the invariance principle does *not* imply that the estimator properties attributed to $\hat{\Theta}$ transfer to the estimator $\mathbf{q}(\hat{\Theta})$ of $\mathbf{q}(\Theta)$. That is, the fact that $\hat{\Theta}$ may be unbiased, MVUE, consistent, asymptotically normal, and/or asymptotically efficient does not necessarily imply that the MLE $\mathbf{q}(\hat{\Theta})$ possesses any of the same properties. The small sample properties of $\mathbf{q}(\hat{\Theta})$ as an estimator of $\mathbf{q}(\Theta)$ would need to be checked on a case-by-case basis. Useful results are not available that delineate general problem conditions under which small sample properties of $\hat{\Theta}$ are transferred to a general function of $\hat{\Theta}$, $\mathbf{q}(\hat{\Theta})$, used to estimate $\mathbf{q}(\Theta)$. However, asymptotic properties of $\hat{\Theta}$ transfer to $\mathbf{q}(\hat{\Theta})$ under fairly general conditions, as the following two theorems indicate.

**Theorem 8.21** *Let $\hat{\Theta}$ be a consistent MLE of $\Theta$, and let $\mathbf{q}(\Theta)$ be a continuous function of $\Theta \in \Omega$. Then the MLE $\mathbf{q}(\hat{\Theta})$ is a consistent estimator of $\mathbf{q}(\Theta)$.*

**Proof** This follows directly from Theorem 5.5, since if $\mathbf{q}(\Theta)$ is a continuous function of $\Theta$, and if $\operatorname{plim} \hat{\Theta} = \Theta$, then $\operatorname{plim} \mathbf{q}(\hat{\Theta}) = \mathbf{q}(\operatorname{plim} \hat{\Theta}) = \mathbf{q}(\Theta)$. ■

**Theorem 8.22** *Let $\hat{\Theta}$ be a consistent, asymptotically normal, and asymptotically efficient MLE of $\Theta$, satisfying the conditions of Theorems 8.17 and 8.18 so that $n^{1/2}(\hat{\Theta} - \Theta_o) \xrightarrow{d} N([0], \mathbf{M}(\Theta_o)^{-1})$. Let $\mathbf{q}(\Theta)$ be continuously differentiable with respect to $\Theta$, and let $\partial \mathbf{q}(\Theta_o)/\partial\Theta'$ have full row rank. Then $\mathbf{q}(\hat{\Theta})$ is a consistent, asymptotically normal, and asymptotically efficient MLE of $\mathbf{q}(\Theta)$.*

**Proof** It follows directly from Theorem 5.40 and Slutsky's theorem that

$$\left[\frac{\partial \mathbf{q}(\Theta_o)'}{\partial\Theta} \mathbf{M}(\Theta_o)^{-1} \frac{\partial \mathbf{q}(\Theta_o)}{\partial\Theta}\right]^{-1/2} n^{1/2}(\mathbf{q}(\hat{\Theta}) - \mathbf{q}(\Theta_o)) \xrightarrow{d} N([0], \mathbf{I}).$$

Using Theorems 8.17 and 8.18 and Slutsky's theorems, it then also follows that

$$\left[\frac{\partial \mathbf{q}(\Theta_o)'}{\partial\Theta} \left[\mathrm{E}\frac{\partial \ln L(\Theta_o;\mathbf{X})}{\partial\Theta}\frac{\partial \ln L(\Theta_o;\mathbf{X})}{\partial\Theta'}\right]^{-1} \frac{\partial \mathbf{q}(\Theta_o)}{\partial\Theta}\right]^{-1/2} \left(\mathbf{q}(\hat{\Theta}) - \mathbf{q}(\Theta_o)\right) \xrightarrow{d} N([0], \mathbf{I}),$$

which by Def. 7.22 justifies the conclusion of the theorem. ■

**Example 8.18** Recall Ex. 8.16. The reader can demonstrate that $\hat{P} = \bar{X}$ is consistent for $p$, $n^{1/2}(\hat{P} - p) \xrightarrow{d} N(0, p(1-p))$, and $\hat{P}$ is asymptotically efficient as an estimator of $p$. Examine the MLE of $\sigma^2 = q(p) = p(1-p)$, which by the invariance principle is given by $\hat{\sigma}^2 = \bar{X}(1 - \bar{X})$.

Since $p(1-p)$ is continuous for $p \in [0, 1]$, $\hat{\sigma}^2$ is consistent for $p(1-p)$ by Theorem 8.21. Now note that $dq(p)/dp = 1 - 2p$ is continuous and $\neq 0$ (and thus "full rank") provided $p \neq .5$. Then $\bar{X}(1 - \bar{X})$ is asymptotically normal and asymptotically efficient by Theorem 8.22, where

$$\hat{\sigma}^2 \overset{a}{\sim} N(\sigma^2, (1 - 2p)^2 p(1 - p)/n).$$

When $p = .5$, the asymptotic distribution of $\hat{\sigma}^2$ can be shown to be equal to the distribution of the random variable $.25(1 - Y/n)$, where $Y \sim \chi_1^2$. See Bickel and Doksum, *Mathematical Statistics*, p. 53. □

**Example 8.19** Recall Ex. 8.17. Assuming $n^{-1}\mathbf{x}'\mathbf{x} \rightarrow \mathbf{Q}$, a symmetric positive definite matrix, it can be shown that the MLE $\left(\begin{smallmatrix}\hat{\beta}\\ \hat{\sigma}^2\end{smallmatrix}\right)$ is consistent, asymptotically normal, and asymptotically efficient as an estimator of $\left[\begin{smallmatrix}\beta\\ \sigma^2\end{smallmatrix}\right]$, where

$$n^{1/2}\left(\begin{bmatrix}\hat{\beta}\\ \hat{\sigma}^2\end{bmatrix} - \begin{bmatrix}\beta_o\\ \sigma_o^2\end{bmatrix}\right) \xrightarrow{d} N\left([0], \left[\begin{array}{c|c}\sigma^2\mathbf{Q}^{-1} & [0]\\ \hline [0] & 2\sigma^4\end{array}\right]\right).$$

Examine the MLE $q_1(\hat{\beta}, \hat{\sigma}^2) = \sum_{i=1}^k \hat{\beta}_i$ of $q_1(\beta, \sigma^2) = \sum_{i=1}^k \beta_i$. Since $q_1(\beta, \sigma^2)$ is continuous in $\beta$ and $\sigma^2$, $q_1(\hat{\beta}, \hat{\sigma}^2)$ is a consistent estimator of $q_1(\beta, \sigma^2)$ by Theorem 8.21. Also, since $\partial q_1/\partial\Theta = [1\ 1\ \ldots\ 1\ 0]' \neq [\mathbf{0}]$ (where $\Theta = \left[\begin{smallmatrix}\beta\\ \sigma^2\end{smallmatrix}\right]$), $q_1(\hat{\beta}, \hat{\sigma}^2)$ is asymptotically normal and asymptotically efficient by Theorem 8.22, where $\sum_{i=1}^k \hat{\beta}_i \overset{a}{\sim} N(\iota'\beta_o, \iota'\sigma^2(\mathbf{x}'\mathbf{x})^{-1}\iota)$ with $\iota$ a $(k \times 1)$ vector of ones.

Now examine the MLE $q_2(\hat{\beta}, \hat{\sigma}^2) = (\hat{\sigma}^2)^{1/2} = \hat{\sigma}$ of $q_2(\beta, \sigma^2) = (\sigma^2)^{1/2} = \sigma$. Since $q_2(\beta, \sigma^2)$ is continuous in $\beta$ and $\sigma^2 > 0, \hat{\sigma}$ is a consistent estimator of $\sigma$ by Theorem 8.21. Also, since $\partial q_2/\partial\Theta = [0\ \ldots\ 0\ ((1/2)(\sigma^2)^{-1/2})]' \neq [\mathbf{0}]$ and the derivative is continuous for $\sigma^2 > 0, \hat{\sigma}$ is asymptotically normal and asymptotically efficient by Theorem 8.21, where $\hat{\sigma} \overset{a}{\sim} N(\sigma, \sigma^2/2n)$. □

**Example 8.20** Recall Ex. 8.4. We know that the MLE of $\theta$ in the exponential population distribution is given by $\hat{\Theta} = \bar{X}$, and the MLE is consistent, asymptotically normal, and asymptotically efficient, where $n^{1/2}(\hat{\Theta} - \theta) \xrightarrow{d} N(0, \theta^2)$. Examine the MLE of the variance of the exponential population distribution, $q(\theta) = \theta^2$. By the invariance principle, the MLE is given by $q(\hat{\Theta}) = \bar{X}^2$. Since $q(\theta)$ is continuous in $\theta$, $\bar{X}^2$ is consistent for $\Theta^2$ by Theorem 8.21. Since $\partial q(\theta)/\partial\theta = 2\theta \neq 0$ and the derivative is continuous $\forall\ \theta > 0$, $q(\hat{\Theta})$ is asymptotically normal and asymptotically efficient by Theorem 8.22, where $\bar{X}^2 \overset{a}{\sim} N(\theta^2, 4\theta^4/n)$. □

### MLE Property Summary

Overall, the maximum likelihood procedure is a focused and relatively straightforward approach to defining point estimators for $\Theta$ and for a myriad of functions of $\Theta$. The procedure can sometimes lead to a MVUE for $\Theta$ or $\mathbf{q}(\Theta)$, and very often in practice the MLE will possess useful asymptotic properties. Specifically, properties of a MLE that we have examined in this section include the following:

1. A MLE is not necessarily unbiased.
2. If an unbiased estimator of $\Theta$ that achieves the CRLB exists, and if the MLE, $\hat{\Theta}$, is defined by solving first-order conditions, then the MLE will be unique and achieve the CRLB.

3. If a MLE is unique, then the MLE is a function of any set of sufficient statistics, including complete sufficient statistics if they exist.
4. If a MLE of $\Theta$ is unique, and complete sufficient statistics exist, then

   a. if the MLE is unbiased for $\Theta$, it is MVUE for $\Theta$;
   b. if a function of the MLE, $\mathbf{q}(\hat{\Theta})$, is unbiased for $\Theta$, then this function is the MVUE for $\Theta$.

5. Under general regularity conditions on the estimation problem (see Theorems 8.14–8.18), the MLE is consistent, asymptotically normal, and asymptotically efficient.
6. If $\hat{\Theta}$ is the MLE of the $(k \times 1)$ parameter vector $\Theta$, then $\mathbf{q}(\hat{\Theta})$ is the MLE of the $(r \times 1)$ vector $\mathbf{q}(\Theta)$, $r \leq k$ (MLE invariance principle).
7. Under general regularity conditions (see Theorems 8.21 and 8.22), if the MLE $\hat{\Theta}$ is consistent, asymptotically normal, and asymptotically efficient for estimating $\Theta$, then the MLE $\mathbf{q}(\hat{\Theta})$ is consistent, asymptotically normal, and asymptotically efficient for estimating $\mathbf{q}(\Theta)$.

## 8.4 The Method of Moments

The method-of-moments (MOM) approach to point estimation begins with the specification of *moment conditions* that must hold for each of the random variables in the random sample $(Y_1, \ldots, Y_n)$, and for the true value of the parameter vector, relating to the statistical model $\{f(\mathbf{y}; \Theta), \Theta \in \Omega\}$. The moment conditions take the general form $\mathrm{E}\mathbf{g}_t(Y_t, \Theta_o) = [\mathbf{0}]$, $t = 1, \ldots, n$, with the dimension of the vector function $\mathbf{g}_t$ being at least as large as the number of elements in the $\Theta_o$ vector. The $Y_t$'s can be either scalar or vector random variables, although in this section we will deal exclusively with the scalar case. An estimate, $\hat{\theta}$, of the parameters of the statistical model is then obtained by choosing the value of $\Theta$ that satisfies, or satisfies as closely as possible, the sample moment counterpart to the moment conditions, $n^{-1} \sum_{t=1}^{n} \mathbf{g}_t(y_t, \Theta) = [\mathbf{0}]$. When the dimension of $\mathbf{g}_t$ and that of $\Theta$ are the same, a MOM estimate of $\Theta_o$ would be defined as $\hat{\theta} = \arg_{\Theta}[n^{-1} \sum_{t=1}^{n} \mathbf{g}_t(y_t, \Theta) = [\mathbf{0}]]$.

For a concrete illustration of the MOM approach, consider a statistical model for a random sample of size $n$ from a Bernoulli population distribution, i.e., $\{f(\mathbf{y}; p), p \in \Omega\}$ where $f(\mathbf{y}; p) = p^{\sum_{t=1}^{n} y_t}(1 - p)^{n - \sum_{t=1}^{n} y_t} \prod_{t=1}^{n} I_{\{0,1\}}(y_t)$ and $\Omega = [0, 1]$. A moment condition that must hold for each $y_t$ when $p = p_o$ in this statistical model is given by $\mathrm{E}g(Y_t, p_o) = \mathrm{E}(Y_t - p_o) = 0$, $t = 1, \ldots, n$. The appropriate sample moment counterpart to this moment condition is $n^{-1} \sum_{t=1}^{n} g(y_t, p) = n^{-1} \sum_{t=1}^{n} (y_t - p) = 0$, which upon solving for $p$ yields the MOM estimate $\hat{p} = \bar{y}$. In this (fortunate) situation, the MOM estimator is unbiased, BLUE, MVUE, consistent, asymptotically normal, and asymptotically efficient for estimating $p$.

The basic rationale underlying MOM estimation is rooted in the laws of large numbers. If $\mathrm{E}\mathbf{g}_t(Y_t, \Theta_0) = [\mathbf{0}]$ defines a vector of $k$ moment conditions involving the $(k \times 1)$ vector $\Theta_0$ that hold $\forall\, t$, and if a law of large numbers applies to the random variables $\mathbf{g}_t(Y_t, \Theta_0)$, $t = 1, \ldots, n$, then $n^{-1} \sum_{t=1}^{n} \mathbf{g}_t(Y_t, \Theta_0) \xrightarrow{p} [\mathbf{0}]$ (weak law) or $n^{-1} \sum_{t=1}^{n} \mathbf{g}_t(Y_t, \Theta_0) \xrightarrow{as} [\mathbf{0}]$ (strong law). Moreover, if $\mathrm{E}\mathbf{g}_t(Y_t, \Theta) = \xi(\Theta) \neq [\mathbf{0}]$ when $\Theta \neq \Theta_0$ so that $n^{-1} \sum_{t=1}^{n} \mathbf{g}_t(Y_t, \Theta) \to \xi(\Theta)$ in probability or almost surely, then intuitively we would expect that solving $n^{-1} \sum_{t=1}^{n} \mathbf{g}_t(y_t, \Theta) = [\mathbf{0}]$ for $\Theta$ should produce a value, $\hat{\theta}$, that becomes very close to $\Theta_0$ as $n$ increases; otherwise, the preceding convergence results will be ultimately contradicted. In fact, under appropriate regularity conditions, this is precisely true and the MOM estimator $\hat{\Theta}$ converges in probability, or almost surely, to $\Theta_0$. It is also possible (as in our preceding Bernoulli example) that $\hat{\Theta}$ is asymptotically normal and efficient based on an application of central limit theory. Related arguments establishing asymptotic properties of MOM estimators can be applied in more general contexts where the dimension of $\mathbf{g}$ exceeds $\Theta$, in which case some measure of the (weighted) distance between $n^{-1} \sum_{t=1}^{n} \mathbf{g}_t(y_t, \Theta)$ and $[\mathbf{0}]$ is minimized. For example,

$$\hat{\Theta} = \arg\min_{\Theta} \left[ n^{-1} \sum_{t=1}^{n} \mathbf{g}_t(Y_t, \Theta) \right]' \mathbf{W} \left[ n^{-1} \sum_{t=1}^{n} \mathbf{g}_t(Y_t, \Theta) \right],$$

with $\mathbf{W}$ being positive definite and symmetric, is one such (generalized) MOM estimator that can be consistent, asymptotically normal, and asymptotically efficient under appropriate regularity conditions.

We will examine the ordinary method-of-moments procedure, which is based on iid sampling and which, for historical reasons, will simply be referred to as the *method of moments*. We will also provide an introduction to the *generalized* method-of-moments (GMM) procedure, which can be applied in much more general sampling contexts and which is recently gaining popularity in econometric applications.

### Method of Moments Estimator

In order to apply the (ordinary) MOM procedure, it is assumed that random sampling is from a population distribution, so that the random sample $(Y_1, \ldots, Y_n)$ is a collection of iid random variables. Given a statistical model for the random sample $\{f(\mathbf{y}; \Theta), \Theta \in \Omega\}$, our initial interest will focus on estimating $\Theta$. We have seen in our study of parametric families of densities that the moments of density functions are functions of the parameters characterizing the parametric family. For example, in the exponential density case, $\mathrm{E}Y_t = \theta$ and $\mathrm{E}Y_t^2 = 2\theta^2$, or in the normal density case, $\mathrm{E}Y_t = \mu$ and $\mathrm{E}Y_t^2 = \sigma^2 + \mu^2$. In general, $\mu_r' = \mathrm{E}Y_t^r = h_r(\Theta)$, i.e., the $r$th moment about the origin for a parametric family of densities $f(\mathbf{y}; \Theta)$, $\Theta \in \Omega$, is a function of the parameter vector $\Theta$.

The MOM procedure for estimating the $(k \times 1)$ vector $\Theta$ is to first define a $(k \times 1)$ vector of invertible moment conditions having the particular form

$$\mathrm{E}\mathbf{g}(Y_t, \Theta) = \mathrm{E}\begin{bmatrix} Y_t - h_1(\Theta) \\ Y_t^2 - h_2(\Theta) \\ \vdots \\ Y_t^k - h_k(\Theta) \end{bmatrix} = [0], \quad t = 1, \ldots, n.$$

The sample moment counterparts to the moment conditions are then specified as

$$n^{-1}\sum_{t=1}^{n} \mathbf{g}(y_t, \Theta) = \begin{bmatrix} m_1' - h_1(\Theta) \\ m_2' - h_2(\Theta) \\ \vdots \\ m_k' - h_k(\Theta) \end{bmatrix} = [0],$$

and the solution for $\Theta$ defines the MOM estimate via the inverse function $h^{-1}$ as $\hat{\theta}_j = h_j^{-1}(m_1', m_2', \ldots, m_k')$, $j = 1, \ldots, k$.[42] Note that an alternative motivation for the MOM estimate in this iid case is to define the sample moment conditions in terms of expectations taken with respect to the empirical distribution function of the $y_i$'s, and then solve for $\Theta$ as $\hat{\theta} = \arg_{\Theta}[\mathrm{E}_{\hat{F}}\mathbf{g}(Y, \Theta) = [0]]$. Some examples of the procedure will illustrate its relative simplicity.

**Example 8.21** Let $(Y_1, \ldots, Y_n)$ be a random sample from a Gamma population distribution of waiting times between customer arrivals in minutes, and suppose it is desired to estimate the parameter vector $(\alpha, \beta)$. In this case the moment conditions can be specified as

$$\mathrm{E}\mathbf{g}(Y_t, \alpha, \beta) = \mathrm{E}\begin{bmatrix} Y_t - \alpha\beta \\ Y_t^2 - \alpha\beta^2(1+\alpha) \end{bmatrix} = [0], \quad t = 1, \ldots, n.$$

The sample moment counterpart to the moment conditions is given by

$$n^{-1}\sum_{t=1}^{n} \mathbf{g}(y_t, \alpha, \beta) = \begin{bmatrix} m_1' - \alpha\beta \\ m_2' - \alpha\beta^2(1+\alpha) \end{bmatrix} = [0],$$

so that the MOM estimate is defined by

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} (m_1')^2/[m_2' - (m_1')^2] \\ [m_2' - (m_1')^2]/m_1' \end{bmatrix}.$$

If $m_1' = 1.35$ and $m_2' = 2.7$, then the MOM estimate is given by $\hat{\alpha}(\mathbf{y}) = 2.0769$ and $\hat{\beta}(\mathbf{y}) = .65$. □

**Example 8.22** Let $(Y_1, \ldots, Y_n)$ be a random sample from a normal population distribution representing the weights, in hundredweights, of steers fed a certain ration of feed

[42] While it is standard practice to use the first $k$ population moments when defining the needed moment conditions, other moments about the origin, or even moments about the mean, can be used in defining the moment conditions. The key requirement is that the moment conditions be invertible so that $\Theta$ can be solved in terms of sample moments.

for six months, and suppose it is desired to estimate the parameter vector $(\mu, \sigma^2)$. The moment conditions in this case are

$$\mathrm{E}g(Y_t, \mu, \sigma^2) = \mathrm{E}\begin{bmatrix} Y_t - \mu \\ Y_t^2 - (\sigma^2 + \mu^2) \end{bmatrix} = [0], \quad t = 1, \ldots, n.$$

The sample moment counterpart to the moment conditions is given by

$$n^{-1}\sum_{t=1}^{n} g(y_t, \mu, \sigma^2) = \begin{bmatrix} m_1' - \mu \\ m_2' - (\sigma^2 + \mu^2) \end{bmatrix} = [0],$$

so that the MOM estimate is defined by

$$\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} = \begin{bmatrix} m_1' \\ m_2' - (m_1')^2 \end{bmatrix}.$$

If $m_1' = 12.3$ and $m_2' = 155.2$, then the MOM estimate is given by $\hat{\mu}(\mathbf{y}) = 12.3$ and $\hat{\sigma}^2(\mathbf{y}) = 3.91$. □

### MOM Estimator Properties

It is very often the case that the inverse function $\hat{\theta} = h^{-1}(m_1', \ldots, m_k')$ is continuous, in which case the MOM estimator inherits consistency from the consistency of $M_r'$ for $\mu_r'$, $\forall\, r$.

**Theorem 8.23** (Consistency of MOM Estimator $\Theta$) *Let the MOM estimator* $\hat{\Theta}_{(k\times 1)} = \mathbf{h}^{-1}(M_1', \ldots, M_k')$ *be such that* $\mathbf{h}^{-1}(\mu_1', \ldots, \mu_k')$ *is continuous* $\forall\, (\mu_1', \ldots, \mu_k') \in \Gamma = \{(\mu_1', \ldots, \mu_k')\colon\ \mu_i' = \mathbf{h}_i(\Theta), i = 1, \ldots, k, \Theta \in \Omega\}$. *Then* $\hat{\Theta} \xrightarrow{p} \Theta$.

**Proof** This is a direct consequence of Theorem 5.5 and the probability limit properties of sample moments about the origin. In particular, given the assumed continuity property of $\mathbf{h}^{-1}$,

$$\operatorname{plim} \hat{\Theta} = \mathbf{h}^{-1}(\operatorname{plim} M_1', \ldots, \operatorname{plim} M_k') = \mathbf{h}^{-1}(\mu_1', \ldots, \mu_k') = \Theta \quad \text{by the invertibility of } \mathbf{h}. \quad \blacksquare$$

**Example 8.23** Reexamine Ex. 8.21 and Ex. 8.22. In Ex. 8.21, $(\alpha, \beta)$ is a continuous function of $\mu_1'$ and $\mu_2'$ for all $(\mu_1', \mu_2') \in \Gamma = \{(\mu_1', \mu_2')\colon\ \mu_1' > 0, \mu_2' > 0, \mu_2' > (\mu_1')^2\}$, which constitutes all admissible values of $(\mu_1', \mu_2')$. (Recall $\sigma^2 = \mu_2' - (\mu_1')^2 = \alpha\beta^2 > 0$ in the Gamma family.) Thus, by Theorem 8.23, the MOM estimator of $(\alpha, \beta)$ is consistent. In Ex. 8.22, $(\mu, \sigma^2)$ is a continuous function of $(\mu_1', \mu_2')$ and so the MOM estimator of $(\mu, \sigma^2)$ is consistent. □

In addition to consistency, the MOM estimator will have an asymptotic normal density if the inverse function $\Theta = \mathbf{h}^{-1}(\mu_1', \ldots, \mu_k')$ is continuously differentiable and its Jacobian matrix has full rank.

**Theorem 8.24** (Asymptotic Normality of MOM Estimator of $\Theta$) *Let the MOM estimator* $\hat{\Theta} = \mathbf{h}^{-1}(M_1', \ldots, M_k')$ *be such that* $\mathbf{h}^{-1}(\mu_1', \ldots, \mu_k')$ *is differentiable* $\forall\, (\mu_1', \ldots,$

*$\mu_k'$) $\in \Gamma = \{(\mu_1',\ldots,\mu_k') : \mu_i' = \mathbf{h}_i(\Theta), i = 1,\ldots,k, \Theta \in \Omega\}$, and let the elements of*

$$\mathbf{A}(\mu_1',\ldots,\mu_k') = \begin{bmatrix} \dfrac{\partial h_1^{-1}(\mu_1',\ldots,\mu_k')}{\partial \mu_1'} & \cdots & \dfrac{\partial h_1^{-1}(\mu_1',\ldots,\mu_k')}{\partial \mu_k'} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial h_k^{-1}(\mu_1',\ldots,\mu_k')}{\partial \mu_1'} & \cdots & \dfrac{\partial h_k^{-1}(\mu_1',\ldots,\mu_k')}{\partial \mu_k'} \end{bmatrix}$$

*be continuous functions with $\mathbf{A}(\mu_1',\ldots,\mu_k')$ having full rank $\forall\ (\mu_1',\ldots,\mu_k') \in \Gamma$. Then*

$$n^{1/2}(\hat{\Theta} - \Theta) \xrightarrow{d} N[0], \mathbf{A}\Sigma\mathbf{A}'), \quad \textit{and } \hat{\Theta} \overset{a}{\sim} N(\Theta, n^{-1}\mathbf{A}\Sigma\mathbf{A}'), \quad \textit{where } \underset{k\times k}{\Sigma} = \mathbf{Cov}(M_1',\ldots,M_k').$$

**Proof** Recall that the sample moments converge to a normal limiting distribution as

$$n^{1/2}\left[\begin{bmatrix} M_1' \\ \vdots \\ M_k' \end{bmatrix} - \begin{bmatrix} \mu_1' \\ \vdots \\ \mu_k' \end{bmatrix}\right] \xrightarrow{d} N([0], \Sigma).$$

Then since the partial derivatives contained in $\mathbf{A}$ are continuous and since $\mathbf{A}$ has full rank, the result of the theorem follows directly from Theorem 5.40. ■

**Example 8.24** Reexamine Ex. 8.21 and 8.22. In Ex. 8.21,

$$\mathbf{A}(\mu_1', \mu_2') = \begin{bmatrix} \dfrac{2\mu_2'\mu_1'}{[\mu_2' - (\mu_1')^2]^2} & \dfrac{-(\mu_1')^2}{[\mu_2' - (\mu_1')^2]^2} \\ \dfrac{-[\mu_2' + (\mu_1')^2]}{(\mu_1')^2} & \dfrac{1}{(\mu_1')} \end{bmatrix},$$

which exists with continuous elements $\forall\ (\mu_1', \mu_2') \in \Gamma$. Furthermore, $\mathbf{A}$ has full rank, since $\det(\mathbf{A}) = (\mu_2' - (\mu_1')^2)^{-1} > 0\ \forall\ (\mu_1', \mu_2') \in \Gamma$. Therefore, by Theorem 8.24, the MOM estimator of $(\alpha, \beta)$ is asymptotically normally distributed. In Ex. 8.22,

$$\mathbf{A}(\mu_1', \mu_2') = \begin{bmatrix} 1 & 0 \\ -2\mu_1' & 1 \end{bmatrix},$$

which exists with continuous elements and has full rank $\forall\ (\mu_1', \mu_2')$, and so the MOM estimator is asymptotically normally distributed, by Theorem 8.24. □

The reader should be warned that although the MOM estimator of $\Theta$ is conceptually straightforward to define and is often quite straightforward to compute, in some cases the inverse function may be difficult to define explicitly, if an explicit representation exists at all. A computer algorithm for solving systems of nonlinear equations might be needed to compute the inverse function values and thereby calculate outcomes of the MOM estimator. As an example of more complicated cases, the reader should consider MOM estimation of the parameter vector $\Theta$ when sampling from a Beta family of densities.

MOM estimators for a function of $\Theta$, say $\mathbf{q}(\Theta)$, can be defined by the corresponding function of the MOM estimator of $\Theta$, as $\mathbf{q}(\hat{\Theta}) = \mathbf{q}(\mathbf{h}^{-1}(M'_1, \ldots, M'_k))$. Under appropriate conditions on the function $\mathbf{q}$, and if the MOM estimator of $\Theta$ satisfies the conditions of Theorem's 8.23 and 8.24, then $\mathbf{q}(\hat{\Theta})$ is consistent and asymptotically normal as an estimator of $\mathbf{q}(\Theta)$.

**Theorem 8.25** (Consistency and Asymptotic Normality of MOM Estimator of $\mathbf{q}(\Theta)$) *Let the MOM estimator* $\hat{\Theta}_{(k\times 1)} = \mathbf{h}^{-1}(M'_1, \ldots, M'_k)$ *of* $\Theta$ *satisfy the conditions of Theorems 8.23 and 8.24. If the function* $\mathbf{q}(\Theta)$ *is continuous for* $\Theta \in \Omega$, *then the MOM estimator* $\mathbf{q}(\hat{\Theta})$ *is consistent for* $\mathbf{q}(\Theta)$. *If* $\partial \mathbf{q}(\Theta)/\partial \Theta'$ *exists, has full row rank, and its elements are continuous functions of* $\Theta$ *for* $\Theta \in \Omega$, *then*

$$n^{1/2}\left(\mathbf{q}(\hat{\Theta}) - \mathbf{q}(\Theta)\right) \xrightarrow{d} N\left([0], \frac{\partial \mathbf{q}(\Theta)'}{\partial \Theta} \Phi \frac{\partial \mathbf{q}(\Theta)}{\partial \Theta}\right),$$

*and*

$$\mathbf{q}(\hat{\Theta}) \overset{a}{\sim} N\left(\mathbf{q}(\Theta), n^{-1}\frac{\partial \mathbf{q}(\Theta)'}{\partial \Theta} \Phi \frac{\partial \mathbf{q}(\Theta)}{\partial \Theta}\right),$$

*where* $\Phi$ *is the covariance matrix of the limiting distribution of* $n^{1/2}(\hat{\Theta} - \Theta)$ *defined in Theorem 8.24.*

**Proof** The proof is analogous to the proofs of Theorem's 8.23 and 8.24 and is left to the reader. ■

Overall, the MOM approach to point estimation of $\Theta$ or $\mathbf{q}(\Theta)$ is conceptually simple, computation of estimates is often straightforward, and the MOM estimator is consistent and asymptotically normal under fairly general conditions. However, the MOM estimator is often not unbiased, BLUE, MVUE, or asymptotically efficient. It has been most often applied in cases where the definition or computation of other estimators, such as the MLE, is extremely complex, or when the sample size is quite large so that the consistency of the estimator can be relied upon to ensure a reasonably accurate estimate. Further discussion of asymptotic efficiency of MOM estimators, in the context of GMM estimators will continue in the next subsection.

### Generalized Method of Moments (GMM) Estimator

In the case of the GMM, the random sample is not restricted to iid random variables, and the moment conditions are not restricted to the particular form presented in the preceding discussion of MOM estimation. In fact, the moment conditions can take a myriad of different forms, depending on the statistical model being analyzed. However, the general estimation principles remain the same, namely, moment conditions $\mathrm{E}\mathbf{g}_t(Y_t, \Theta_o) = [0]$, $t = 1, \ldots, n$ are specified pertaining to the statistical model $\{f(\mathbf{y}; \Theta), \Theta \in \Omega\}$, sample moment counterparts are specified as $n^{-1}\sum_{t=1}^{n} \mathbf{g}_t(y_t, \Theta) = [0]$, and then the GMM estimate is defined as the value of $\Theta$ that satisfies the sample moment conditions as closely as possible—exactly if $\mathbf{g}$ is of the same dimension as $\Theta$.

The following examples demonstrate how the GMM estimator can be formulated to subsume the least-squares estimator and the MLE as special cases.

**Example 8.25** Least-Squares Estimator as a GMM Estimation

Let the statistical model of a random sample of the yields/acre of a particular type of corn be (incompletely) specified by the linear model $\mathbf{Y} = \mathbf{x}\beta + \varepsilon$, where the classical GLM assumptions are assumed to apply and $\beta$ is a $k \times 1$ vector of parameters indicating the responsiveness of yield to various inputs. Moment conditions for the statistical model can be specified by the $(k \times 1)$ vector function

$$\mathrm{E}\mathbf{g}_t(Y_t, \beta) = \mathrm{E}\mathbf{x}'_{t.}(Y_t - \mathbf{x}_{t.}\beta) = \mathrm{E}\mathbf{x}'_{t.}\varepsilon_t = [\mathbf{0}], \quad t = 1, \ldots, n.$$

Corresponding sample moment conditions can be defined as

$$n^{-1}\sum_{t=1}^{n} \mathbf{g}_t(y_t, \beta) = n^{-1}\sum_{t=1}^{n} \mathbf{x}'_{t.}(y_t - \mathbf{x}_{t.}\beta) = n^{-1}[\mathbf{x}'\mathbf{y} - \mathbf{x}'\mathbf{x}\beta] = [\mathbf{0}],$$

which when solved for $\beta$ defines the GMM estimate of $\beta$ as $\hat{\beta}(\mathbf{y}) = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$. Thus, the least-squares estimator is a GMM estimator. □

**Example 8.26** MLE as a GMM Estimator

Let $\{f(\mathbf{y}; \Theta), \Theta \in \Omega\}$ be the statistical model for a random sample for which the $Y_i$'s are not necessarily iid, and represent the joint density of the random sample as $f(\mathbf{y}; \Theta) = m(y_1; \Theta)\prod_{t=2}^{n} f(y_t \mid y_{t-1}, \ldots, y_1; \Theta)$. Assume that differentiation under the integral or summation sign is possible (recall CRLB regularity condition (4) of Def. 7.21) so that moment conditions can be defined as[43]

$$\mathrm{E}\mathbf{g}_1(Y_1, \Theta_o) = \mathrm{E}\left[\frac{\partial \ln m(Y_1; \Theta_o)}{\partial \Theta}\right] = [\mathbf{0}],$$

$$\mathrm{E}\mathbf{g}_t(Y_t, \Theta_o) = \mathrm{E}\left[\frac{\partial \ln f(Y_t | y_{t-1}, \ldots, y_1; \Theta_o)}{\partial \Theta}\right] = [\mathbf{0}] \quad \text{for } t = 2, \ldots, n.$$

The sample moment conditions are then

$$n^{-1}\sum_{t=1}^{n} \mathbf{g}_t(y_t, \Theta_o) = n^{-1}\left[\frac{\partial \ln m(y_1; \Theta)}{\partial \Theta} + \sum_{t=2}^{n} \frac{\partial \ln f(y_t \mid y_{t-1}, \ldots, y_1; \Theta)}{\partial \Theta}\right]$$

$$= n^{-1}\frac{\partial \ln L(\Theta; \mathbf{y})}{\partial \Theta} = [\mathbf{0}],$$

[43] Under the assumed conditions, for each density (conditional or not),

$$\mathrm{E}_{\Theta_o}\frac{\partial \ln f(Z; \Theta_o)}{\partial \Theta} = \int_{-\infty}^{\infty} \frac{\partial f(z; \Theta_o)}{\partial \Theta} dz = \frac{\partial \int_{-\infty}^{\infty} f(z; \Theta_o) dz}{\partial \Theta} = [0]$$

in the continuous case, and likewise in the discrete case.

which when solved for $\Theta$ defines the GMM estimate of $\Theta$ as $\hat{\theta} = \arg_{\Theta}[\partial \ln L(\Theta; y)/\partial\Theta = [0]]$. Assuming the MLE of $\Theta$ is defined as the solution to first-order conditions, the MLE is a GMM estimator. □

In later studies of econometrics and/or advanced statistics, the reader will come to find that the GMM subsumes a large number of the estimation procedures used in practice. For example, *instrumental variable techniques, two- and three-stage least squares,* and *quasi- or pseudo-maximum likelihood estimation techniques* can all be interpreted as GMM procedures. The interested reader is referred to A. R. Gallant (1987), *Nonlinear Statistical Models.* New York: John Wiley, Chapter 4, and to R. Davidson and J. G. MacKinnon (1993), *Estimation and Inference in Econometrics,* New York, Oxford Univ. Press, Chapter 17, for additional details on applying the GMM approach to more general statistical models.

It is possible to apply the GMM procedure in cases where the vector functions defining the moment conditions, $\mathbf{g}_t(y_t, \Theta)$, are of larger dimension than the dimension of the vector $\Theta$. In these cases the GMM estimate is defined to be the value of $\Theta$ that minimizes a weighted measure of the distance between the sample moments and the zero vector, as

$$\hat{\theta} = \arg\min_{\Theta\in\Omega}\left[\left[n^{-1}\sum_{t=1}^{n}\mathbf{g}_t(y_t,\Theta)\right]' \mathbf{W}_n(\mathbf{y})\left[n^{-1}\sum_{t=1}^{n}\mathbf{g}_t(y_t,\Theta)\right]\right].$$

The matrix $\mathbf{W}_n(\mathbf{y})$ is a symmetric, positive definite, conformable weighting matrix that may or may not depend on $\mathbf{y}$ and is such that $\mathbf{W}_n \xrightarrow{p} \mathbf{w}$, with $\mathbf{w}$ being a nonrandom, symmetric, and positive definite matrix. The previous case where $\mathbf{g}_t$ and $\Theta$ are of equal dimension is subsumed within the current context by defining $\mathbf{W}_n = \mathbf{I}$, in which case the minimum of the distance measure occurs at the value of $\Theta$ for which $n^{-1}\sum_{t=1}^{n}\mathbf{g}_t(y_t,\Theta) = [\mathbf{0}]$. The GMM estimator is consistent and asymptotically normal quite generally with respect to choices of the weighting matrix, although asymptotic efficiency considerations require a specific choice of $\mathbf{W}_n$. These issues will be considered in the next section.

### GMM Properties

As was the case for the MOM procedure, general statements regarding the properties of GMM estimators are relegated to the asymptotic variety. We will examine a set of *sufficient* conditions for the consistency and asymptotic normality of the GMM estimator as well as a sufficient condition for asymptotic efficiency within the class of estimators based on a particular set of moment conditions. The reader can refer to Gallant, *Nonlinear Models,* and L. P. Hansen (1982), "Large sample properties of generalized method of moments estimators," *Econometrica,* pp. 1029–1054, for alternative conditions that lead to asymptotic properties of the GMM estimator. Throughout this section we will be referring to the GMM estimator defined in terms of the minimization of the weighted distance measure presented at the end of the previous section, which subsumes the case where the dimensions of $\mathbf{g}_t$ and $\Theta$ are equal.

*Consistency* Sufficient conditions under which the GMM estimator is consistent are given in the following theorem. Recall that $E_\Theta$, $P_\Theta$, and $\text{plim}_\Theta$, respectively, denote an expectation, a probability, and a probability limit taken with respect to a probability distribution characterized by the parameter value $\Theta$.

**Theorem 8.26** (Consistency of GMM Estimator) *Let $\{f(\mathbf{y};\Theta), \Theta \in \Omega\}$ be the statistical model for the random sample* $\mathbf{Y}$, *where $\Theta$ is $(k \times 1)$. Let $\mathrm{E}_\Theta \mathbf{g}_t(Y_t,\Theta) = [0]$ be an $(m \times 1)$ moment condition for $t = 1, \ldots, n$ with $m \geq k$. Define the GMM estimator of $\Theta$ as*

$$\hat{\Theta} = \arg\min_{\Theta\in\Omega} Q_n(\mathbf{Y},\Theta) = \arg\min_{\Theta\in\Omega}\left[\left[n^{-1}\sum_{t=1}^{n}\mathbf{g}_t(Y_t,\Theta)\right]' \mathbf{W}_n(\mathbf{Y})\left[n^{-1}\sum_{t=1}^{n}\mathbf{g}_t(Y_t,\Theta)\right]\right],$$

*where* $\text{plim}_{\Theta_o}\mathbf{W}_n(\mathbf{Y}) = \mathbf{w}$, *a nonrandom, positive definite, and symmetric matrix. Assume that*

**a.** *$\Omega$ is a closed and bounded rectangle,*
**b.** *$\hat{\Theta}$ is unique,*
**c.** $\text{plim}_{\Theta_o}[n^{-1}\sum_{t=1}^{n}\mathbf{g}_t(Y_t,\Theta)] = \mathbf{G}(\Theta)$, *a continuous, nonstochastic, $(m \times 1)$ vector function of $\Theta$ for which $\mathbf{G}(\Theta) = [0]$ iff $\Theta = \Theta_o$,*
**d.** $\lim_{n\to\infty}P_{\Theta_o}(\max_{\Theta\in\Omega}|Q_n(\mathbf{y},\Theta) - h(\Theta)| < \varepsilon) = 1, \forall\, \varepsilon > 0,$

*where $h(\Theta) = \mathbf{G}(\Theta)'\mathbf{w}\mathbf{G}(\Theta)$. Then $\hat{\Theta} \xrightarrow{p} \Theta_o$.*

***Proof** The function $h(\Theta) = \mathbf{G}(\Theta)'\mathbf{w}\mathbf{G}(\Theta)$ is a continuous function of $\Theta$ that is uniquely minimized at $\Theta = \Theta_o$ since $w$ is positive definite and $\mathbf{G}(\Theta) = [0]$ iff $\Theta = \Theta_o$. Define $N(\varepsilon) = \{\Theta\colon d(\Theta,\Theta_o) < \varepsilon\}$, which is an open $\varepsilon$-neighborhood of the true parameter value $\Theta_o$, and let $\xi(\varepsilon) = \min_{\Theta\in\Omega\cap\overline{N(\varepsilon)}} h(\Theta) - h(\Theta_o)$ (where the minimum exists by Weierstrass's theorem because $h(\Theta)$ is continuous and $\Omega \cap \overline{N(\varepsilon)}$ is closed and bounded). Let the event $A_n(\varepsilon)$ be defined as

$$A_n(\varepsilon) = \{\mathbf{x}\colon \max_{\Theta\in\Omega}|Q_n(\mathbf{y},\Theta) - h(\Theta)| < \xi(\varepsilon)/2\} \quad \text{for } \varepsilon > 0.$$

Letting $\hat{\Theta}$ represent the unique GMM estimate, it follows that

1. $\mathbf{x} \in A_n(\varepsilon) \Rightarrow h(\hat{\Theta}) < Q_n(\mathbf{y},\hat{\Theta}) + \xi(\varepsilon)/2,$
2. $\mathbf{x} \in A_n(\varepsilon) \Rightarrow \mathbf{Q}_n(\mathbf{y},\Theta_o) < h(\Theta_o) + \xi(\varepsilon)/2.$
   By the definition of the GMM, $Q_n(\mathbf{y},\hat{\Theta}) \leq \mathbf{Q}_n(\mathbf{y},\Theta_o)$, so it follows from (1) that
3. $\mathbf{x} \in A_n(\varepsilon) \Rightarrow h(\hat{\Theta}) < Q_n(\mathbf{y},\Theta_o) + \xi(\varepsilon)/2.$
   Substituting (2) into the right side of the inequality in (3) yields
4. $\mathbf{x} \in A_n(\varepsilon) \Rightarrow h(\hat{\Theta}) < h(\Theta_o) + \xi(\varepsilon).$
   Then substituting the definition of $\xi(\varepsilon)$ into (4) yields
5. $\mathbf{x} \in A_n(\varepsilon) \Rightarrow h(\hat{\Theta}) < \min_{\Theta\in\Omega\cap\overline{N(\varepsilon)}} h(\Theta) \Rightarrow \hat{\Theta} \in N(\varepsilon).$

Since by condition (d) of the theorem $\lim_{n\to\infty} P(A_n(\varepsilon)) = 1 \ \forall\ \varepsilon > 0$, then $\lim_{n\to\infty} P(\hat{\Theta} \in N(\varepsilon)) = 1 \ \forall\ \varepsilon > 0$, and thus $\hat{\Theta} \xrightarrow{p} \Theta_o$. ■

In practice, the GMM estimator is often represented by an *explicit* function of the random sample, in which case it is often more straightforward to verify consistency by taking probability limits directly. Sufficient conditions, such as those in Theorem 8.26, are generally needed when the GMM estimator is only an *implicit* function of the random sample. In the following example, we illustrate both approaches to verifying consistency.

***Example 8.27** (Instrumental Variables Estimator) Let the aggregate demand for a given commodity be approximated by the linear function $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{V}$, where the $V_t$'s are iid with $EV_t = 0$ and $\text{var}(V_t) = \sigma^2, \ \forall\ t$. Suppose that the random matrix $\mathbf{X}$ contains factors affecting demand, including some commodity prices that are simultaneously determined with quantity demanded, $\mathbf{Y}$, so that $E(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V} \neq [\mathbf{0}]$ (recall Table 8.2). Thus, the OLS estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, will be biased and inconsistent.

Suppose there existed a conformable $(n \times k)$ matrix of variables, $\mathbf{z}$ (called *instrumental variables* in the literature), such that

$$\mathrm{E}_{\boldsymbol{\beta}}\mathbf{g}_t(Y_t, \boldsymbol{\beta}) = \mathrm{E}_{\boldsymbol{\beta}}\mathbf{z}'_{t.}(Y_t - \mathbf{X}_{t.}\boldsymbol{\beta}) = \mathrm{E}_{\boldsymbol{\beta}}\mathbf{z}'_{t.}V_t = [\mathbf{0}] \quad \text{for } t = 1, \ldots, n,$$

where $n^{-1}\mathbf{z}'\mathbf{z} \to \mathbf{A}_{zz}$, a finite, positive definite matrix, and $n^{-1}\mathbf{z}'\mathbf{X} \xrightarrow{p} \mathbf{A}_{zx}$, a finite, nonsingular, square matrix. The sample moment conditions are

$$n^{-1}\sum_{t=1}^{n} \mathbf{g}_t(y_t, \boldsymbol{\beta}) = n^{-1}\sum_{t=1}^{n} \mathbf{z}'_{t.}(y_t - \mathbf{x}_{t.}\boldsymbol{\beta}), = n^{-1}[\mathbf{z}'\mathbf{y} - \mathbf{z}'\mathbf{x}\boldsymbol{\beta}] = [\mathbf{0}].$$

Assuming that $\mathbf{z}'\mathbf{x}$ is nonsingular, the sample moment conditions have a *unique* solution defining the GMM estimate as $\hat{\boldsymbol{\beta}} = (\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y}$, which is also referred to in the literature as the **instrumental variables estimate**. Thus condition (b) of Theorem 8.26 is satisfied.

Note that $\text{plim}_{\boldsymbol{\beta}_o}[n^{-1}\sum_{t=1}^{n} \mathbf{g}_t(Y_t, \boldsymbol{\beta})] = \mathbf{A}_{zx}[\boldsymbol{\beta}_o - \boldsymbol{\beta}] = \mathbf{G}(\boldsymbol{\beta})$, which equals $[\mathbf{0}]$ iff $\boldsymbol{\beta} = \boldsymbol{\beta}_o$, satisfying condition (c) of Theorem 8.26. Assume that $\Omega$ is a closed and bounded rectangle. Representing the GMM estimate equivalently as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q_n(\mathbf{y}, \boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}\in\Omega}\left[n^{-1}[\mathbf{z}'\mathbf{y} - \mathbf{z}'\mathbf{x}\boldsymbol{\beta}]'\mathbf{I}[\mathbf{z}'\mathbf{y} - \mathbf{z}'\mathbf{x}\boldsymbol{\beta}]n^{-1}\right],$$

note that $\max_{\boldsymbol{\beta}\in\Omega} |Q_n(\mathbf{y}, \boldsymbol{\beta}) - h(\boldsymbol{\beta})|$, where $h(\boldsymbol{\beta}) = (\boldsymbol{\beta}_o - \boldsymbol{\beta})'\mathbf{A}'_{zx}\mathbf{A}_{zx}(\boldsymbol{\beta}_o - \boldsymbol{\beta})$, will exist by Weierstrass's theorem, and the maximum is a continuous function of the elements in $n^{-1}\mathbf{z}'\mathbf{y}$ and $n^{-1}\mathbf{z}'\mathbf{x}$ by the theorem of the maximum. It follows that

$$\underset{\boldsymbol{\beta}_o}{\text{plim}}\left[\max_{\boldsymbol{\beta}\in\Omega} |Q_n(\mathbf{Y}, \boldsymbol{\beta}) - h(\boldsymbol{\beta})|\right] = \max_{\boldsymbol{\beta}\in\Omega} |\,\text{plim}\, Q_n(\mathbf{Y}, \boldsymbol{\beta}) - h(\boldsymbol{\beta})| = [0]$$

so that condition (d) of Theorem 8.26 is satisfied. Therefore, $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

Given the explicit functional definition of $\hat{\beta}$, the consistency of the GMM estimator can be proven more directly by noting that

$$\begin{aligned}\text{plim}\,\hat{\beta} &= \text{plim}(n^{-1}\mathbf{z}'\mathbf{X})^{-1}(n^{-1}\mathbf{z}'\mathbf{Y})\\ &= \text{plim}(n^{-1}\mathbf{z}'\mathbf{X})^{-1}\,\text{plim}(n^{-1}(\mathbf{z}'\mathbf{X}\beta+\mathbf{z}'\mathbf{V}))\\ &= \mathbf{A}_{zx}^{-1}(\mathbf{A}_{zx}\beta)=\beta,\end{aligned}$$

where $n^{-1}\mathbf{z}'\mathbf{V} \xrightarrow{p} [0]$ since $\text{var}(n^{-1}\mathbf{z}'\mathbf{V}) = (\sigma^2/n)(n^{-1}\mathbf{z}'\mathbf{z}) \to [0]$. □

*Asymptotic Normality and Efficiency in the GMM Class* Under additional assumptions relating to the moment conditions, it can be shown that GMM estimators are asymptotically normally distributed.[44]

**Theorem 8.27** (Asymptotic Normality of GMM Estimator) *Assume the conditions of Theorem 8.26. In addition, assume that*

**a.** $\partial^2 Q_n(\mathbf{y},\Theta)/\partial\Theta\partial\Theta'$ *exists and is continuous* $\forall\, \Theta \in \Omega$,

**b.** *for any sequence* $\Theta_n^*$ *for which* $\text{plim}\,\Theta_n^* = \Theta_o$, $\partial^2 Q_n(\mathbf{Y},\Theta_n^*)/\partial\Theta\partial\Theta' \xrightarrow{p} \mathbf{D}(\Theta_o)$, *a finite, symmetric, nonsingular matrix, and*

**c.** $n^{1/2}(\partial Q_n(\mathbf{Y},\Theta_o)/\partial\Theta) \xrightarrow{d} \mathbf{Z} \sim N([0],\mathbf{C}(\Theta_o))$.

*Then*

$$n^{1/2}(\hat{\Theta}-\Theta_o) \xrightarrow{d} N([0],\mathbf{D}(\Theta_o)^{-1}\mathbf{C}(\Theta_o)\mathbf{D}(\Theta_o)^{-1})$$

*and* $\hat{\Theta} \overset{a}{\sim} N(\Theta_o, n^{-1}\mathbf{D}(\Theta_o)^{-1}\mathbf{C}(\Theta_o)\mathbf{D}(\Theta_o)^{-1})$.

**Proof** By a first-order Taylor series representation,

$$\frac{\partial Q_n(\mathbf{Y},\hat{\Theta})}{\partial\Theta} = \frac{\partial Q_n(\mathbf{Y},\Theta_o)}{\partial\Theta} + \frac{\partial^2 Q_n(\mathbf{Y},\Theta_*)}{\partial\Theta\partial\Theta'}(\hat{\Theta}-\Theta_o),$$

where $\Theta_* = \lambda\hat{\Theta} + (1-\lambda)\Theta_o$ and $\lambda \in [0,1]$.[45] By the first-order conditions to the minimization problem defining the GMM estimator, $\partial Q_n(\mathbf{y};\hat{\Theta})/\partial\Theta = [0]$, so that by Slutsky's theorems and the consistency of $\hat{\Theta}$,

$$\begin{aligned}n^{1/2}(\hat{\Theta}-\Theta_o) &= -\left[\frac{\partial^2 Q_n(\mathbf{Y},\Theta_*)}{\partial\Theta\partial\Theta'}\right]^{-1} n^{1/2}\frac{\partial Q_n(\mathbf{Y},\Theta_o)}{\partial\Theta}\\ &\xrightarrow{d} -\mathbf{D}(\Theta_o)\mathbf{Z} \sim N([0],\mathbf{D}(\Theta_o)^{-1}\mathbf{C}(\Theta_o)\mathbf{D}(\Theta_o)^{-1}),\end{aligned}$$

and the asymptotic distribution follows from Definition 5.10. ■

**Example 8.28** Revisit the instrumental variable estimator of Ex. 8.27. Note that $\partial^2 Q_n(\mathbf{y},\beta)/\partial\beta\partial\beta' = 2n^{-2}\mathbf{x}'\mathbf{z}\mathbf{z}'\mathbf{x}$, which exists and is (trivially) continuous for $\beta \in \Omega$. Also,

[44] See T. Amemiya, *Advanced Econometrics*, pp. 111–112, for a closely related theorem relating to extremum estimators. The assumptions of Theorem 8.26 can be weakened to assuming uniqueness and consistency of $\hat{\Theta}$ for use in this theorem.

[45] We are suppressing the fact that each row of the matrix $\partial^2 Q_n(\mathbf{Y},\Theta_*)/\partial\Theta\partial\Theta'$ will generally require a different value of $\Theta_*$ defined by a different value of $\lambda$ for the representation to hold. The conclusions of the argument will remain the same.

$\partial^2 Q_n(\mathbf{Y}, \boldsymbol{\beta}_*)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}' = 2(n^{-1}\mathbf{X}'\mathbf{z})(\mathbf{z}'\mathbf{X}n^{-1}) \xrightarrow{p} \mathbf{D}(\boldsymbol{\beta}_o) = 2\mathbf{A}'_{zx}\mathbf{A}_{zx}$, which is a finite, nonsingular matrix regardless of the sequence $\{\boldsymbol{\beta}_*\}$ for which $\boldsymbol{\beta}_* \xrightarrow{p} \boldsymbol{\beta}_o$. Finally, note that $n^{1/2}\partial Q_n(\mathbf{Y}, \boldsymbol{\beta}_o)/\partial\boldsymbol{\beta} = -2(n^{-1}\mathbf{X}'\mathbf{z})(n^{-1/2}\mathbf{z}'\mathbf{V}) \xrightarrow{d} -2\mathbf{A}'_{zx}\mathbf{T} \sim N([\mathbf{0}], 4\sigma^2\mathbf{A}'_{zx}\mathbf{A}_{zz}\mathbf{A}_{zx})$, where $n^{-1/2}\mathbf{z}'\mathbf{V} \xrightarrow{d} \mathbf{T} \sim N([\mathbf{0}], \sigma^2\mathbf{A}_{zz})$ by an application of the multivariate Lindberg–Levy CLT. Then, from Theorem 8.27,

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) \xrightarrow{d} N([\mathbf{0}], \sigma^2(\mathbf{A}_{zx})^{-1}\mathbf{A}_{zz}(\mathbf{A}'_{zx})^{-1}), \qquad \text{and}$$

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}_o, (\sigma^2/n)(\mathbf{A}_{zx})^{-1}\mathbf{A}_{zz}(\mathbf{A}'_{zx})^{-1}).$$

An *estimate* of the asymptotic covariance matrix can be defined as

$$\hat{\sigma}^2(\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{z}(\mathbf{x}'\mathbf{z})^{-1}, \quad \text{where } \hat{\sigma}^2 = n^{-1}(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}). \qquad \square$$

Regarding the efficiency of the GMM estimator, L. P. Hansen, *Moments Estimators*, has provided conditions that characterize the optimal choice of $\mathbf{W}_n(\mathbf{y})$ in the definition of the GMM estimator

$$\hat{\boldsymbol{\beta}}_{\mathbf{W}_n} = \arg\min_{\Theta}\left[\left[n^{-1}\sum_{t=1}^{n}\mathbf{g}_t(Y_t, \Theta)\right]' \mathbf{W}_n(\mathbf{Y})\left[n^{-1}\sum_{t=1}^{n}\mathbf{g}_t(Y_t, \Theta)\right]\right].$$

In particular, setting $\mathbf{W}_n(\mathbf{Y})$ equal to the inverse of the asymptotic covariance matrix of $n^{-1/2}\sum_{t=1}^{n}\mathbf{g}_t(Y_t, \Theta_o)$ is the choice that defines the GMM estimator that is asymptotically most efficient relative to all choices of the weighting matrix. Note that this result establishes asymptotic efficiency within the class of GMM estimators based on a *particular* set of moment conditions. Results relating to the optimal choice of moment conditions with which to define the GMM estimator in the general case are not available, although a number of special cases have been investigated. The interested reader can consult Davidson and MacKinnon, *Estimation and Inference*, Chapter 17 for additional reading and references.

**Example 8.29** Revisit Ex. 8.28 concerning the instrumental variable estimator. Note that

$$n^{-1/2}\sum_{t=1}^{n}\mathbf{g}_t(Y_t, \boldsymbol{\beta}_o) = n^{-1/2}[\mathbf{z}'\mathbf{Y} - \mathbf{z}'\mathbf{X}\boldsymbol{\beta}_o] = n^{-1/2}[\mathbf{z}'\mathbf{V}] \xrightarrow{d} N([\mathbf{0}], \sigma^2\mathbf{A}_{zz}),$$

as argued in Ex. 8.28. Thus, the optimal GMM estimator, which is based on $\mathbf{W}_n(\mathbf{Y}) = (\sigma^2\mathbf{A}_{zz})^{-1}$, is defined by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\left[n^{-2}(\mathbf{z}'\mathbf{Y} - \mathbf{z}'\mathbf{X}\boldsymbol{\beta})'(\sigma^2\mathbf{A}_{zz})^{-1}(\mathbf{z}'\mathbf{Y} - \mathbf{z}'\mathbf{X}\boldsymbol{\beta})\right] = (\mathbf{z}'\mathbf{X})^{-1}\mathbf{z}'\mathbf{Y}. \qquad \square$$

In practice, if the solution for $\hat{\Theta}$ depends on a weighting matrix that contains unknown elements (in Ex. 8.29, the fact that $\mathbf{A}_{zz}$ was unknown is irrelevant—only its positive definiteness mattered), an estimator that converges in probability to the asymptotically efficient GMM estimator can be constructed using any weighting matrix that converges in probability to the inverse of the appropriate asymptotic covariance matrix of $n^{-1/2}\sum_{t=1}^{n}\mathbf{g}_t(Y_t, \Theta_o)$.

## Key Words, Phrases, and Symbols

general linear model (GLM)
dependent variable
explanatory (or independent) variables
disturbance, error, or residual vector
classical assumptions of the GLM
design matrix
homoskedasticity
zero autocovariance or zero autocorrelation
method of least squares
coefficient of determination, $R^2$
Gauss–Markov's theorem
least-squares estimator
specification error
heteroskedastic
autocorrelated
generalized least-squares estimator
perfect multicollinearity
econometrics, sociometrics
least squares under normality
likelihood function, $L(\Theta; \mathbf{x})$
maximum likelihood (ML) estimate
maximum likelihood estimator (MLE)
$\propto$ (proportional to)
bias-adjusted MLE
reparameterization of the density family
parameter space induced by $\mathbf{q}(\Theta)$
likelihood function induced by $\mathbf{q}(\Theta)$
invariance principle
method of moments (MOM)
generalized method of moments (GMM)
instrumental variables estimator

## Problems

**1.** The daily production of electricity generated by a coal-fired power plant operating in the midwest can be represented as

$$Y_i = \beta_1 \ell_i^{\beta_2} m_i^{\beta_3} e^{\varepsilon_i}$$

where

$Y_i$ = quantity of electricity produced on day $i$, measured in megawatts,
$\ell_i$ = quantity of labor input used on day $i$, measured in hundreds of hours,
$m_i$ = units of fossil fuel input on day $i$, and
$\varepsilon_i \sim N(0, \sigma^2)$, the $\varepsilon_i$'s being iid.

One hundred days of observations on the levels of inputs used and the quantity of electricity generated were collected. The following information is provided to you by the utility company, where $\mathbf{y}_* = \ln(\mathbf{y})$, and $\mathbf{x}_*$ is a matrix consisting of a vector of 1's followed by column vectors corresponding to the natural logarithms of labor and fuel levels.

$$\mathbf{y}_*'\mathbf{y}_* = 288.93382, \qquad \mathbf{x}_*'\mathbf{y}_* = \begin{bmatrix} 165.47200 \\ 180.32067 \\ 122.90436 \end{bmatrix},$$

$$(\mathbf{x}_*'\mathbf{x}_*)^{-1} = \begin{bmatrix} .06156 & -.05020 & -.00177 \\ & .09709 & -.07284 \\ \text{(symmetric)} & & .11537 \end{bmatrix}.$$

a. Transform the production function into a form in which parameters can be estimated using the least-squares estimator. Estimate the parameters of the transformed model.

b. Is the estimator you used in part (a) to estimate $\mathbf{q}(\boldsymbol{\beta})$ (1) unbiased, (2) asymptotically unbiased, (3) BLUE, (4) MVUE, (5) consistent, and/or (6) normally distributed?

c. Is the estimator you used to estimate $\sigma^2$ (1) unbiased, (2) asymptotically unbiased, (3) BLUE, (4) MVUE, (5) consistent, and/or (6) Gamma distributed?

d. Define the MVUE for the degree of homogeneity of the production function (i.e., define the MVUE for $q_*(\boldsymbol{\beta}) = \beta_2 + \beta_3$. Estimate the degree of homogeneity of the production function using the MVUE. Is the MVUE a consistent estimator of the degree of homogeneity? Is the MVUE normally distributed?

e. Define an MVUE estimator for the covariance matrix of the least-squares estimator. Estimate the covariance matrix of the least-squares estimator.

**2.** Smith's Dairy is contemplating the profitability of utilizing a new bovine growth hormone for increasing the milk production of its cows. Smith's randomly selects cows and administers given dosages of growth hormone at regular intervals. All animals are cared for identically except for the levels of hormone administered. A random sample of size $n$ from the composite experiment measuring the total milk production over the lactation of each animal is represented by

$$Y_t = \beta_1 + \beta_2 x_t + \varepsilon_t, \quad t = 1, \ldots, n,$$

where the $\varepsilon_t$'s are presumed to be iid with a common marginal density function

$$f(z; a, b) = \frac{1}{b-a} I_{(a,b)}(z)$$

with $a = -b$, and $b$ is some positive number.

The $x_t$ values are the dosages of the growth hormone, measured in cc units, at levels defined by

$$x_t = t - 10\,\text{trunc}\left(\frac{t-1}{10}\right), \quad t = 1, 2, 3, \ldots,$$

and $y_t$ is the milk production of the $t$th dairy cow, measured in hundredweights.

a. Is the least-squares estimator of $\beta$ (1) unbiased, (2) BLUE, (3) consistent, and/or (4) asymptotically normal? Justify your answer.

b. Is the estimator $S^2$ of $\sigma^2$ (1) unbiased, (2) consistent, and/or (3) asymptotically normal? Justify your answer.

(Hint: The $x_t$'s occur in a repeating sequence of the numbers 1, 2, 3, ..., 10. It follows that $\sum_{t=1}^n x_t/n \to 5.5$ and $\sum_{t=1}^n x_t^2/n \to 38.5$.)

c. In the outcome of a random sample of size 100, Smith's found that $\mathbf{y}'\mathbf{y} = 4{,}454{,}656.3$ and $\mathbf{x}'\mathbf{y} = \begin{bmatrix} 21{,}097.673 \\ 117{,}570.78 \end{bmatrix}$. Estimate the parameters of the linear model using $\hat{\beta}$ and $S^2$.

d. Given that raw milk sells for \$10 per hundredweight, define a BLUE for the expected marginal revenue per cow obtained from administering the growth hormone. If the total cost of the hormone treatment per cow over the entire lactation is \$12 per cc, and if the maximum allowable dose is 10 cc, is it profitable for Smith's to utilize the hormone, based on the BLUE estimate of marginal revenue calculated above? Why or why not? If so, determine what level should be administered, and provide a BLUE estimate of the gain in profitability.

e. Regarding Smith's linear model assumptions, are there other distributional assumptions that you would suggest for consideration besides the uniform distribution? Would this change any of your answers above? Would you suggest examining alternative functional forms for the relationship between milk production and hormone treatment? If so, would this affect your answers to the questions above?

**3.** Suppose $\mathbf{Y}_{(n\times 1)} = \mathbf{x}_{(n\times k)}\beta_{(k\times 1)} + \varepsilon_{(n\times 1)}$, where $\varepsilon \sim N([\mathbf{0}], \sigma^2\mathbf{I})$ and $\mathbf{x}$ has full column rank.

a. Show that

$$T = \frac{\ell'(\hat{\beta} - \beta)}{(S^2\ell'(\mathbf{x}'\mathbf{x})^{-1}\ell)^{1/2}},$$

for any conformable $\ell \neq [\mathbf{0}]$, has a $t$-distribution with $(n-k)$ degrees of freedom. (Hint: Transform $T$ so that it is expressed as a ratio of two independent random variables, with an $N(0,1)$ random variable in the numerator, and the square root of a $\chi^2$ random variable divided by its degrees of freedom in the denominator.)

b. Using the fact that $T$ has a $t$-distribution with $(n-k)$ degrees of freedom, define a random interval $(Z_1, Z_2)$ that satisfies $P(\ell'\beta \in (z_1, z_2)) = .95$ when $n-k = 25$. (Hint: Define your $z_1$ and $z_2$ variables as appropriate functions of $(\hat{\beta}, S^2)$, which will be suggested by a transformation of $P(t \in (t_\ell, t_h)) = .95$), for outcomes $t$ of $T$.

**4.** An economist is analyzing the relationship between disposable income and food expenditure for the citizens of a developing country. A survey of the citizens has been taken, and data have been collected on income and food expenditures. The following statistical model is postulated for the data:

$$Y_i = \beta_1 z_i^{\beta_2} e^{\varepsilon_i},$$

where

$y_i$ = expenditure on food for the $i$th household,

$Z_i$ = disposable income for the $i$th household,

$\varepsilon_i's \sim$ iid $N(0, \sigma^2)$.

a. Transform the model into the GLM form. What parameters or functions of parameters are being estimated by the least-squares estimator applied to the transformed model?

b. Is the least-squares estimator unbiased, the BLUE, and/or the MVUE for the parameters or functions of parameters being estimated?

c. The actual survey consisted of 5,000 observations, and the following summary of the data is available:

$$(\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} .17577542 & -.019177442 \\ -.019177442 & .0020946798 \end{bmatrix},$$

$$\mathbf{x}'\mathbf{y} = \begin{bmatrix} 10759.646 \\ 98598.324 \end{bmatrix},$$

$$\mathbf{y}'\mathbf{y} = 149965.67,$$

where the **x** matrix is a $(5,000 \times 2)$ matrix consisting of a column of 1's and a column representing the natural logarithms of the observations on income, and the **y**-vector refers to the corresponding natural logarithms of food expenditures. Calculate the least-squares estimate of the parameters of the transformed model. What is your estimate of the elasticity of food expenditure with respect to income?

d. What is the probability distribution of $\hat{\beta}_2$, the least-squares estimator of $\beta_2$? Generate the MVUE estimates of the mean and variance of this distribution. (You might find it useful to know that $(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y})/(4998) = 25.369055$.) Given the assumptions of the model, and using the MVUE estimates of mean and variance, what is the (estimated) probability that the income elasticity estimated by the least-squares approach will be within $\mp.2$ of the true income elasticity?

e. Discuss any alterations to the specification of the relationship between food expenditure and disposable income that you feel is appropriate for this problem.

**5.** The research department of the *Personal Computer Monthly* magazine is analyzing the operating life of computer chips that are produced by a major manufacturer located in Silicon Valley. The research staff postulates that a random sample of lifetimes being analyzed adheres to the statistical model

$$\mathbf{X} \sim \theta^{-n} e^{-\sum_{i=1}^{n} x_i/\theta} \prod_{i=1}^{n} I_{(0,\infty)}(x_i), \quad \text{where } \theta > 0.$$

The magazine publishes an index defined by $\beta = 1/\theta$ to measure the quality of computer chips, where the closer $\beta$ is to zero, the better the computer chip. The joint density of **X** is reparameterized so that the density function is parameterized by $\beta$, as

$$\mathbf{X} \sim \beta^{n} e^{-\beta \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} I_{(0,\infty)}(x_i), \quad \text{where } \beta > 0.$$

a. Does the reparameterized family of density functions belong to the exponential class of density functions?

b. Define a set of minimal sufficient statistics for the reparameterized density function. Are the minimal sufficient statistics complete sufficient statistics?

c. Does there exist an unbiased estimator of $\beta$ whose variance achieves the Cramér–Rao lower bound?

d. Define the maximum likelihood estimator for the parameter $\beta$. Is the MLE a function of the complete sufficient statistic? Is the MLE a consistent estimator of $\beta$?

*e. Is the MLE the MVUE of $\beta$? Is the MLE asymptotically normally distributed? Asymptotically efficient?

**6.** A large commercial bank intends to analyze the accuracy with which its bank tellers process cash transactions. In particular, it desires an estimate of the expected proportion of daily cash transactions that the bank tellers process correctly. It plans to analyze 200 past observations on the daily proportion of correct cash transactions by the tellers, and specify the statistical model underlying the daily observations as

$$f(z;\alpha) = \alpha z^{\alpha-1} I_{(0,1)}(z), \quad \alpha \in (0,\infty)$$

(i.e., $f(z;\alpha)$ is a beta density function with $\beta = 1$).

a. Define the maximum likelihood estimator of $\alpha$.

b. Show that the MLE is a function of the complete sufficient statistic for this problem.

c. Is the MLE of $\alpha$ a consistent estimator?

d. It can be shown (*you* don't have to) that

$$\mathrm{E}\left(\sum_{i=1}^{n} \ln X_i\right)^{-1} = -\alpha/(n-1).$$

(See W. C. Guenther (1967), "A best statistic with variance not equal to the Cramér–Rao lower bound." *American Mathematical Monthly*, 74, pp. 993–994, or else you can derive the density of $(\sum_{i=1}^{n} \ln X_i)^{-1}$ and find its expectations—the density of $\sum_{i=1}^{n} \ln X_i$ is the mirror image (around the vertical axis at zero) of a Gamma density.) Is the MLE the MVUE for $\alpha$? If not, is there a function of the MLE that is MVUE for $\alpha$?

*e. Show that the MLE is asymptotically normal and asymptotically efficient, where

$$n^{1/2}(\hat{\alpha} - \alpha) \xrightarrow{d} N(0, \alpha^2).$$

f. Define the MLE of $q(\alpha) = \alpha/(\alpha + 1)$, which is the expected proportion of correct cash transactions.

g. Is the MLE of $q(\alpha)$ a consistent estimator?

h. Is the MLE of $q(\alpha)$ asymptotically normal and asymptotically efficient? If so, define the asymptotic distribution of the MLE estimator.

i. The outcome of the sufficient statistic was

$$\sum_{i=1}^{n} \ln x_i = -9.725.$$

Calculate the maximum likelihood estimate of $\alpha$. Calculate the estimate of $\alpha$ using the MVUE of $\alpha$.

j. Calculate the maximum likelihood estimate of $q(\alpha) = \alpha/(\alpha + 1)$, the expected proportion of correct transactions.

k. Calculate the maximum likelihood estimate of the probability that there will be greater than or equal to 97% correct cash transactions on a given day. (Hint: Determine the appropriate function of $\alpha$ in this case, and use the invariance principle.)

**7.** The personnel department of the ACME Textile Co. administers an aptitude test to all prospective assembly-line employees. The average number of garments per hour that an employee can produce is approximately proportional to the score received on the aptitude test. In particular, the relationship is represented by

$$Y_i = x_i\beta + \varepsilon_i$$

where

$y_i$ = average number of garments per hour produced by employee $i$,

$x_i$ = score of employee $i$ on aptitude test,

$\beta$ = proportionality factor,

$\varepsilon_i$ = error term,representing the deviation between actual average number of garments per hour produced by employee $i$ and the production level implied by $x_i\beta$.

You may assume that the $\varepsilon_i$'s are independent, and you may also assume that $\mathrm{E}\varepsilon_i = 0$ and $\mathrm{E}\varepsilon_i^2 = \sigma^2, \forall\, i$. Suppose you had the outcome, $\{y_1, y_2, \ldots, y_n\}$, of a random sample of average production rates for $n$ employees, together with their associated scores, $\{x_1, x_2, \ldots, x_n\}$, on the aptitude test.

a. Should the random sample $\{Y_1, \ldots, Y_n\}$ be interpreted as a random sample from some population distribution, or should it be interpreted as a random sample generated from a composite experiment? (Note: It cannot be expected that the aptitude scores will all be the same.)

b. Derive the functional form of the least-squares estimator of the proportionality factor, $\beta$. Is the least-squares estimator BLUE in this case? Is it the MVUE of $\beta$?

c. Presuming that you could increase the sample size without bound, is the estimator you derived in (b) a consistent estimator of $\beta$? Is it asymptotically normally distributed? Justify your answer, being explicit about any assumptions you have made about the behavior of the $x_i$ values and/or the $\varepsilon_i$ values.

d. From a sample of size $n = 100$, the sample outcome resulted in

$$\sum_{i=1}^{100} x_i y_i = 92{,}017 \quad \text{and} \quad \sum_{i=1}^{n} x_i^2 = 897{,}235.$$

Use the estimator you derived in (b) to generate an estimate of $\beta$.

**8.** A business consultant to the ACME Textile Co. suggests that the estimator

$$\hat{\beta}^* = (\mathbf{x}'\mathbf{x} + k)^{-1}\mathbf{x}'\mathbf{Y}$$

might be useful to consider as an alternative to the least-squares estimator of $\beta$ in the preceding problem (the estimator $\hat{\beta}^*$ is a special case of the so-called "ridge regression" estimator in the statistics literature). In this case,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

and $k$ is some positive constant.

a. Is the estimator unbiased? If the estimator is not unbiased, derive an expression for the bias.

b. Derive an expression for the variance of this estimator.

c. Is the estimator a consistent estimator of $\beta$? Justify your answer, being explicit about any assumptions you have made about the behavior of the $x_i$ values.

d. Compare the mean square errors of the least-squares estimator and the estimator $\hat{\beta}^*$. Is one estimator superior in MSE to the other $\forall\ \beta$ and $\sigma^2$? If not, can you characterize the problem conditions under which each estimator would be superior in terms of MSE?

**9.** Henri Theil, a famous economist/econometrician, analyzed the demand for textiles in the Netherlands during the period 1923–1939, using a general linear model framework. In particular, he specified the relationship between per-capita textile consumption, real price of textiles, and per-capita real income as

$$Y_t = \beta_1 p_t^{\beta_2} i_t^{\beta_3} e_t^{\varepsilon_t}, \quad t = 1923, \ldots, 1939,$$

where

$Y_t$ = per capita textile consumption in year $t$, represented as an index with base year 1925;

$p_t$ = retail price of clothing divided by a general cost-of-living index in year $t$,represented as an index with base year 1925,

$i_t$ = real income per capita in year $t$,defined as the total money income of private consumers divided by the population size and a general cost of living index, represented as an index with base year 1925,

$\varepsilon_t \sim \text{iid}\, N(0, \sigma^2), \quad t = 1923, \ldots, 1939.$

The data Theil used in estimating the demand relationship is presented in the following table.

| YEAR | $Y$ | $i$ | $p$ |
|---|---|---|---|
| 1923 | 99.2 | 96.7 | 101.0 |
| 1924 | 99.0 | 98.1 | 100.1 |
| 1925 | 100.0 | 100.0 | 100.0 |
| 1926 | 111.6 | 104.9 | 90.6 |
| 1927 | 122.2 | 104.9 | 86.5 |
| 1928 | 117.6 | 109.5 | 89.7 |
| 1929 | 121.1 | 110.8 | 90.6 |
| 1930 | 136.0 | 112.3 | 82.8 |
| 1931 | 154.2 | 109.3 | 70.1 |
| 1932 | 153.6 | 105.3 | 65.4 |
| 1933 | 158.5 | 101.7 | 61.3 |
| 1934 | 140.6 | 95.4 | 62.5 |
| 1935 | 136.2 | 96.4 | 63.6 |
| 1936 | 168.0 | 97.6 | 52.6 |
| 1937 | 154.3 | 102.4 | 59.7 |
| 1938 | 149.0 | 101.6 | 59.5 |
| 1939 | 165.5 | 103.8 | 61.3 |

a. Present the statistical model in a form that is consistent with the general linear model framework (variables should be measured in a way that coincides with the way they are used in the GLM specification). Can the random sample $(Y_1, \ldots, Y_n)$ be interpreted as a random sample from a population distribution? Why or why not?

b. Define complete (and minimal) sufficient statistics for the parameters $(\beta, \sigma)$.

c. Define the BLUE estimator of $(\ln \beta_1, \beta_2, \beta_3)$. Generate a BLUE estimate for this vector.

d. Define the MVUE for the vector $(\ln \beta_1, \beta_2, \beta_3)$. Justify that your estimator is in fact the MVUE. Generate a MVUE estimate for the vector.

e. What is the MVUE for $(\beta_2 + \beta_3)$, i.e., what is the MVUE for the degree of homogeneity of the demand function in terms of relative prices and real income? Justify that your estimator is the MVUE for $(\beta_2 + \beta_3)$. Generate an MVUE estimate of $(\beta_2 + \beta_3)$.

f. Define the probability distribution of the MVUE for $(\beta_2 + \beta_3)$. What is the probability distribution of the MVUE for $(\ln \beta_1, \beta_2, \beta_3)$?

g. Present conditions under which the MVUE of $(\ln \beta_1, \beta_2, \beta_3)$ would be: (1) a consistent estimator, and (2) an asymptotically normally distributed estimator.

h. Define the MVUE for $\sigma^2$. Justify that the estimator is, in fact, the MVUE. Generate an MVUE estimate of $\sigma^2$.

i. Present conditions under which the MVUE for $\sigma^2$ would be a consistent estimator.

j. Estimate the probability that your MVUE estimator of the price elasticity of demand will generate an estimate that is within ±.15 of the true price elasticity. You may use estimates of unknown parameters in generating this probability estimate.

k. Is $\hat{\text{E}}Y_t = e^{\hat{\beta}_1^*} p_t^{\hat{\beta}_2} i_t^{\hat{\beta}_3}$, where $(\hat{\beta}_1{}^*, \hat{\beta}_2, \hat{\beta}_3)$ is the BLUE estimator of $(\ln \beta_1, \beta_2, \beta_3)$, a consistent estimator of $\text{E}Y_t$ for given values of $p_t$ and $i_t$?

*l. Is the estimator in (k) BLUE? MVUE?

**10.** In each case ahead, determine whether, and under what assumptions, the stated relationship between $Y$ and $x$ can be represented in general linear model form such that the least-squares estimator will provide a BLUE estimator of the $\beta$ parameters.

a. $Y_i = \exp(\sum_{j=1}^{n} x_{ij}\beta_j + \varepsilon_i)$

b. $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$

c. $Y_i = \beta_0 \prod_{j=1}^{k} x_{ij}^{\beta_j} + \varepsilon_i$

d. $Y_i = x_{2i}/(1 + e^{x_{1i}/\beta + \varepsilon_i})$

**11.** The following statistical model is postulated for representing the relationship between real aggregate disposable income and real aggregate expenditure on nondurable goods:

$$Y_i = \exp(\beta_1 + \beta_2 x_i + \varepsilon_i)$$

where

$y_i$ = real aggregate expenditure on nondurables in period $i$ measured in billions of dollars,

$x_i$ = real aggregate disposable income in period $i$, measured in billions of dollars, and

$\varepsilon_i'$s are iid, with $\varepsilon_i + ab \sim \dfrac{1}{b^a\Gamma(a)}\varepsilon_i^{a-1}e^{-\varepsilon_i/b}I_{(0,\infty)}(\varepsilon_i)$.

There are 100 observations, $(y_1, x_1), \ldots, (y_{100}, x_{100})$ available to estimate the values of $\beta_1$, $\beta_2$, and $\sigma^2 = var(\varepsilon_i)$. Assuming the model specification is correct, answer the following questions:

a. Transform the model into GLM form. What parameters or functions of parameters are being estimated by the least-squares estimator applied to the transformed model?

b. Is the least-squares estimator unbiased? BLUE? Asymptotically unbiased?

c. Letting

$$\mathbf{x} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \text{if } (\mathbf{x}'\mathbf{x})^{-1} \to [\mathbf{0}] \quad \text{and } n^{-1}\mathbf{x}'\mathbf{x} \to \mathbf{Q}$$

(a symmetric, positive definite matrix) as $n \to \infty$, would it follow that the least-squares estimator is consistent and asymptotically normally distributed? Why or why not?

d. Letting

$$\mathbf{y}_* = \begin{bmatrix} \ln y_1 \\ \vdots \\ \ln y_n \end{bmatrix}$$

and $\hat{\boldsymbol{\beta}}$ be the BLUE of $\boldsymbol{\beta}$, is it true that

$$S^2 = \left(\mathbf{Y}_* - \mathbf{x}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{Y}_* - \mathbf{x}\hat{\boldsymbol{\beta}}\right) \Big/ (n-2)$$

is an unbiased and consistent estimator of $ab^2$?

**12.** In each case, indicate whether the statement regarding the relationship $\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is true or false, and justify your answer.

a. Let the random $(n \times 1)$ vector $\mathbf{Y}$ represent a random sample from some composite experiment, where $\mathrm{E}\boldsymbol{\varepsilon} = [\mathbf{0}]$, and $\mathrm{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \sigma^2\mathbf{I}$. Suppose the $\mathbf{x}$-matrix has full column rank, but that the first and second columns of $\mathbf{x}$ are nearly linearly dependent and, as a result, the determinant of $\mathbf{x}'\mathbf{x}$ is near zero, equaling $.273 \times 10^{-7}$. In this case, although $\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ is still an unbiased estimator, it is no longer BLUE (i.e., it loses its "best" property of having the smallest covariance matrix in the linear unbiased class of estimators).

b. The $\varepsilon_i$'s are homoskedastic and independent with $\mathrm{E}\varepsilon_i = \delta \neq 0 \; \forall \; i$. Also, $\mathbf{x}\boldsymbol{\beta} = \beta_1\boldsymbol{\iota} + \beta_2\mathbf{Z}$, where $\boldsymbol{\iota}$ is an $(n \times 1)$ column vector of 1's, and $\mathbf{Z}$ is an $(n \times 1)$ column vector of explanatory variable values. Then if $\hat{\boldsymbol{\beta}}$ is the least-squares estimator of $\boldsymbol{\beta}$, $\hat{\beta}_2$ is the BLUE of $\beta_2$.

c. The disturbance terms are related as $\varepsilon_t = \rho\varepsilon_{t-1} + V_t$, where the $V_t$'s are iid with $\mathrm{E}V_t = 0$ and $\mathrm{var}(V_t) = \sigma^2 \; \forall \; t$, and $|\rho| < 1$. The least-squares estimator is both BLUE and consistent.

**13.** Your company markets a disposable butane lighter called "surelight." In your product advertising, you use the slogan "lights on the first try—everytime!" As a quality check, you intend to examine a random sample of 10,000 lighters from the assembly line and observe for each lighter the number of trials required for the lighter to light. Your assistant obtains the random-sample outcome and reports to you that a total of 10,118 trials was required to get all of the lighters to light. She did *not* record the 10,000 individual outcomes of how many trials were required for each lighter to light. You are interested in estimating both the expected number of trials needed for a lighter to light and the probability, $p$, that the lighter lights on any given trial.

a. Define an appropriate statistical model for the 10,000 outcomes of how many trials were required for each lighter to light.

b. Define the MLE for the expected number of trials needed for a lighter to light. Is the estimator the MVUE? Is it consistent? Is it asymptotically normal? Is it asymptotically efficient?

c. Define the MLE for the probability that the lighter lights on any given trial. *Is the estimator the MVUE? Is it consistent? Is it asymptotically normal? Is it asymptotically efficient? (Hint: Can you show that $t(\mathbf{X}) = (n-1)/((\sum_{i=1}^{n} X_i) - 1)$ has an expectation equal to $p$?)

d. Provide MLE estimates and MVUE estimates of both the expected number of trials needed for the first light and the probability that the lighter lights on any given trial.

**14.** A regional telephone company is analyzing the number of telephone calls that are connected to wrong numbers at its telephone exchange. It collects the number of wrong telephone connections on each of 200 days and treats the observations as the outcome of a random sample of size 200 from a Poisson population distribution:

$$f(z;\lambda) = \frac{e^{-\lambda}\lambda^z}{z!} I_{\{0,1,2,\ldots\}}(z).$$

a. Define the MLE of $\lambda$, the expected number of wrong connections per day.

b. Is the MLE the MVUE for $\lambda$? Is it consistent? Asymptotically normal? Asymptotically efficient?

c. If $\sum_{i=1}^{200} x_i = 4,973$, what is the ML estimate of the expected number of wrong connections? If each wrong connection costs the company \$0.70, define a MLE for the expected daily cost of wrong connections, and generate a ML estimate of this cost.

d. Define a MLE for the standard deviation of the daily number of wrong connections. Is the MLE consistent? Is it asymptotically normal? Generate a ML estimate of the standard deviation.

**15.** The number of minutes past the scheduled departure time that jets with no mechanical problems leave the terminal in an overcrowded airport in the northeast are iid outcomes from a uniform population distribution of the form $f(z;\Theta) = \Theta^{-1} I_{[0,\Theta]}(z)$. A random sample of 1,000 departures is to be used to estimate the parameter $\Theta$ and the expected number of minutes past the scheduled departure time that a jet will leave the terminal. Summary statistics from the outcome of the random sample include $\min(\mathbf{x}) = .1$, $\max(\mathbf{x}) = 13.8$, $\bar{x} = 6.8$, $s^2 = 15.9$.

a. Define a MLE for $\Theta$ and for the expected number of minutes past the scheduled departure time that a jet will leave the terminal. Are these MLEs functions of minimal sufficient statistics?

b. Use the MLEs you defined above to generate ML estimates of the respective quantities of interest.

c. Are the estimators in (a) unbiased? consistent? (Hint: $\mathrm{E}\max(X) = \Theta[n/(n+1)]$ and $\mathrm{E}(\max(X))^2 = \Theta^2[n/(n+2)]$.)

*d. Are the estimators in (a) MVUES?

**16.** Define MLEs for the following problems:

a. Estimating the $p_i$'s based on a random sample from a multinomial population distribution.

b. Estimating the unknown parameters in the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ based on a random sample from a bivariate normal distribution.

c. Estimating $\beta$ in the population distribution

$$f(x;\beta) = \frac{2x}{\beta}\exp\left[\frac{-x^2}{\beta}\right] I_{(0,\infty)}(x).$$

d. Estimating the parameter $p$ in a negative binomial population distribution, where $r$ is known.

**17.** Define MOM estimators in each of the following cases. Are estimators consistent? Are they asymptotically normal?

a. Estimating $\mu$ and $\sigma^2$ based on a random sample from a normal population distribution.

b. Estimating $\alpha$ and $\beta$ based on a random sample from a Beta population distribution.

c. Estimating $p$ based on a random sample from a geometric population distribution.

d. Estimating $a$ and $b$ based on a random sample from a continuous uniform population distribution.

***18.** (*Generalized least squares estimator*) Consider the linear model $\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ in which all of the classical assumptions apply, except that $\mathrm{E}\boldsymbol{\varepsilon\varepsilon}' = \boldsymbol{\Phi} \neq \sigma^2\mathbf{I}$. Consider a GMM estimator of the parameter vector $\boldsymbol{\beta}$ based on the moment conditions

$$\mathrm{E}g_t(Y_t,\boldsymbol{\beta}) = \mathrm{E}\mathbf{z}'_{t.}(Y_t - \mathbf{x}_{t.}\boldsymbol{\beta}) = [\mathbf{0}],$$

where $\mathbf{z} = \boldsymbol{\Phi}^{-1}\mathbf{x}$.

a. Assuming temporarily that $\boldsymbol{\Phi}$ were known, identify the sample moment conditions that would be used to define the GMM estimator. Solve the moment conditions to provide an explicit functional representation of the estimator (called the *generalized least squares estimator* in the literature).

b. Discuss conditions under which the estimator you have defined in (a) is consistent, asymptotically normal, and asymptotically efficient.

c. For the general case where $\boldsymbol{\Phi}$ is unknown, discuss how you would define an operational version of the GMM estimator, and discuss its relationship with the estimator in (a).

# 9 Elements of Hypothesis-Testing Theory

## 9.1 Introduction

A primary goal of scientific research often concerns the verification or refutation of assertions, conjectures, currently accepted laws, or descriptions relating to a given economic, sociological, physical, or biological process or population. Statistical hypothesis testing concerns the use of a random sample of observations from the process or population being studied, together with probability and mathematical statistics principles, to judge the validity of stated assertions, conjectures, laws, or descriptions in such a way that the probability of making incorrect decisions can be controlled. The precise nature of the types of errors that can be made and how, and in what sense, the probabilities of such errors can be controlled is the subject of this chapter.

We will henceforth use the word *hypothesis* generically to refer to any statement relating to a process or population under study for which a researcher wishes verification or refutation. Examples of hypotheses that might be tested include statements such as the following:

1. The demand for fresh asparagus in the United States is price elastic;
2. A newly formulated feeding ration results in higher milk production per cow than the formulation currently being fed to dairy cows in a commercial dairy operation;
3. A new computer chip has an expected usable life span of over 50,000 hours; and

4. Fewer than 1 in 10,000 college-educated individuals engages in violent criminal activity in a given year in North America.

## 9.2 Statistical Hypotheses

In order to apply probability and mathematical statistics principles to the problem of hypothesis testing, it is required that a hypothesis be translated into statements relating to the characteristics of the probability space associated with a random sample from the random process or population being analyzed. In particular, a *statistical* hypothesis is a set of potential probability distributions for the random sample **X**. The statistical hypothesis is defined by statements of the characteristics asserted to be true about the probability space associated with **X**. These statements are often in the form of assertions about the values of parameters or functions of parameters relating to the true PDF of **X**. Hypotheses will be *statistically testable* only if they can be represented in terms of a statistical hypothesis.

**Definition 9.1**
***Statistical hypothesis***

A set of potential probability distributions for a random sample from a process or population is called a **statistical hypothesis.**

If a statistical hypothesis consists of a single element such that a specific probability distribution for **X** is completely and uniquely identified, the statistical hypothesis will be called a **simple hypothesis.** If the statistical hypothesis is not simple, then it is called a **composite hypothesis.** A composite hypothesis is a statistical hypothesis containing two or more potential probability distributions for the outcomes of the random sample.

It is customary to represent the set of probability distributions that defines a statistical hypothesis by the capital letter $H$ and to use subscripts, when needed, to distinguish between various statistical hypotheses under investigation, such as $H_0$, $H_a$, or $H_i$. Regarding the set-defining conditions that are used to represent a statistical hypothesis, the researcher is responsible for identifying the characteristics of the probability space of **X** that are of interest in any given problem setting. The following examples illustrate the representation of statistical hypotheses.

**Example 9.1** A manufacturer of metal hardware claims that the percentage of defective bolts in a shipment of 1,000 bolts is no more than 2 percent. The receiver of the shipment intends to use a random sample, with replacement, of size $n$ from the bolt shipment to assess the validity of the manufacturer's assertion. The hypothesis of the manufacturer, translated into a statistical hypothesis stated in terms of a random sample of size $n$ from the bolt population, (where $x_i = 1$

indicates a defective bolt, and $x_i = 0$ represents a nondefective bolt), is

$$H = \left\{ f(\mathbf{x}; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} I_{\{0,1\}}(x_i), \quad p \in [0, .02] \right\}.$$

This statistical hypothesis is a *composite* hypothesis and specifies that the probability distribution of the random sample of size $n$ from the bolt population is the product of $n$ Bernoulli distributions, with the probability of observing a defective bolt in any Bernoulli trial being $p \leq .02$. The statistical hypothesis implies a *range* of potential probability distributions for defective bolts. □

**Example 9.2** It is hypothesized that the expected life of a computer chip manufactured by a certain manufacturing process exceeds the industry average of 20,000 hours, and it is further hypothesized that the lifetimes of computer chips manufactured by this process are distributed according to the exponential family of distributions. The statistical hypothesis, stated in terms of a random sample of size $n$ from the population distribution of computer chip lifetimes (assuming elements of $\mathbf{X}$ are measured in thousands of hours), is

$$H = \left\{ f(\mathbf{x}; \theta) = \theta^{-n} \exp\left(-\sum_{i=1}^{n} x_i/\theta\right) \prod_{i=1}^{n} I_{(0,\infty)}(x_i), \theta \in (20, \infty) \right\}.$$

This statistical hypothesis is *composite* since a single probability distribution is not completely and uniquely defined by $H$—the mean life measured in thousands of hours, $\theta$, is any number larger than 20, according to $H$. Thus, the statistical hypothesis implies a *range* of potential probability distributions for the random sample of computer chip lifetimes. □

**Example 9.3** Suppose in Ex. 9.1 that the hypothesis was that there are *no* defective bolts in the 1,000-bolt shipment. Then the statistical hypothesis could be specified in terms of a random sample of size $n$ from the bolt population as

$$H = \left\{ f(\mathbf{x}) = \prod_{i=1}^{n} I_{\{0\}}(x_i) \right\}.$$

This is a *simple* statistical hypothesis indicating that the probability distribution is degenerate with the outcome $\mathbf{x} = [\mathbf{0}]$ (all $x_i$'s are nondefective) having probability 1. The statistical hypothesis could have been equivalently represented as in 9.1, but with $p = 0$ instead of $p \in [0, .02]$. □

**Example 9.4** The average yield per acre for a particular variety of wheat grown in a certain region of the country is represented by

$$Y_i = \beta_0 + \beta_1 f_i + \beta_2 r_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

where $y_i$ is the average yield per acre measured in bushels in year $i$, $f_i$ is the average pounds of fertilizer applied per acre, and $r_i$ is the average rainfall per acre measured in inches. Independent observations on 30 years' worth of values for $f_i$ and $r_i$ and corresponding outcomes of $Y_i$ are available. A researcher wishes to assess the hypothesis that a 1 pound per acre increase in fertilizer applied to the crop generates an expected wheat yield increase of .25 bushels per acre. Note that the implied probability distribution for $Y_i$ is given by $N(\beta_o+\beta_1 f_i+\beta_2 r_i, \sigma^2)$. Also note that the observations on wheat yields can be conceptualized as the vector outcome of a random sample generated via a composite experiment, with the joint probability distribution of $\mathbf{Y} = [Y_1, \ldots, Y_{30}]'$ being given by

$$\mathbf{Y} \sim \prod_{i=1}^{30} N(y_i; \beta_0 + \beta_1 f_i + \beta_2 r_i, \sigma^2).$$

Translating the hypothesis of the researcher into a statistical hypothesis yields

$$H = \{f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{30} N(y_i; \beta_0 + \beta_1 f_i + \beta_2 r_i, \sigma^2), \beta_1 = .25\}.$$

The statistical hypothesis is *composite*, since a unique probability distribution for $\mathbf{Y}$ has not been completely identified because $\beta_0$, $\beta_2$, and $\sigma^2$ were left unspecified and thus can take any values that result in a legitimate normal density function. (Normal distributions are clearly only approximations in this case, since yields cannot be negative; it is assumed that $P(y_i < 0)$ is negligible for the relevant $f_i$ and $r_i$ values being analyzed.) □

**Example 9.5** The average annual salary of accountants in a certain region of the county is \$38,250. An allegation is made that male accountants in this region have lower than average salaries, and a researcher wishes to assess the allegation. It is not known which parametric family of PDFs contains the distribution of male accountants' salary in the region. Translating the allegation into a statistical hypothesis concerning a random sample of size $n$ from the population distribution of male accountants in the region yields

$$H = \{f(\mathbf{x}; \Theta) = \prod_{i=1}^{n} m(x_i; \Theta), \quad \mu = \mathrm{E}X_i = h(\Theta) < 38,250 \quad \forall\, i\}.$$

The statistical hypothesis is *composite*, since $H$ contains *all* PDFs for $\mathbf{X}$ that have $\mathrm{E}X_i < 38,250\ \forall\, i$. □

When defining $H$, choices of functional forms for $f(\mathbf{x})$ are often suggested by the nature of the random process or population being analyzed. For example, in Ex. 9.1, the Bernoulli family of density functions was dictated by the fact that the random sample was defined via random sampling, with replacement, from a finite population of defective/nondefective bolts. In Ex. 9.2, the exponential family of distributions was presumably motivated by certain physical features of the computer chips under study (nonnegativity of lifetimes is obvious; the exponential family of density functions might be suggested by a "good as new while functioning" characteristic of the chip—recall the "memoryless property" of the exponential family of distributions). In Ex. 9.3, the stated hypothesis directly implied that only one outcome of $\mathbf{X}$ was possible, so that the probability distribution of **X** must be degenerate. In Ex. 9.4, the normal family of distributions might have been motivated via an appeal to a central limit theorem, where the error term, $\varepsilon_i$, would be assumed to represent the additive effect of a large number of random influences not specifically accounted for in the simple linear relationship specified to represent $\mathrm{E}Y_i$. In Ex. 9.5, there was no indication of the functional form of the probability distribution except for the range of possible values for its mean.

When one represents statistical hypotheses, a simplification is often introduced when the functional form of the joint density of **X** is taken to be known except for the value of some parameter vector $\Theta$, as in Ex. 9.1 through Ex. 9.4. In this case, $H$ is often represented as simply a set of *parameter values*, the idea being that if $f(\mathbf{x}; \Theta)$ is known except for the value of $\Theta$, then $\Theta \in H$ will characterize a set of potential probability distributions for **X** perfectly well and thus provide an equivalent representation of a statistical hypothesis.

**Example 9.6** Revisit Ex. 9.1 through Ex. 9.4. Assuming the functional form for the joint density of **X** to be given as stated in each example, the statistical hypotheses can be represented in *abbreviated* form as $H = \{p\colon\ 0 \leq p \leq .02\}$, $H = \{\theta\colon\ \theta > 20\}$, $H = \{p\colon\ p = 0\}$, and $H = \{(\beta, \sigma^2)\colon\ \beta_1 = .25\}$, respectively. □

The representation of statistical hypotheses in the preceding case is often abbreviated still further, using general notation of the form ***H*: set-defining conditions**, meaning that $H$ is a set of parameter values defined by whatever set-defining conditions are stated following the colon. This further abbreviation of the representation of $H$ is illustrated below for the case of Ex. 9.1 through Ex. 9.4.

**Example 9.7** Revisit Ex. 9.6. Alternative abbreviated representations for the statistical hypotheses of Ex. 9.1 through Ex. 9.4 are given by $H\colon\ 0 \leq p \leq .02$, $H\colon\ \theta > 20$, $H\colon\ p = 0$, and $H\colon\ \beta_1 = .25$, respectively. □

In general, the use of abbreviated representations of $H$ are acceptable so long as the context of the problem being analyzed results in no ambiguity in defining the set of potential probability distributions for **X** constituting the statistical hypothesis. The notation is quite prevalent in cases of *parametric hypothesis testing*, where it is assumed at the outset that the probability distribution of **X** is a member of a given parametric family of PDFs. We will examine parametric hypothesis testing in considerable detail beginning in Section 9.4.

For later reference, note that the specification of a statistical hypothesis $H$ defines concurrently a set of potential outcomes for the random sample. That is, each PDF in $H$ will imply a particular range of possible outcomes for the random sample **X**, and then the set of possible random-sample outcomes implied by the *entire collection* of PDFs in $H$ is logically the union of all of the respective ranges for **X**. The idea is that since any of the PDFs in $H$ is asserted to be a candidate for the true PDF of **X**, then any of the associated ranges of **X** is also a candidate for $R(\mathbf{X})$. Assuming henceforth that the range of **X** is synonymous with the support of **X**'s PDF, we can define the set of potential random sample outcomes implied by $H$ as

$$R(\mathbf{X} \mid H) = \{\mathbf{x}: f(\mathbf{x}; \Theta) > 0 \quad \text{and} \quad f(\mathbf{x}; \Theta) \in H\}.$$

We will refer to $R(\mathbf{X} \mid H)$ as **the range of $X$ over $H$.**[1] Analogous definitions for the range of **X** over $\bar{H}$ and over $H \cup \bar{H}$ can be given as

$$R(\mathbf{X} \mid \bar{H}) = \{\mathbf{x}: f(\mathbf{x}; \Theta) > 0 \quad \text{and} \quad f(\mathbf{x}; \Theta) \in \bar{H}\},$$

$$R(\mathbf{X} \mid H \cup \bar{H}) = \{\mathbf{x}: f(\mathbf{x}; \Theta) > 0 \quad \text{and} \quad f(\mathbf{x}; \Theta) \in H \cup \bar{H}\}.$$

For now, it will suffice to interpret the complement of $H$, $\bar{H}$, as the collection of all probability distributions not specified in $H$ that are potential candidates for the true PDF of the random sample under study. Thus, the complement operation is being applied in the context of a universal set of all potential probability distributions relevant for the random sample being examined. Later we will be more specific about the probability distributions represented by $\bar{H}$. Then $R(\mathbf{X} \mid H \cup \bar{H})$ can be interpreted as the set of *all* potential outcomes of the random sample.

It can be the case that $R(\mathbf{X} \mid H) = R(\mathbf{X} \mid \bar{H}) = R(\mathbf{X} \mid H \cup \bar{H})$, which occurs when all of the supports of the PDFs in $H$ and $\bar{H}$ are the same. Such is the case in Ex. 9.1, where all three sets equal $\times_{i=1}^{n}\{0, 1\}$. Alternatively, the sets can differ, as in Ex. 9.3, where these sets are respectively $\{\mathbf{0}\}$, $\times_{i=1}^{n}\{0, 1\}$, and $\times_{i=1}^{n}\{0, 1\}$. The nature of the ranges of **X** over $H$, $\bar{H}$, and $H \cup \bar{H}$ will play pivotal roles in determining whether an ideal statistical test of $H$ exists (i.e., a test that makes no errors in deciding the validity of $H$) and are also useful for motivating why one cannot generally expect statistical tests of the truth of $H$ to be errorfree. We examine these ideas in the next section.

## 9.3 Basic Hypothesis-Testing Concepts

In this section we define the concept of a statistical hypothesis test, we identify two types of errors such a test can make, and we indicate in what sense such errors can be controlled. We also point out the general fact that the incidence of one type of test error can be lessened only at the expense of increasing the incidence of the other if the size of the random sample on which the test is based is held constant.

### Statistical Hypothesis Tests

The objective of a *test* of a statistical hypothesis is to assess the validity of the statistical hypothesis. In practice, a statistical hypothesis is tested by first obtaining the outcome of a random sample from the process or population being

[1] Compare this set to the *range of* X *over* Ω introduced in our discussion of minimal sufficient statistics, Section 7.4.

analyzed and then using the observed outcome of the random-sample to assess whether it is reasonable to conclude that one of the probability distributions for the random sample implied by the statistical hypothesis could have governed the outcome behavior of the random sample. Deciding when an outcome of a random sample will cause rejection of $H$ and when it will not is accomplished by a rule or procedure that partitions the range of potential random-sample outcomes into two subsets. The hypothesis is then rejected or not rejected depending on to which subset the outcome of the random sample belongs. In its most general sense, a *test of a statistical hypothesis* can be defined as follows.

**Definition 9.2**
***Test of a statistical hypothesis***

> This is a rule or procedure, based on the outcome of a random sample from the process or population under study, used to decide whether to reject a statistical hypothesis.

A test of a statistical hypothesis is often referred to simply as a **statistical test**. It follows from Def. 9.2 that a valid functional representation of a statistical test would be in the form of an indicator function whose domain elements are the potential outcomes of the random sample and whose range elements indicate whether or not the statistical hypothesis is rejected. A statistical test is fully defined when the set of potential random-sample outcomes is partitioned into two disjoint subsets, one called the **rejection** or **critical region** and the other called the **acceptance region**. Letting the set of potential random-sample outcomes be represented by $C_r \cup C_a$, where $C_r$ is the rejection (or critical) region, $C_a$ is the acceptance region, and $C_r \cap C_a = \emptyset$, then

$$I_{C_r}(\mathbf{x}) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} \text{reject } H \\ \text{accept } H \end{bmatrix}$$

is a fully defined rule for deciding whether to reject the statistical hypothesis (see Figure 9.1).

In the remainder of our discussion of hypothesis-testing concepts and methods, if an explicit definition of the acceptance region, $C_a$, is not given in any hypothesis-testing situation, it will always be tacitly assumed that $C_a = \bar{C}_r$. The complement operation will be interpreted to occur within a universal set defined by $R(\mathbf{X} \mid H \cup \bar{H})$, which represents all possible outcomes of the random sample under PDFs in either $H$ or $\bar{H}$. With this convention in mind, a statistical test can be defined completely in terms of a critical region, $C_r$, as was the case in

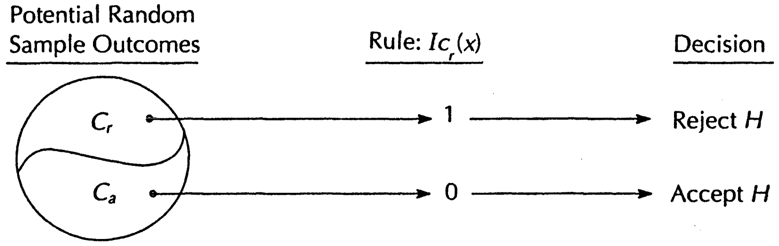


**Figure 9-1**
Test of a statistical hypothesis.

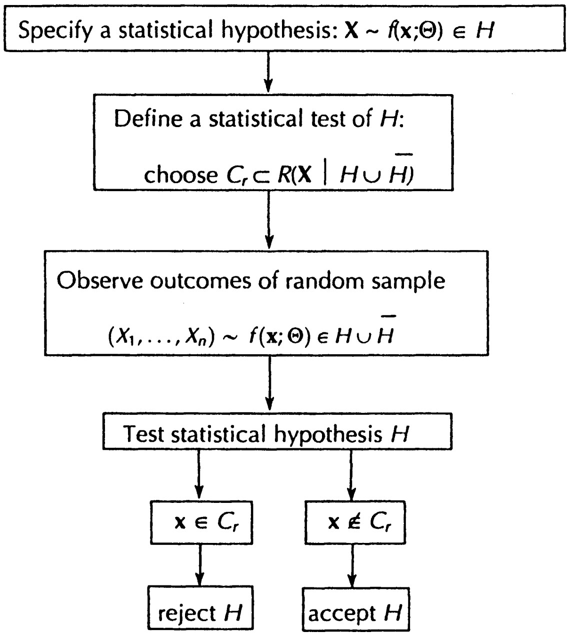

Specify a statistical hypothesis: $\mathbf{X} \sim f(\mathbf{x};\Theta) \in H$
Define a statistical test of $H$:
choose $C_r \subset R(\mathbf{X} \mid H \cup \bar{H})$
Observe outcomes of random sample
$(X_1,\ldots,X_n) \sim f(\mathbf{x};\Theta) \in H \cup \bar{H}$
Test statistical hypothesis $H$
$\mathbf{x} \in C_r$
$\mathbf{x} \notin C_r$
reject $H$
accept $H$

**Figure 9-2**
General hypothesis testing procedure

our indicator function representation above. Furthermore, we will sometimes use the phrase **statistical test defined by $C_r$** to mean a statistical test defined as $\mathbf{x} \in C_r \Rightarrow$ reject $H$, $\mathbf{x} \notin C_r$ (or equivalently, $\mathbf{x} \in \bar{C}_r = C_a$) $\Rightarrow$ accept $H$.

A schematic overview of the general context of a hypothesis-testing problem appears in Figure 9.2. The design or choice of a statistical test is seen to be equivalent to the design or choice of the partition of the set of potential random-sample outcomes into rejection and acceptance regions. This begs the question of how the partition should be chosen so as to design a test that is "good" in some appropriate sense. Operationally, this will amount to designing tests that minimize the probability of making incorrect decisions regarding the validity of hypotheses. The types of incorrect decisions that can be made using statistical tests are examined next.

### Type I Error, Type II Error, and Ideal Statistical Tests

Given a situation where the researcher must decide whether to accept or reject a statistical hypothesis, $H$, and given that in reality the statistical hypothesis is either true or false, then there are four possible states of affairs with respect to test decisions and their validity, as summarized in Figure 9.3.

The two potential error situations that can occur when deciding to accept or reject $H$ are given distinct, albeit somewhat uninspired, names so that they can be clearly distinguished.

**Definition 9.3**
***Type I and type II error***

Let $H$ be a statistical hypothesis being tested for acceptance or rejection. Then the two types of errors that can be made by the statistical test are

1. **type I error**: rejecting $H$ when $H$ is true,
2. **type II error**: accepting $H$ when $H$ is false.

In other words, a type I error is committed when the statistical test mistakenly indicates that $H$ should be rejected (i.e., the true probability distribution of $\mathbf{X}$ is an element of $H$, but the test outcome indicates that the true probability distribution is not in $H$). A type II error is committed when the statistical test mistakenly indicates that $H$ should be accepted (i.e., the true probability distribution of $\mathbf{X}$ is not an element of $H$, but the test outcome nonetheless indicates that the true probability distribution is in $H$).

Clearly, the **ideal statistical test** would be such that once an outcome of the random sample were observed, the hypothesis would *always* be correctly identified as being either true or false, and thus *no errors* would be made. For such an ideal statistical test to exist, it must be possible to partition the range of potential random-sample outcomes a priori in such a way that outcomes in the acceptance region, $C_a$, would occur iff $H$ were true and outcomes in the rejection region, $C_r$, would occur iff $H$ were false. To define such a partition, it is clear that $R(\mathbf{X} \mid H)$ and $R(\mathbf{X} \mid \bar{H})$ need to be disjoint so that, *with certainty*, $\mathbf{x} \in C_a = R(\mathbf{X} \mid H) \Rightarrow H$ is true and $\mathbf{x} \in C_r = R(\mathbf{X} \mid \bar{H}) \Rightarrow H$ is false. The following example illustrates a case where an ideal statistical test can be defined.

**Example 9.8** An envelope manufacturing machine is such that when all parts are functioning properly, the machine produces boxes containing 1,000 envelopes that never contain more than 2 defective envelopes, with the probability distribution of $X$, the number of defectives per box, being of the binomial form

$$f(x; p) = (2!/[x!(2-x)!])p^x(1-p)^{2-x} I_{\{0,1,2\}}(x),$$

for $p \in (0, 1)$. The machine will continue to operate if one of the teeth on the main drive gear breaks, but then the distribution of defectives per box of 1,000 envelopes changes such that the box will always contain at least 6 defective envelopes, with the probability distribution of the number of defectives per box

| | | True Probability Distribution | |
|---|---|---|---|
| | | $H$ | $\bar{H}$ |
| Test Decision | $H$ | Correct | Type II error |
| | $\bar{H}$ | Type I error | Correct |

**Figure 9-3**
Potential outcomes of statistical test relative to true probability distribution

changing to the distribution of $Y = X + 6$, i.e.,

$$g(y; p) = (2!/[(y-6)!(8-y)!])p^{y-6}(1-p)^{8-y}I_{\{6,7,8\}}(y)$$

for $p \in (0, 1)$. Any other machine problem will cause the machine to shut down, so that no envelopes will be produced. A quality-control engineer wants to test the hypothesis that all parts of the envelope machine are working properly. Her statistical hypothesis, in terms of a random sample of one box of envelopes, is

$$H = \{f(x; p), p \in (0, 1)\}.$$

The engineer designs the statistical test as follows:

$$x \in C_a = \{0, 1, 2\} \Rightarrow \text{accept } H \quad \text{and} \quad x \in C_r = \{6, 7, 8\} \Rightarrow \text{reject } H.$$

Note that the test is an *ideal statistical test* which produces no decision error of either type. For example, if the engineer were to observe that $x = 1$ in a box of envelopes she randomly chose to examine, she would conclude that $H$ is true; all parts of the machine are working properly. She is certain the assessment is correct, since the statistical test used in making the decision is an ideal statistical test. □

If $R(\mathbf{X} \mid H) \cap R(\mathbf{X} \mid \bar{H}) \neq \emptyset$, which is virtually always the case in practice, then there are potential outcomes of the random sample that reside simultaneously in the supports of some PDFs in $H$ and some PDFs in $\bar{H}$. A random-sample outcome equal to one of the points in $R(\mathbf{X} \mid H) \cap R(\mathbf{X} \mid \bar{H})$ cannot be used to resolve *with certainty* whether or not $H$ is true, since such points are not uniquely associated with either $H$ or $\bar{H}$. It follows that there would not exist a dichotomous partition of $R(\mathbf{X} \mid H \cup \bar{H})$ that would define a statistical test that, with a priori certainty, could always correctly decide the truth of $H$, since any dichotomous partition would include some points that were in both $R(\mathbf{X} \mid H)$ and $R(\mathbf{X} \mid H)$.

If it were true that $R(\mathbf{X} \mid H) \neq R(\mathbf{X} \mid \bar{H})$, then a *trichotomous* partition of $R(\mathbf{X} \mid H \cup \bar{H})$ can be devised which leads to a decision rule that makes no errors in deciding the truth of $H$, but the rule will also result in situations in which random-sample outcomes lead to no decision whatsoever. The rule could be based on the sets $C_n = R(\mathbf{X} \mid H) \cap R(\mathbf{X} \mid \bar{H})$, $C_a = R(\mathbf{X} \mid H) - C_n$, and $C_r = R(\mathbf{X} \mid \bar{H}) - C_n$, as

$$\mathbf{x} \in C_a \Rightarrow \text{accept } H, \qquad \mathbf{x} \in C_r \Rightarrow \text{reject } H, \qquad \text{and } \mathbf{x} \in C_n \Rightarrow \text{no decision.}$$

While making no errors, such a rule can be wasteful of sample information, especially when outcomes in the set $C_n$ have a high probability of occurring. Furthermore, in the majority of hypothesis-testing applications, $R(\mathbf{X} \mid H) = R(\mathbf{X} \mid \bar{H})$, so that not only is an ideal statistical test not available, but the preceding trichotomous partition of $R(\mathbf{X} \mid H \cup \bar{H})$ is not available either. In practice, it is therefore generally the case that errorfree statistical tests *cannot* be defined. Instead, one must seek to define statistical tests that control the incidence of errors such that they occur at "acceptable levels." We examine this idea next.

### Controlling Type I and II Errors

In applications, a statistical test can often be designed to *control* type I and type II errors. By "controlling" the errors, we mean that $C_a$ and $C_r$ are chosen in a way such that the error *probabilities* are known, and thus *controlled*, in advance of performing the test and are at levels that are, in some sense, acceptable to the researcher.

In order to illustrate how a statistical test can be designed to "control" errors, recall Ex. 9.3. In this case, $R(\mathbf{X} \mid H) = \{\mathbf{0}\}$, $R(\mathbf{X} \mid \bar{H}) = \times_{i=1}^{n}\{0, 1\}$, and thus $R(\mathbf{X} \mid H \cup \bar{H}) = \times_{i=1}^{n}\{0, 1\}$, where the definition of $R(X \mid \bar{H})$ follows from the fact that if $\bar{H}$ is true, then the population distribution for the problem is a Bernoulli density with $p > 0$, and so both 0 and 1 are potential outcomes of each $X_i$. Also note that $R(\mathbf{X} \mid H) \cap R(\mathbf{X} \mid \bar{H}) = \{\mathbf{0}\} \neq \emptyset$, so that no ideal statistical test exists for this problem.

Suppose a statistical test of $H$ is defined by the following critical region for a random sample of size 100: $C_r = \{\mathbf{x}: \sum_{i=1}^{100} x_i > 0\}$. In other words, if the random sample of size 100 results in no defectives, we will decide $H$ is true, while if any defectives are observed in the outcome of the random sample, $H$ will be declared false. Using this test, how probable are type I and type II errors? Determining the probability of type I error is straightforward in this case, since *if $H$ is* true, then $P(\mathbf{x} \notin C_r) = P(\mathbf{x} = \{\mathbf{0}\}) = 1$, which follows from the fact that $P(x_i = 0) = 1\ \forall\ i$, so that $P(\mathbf{x} = \{\mathbf{0}\}) = P(x_i = 0, \forall\ i) = \prod_{i=1}^{100} P(x_i = 0) = 1$ (recall that the $X_i$'s are independent random variables). It follows that if $H$ is true, the probability of rejecting $H$ is $P(\text{type I error}) = P(\mathbf{x} \notin C_a) = P(\mathbf{x} \in C_r) = 0$. Note that having "controlled" the type I error to have probability zero, we have substantial confidence that *$H$ is false* if the test *rejects $H$*, since the probability is 0 that the test would mistakenly reject $H$.

Now examine the probability of type II error. Assuming $H$ is false, then $X_i$ has a Bernoulli distribution with the probability, $p$, of observing a defective bolt in any Bernoulli trial being an element in the set $\{.001, .002, \ldots, 1.000\}$ (why?). Thus, the probability of type II error depends on *which* probability distribution in $\bar{H}$ is assumed to be the correct one. Since we do not know which is correct (or else we would have no need to test a statistical hypothesis in the first place), we examine the range of possibilities for type II error as a function of $p$. In this case the parameter $p$ serves as an index for the 1,000 possible alternative probability distributions for $\mathbf{X}$ that are contained in $\bar{H}$. Figure 9.4 provides a partial graph of the functional relationship between $p$ and the probability of type II error (see graph labeled $n = 100$).[2] Note that in this case, a probability of type II error is equal to $P(\mathbf{x} \notin C_r) = P(\sum_{i=1}^{100} x_i = 0)$ based on a binomial distribution for $\sum_{i=1}^{100} X_i$ with a given value of $p \in \bar{H}$ and $n = 100$. This definition of type II error probability is motivated from the fact that the event $\sum_{i=1}^{n} x_i = 0$ leads to the acceptance of $H$, and $p \in \bar{H}$ indicates that $H$ is false, so that together these two conditions define a situation where a false $H$ is accepted, i.e., a type II error is

[2] For convenience, we have chosen to "connect the dots" and display the graph as a continuous curve. We will continue with this practice wherever it is convenient and useful.

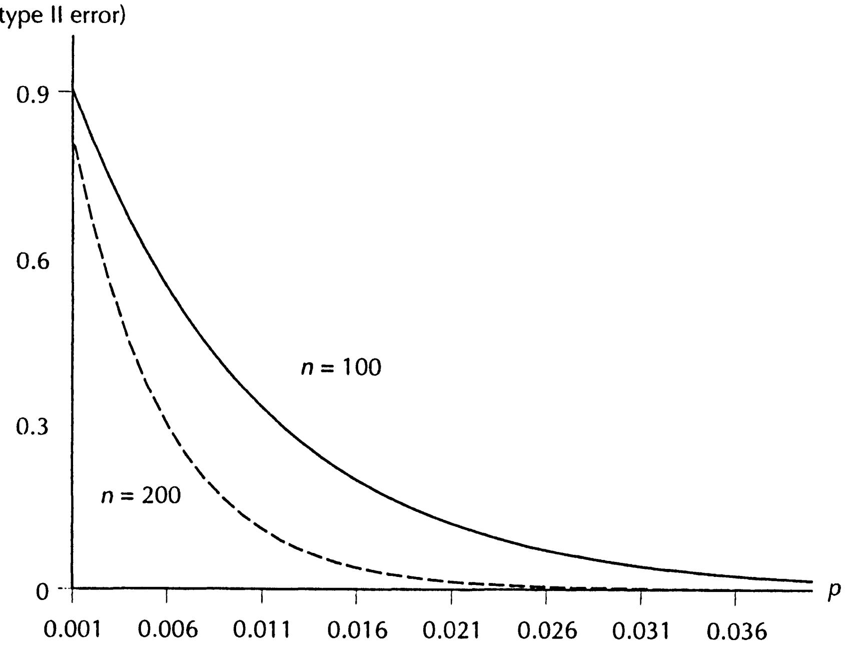


**Figure 9-4**
Probabilities of type II error for sample sizes: $n = 100$ and $n = 200$

committed. The functional relationship between $p$ and the level of type II error, which indicates the probability levels at which type II error has been controlled for the various values of $p \in \bar{H}$, is given by

$$P(\text{type II error}) = h(p) = (1 - p)^{100} \quad \text{for } p \in \{.001, .002, \ldots, 1.000\}.$$

Figure 9.4 indicates that the probability of committing a type II error declines rapidly as $p$ increases. When $p$ is only .03, the probability of type II error is less than .05. Note, however, that when $p$ is very close to its hypothesized value in $H$, i.e., $p = 0$, the probability of type II error can be quite high. For example, if $p = .001$, then $P(\text{type II error}) = .905$, or if $p = .002$, then $P(\text{type II error}) = .819$. Thus, in practice, if the bolt shipment has very few defectives (implying a very small $p$), then the proposed statistical test will not be very effective in rejecting the assertion that the shipment has no defectives, i.e., the probability of rejection will not be high.

In general, the probabilities of type I error are equal to the values of $P(\mathbf{x} \in C_r)$ for each of the PDFs in $H$. The probabilities of type II error equal the values of $P(\mathbf{x} \notin C_r)$ for each of the PDFs in $\bar{H}$. If the error probability characteristics of a given statistical test are unacceptable, the researcher generally has two options, which can also be pursued simultaneously—she can *choose a different critical region*, $C_r$, and thus define a different statistical test, or she can *alter the size of the random sample* on which to base the test. We continue examining the hypothesis in Ex. 9.3 to illustrate these options. First, suppose that rather than using a random sample size of 100, we increase the sample size to 200 but otherwise use a test rule analogous to the preceding rule to test the hypothesis that there are no defectives in the 1,000-bolt shipment. Stated in terms of the parameter $p$, the statistical hypothesis, $H$, remains precisely as before, $H: p = 0$. The probability of type I error remains equal to zero, since *if H were true*, then $P(\text{type I error}) = 1 - P(\mathbf{x} = [\mathbf{0}]) = 1 - P(\sum_{i=1}^{200} x_i = 0) = 0$, which can be demonstrated following the same reasoning as before except that now we are dealing with $n = 200$ independent random variables as opposed to $n = 100$.

The probability of type II error as a function of $p$ is now given by $P(\sum_{i=1}^{200} x_i = 0)$ calculated using the binomial distribution of $\sum_{i=1}^{200} X_i$ with a given value of $p > 0$ and $n = 200$. The functional relationship between $p$ and the level of type II error can be represented by

$$P(\text{type II error}) = h(p) = (1-p)^{200} \quad \text{for } p \in \{.001, .002, \ldots, 1.000\},$$

and a partial graph of $P(\text{type II error})$ as a function of $p$ is given in Figure 9.4 (see the graph labeled $n = 200$). It is seen that the probability of type II error has been uniformly lowered for all $p > 0$. For example, the probability of type II error is less than .05 when $p$ is only .015. The probability of type II error is appreciably reduced even for very small values of $p$, although the probability remains high for values of $p$ close to the hypothesized value of 0 (when $p = .001$ or .002, the probability is .819 or .670, respectively).

If the sample size were increased further, the probabilities of type II error would continue to decline uniformly as a function of $p$. This becomes clear upon examining the functional relationship between $p$ and the probability of type II error for an unspecified sample size of $n$, as

$$P(\text{type II error}) = h(p,n) = (1-p)^n \quad \text{for } p \in \{.001, .002, \ldots, 1.000\}.$$

Note that $dh(p,n)/dn = (1-p)^n \ln(1-p) < 0 \; \forall \; p \in (0,1)$ since $\ln(1-p) < 0$, and so the probability of type II error is a decreasing function of $n \; \forall \; p \in (0,1)$ (when $p = 1$, the probability of type II error is $0 \; \forall \; n$).[3] Note also that $P(\text{type II error}) = h(p,n) \rightarrow 0 \; \forall \; p \in (0,1)$ as $n \rightarrow \infty$. These results illustrate a generally valid principle for "good" statistical tests: The more sample information one has, the more accurate a statistical test will be.

Suppose it were too costly, or impossible, to increase the sample size beyond $n = 100$. Consider an alternative critical region in order to alter the error probabilities of the associated statistical test. In particular, suppose we define $C_r = \{\mathbf{x}: \sum_{i=1}^{100} x_i \geq 2\}$ and make test decisions accordingly. Then the probability of type I error is still zero, since *if* $H$ were true, then $P(\sum_{i=1}^{100} x_i = 0) = 1$, and so long as the point $\mathbf{x} = [\mathbf{0}] \notin C_r$, then $P(\mathbf{x} \notin C_r) = 1$ when $H$ is true, and so the probability of type I error remains zero. The probability of type II error as a

[3] Recall the general differentiation result

$$da^{u(x)}/dx = a^{u(x)} \ln(a)(du/dx).$$

In the case at hand, $u(x) = x = n$, and $a = (1-p)$.

function of $p$, based on the binomial distribution of $\sum_{i=1}^{n} X_i$, is now given by

$$\begin{aligned} P(\text{type II error}) &= P\Big(\sum_{i=1}^{100} x_i = 0 \text{ or } 1\Big) \\ &= (1-p)^{100} + 100p(1-p)^{99} \quad \text{for } p \in \{.001, .002, \ldots, 1.000\}. \end{aligned}$$

It is evident that the probabilities of type II error are uniformly *higher* for this statistical test as compared to the previous test (both tests using $n = 100$).

The reader might have surmised by now that the original statistical test that we examined in Ex. 9.3 is, for any given sample size, the best one can do with respect to minimizing the probability of type II error while maintaining the probability of type I error at *zero*. This follows from the fact that the type I error probability will be zero iff $C_r$ does not contain the element $\mathbf{x} = [\mathbf{0}]$, and any element removed from $C_r$, while not affecting the probability of type I error, will increase the probability of type II error $\forall\, p \in (0, 1)$. Note further that the only other choice of type I error probability in this statistical hypothesis-testing situation is the value 1, which would be the case for any critical region (i.e., for any statistical test) for which $[\mathbf{0}] \in C_r$. A statistical test having $P(\text{type I error}) = 1$ would clearly be absurd since *if H were* true, one would be essentially certain to *reject H*! Thus, for the case in Ex. 9.3, the statistical test based on $C_r = \{\mathbf{x}: \sum_{i=1}^{n} x_i > 0\}$ is effectively the best one can do with respect to controlling type I and type II error probabilities for any sample size.

In general hypothesis-testing applications, a primary concern is the choice of a statistical test that provides acceptable levels of control on both type I and type II errors. Unfortunately, these are most often conflicting objectives, as we now discuss.

### Type I/Type II Error Tradeoff

The choice of values at which type I error probability can be controlled is generally not as limited as in the case discussed in the previous section. It is more typical for there to be a large number, or even a continuum, of possible choices for the type I error probability of a statistical test of a simple hypothesis. Furthermore, there is typically also a tradeoff between the choice of type I and type II error probabilities such that the probability of one type of error can be decreased (increased) only at the expense of increasing (decreasing) the other. When $H$ is composite, there is also a range of type I error probabilities to consider, which results from the fact that $H$ then contains a range of potential probability distributions for $\mathbf{X}$. We will encounter a large number of examples of these more typical situations in the remainder of this chapter, as well as in Chapter 10.

In order to illustrate some of the more general characteristics of statistical hypothesis tests, reexamine the hypothesis concerning the bolt shipment, as described in Ex. 9.1. In this case,

$$R(\mathbf{X} \mid H \cup \bar{H}) = R(\mathbf{X} \mid H) = R(\mathbf{X} \mid \bar{H}) = \times_{i=1}^{n}\{0, 1\},$$

and thus no ideal statistical test exists to test the statistical hypothesis that the probability distribution of the random sample is defined via a product of identical Bernoulli densities with $p \leq .02$. How then should we choose a statistical test of $H$? Suppose we want the probability of type I error to equal 0. In the current testing situation, this would mean that if the true PDF for the random sample were a product of identical Bernoulli densities having any value of $p \in \{0, .001, .002, \ldots, .020\}$, we want our test to accept $H$ with probability 1. Unfortunately, the only choice for $C_r \subset R(\mathbf{X} \mid H \cup \bar{H})$ that will ensure that $P(\mathbf{x} \notin C_r) = 1$ when $H$ is true is $C_r = \emptyset$. This follows because $\forall\, p \in \{.001, .002, \ldots, .020\}$, each and every point in $\times_{i=1}^{n}\{0, 1\}$ is assigned a positive probability value. Then the statistical test would have the characteristic that $P(x \in C_r) = 0$ *whether or not H is true*. The test would *always* accept $H$, no matter what the outcome of the random sample. Thus to obtain a probability of type I error equal to 0, we have to accept a probability of type II error equal to 1. This is clearly unacceptable—we would never reject $H$, even if it were profoundly false (even if *all* of the bolts in the shipment were defective!).

Given the preceding results, it is clear that to choose a useful statistical test in the current situation requires that we be willing to accept some positive level of type I error probability. Suppose we agree that the maximum probability of type I error that we are willing to accept is .05, i.e., we are implicitly stating that $P(\text{type I error}) \leq .05$ provides sufficient **protection against type I error**. The interpretation of "protection" is that if $P(\text{type I error}) \leq .05$, then we know that the test will *mistakenly* reject a true $H$ no more than 1 time in 20, on the average, in a repeated sampling context. If the test actually rejects $H$ for a given random-sample outcome, it is much more likely that the test is rejecting $H$ because $H$ is false, rather than because the test has mistakenly rejected a true $H$. The level of confidence we have that $H$ is false when $H$ is rejected by a statistical test is thus derived from the level of protection against type I error that the test provides.

How should $C_r$ be chosen in order to define a test with the desired control of type I error? We need to find a set $C_r \subset R(\mathbf{X} \mid H \cup \bar{H})$ such that $P(\mathbf{x} \in C_r) \leq .05$ no matter which probability distribution in $H$ were true, i.e., no matter which $p \in \{0, .001, \ldots, .020\}$ were true. In addition, we want $C_r$ to be such that $P$(type II error) is as small as possible no matter which probability model in $\bar{H}$ were true, i.e., no matter which $p > .02$ were true. Intuitively, it would make sense that if $P$(type II error) is to be minimized for a given level of control for type I error, then we should attempt to choose points for the rejection region, $C_r$, that would have their *highest* probability of occurrence if $\bar{H}$ were true. In the case at hand, this corresponds to choosing sample outcomes for inclusion in $C_r$ that represent a higher number of defectives than lower, since $\bar{H}$ corresponds to bolt populations with more defectives than is the case for $H$.

Suppose that a random sample of size $n = 200$ will be used to make the test. Examine the following critical region contained in $R(\mathbf{X} \mid H \cup \bar{H}) = \times_{i=1}^{200}\{0, 1\}$: $C_r = \{\mathbf{x}\colon \sum_{i=1}^{200} x_i > 7\}$. In order to verify that the statistical test implied by $C_r$ has a probability of type I error that does not exceed .05, we must verify that $P(\mathbf{x} \in C_r) \leq .05$ no matter which probability distribution in $H$ were true. In other words, we must verify that the outcome of the binomial random variable $\sum_{i=1}^{200} X_i$ is such that

$$P\left(\sum_{i=1}^{200} x_i > 7\right) = \sum_{j=8}^{200} \frac{200!}{j!(200-j)!} p^j (1-p)^{200-j} \leq .05 \quad \forall p \in \{0, .001, \ldots, .020\}.$$

The industrious reader can verify that $P(\sum_{i=1}^{200} x_i > 7)$ in fact achieves a maximum of .049 when $p = .02$, and so the inequality is met for all choices of $p$ implied by $H$. Thus, the probability of type I error is upper bounded by .05. We emphasize that in this case, since $H$ was composite, the probability of type I error is a function of $p$, with the range of probabilities upper bounded by .05. The largest value of type I error probability over all PDFs in $H$ is generally referred to as the *size of the statistical test* and is a measure of the minimum degree of protection against type I error provided by the test. We will examine this concept in more detail in Section 9.4.

Now let's examine the probability of type II error as a function of $p$. The various probabilities of type II error are given by $P(x \notin C_r) = P(\sum_{i=1}^{200} x_i \leq 7)$ evaluated with respect to binomial distributions for $\sum_{i=1}^{200} X_i$ that have $p > .02$ (which are the values of $p$ implied by $\bar{H}$). Figure 9.5 presents a partial graph of the probability of type II error as a function of $p$ for this statistical test. It is seen from Figure 9.5 that the probability of type II error decreases rapidly as $p$ increases, with $P$(type II error) $\leq .05$ when $p \geq .065$. However, as we have seen previously, $P$(type II error) is quite high for values of $p$ close to those implied by $H$, e.g., if $p = .025$ or .03, then $P$(type II error) = .87 or .75, respectively. Thus, if the test does *not* reject $H$, the behavior of $P$(type II error) as a function of $p$ suggests that we would have substantial confidence that $p < .065$, say,

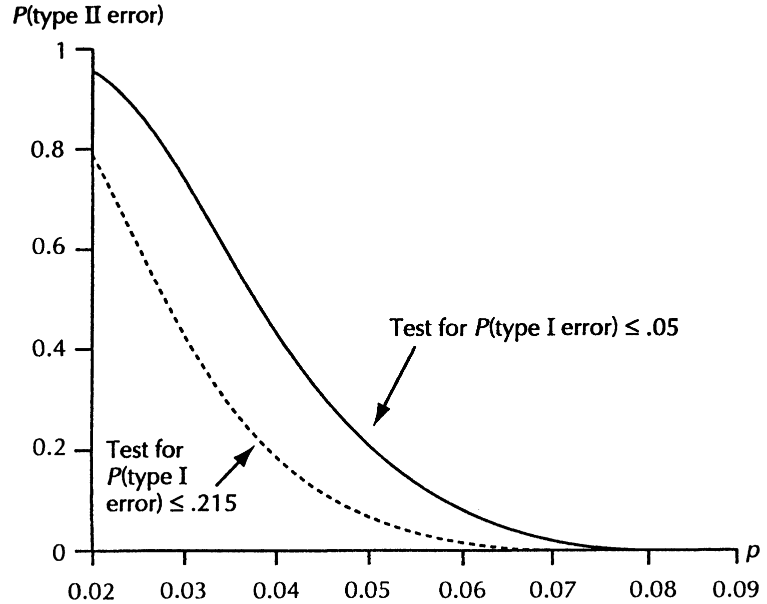


**Figure 9-5**
Probabilities of type II error for tests of $p \leq .02$

since otherwise it would have been highly probable that the test would have rejected $H$ (with a probability of .95 or greater). However, we do not have a great deal of confidence that $H$ is *literally true*, since the test does not have a high probability of rejecting $H$ if $p > .02$ but $p$ is near .02. In this example, if $H$ is accepted, it would be best to interpret this action as meaning that ***H* is at least approximately true**. In fact, we will see that this situation is typical of statistical tests in practice, so that in accepting $H$, it is best not to interpret the term *accept* in the *literal* sense.[4]

If the relationship depicted in Figure 9.5 for the case where $P$(type I error) $\leq .05$ were deemed unacceptable, the researcher has two basic options, as we discussed previously. One option would be to increase the random-sample size to some value larger than 200 and apply an analogous test rule to the larger random sample. This would decrease $P$(type II error) uniformly for all $p > .02$ assuming $C_r$ were chosen to minimize $P$(type II error) while maintaining $P$(type I error) $\leq .05$. The other option would be to accept a larger probability bound on type I error (i.e., define a different test based on the same sample size), which would also result in a uniform reduction in $P$(type II error). To illustrate the tradeoff between the two types of errors that occurs when pursuing the latter option, suppose we redesign the statistical test using the following critical region: $C_r = \{\mathbf{x} : \sum_{i=1}^{200} x_i > 5\}$. Using the binomial distribution for $\sum_{i=1}^{200} X_i$, it can be shown that $P(\sum_{i=1}^{200} x_i > 5) \leq .215 \ \forall\, p \in \{0, .001, \ldots, .02\}$, and so $P$(type I error) is now upper bounded by .215. Thus, if $H$ were true, then approximately 1 out of 5 times, in a repeated sampling context, the test would nonetheless reject $H$. Note, using the redesigned test, that there is notably *less* confidence that $H$ is actually false when the test rejects $H$ than was the case with the previous test since the test will mistakenly reject a true $H$ 21.5 percent of the time, on the average. However, $P$(type II error) is also notably decreased using the redesigned test, as illustrated in Figure 9.5. Now if $p = .025$ or .03, then $P$(type II error) is reduced to .62 or .44, respectively, and $P$(type II error) $\leq .062 \ \forall\, p > .05$.

In any application, the researcher must decide the appropriate degree of tradeoff between the type I and type II error probabilities of the test, as well as the appropriate sample size to use in defining the test. An acceptable choice of the level of protection against test decision errors will generally depend on the nature of the consequences of making a decision error, and on the cost and/or feasibility of changing the sample size. For example, if the hypothesis was that the level of residual pesticide that appears in fruit sold at retail exceeds levels that are safe for human consumption, it would seem that type I error would be extremely serious—we want a great deal of protection against making an incorrect decision that the level of residual pesticide is at a safe level. Thus, we would need to design a test with an extremely low probability of type I error. Type II error in this case would be a situation where a safe level of pesticide was declared unsafe, potentially resulting in increased processing and growing

[4] In recognition of this difficulty, some authors advocate that one either reject or *not reject* a statistical hypothesis, the latter phrase taken to mean that the hypothesis is neither rejected nor accepted, i.e., one "reserves judgment."

costs, loss of sales by the pesticide manufacturer, and the loss of an input to the agricultural sector that could potentially lower the cost of food production. Thus, one would also need to strive for a low $P$(type II error) in fairness to the pesticide manufacturer, farmers, and consumers. The precise choice of the type I/type II error tradeoff must be addressed on a case-by-case basis. The reader should consider what the appropriate tradeoff might be for testing the bolt manufacturer's hypothesis examined previously.[5]

### Test Statistics

It should be noted that the critical regions of the statistical tests defined in the previous two subsections could all be defined in terms of outcomes of a scalar statistic. In particular, the statistic used was of the form $T = t(\mathbf{X}) = \sum_{i=1}^{n} X_i$, and the set-defining conditions for the critical regions were of the general form $t(\mathbf{x}) > c$. A scalar statistic whose outcomes are used to define critical regions of statistical tests is called a *test statistic.*

**Definition 9.4**
***Test statistic***

Let $C_r$ define the critical region associated with a statistical test of the hypothesis $H$ versus $\bar{H}$. If $T = t(\mathbf{X})$ is a scalar statistic such that $C_r = \{\mathbf{x} : t(\mathbf{x}) \in C_r^T\}$, i.e., the critical region can be defined in terms of outcomes, $C_r^T$, of the statistic $T$, then $T$ is referred to as a *test statistic* for the hypothesis $H$ versus $\bar{H}$. The set $C_r^T$ will be referred to as the *critical* (*or rejection*) *region of the test statistic, T.*

The use of test statistics can simplify in at least two significant ways the problem of testing statistical hypotheses. First of all, it allows one to check whether a random-sample outcome is an element of the critical region of a statistical test by examining whether a scalar outcome of the test statistic resides in a set of scalars (the set $C_r^T$ in Def. 9.4). This eliminates the need for dealing with $n$-dimensional outcomes and $n$-dimensional critical regions. In effect, the critical region of the statistical test is alternatively represented as a unidimensional set of real numbers. Second, test statistics can facilitate the evaluation of type I and type II error probabilities of statistical tests if the PDF of the test statistic can be both identified and tractably analyzed. Relatedly, if an asymptotic distribution for a test statistic can be identified, then it may be possible to approximate type I and type II error probabilities based on this asymptotic distribution even when exact error probabilities for the statistical test cannot be established. We will examine the notion of *asymptotic tests* in Chapter 10.

In practice, the vast majority of statistical hypothesis testing is conducted using test statistics, and the statistical test itself is generally defined in terms

[5] If the magnitudes of the costs or losses incurred when errors are comitted can be expressed in terms of a loss function, then a formal analysis of expected losses can lead to a choice of type I and type II error probabilities. For an introduction to the ideas involved, see A. Mood, F. Graybill, and D. Boes (1974), *Introduction to the Theory of Statistics*, 3rd ed., New York: McGraw-Hill, pp. 414–418.

of the critical region of the test statistic. That is, a statistical test of $H$ versus $\bar{H}$ will generally be defined as

$t(\mathbf{x}) \in C_r^T \Rightarrow$ reject $H$ and $t(\mathbf{x}) \notin C_r^T \Rightarrow$ accept $H$.

We will henceforth endeavor to express statistical tests in terms of test statistics whenever it is useful to do so.

It should be noted that test statistics are *not* unique. For example, the critical regions of the statistical tests in the preceding two subsections can be represented in terms of the test statistics $\sum_{i=1}^n X_i$, $\bar{X}$, or $\ln \bar{X}$ (among a myriad of other possibilities), with corresponding critical regions for the test statistics given by $(c, n]$, $(c/n, 1]$, and $(\ln(c/n), 0]$, respectively. The choice of $\sum_{i=1}^n X_i$ was particularly attractive in the preceding application because its PDF (binomial) was easily identified and tractable to work with. In general, the choice of functional form for a test statistic is motivated by the ease with which the outcomes of the statistic can be obtained *and* by the tractability of the test statistic's PDF.

### Null and Alternative Hypotheses

The terminology *null* and *alternative* hypotheses is sometimes used in the context of testing statistical hypotheses. Historically, a **null hypothesis** was a hypothesis interpreted as characterizing no change, no difference, or no effect. For example, if the average life of a "typical" 60-watt incandescent light bulb sold in the United States is 1,000 hours, then a test of the hypothesis that a bulb manufactured by General Electric Co. has a life equal to the average life, i.e., $H: \mu = 1,000$, could be labeled a null hypothesis and interpreted as representing a situation where there is "no difference" between General Electric's bulb and the average life of a typical bulb. The hypothesis that is accepted if the null hypothesis is rejected is referred to as the **alternative hypothesis**. In current usage, the term "null hypothesis" is used more generally to refer to any hypothesis whose mistaken rejection is assumed to characterize a type I error. Thus, a type I error refers to the mistaken rejection of a null hypothesis, while a type II error refers to a mistaken acceptance of the null hypothesis.

Having examined a number of basic conceptual issues related to statistical hypothesis testing, we henceforth focus our attention on two specific classes of hypothesis-testing situations: (1) parametric hypothesis-testing, and (2) testing distributional assumptions. The next two sections examine parametric hypothesis testing in some detail. We defer discussing the testing of distributional assumptions until Chapter 10.

## 9.4 Parametric Hypothesis Tests and Test Properties

In *parametric hypothesis testing*, it is assumed at the outset that the probability distributions in both $H$ and $\bar{H}$ are all characterized by members of a known or given parametric family of distributions. Furthermore, it is assumed that the only unknown component of the true probability distribution associated with random-sample outcomes is the true value of the parameter vector. In this case, the statistical hypothesis and its complement can be defined entirely in terms of sets of parameter values since such sets correspond directly to sets of probability distributions via parameter-indexing of specific members of the known or given parametric family of distributions. In the case of *parametric* hypothesis testing, one is then effectively testing hypotheses about *parameters* or *functions of parameters*. Example 9.1 and Ex. 9.3 are examples of parametric hypothesis tests, where the known parametric family of distributions underlying the definitions of the statistical hypotheses and their complements is the Bernoulli distribution. Once it is understood that all probability distributions will be defined in terms of a Bernoulli distribution with parameter $p$, then the statistical hypotheses under consideration can be compactly represented in terms of statements about the value of $p$ or about the value of a function of $p$.

In the parametric hypothesis-testing case, we will most often utilize the abbreviated notation $H$: *set-defining conditions* introduced in Section 9.2 to denote the statistical hypothesis. In this case, the *set-defining conditions* will denote that a parameter vector, or a function of the parameter vector, is contained in a certain set of values (recall Ex. 9.7). We will adopt a simplification that is made possible by the parametric context and interpret $H$ as a set of parameter values, and thus the notation $\Theta \in H$ will mean that the parameter (vector) is an element of the set of hypothesized values. We emphasize that in the parametric hypothesis-testing context, $H$ still concurrently identifies a set of probability distributions for **X**.

Whether the hypothesis-testing situation in Ex. 9.2 could be viewed as a parametric hypothesis-testing problem depends on the nature of $\bar{H}$. If it were known or given that the exponential family of distributions characterized the probability distributions in $\bar{H}$ as well as in $H$, then we could represent $H$ in abbreviated form as $H: \Theta > 20$, and $\bar{H}$ could be represented as $\bar{H}: \Theta \in (0, 20]$. The problem would then be one of parametric hypothesis testing. However, if the family of probability distributions characterizing the probability models in $\bar{H}$ were left unspecified, then $H$ and $\bar{H}$ could no longer be represented in terms of two sets of $\Theta$ values, since there would be a question regarding the functional form of the probability distributions associated with the random sample if $\bar{H}$ were true, and thus it would not be known to what $\Theta$ referred in the set $\bar{H}$.

The real-world meaning of a statistical hypothesis concerning the values of parameters or functions of parameters depends on the characteristics of the real-world population or process being analyzed and on the interpretation of what the parameters represent or measure in the context of the specified family of PDFs. For example, the exponential family of densities used in Ex. 9.2 leads to the interpretation of $H: \Theta > 20$ as being an assertion about the *mean* of the distribution of computer lifetimes, while the Bernoulli family of densities used in Ex. 9.1 leads to the interpretation of $H: p \leq .02$ as an assertion about the *proportion* of defectives in a shipment of bolts.

### Maintained Hypothesis

A parametric hypothesis-testing situation will be characterized as having two major components: (1) the *maintained hypothesis,* and (2) the statistical hypothesis.

**Definition 9.5**
***Maintained hypothesis***

All facts, assertions, or assumptions about the probability space of a random sample that are in common to $H$ and $\bar{H}$, and are thus maintained to be true regardless of the outcome of a statistical test of $H$ versus $\bar{H}$, are collectively called the *maintained hypothesis.*

In a parametric hypothesis-testing situation, whatever parametric family of distributions is assumed to characterize the probability distributions in $H$ and $\bar{H}$ is then part of the maintained hypothesis. For example, in Ex. 9.1, the Bernoulli family of distributions is part of the maintained hypothesis, while in Ex. 9.2, the exponential family of distributions could be part of the maintained hypothesis if the distributions in $\bar{H}$ were also assumed to be in the exponential family. In addition to an assertion about the parametric family of distributions, a number of other "facts" about the probability space can be contained in the maintained hypothesis. For example, with reference to Ex. 9.4, in examining the hypothesis $H$: $\beta_1 = .25$ versus $\bar{H}$: $\beta_1 \neq .25$, the maintained hypothesis includes assertions (explicitly or implicitly) that the normal family of distributions characterizes the probability distributions for the $\varepsilon_i$'s, the $\varepsilon_i$'s are independent random variables with $\mathrm{E}\varepsilon_i = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma^2 \; \forall \; i$, and the expected value of wheat yield in any given time period is a linear function of both the level of fertilizer applied and the level of rainfall.

The facts contained in the maintained hypothesis can be thought of as contributing to the definition of the specific context in which a statistical test of any hypothesis is to be applied and interpreted. Indeed, the probabilities of type I and type II errors for a given statistical test of an assertion like $H: \Theta > 20$, and in fact the fundamental meaning of the statistical hypothesis itself, will generally depend on the context of the parametric hypothesis-testing problem defined by the maintained hypothesis.

Regarding the genesis of the maintained hypothesis, the facts that are included in the maintained hypothesis might have been supported by the results of previous statistical tests of hypotheses. Alternatively, the facts may be derived from laws or accepted theories about the process or population being studied. Finally, the facts may be largely unsubstantiated conjectures about the population or process under study that are tentatively accepted as truths for the purpose of conducting a parametric hypothesis test. In any case, the interpretation of a statistical test can depend critically on facts stated in the maintained hypothesis, and the result of a statistical test must be interpreted as being conditional on these facts. If a component of the maintained hypothesis is actually false, the interpretation of the type I and type II error probabilities of a given statistical test, and the interpretation of a statistical hypothesis itself, can be completely changed and/or disrupted. Thus, in a parametric hypothesis-testing situation, one must be careful that the facts stated in the maintained hypothesis are defensible. Facts that are "tentative" will generally require statistical testing themselves when convincing supporting evidence is lacking. At the least, the researcher is obliged to point out any tentative assumptions contained in the maintained hypothesis when reporting the results of a statistical test.

### Power Function

Conditional on the validity of the maintained hypothesis, the *power function* of a statistical test can be constructed in a parametric hypothesis-testing situation. The power function completely summarizes all of the operating characteristics of a statistical test with respect to probabilities of making correct/incorrect decisions about $H$. The power function is useful in comparing alternative statistical tests of a particular parametric hypothesis. The power function is defined as follows.

**Definition 9.6**
***Power function of a statistical test***

Let a parametric statistical hypothesis be defined by $H\colon\ \Theta \in \Omega_H$, and let its complement be defined by $\bar{H}\colon\ \Theta \in \Omega_{\bar{H}}$. Let the critical region $C_r$ define a statistical test of $H$. Finally let $f(\mathbf{x};\Theta), \Theta \in \Omega = H \cup \bar{H}$, represent the parametric family of density functions indexed by $H$ and $\bar{H}$. Then the **power function** of the statistical test is the function of $\Theta \in H \cup \bar{H}$ defined by

$$\pi(\Theta) = P(x \in C_r; \Theta) \equiv \underset{\mathbf{x}\in C_r}{\int \cdots \int} f(\mathbf{x};\Theta)dx \qquad \text{(continuous case)},$$

$$\equiv \underset{x\in C_r}{\sum \cdots \sum} f(\mathbf{x};\Theta) \qquad \text{(discrete case)}.$$

In words, the power function indicates the probability of rejecting $H$ for every possible value of $\Theta$. The value of the function $\pi$ at a particular value of the parameter vector, $\Theta$, is called the **power of the test at** $\Theta$, which is the probability of rejecting $H$ if $\Theta$ were the true value of the parameter vector. From the definition of the power function, it follows that $\pi(\Theta)$ is identically a probability of type I error if $\Theta \in H$, or else a probability of *not* making a type II error if $\Theta \in \bar{H}$. Then $[1 - \pi(\Theta)]$ is identically the probability of type II error if $\Theta \in \bar{H}$ and is the probability of *not* making a type I error if $\Theta \in H$. The graph in Figure 9.6 represents a partial graph of the power function for the test of the hypothesis $H\colon\ p \leq .02$ from Ex. 9.1 using the statistical test based on $C_r = \{\mathbf{x}\colon \sum_{i=1}^{200} x_i > 5\}$. We know from before that $P(\text{type I error}) \leq .215$ for this test. As indicated in Figure 9.6, all relevant operating characteristics of the statistical test can be deduced from the graph of the power function. In fact, some statisticians utilize a function of $\Theta$ defined by $OC(\Theta) = 1 - \pi(\Theta)$, called

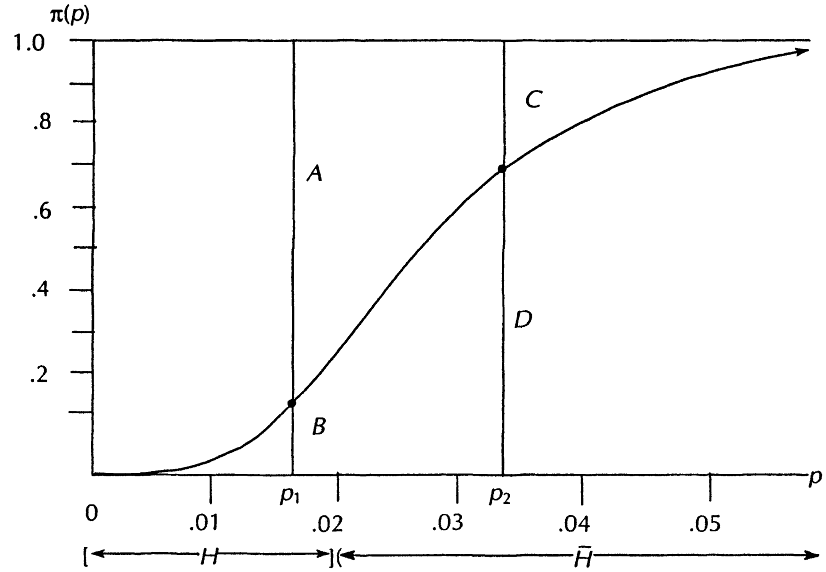


**Figure 9-6**
Power function of test for $H: p \leq .02$ (Notes: $A = 1 - P$(type I error) $= P$(Correct Decision) for $p \in H$; $B = P$(type I error) for $p \in H$; $C = P$(type II error) for $p \in \bar{H}$; $D = 1 - P$(type II error) $= P$(Correct Decision) for $p \in \bar{H}$.)

the **operating characteristic function** of a statistical test, to portray all of the relevant probabilities of correct and incorrect decisions associated with a statistical test. The two functions clearly can be used interchangeably to portray precisely the same information about the performance of a statistical test.

The closer the power function of a test is to the theoretical **ideal power function** for the hypothesis testing situation, the better the test. The ideal power function can be compactly defined as $\pi_a(\Theta) = I_{\bar{H}}(\Theta)$ (why?) and would correspond to the *ideal statistical test* for a given $H$, were such a test to exist. When comparing two statistical tests of a given $H$, a test is better if it has higher power for $\Theta \in \bar{H}$ and lower power for $\Theta \in H$, which implies that the better test will have lower probabilities of both type I and type II error. Figure 9.7 provides graphs of hypothetical power functions for two tests of the hypothesis that $H: \ \Theta \leq c$ and a graph of the ideal power function for the test situation, where $\Omega = H \cup \bar{H} = [0, k]$. From Figure 9.7, it is apparent that the statistical test associated with the power function $\pi_2(\Theta)$ is the better test, having probabilities of type I and type II error everywhere less than or equal to the corresponding probabilities associated with the alternative test. The power function $\pi_2(\Theta)$ is also seen to be closer to the ideal power function, $\pi_*(\Theta)$.

The power function of a statistical test can be used in defining additional properties of a test, including the *size* and *significance level of the test*, and whether a test is *unbiased, consistent*, and/or *uniformly most powerful*. We examine each of these test properties ahead.

### Properties of Statistical Tests

*Test Size and Significance Level* A number of test properties can be examined to assess whether a given statistical test is appropriate in a given problem context

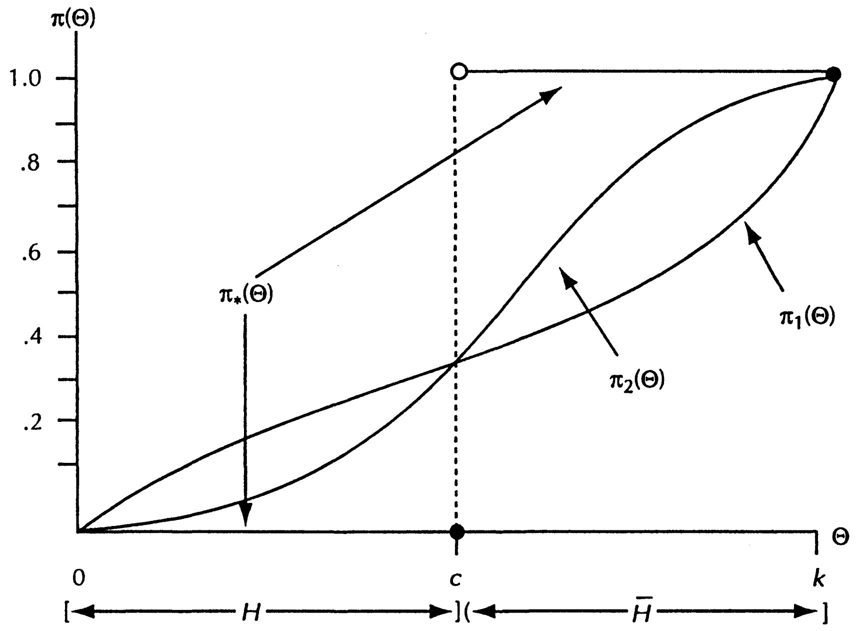


**Figure 9-7**
Two power functions and the ideal power function for testing $H: \Theta \leq c$

and/or to compare a test rule to competing test rules.[6] One such property, called the *size of the test*, is a measure of the minimum level of protection against type I error that a given statistical test provides.

**Definition 9.7**
***Size of statistical test***

Let $\pi(\Theta)$ be the power function of a test rule for testing the hypothesis $H$. Then $\alpha = \sup_{\Theta \in H} \pi(\Theta)$ is called the **size of the statistical test.**[7]

The size of the statistical test is essentially the maximum probability of type I error associated with a given test rule. The lower the size of the test, the lower the maximum probability of mistakenly rejecting $H$.

An alternative concept that is used as a measure of protection against type I error is the **significance level**, or simply, the **level** of statistical tests. A **level-$\alpha$ test** of $H$ is any test for which $P_\Theta$(type I error) $\leq \alpha \; \forall \; \Theta \in H$. Thus, the level is an *upper bound* to the type I error probability of a statistical test. The key difference between the concepts of *size* and *level* of a statistical test is that the former represents the maximum (or at least the supremum) value of $P_\Theta$(type I error) for $\Theta \in H$ while the latter is only a bound that might not equal $P_\Theta$(type I error) for any $\Theta \in H$ nor equal the supremum of $P_\Theta$(type I error) for $\Theta \in H$. Thus, a statistical test of $H$ having size $\gamma$ is an $\alpha$-level test for any $\alpha \geq \gamma$.

[6] The properties we will examine do not exhaust the possibilities. See E. Lehmann (1986), *Testing Statistical Hypotheses*, New York: John Wiley.

[7] Recall that $\sup_{\Theta \in H} \pi(\Theta)$ denotes the smallest upper bound to the values of $\pi(\Theta)$ for $\Theta \in H$ (i.e., the supremum). If the maximum of $\pi(\Theta)$ for $\Theta \in H$ exists, then "sup" can be replaced by "max."

In applications when a hypothesis $H$ is (not) rejected, it is often stated that **H is (not) rejected at the $\alpha$-level of significance**, which is intended to indicate the bound on the level of protection against type I error that was used when conducting the test. Customarily, the phrase *level of significance* is often taken to be synonymous with *size* in applications, although this need not be the case and care should be exercised to eliminate ambiguity relating to whether the level of significance is actually achieved by the statistical test used. A more definitive statement regarding level of protection against type I error would be that **H is (not) rejected using a size-$\alpha$ test**. As an illustration of terminology, the power function in Figure 9.6 refers to a test that has a *size* of .215, and the *significance level of the test* is .215 (or any number larger than .215). If the hypothesis $H: p \leq .02$ is (not) rejected, then one could state that $H: p \leq .02$ *is (not) rejected at the .215 level of significance,* or that $H: p \leq .02$ *is not rejected using a size-.215 test.*

*Unbiasedness* The concept of *unbiasedness* within the context of hypothesis testing refers to a statistical test that has a smaller probability of rejecting the null hypothesis, $H$, when $H$ is indeed true compared to when $H$ is false.

**Definition 9.8**
***Unbiasedness of a statistical test***

Let $\pi(\Theta)$ be the power function of a statistical test of $H$. The statistical test is called unbiased iff $\sup_{\Theta \in H} \pi(\Theta) \leq \inf_{\Theta \in \bar{H}} \pi(\Theta)$.[8]

The property of unbiasedness makes intuitive sense as a desirable property of a test rule since we would generally prefer to have a higher probability of rejecting $H$ when $H$ is false then when $H$ is true. As indicated in Def. 9.8, whether a test rule is unbiased can be determined from the behavior of its power function. In particular, if the height of the power function graph is everywhere lower for $\Theta \in H$ than for $\Theta \in \bar{H}$, the statistical test is *unbiased.* The tests associated with the power functions in Figures 9.6 and 9.7 are unbiased tests.

*More Powerful, Uniformly Most Powerful, Admissibility* Another desirable property of a statistical test is that, for a given level, $\alpha$, the test exhibits the highest probability of rejecting $H$ when $H$ is false compared to all other competing tests of $H$ having level $\alpha$. Such a test is called uniformly most powerful of level $\alpha$, where the term *uniformly* refers to the test being the most powerful (i.e., having highest power) for *each and every* $\Theta$ in $\bar{H}$ (i.e., *uniformly* in $\bar{H}$). In the formal definition of the concept presented below, we introduce the notation $\pi_{C_r}(\Theta)$ to indicate the power function of a test of $H$ when the test is defined by the critical region $C_r$.

[8] Recall the previous footnote and the fact that $\inf_{\Theta \in \bar{H}} \pi(\Theta)$ denotes the largest lower bound to the values of $\pi(\Theta)$ for $\Theta \in \bar{H}$ (i.e., the infimum). The sup and inf of $\pi(\Theta)$ can be replaced by max and min, respectively, when the maximum and/or minimum exists.

**Definition 9.9** *Uniformly most powerful (UMP) level-$\alpha$ test*

Let $\Xi = \{C_r\colon \sup_{\Theta \in H} \pi_{C_r}(\Theta) \leq \alpha\}$ be the set of all critical regions of level $\alpha$ for testing the hypothesis $H$ based on a random sample $\mathbf{X} \sim f(\mathbf{x}; \Theta)$ of size $n$. The critical region $C_r^* \in \Xi$, and the associated statistical test of $H$ that $C_r^*$ defines, are each called **uniformly most powerful of level $\alpha$** iff $C_r^*$ has level $\alpha$ and $\pi_{C_r^*}(\Theta) \geq \pi_{C_r}(\Theta)\ \forall\ \Theta \in \bar{H}$ and $\forall\ C_r \in \Xi$.

In the case where $\bar{H}$ is *simple,* the critical region, $C_r^*$, and the associated test are also referred to as being **most powerful** (the adverb "uniformly" being effectively redundant in this case). The UMP test of $H$ is seen to be the test of $H$, among all possible tests of $H$ providing protection against type I error equal to $P$(type I error) $\leq \alpha$, that has the highest probability of rejecting $H$ when $H$ is false (equivalently, it is the level-$\alpha$ test having the most power uniformly in $\Theta$ for $\Theta \in \bar{H}$). Such a test of $H$ is effectively the "best one can do" with respect to minimizing the probability of type II error, *given* that protection against type I error is at level $\alpha$. As implied by Def. 9.9, if one could plot the power function graphs of all tests of level $\alpha$, the power function of the UMP level-$\alpha$ test of $H$ would lie on or above the power functions of all other level-$\alpha$ tests of $H\ \forall\ \Theta \in \bar{H}$. The reader is encouraged to reexamine Ex. 9.3, and the subsequent discussion of the example in the text, to deduce that $C_r = \{\mathbf{x}\colon \sum_{i=1}^{n} x_i = 0\}$ defines the UMP test of level $\alpha = 0$ for the null hypothesis $H\colon p = 0$. We will examine general methods for finding UMP tests of $H$ (when such tests exist) in Section 9.5. Once we have introduced the Neyman–Pearson lemma in Section 9.5, we will also show that a UMP level-$\alpha$ test is also generally an unbiased test.

Unfortunately, in a significant number of cases of practical interest, UMP tests do not exist. We will see later that it is sometimes possible to restrict attention to the unbiased class of tests and define a UMP test within this class, but even this approach fails in important cases. In practice, one then often resorts to the use of statistical tests that have at least acceptable (to the analyst and to those she wishes to convince regarding the validity of her hypotheses) power function characteristics. Assuming that a test of level $\alpha$ is deemed to provide sufficient protection against type I error, then in ranking two competing level-$\alpha$ tests of $H$, the test that is **more powerful** is preferred.

**Definition 9.10** *More powerful level $\alpha$ test*

Let $C_r$ and $C_r^*$ represent two level-$\alpha$ statistical tests of $H$. The test based on $C_r^*$ is said to be *more powerful* than the test based on $C_r$ if $\pi_{C_r^*}(\Theta) \geq \pi_{C_r}(\Theta)$ $\forall\ \Theta \in \bar{H}$, with strict inequality for at least one $\Theta \in \bar{H}$.

The reason for preferring a *more powerful test* is clear—if both tests provide the same desired level of protection against type I error, then the one that provides more protection against type II error is better.

While in comparing level-$\alpha$ tests, the more powerful test is preferred, the less powerful test may have some redeeming qualities. In particular, the latter test may provide better protection against type I error for some or all $\Theta \in H$, and the test may then be appropriate in cases where a smaller level of significance is desired. Tests with no such possible redeeming qualities are *inadmissible* tests, formally defined as follows.

**Definition 9.11**
***Admissibility of statistical test***

Let $C_r$ represent a statistical test of $H$. If there exists a critical region $C_r^*$ such that $\pi_{C_r^*}(\Theta)\left\{\begin{smallmatrix}\geq\\\leq\end{smallmatrix}\right\}\pi_{C_r}(\Theta)\ \forall\ \Theta \in \left\{\begin{smallmatrix}\bar{H}\\H\end{smallmatrix}\right\}$, with strict inequality holding for some $\Theta \in H \cup \bar{H}$, then the test based on $C_r$ is **inadmissible**. Otherwise, the test is **admissible**.

From the definition, it follows that an inadmissible test is one that is dominated by another test in terms of protection against both types of test errors. Thus, there is no $\alpha$-level at which an inadmissible test would be the preferred test. Inadmissible tests can be eliminated from consideration in any hypothesis-testing application.

*Consistency* The limiting behavior of the power functions of a sequence of level-$\alpha$ tests as the size of the random sample on which the tests are based increases without bound relates to the property of *consistency*. In particular, a sequence of level-$\alpha$ tests of $H$ for which the probability of type II error $\to 0$ as $n \to \infty$ will be called a *consistent sequence of level-$\alpha$ tests.* Since the definition of the critical region of a test will generally change as the sample size, $n$, changes, we will introduce the notation $C_{rn}$ to indicate the dependence of the critical region on sample size.

**Definition 9.12**
***Consistent sequence of level-$\alpha$ tests***

Let $\{C_{rn}\}$ represent a sequence of tests of $H$ based on a random sample $(X_1,\ldots,X_n) \sim f(x_1,\ldots,x_n;\Theta)$ of increasing size, $n$, and let the level of the test defined by $C_{rn}$ be $\alpha\ \forall\ n$. Then the sequence of tests of level $\alpha$ is said to be **consistent** iff

$$\lim_{n\to\infty} \pi_{C_{rn}}(\Theta) = 1 \quad \forall\Theta \in \bar{H}.$$

Thus, a consistent sequence of tests of level $\alpha$ is such that, in the limit, the probability is 1 that $H$ will be rejected whenever $H$ is false. Then, for a large enough sample size $n$, the $n$th test in a consistent sequence of tests provides a level of protection against type I error given by a probability $\leq \alpha$ and is essentially certain to reject $H$ if $H$ is false.

### *P* Values

Different analysts and/or reviewers of statistical analyses do not always agree on what the appropriate significance level of statistical tests should be when testing statistical hypotheses. In order to accommodate such differences of opinion, it is becoming frequent in applications to report *P* **values** (an abbreviation for "significance probability values") for each statistical test conducted. A

$P$ value is effectively the *marginal size test* at which a given hypothesis would be rejected based on the observed outcome of $\mathbf{X}$.

In order for the idea of a *marginal size* to make sense, there should exist a prespecified family of *nested* critical regions, and thus an associated family of statistical tests, that correspond to increasing type I error probability values, $\alpha$, for testing the null hypothesis under consideration. By *nested* critical regions corresponding to increasing values of $\alpha$, we mean that $C_r(\alpha_1) \subset C_r(\alpha_2)$ for $\alpha_1 < \alpha_2$, where $C_r(\alpha)$ is the critical region corresponding to $P(\text{type I error}) = \alpha$. Then upon observing $\mathbf{x}$, one calculates the $P$ value as being equal to the smallest $\alpha$-level corresponding to the smallest critical region that contains $\mathbf{x}$, i.e., $P$ value $= \arg\min_\alpha[\mathbf{x} \in C_r(\alpha)] = \min_\alpha[\sup_{\Theta \in H} P(\mathbf{x} \in (C_r(\alpha); \Theta)$ such that $\mathbf{x} \in C_r(\alpha)]$. The $P$ value can thus be interpreted as the smallest size at which the null hypothesis would be rejected based on the nested set of critical regions or statistical tests.

Assuming there is general agreement on the nested set of critical regions and associated statistical tests to be used for testing a particular $H$, then once the analyst reports the $P$ value, a reviewer of the statistical analysis will know whether she would reject $H$ based on her own choice of test size within the set of possible size values.[9] In particular, if the reviewer's choice of size were less than (greater than) the reported $P$ value, then she would not (would) reject $H$ since her critical region would be smaller than (larger than) the one associated with the reported $P$ value and would not (would) contain $x$. The $P$ value is also often interpreted as indicating the *strength of evidence* against the null hypothesis, where the smaller the $P$ value, the greater the evidence against $H$. The idea is that if the $P$ value is small, then the protection against type I error can be set high and $H$ is still rejected by the test.

An accepted set of nested critical regions and statistical tests necessary for calculating $P$ values often exists in practice, especially when the critical regions can be defined in terms of a test statistic. We illustrate the use of both $P$ values and other test properties in the following example.

**Example 9.9** A random sample of $n = 100$ observations on the miles per gallon achieved by a new model pickup truck manufactured by a major Detroit manufacturer is going to be used to test the null hypothesis that the expected miles per gallon achieved by the truck model is less than or equal to 15, the alternative hypothesis being that expected miles per gallon $> 15$. It can be assumed that miles per gallon measurements are independent and that the population is (approximately) normally distributed with a standard deviation equal to .1.

The null and alternative hypotheses in this parametric hypothesis-testing problem can be represented by $H_0\colon \mu \leq 15$ and $H_a\colon \mu > 15$, respectively. Suppose we desire to assess the validity of the null hypothesis using a statistical test of size .05. Examine a test based on the following rejection region: $C_{rn} = \{\mathbf{x}\colon 10n^{1/2}(\bar{x}_n - 15) \geq 1.645\}$. To see that the rejection region implies a *test of*

[9] In the case of continuous $\mathbf{X}$, the choice of size is generally a continuous interval contained in $[0, 1]$. If $\mathbf{X}$ is discrete, the set of choices for size is generally finite, as previous examples have illustrated.

*size .05*, first note that

$$\frac{\bar{X}_n - 15}{\sigma/\sqrt{n}} = \frac{(\bar{X}_n - 15)}{(.1/\sqrt{n})} = 10n^{1/2}(\bar{X}_n - 15)$$

$$= \underbrace{10n^{1/2}(\bar{X}_n - \mu)}_{\sim N(0,1)} + 10n^{1/2}(\mu - 15) \sim N(10n^{1/2}(\mu - 15), 1).$$

It follows that the *power function* of the test can be defined as

$$\pi_n(\mu) = \int_{1.645}^{\infty} N(z; 10n^{1/2}(\mu - 15), 1)dz = \int_{1.645-10n^{1/2}(\mu-15)}^{\infty} N(z; 0, 1)dz,$$

so that $\sup_{\mu \leq 15} \pi_n(\mu) = .05$ and the test is of size .05.[10] The test is also *unbiased*, since it is evident that $\sup_{\mu \leq 15} \pi_n(\mu) \leq \inf_{\mu > 15} \pi_n(\mu) = .05$.[11] The *consistency* of the test can be demonstrated by noting that $\forall\ \mu \in H_a$,

$$\lim_{n\to\infty} \pi_n(\mu) = \lim_{n\to\infty} \int_{1.645-10n^{1/2}(\mu-15)}^{\infty} N(z; 0, 1)dz = \int_{-\infty}^{\infty} N(z; 0, 1)dz = 1,$$

and so the test is a consistent test.

The test is also a *uniformly most powerful* test for testing the null hypothesis $H_0: \mu \leq 15$ versus $H_a: \mu > 15$, i.e., there is no other test rule with $\pi(\mu) \leq .05$ $\forall\ \mu \in H_0$ that has higher power for any $\mu \in H_a$. The rationale in support of the UMP property will be provided in Section 9.5. A graph of the power function for this test is given in Figure 9.8.

Upon examination of the power function of the test, it is clear that protection against type II error increases rapidly for values of $\mu > 15$. In particular, it would seem reasonable to conclude that if $H_0$ were accepted, then the true value of $\mu \leq 15.03$, say, since the probability of accepting $H_0$ when $\mu > 15.03$ is $< .088$.

Suppose that the outcome of the random sample of 100 mile-per-gallon measurements yielded $\bar{x} = 15.02$. It follows that $\mathbf{x} \in C_{rn}$ since $10n^{1/2}(15.02 - 15) = 2 \geq 1.645$ when $n = 100$, and thus $H_0: \mu \leq 15$ is rejected using a size .05 test.

Regarding the $P$ value for this test where $n = 100$, first note that

$$C_r(\alpha) = \{\mathbf{x}:\ 100(\bar{x} - 15) \geq k(\alpha)\}$$

forms a nested set of critical regions such that for each value of size $\alpha \in (0, 1)$, there exists a constant $k(\alpha)$ such that $\max_{\mu \leq 15} P(\mathbf{x} \in C_r(\alpha); \mu) = P(\mathbf{x} \in C_r(\alpha); \mu = 15) = \alpha$. The $P$ value for the test is then

$$P \text{ value} = \arg\min_{\alpha} [100(15.02 - 15) \geq k(\alpha)]$$

[10] The maximum is achievable in this case, and equals .05 when $\mu = 15$.

[11] Note that $\min_{\mu > 15} \pi_n(\mu)$ does not exist in this case. The largest possible lower bound (i.e., the infimum) is .05, which $< \pi_n(\mu)$, $\forall\ \mu > 15$.

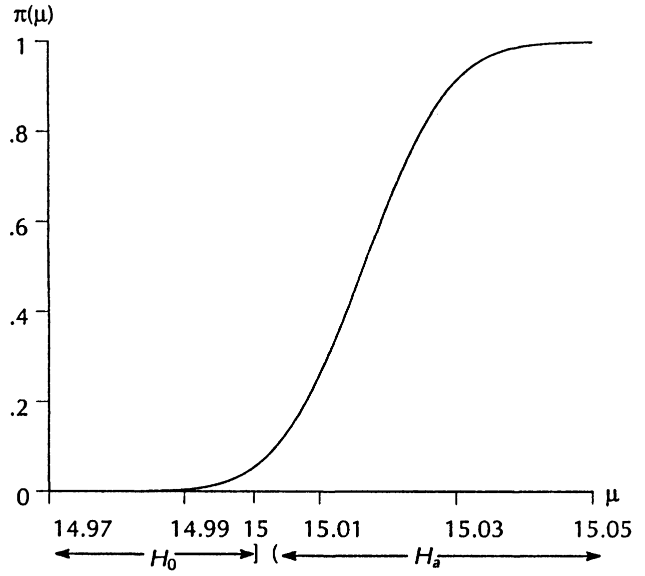


**Figure 9-8** Power function for testing $H_0: \mu \leq 15$ versus $H_a: \mu > 15$, $n = 100$.

$$= \min_{\alpha} [\max_{\mu \leq 15} P(\mathbf{x} \in C_r(\alpha); \mu) \text{ such that } k(\alpha) \leq 2]$$

$$= \min_{\alpha} [P(\mathbf{x} \in C_r(\alpha); \mu = 15) \text{ such that } k(\alpha) \leq 2]$$

$$= \min_{\alpha} \left[ \int_{k(\alpha)}^{\infty} N(z; 0, 1) dz \text{ such that } k(\alpha) \leq 2 \right]$$

$$= \int_{2}^{\infty} N(z; 0, 1) dz = .023.$$

It follows that $H_0$ would *not* be rejected at $\alpha = .01$ and would be rejected at $\alpha = .05$, say. □

### Asymptotic Tests

As we have alluded to previously, it is sometimes difficult or intractable in practice to determine the exact probabilities of type I and type II errors for a given statistical test. Thus the determination of the exact power function of a test may not be possible. Similar to the case of point estimation, one must then rely on asymptotic properties to assess the efficacy of a given test rule.

In the majority of hypothesis-testing applications, the critical regions of statistical tests are defined in terms of test statistics. It is often the case that laws of large numbers and central limit theory can be utilized to establish various convergence properties of a test statistic. In particular, asymptotic distributions of test statistics can be used to approximate power functions of tests, in which case the efficacy of a given test procedure can be assessed based on the power function approximation. In practice, the choice of a particular statistical test for a given hypothesis is often motivated by its adequacy vis-à-vis type I and type II error protection as assessed through an examination of its *approximate* rather than its *exact* power function characteristics. We will examine the asymptotics of some important general classes of statistical tests, including generalized likelihood ratio, Lagrange multiplier, and Wald tests, in Chapter 10.

## 9.5 Results on UMP Tests

In this section we examine results that can be used to facilitate the discovery of a uniformly most powerful test of a given statistical hypothesis. However, it should be noted at the outset that UMP tests of hypotheses do not always exist. Whether a UMP test exists depends very much on the type of hypothesis being tested and on the characteristics of the joint density of the random sample under study. In a case where the results of this section cannot be applied, or in cases where UMP tests simply do not exist, it is nonetheless often possible to define statistical tests with *acceptable* exact or asymptotic power function characteristics, and we will examine specific procedures for defining such tests in Chapter 10.

We will examine four basic approaches for attempting to find UMP tests: Neyman–Pearson, monotone likelihood, exponential class, and conditioning approaches. While these approaches have distinct labels, they are all interrelated to some degree as will become evident in our discussion of them. Most of our discussion will focus on scalar parameter cases, although some important multiparameter results will be examined. This scalar orientation is a practical one—UMP tests do not exist for most multiparameter hypothesis-testing contexts. In cases where a UMP test does not exist, it is sometimes still possible to define a UMP test within the class of *unbiased* tests, as we shall see.

### Neyman–Pearson Approach

We begin here with the case where both the null and alternative hypotheses are *simple*. This situation is not representative of the typical hypothesis-testing problem in which at least one of the hypotheses is composite, but it provides a simplified context that can be used to motivate some of the basic principles underlying the definition of a UMP test. We note in this case that the alternative hypothesis consists of a single point. We could thus drop the adverb "uniformly" and characterize our objective as one of finding the *most powerful* test of the null hypothesis versus the alternative hypothesis.

*Simple Hypotheses* Given that both the null and alternative hypotheses are simple, the hypotheses can be represented as $H_0: \Theta = \Theta_0$ and $H_a: \Theta = \Theta_a$, where $\Theta_0$ and $\Theta_a$ are parameter vectors (or scalars). Under these conditions, there is a well-defined procedure due to Neyman and Pearson[12] for defining a most powerful test of $H_0$. The procedure is motivated by the following theorem, which is generally referred to in the statistical literature as the *Neyman–Pearson lemma.* Recall from Def. 9.6 that the notation $P(\mathbf{x} \in A; \Theta)$ represents the probability that $\mathbf{x} \in A$ when $\Theta$ is the value of the parameter vector.

**Theorem 9.1** (Neyman–Pearson Lemma) *A critical region of the form $C_r(k) = \{\mathbf{x} : f(\mathbf{x};\Theta_0) \leq kf(\mathbf{x};\Theta_a)\}$, for $k > 0$, is a most powerful critical region of level $\alpha = P(\mathbf{x} \in C_r(k);\Theta_0) \in (0,1)$ for testing the hypothesis $H_0 : \Theta = \Theta_0$ versus $H_a : \Theta = \Theta_a$. Furthermore, $C_r(k)$ is the unique (with probability 1) most powerful critical region of size $\alpha$.*

**Proof** (Most powerful level $\alpha$): Let $C_r^*$ refer to any other critical region for which $P(\mathbf{x} \in C_r^*; \Theta_0) \leq P(\mathbf{x} \in C_r(k); \Theta_0) = \alpha$, so that the test of $H_0$ versus $H_a$ based on $C_r^*$ has a size no larger than the size of the statistical test based on $C_r$. Note that

$$I_{C_r(k)}(\mathbf{x}) - I_{C_r^*}(\mathbf{x}) = \begin{bmatrix} 1 - I_{C_r^*}(\mathbf{x}) \geq 0, \forall \mathbf{x} \in C_r(k) \\ 0 - I_{C_r^*}(\mathbf{x}) \leq 0, \forall \mathbf{x} \notin C_r(k) \end{bmatrix},$$

since the range of $I_{C_r^*}(\mathbf{x})$ is $\{0, 1\}$. Also note that

$$f(\mathbf{x}; \Theta_a) \begin{Bmatrix} \geq \\ < \end{Bmatrix} k^{-1} f(\mathbf{x}; \Theta_0) \quad \text{if} \begin{bmatrix} \mathbf{x} \in C_r(k) \\ \mathbf{x} \notin C_r(k) \end{bmatrix},$$

which follows directly from the definition of $C_r(k)$. The preceding results together imply that

$$\left[I_{C_r(k)}(\mathbf{x}) - I_{C_r^*}(\mathbf{x})\right] f(\mathbf{x}; \Theta_a) \geq k^{-1} \left[I_{C_r(k)}(\mathbf{x}) - I_{C_r^*}(\mathbf{x})\right] f(\mathbf{x}; \Theta_0), \quad \forall\, \mathbf{x}.$$

Integrating both sides of the inequality over $\mathbf{x} \in R^n$ if $\mathbf{X}$ is continuous, or summing over all $\mathbf{x}$ values for which $f(\mathbf{x}; \Theta_0) > 0$ or $f(\mathbf{x}; \Theta_a) > 0$ if $\mathbf{X}$ is discrete, obtains

$$P(\mathbf{x} \in C_r(k); \Theta_a) - P(\mathbf{x} \in C_r^*; \Theta_a) \geq k^{-1} \left[P(\mathbf{x} \in C_r(k); \Theta_0) - P(\mathbf{x} \in C_r^*; \Theta_0)\right] \geq 0.$$

The right-hand side of the first inequality is nonnegative because $k^{-1} > 0$ and because the bracketed probability difference multiplying $k^{-1}$ is nonnegative (recall that the size of the test based on $C_r^*$ is no larger than that of the test based on $C_r(k)$). Since the probabilities on the left-hand side of the first inequality represent the powers of the respective tests when $\Theta = \Theta_a$, the test based on $C_r(k)$ is the most powerful test.

(Unique most powerful size $\alpha$): We discuss the proof for the discrete case. The continuous case is proven similarly by replacing summation with integration. Let the critical region $C_r^*$ define any other most powerful size $\alpha$ test of $H_0 : \Theta = \Theta_0$ versus $H_a : \Theta = \Theta_a$. Letting $\mathrm{E}_{\Theta_0}$ and $\mathrm{E}_{\Theta_a}$ denote expectations taken using the densities $f(\mathbf{x}; \Theta_0)$ and $f(\mathbf{x}; \Theta_a)$, respectively, note that

[12] J. Neyman, and E. S. Pearson (1933), "On the problem of the most efficient tests of statistical hypotheses," *Phil. Trans.*, A, 231, p. 289.

$\alpha = \mathrm{E}_{\Theta_0} I_{C_r(k)}(\mathbf{X}) = \mathrm{E}_{\Theta_0} I_{C_r^*}(\mathbf{X})$, and $\mathrm{E}_{\Theta_a} I_{C_r(k)}(\mathbf{X}) = \mathrm{E}_{\Theta_a} I_{C_r^*}(\mathbf{X})$, since by assumption both tests have size $\alpha$, and both tests are most powerful tests. It follows immediately that

$$\sum \cdots \sum_{\mathbf{x} \in R(\mathbf{x})} [I_{C_r(k)}(\mathbf{x}) - I_{C_r^*}(\mathbf{x})]\,[f(\mathbf{x};\Theta_0) - kf(\mathbf{x};\Theta_a)]$$
$$= \mathrm{E}_{\Theta_0}\left[I_{C_r(k)}(\mathbf{X}) - I_{C_r^*}(\mathbf{X})\right] - k\mathrm{E}_{\Theta_a}\left[I_{C_r(k)}(\mathbf{X}) - I_{C_r^*}(\mathbf{X})\right] = 0.$$

Note further that the summand in the preceding summation is *nonpositive* valued $\forall \mathbf{x}$, as was shown in the proof of sufficiency above. Then the sum itself can be zero-valued only if the summand equals zero $\forall\ \mathbf{x}$. This in turn implies that $I_{C_r(k)}(\mathbf{x}) = I_{C_r^*}(\mathbf{x})\ \forall\ x \in \Gamma = \{\mathbf{x}: f(\mathbf{x};\Theta_0) \neq kf(\mathbf{x};\Theta_a)\}$. Thus *both* $C_r(k)$ and $C_r^*$ contain the set of x's, $\Gamma$, for which $f(\mathbf{x};\Theta_o) \neq kf(\mathbf{x};\Theta_a)$.

Regarding when $f(\mathbf{x};\Theta_o) = kf(\mathbf{x};\Theta_a)$ note that $\mathbf{x} \in C_r(k)\ \forall\ \mathbf{x} \in K = \{\mathbf{x}: f(\mathbf{x};\Theta_0) = kf(\mathbf{x};\Theta_a)\}$. Let $A = K - C_r^*$ be the set of $\mathbf{x}$ values satisfying $f(\mathbf{x};\Theta_0) = kf(\mathbf{x};\Theta_a)$ that are *not* in $C_r^*$. Then $P(\mathbf{x} \in A;\Theta_0) = P(\mathbf{x} \in A;\Theta_a) = 0$, for consider the contrary. If $P(\mathbf{x} \in A;\Theta_0) > 0$, then since $C_r(k)$ is an $\alpha$-level test and $C_r^* = C_r(k) - A$, it follows that $P(\mathbf{x} \in C_r^*;\Theta_0) < \alpha = P(\mathbf{x} \in C_r(k);\Theta_0)$, contradicting that $C_r^*$ is an $\alpha$-size critical region. If $P(\mathbf{x} \in A;\Theta_a) > 0$, then since $C_r^*$ does not contain $A$, $P(\mathbf{x} \in C_r^*;\Theta_a) < P(\mathbf{x} \in C_r(k);\Theta_a)$, contradicting that $C_r^*$ is a most powerful critical region. Thus, the most powerful critical region for testing $H_0: \Theta = \Theta_0$ versus $H_a: \Theta = \Theta_a$ has the form given in the statement of the theorem with probability 1. ■

Application of the Neyman–Pearson lemma for constructing most powerful tests of $H_0: \Theta = \Theta_0$ versus $H_a: \Theta = \Theta_a$ is relatively straightforward, at least in principle. In particular, one chooses the value of $k$ that defines the critical region in Theorem 9.1 having the desired size $\alpha = P(\mathbf{x} \in C_r(k);\Theta_0)$, and then the most powerful statistical test of having level $\alpha$ is defined by

$$\mathbf{x} \in C_r(k) \Rightarrow \text{reject } H_0 \quad \text{and} \quad \mathbf{x} \notin C_r(k) \Rightarrow \text{accept } H_0.$$

Note that since the probability calculation for determining the size of the critical region $C_r(k)$ is based on the density $f(\mathbf{x};\Theta_o)$, values of $\mathbf{x}$ for which $f(\mathbf{x};\Theta_o) = 0$ are irrelevant and can be ignored. Then an equivalent method of finding an $\alpha$-size most powerful test via the Neyman–Pearson lemma is to find the value of $k$ for which $P(f(\mathbf{x};\Theta_a)/f(\mathbf{x};\Theta_o) \geq k^{-1}) = \alpha$, and define $C_r(k)$ accordingly. Likewise, if the supports[13] of the densities $f(\mathbf{x};\Theta_o)$ and $f(\mathbf{x};\Theta_a)$ are the same, then another equivalent method is to find the value of $k$ for which $P(f(\mathbf{x};\Theta_o)/f(\mathbf{x};\Theta_a) \leq k) = \alpha$ and use it in the definition of $C_r(k)$. We illustrate the procedure in the following examples. Note that in the case where $\mathbf{X}$ is discrete, the feasible choices of $\alpha$ can be somewhat limited, as will be illustrated ahead.[14]

[13] Recall that the support of a density function is the set of $\mathbf{x}$ values for which $f(\mathbf{x};\Theta_0) > 0$, i.e., $\{\mathbf{x}: f(\mathbf{x};\Theta_0) > 0\}$ is the support of the density $f(\mathbf{x};\Theta_0)$.

[14] This limitation can be overcome, in principle, by utilizing what are known as **randomized tests**. Essentially, the test rule is made

**Example 9.10** Your company manufactures a hair-restoring chemical for middle-aged balding men and claims that the chemical will stimulate hair growth in 80 percent of the men who use it. A competitor of the company claims that your scientists must be referring to the complementary event by mistake—the counter claim is that your chemical stimulates hair growth in only 20 percent of the men who use it.

A government regulatory agency steps in to settle the dispute. It intends to test the hypothesis that the chemical is 20 percent effective versus the alternative hypothesis that it is 80 percent effective. A random sample of 20 middle-aged balding men received your company's hair-restoring treatment. The success ($x_i = 1$) or failure ($x_i = 0$) of the chemical in stimulating hair growth for any given individual is viewed as the outcome of a Bernoulli trial, where

$$X_i \sim p^{x_i}(1-p)^{1-x_i} I_{\{0,1\}}(x_i), \quad p \in \Omega = \{.2, .8\}.$$

The collection of 20 trials is viewed as a random sample from a Bernoulli population distribution, and letting $(X_1, \ldots, X_{20})$ represent the random sample of size 20 from a Bernoulli distribution, we then have that

$$f(\mathbf{x}; p) = p^{\sum_{i=1}^{20} x_i}(1-p)^{20-\sum_{i=1}^{20} x_i} \prod_{i=1}^{20} I_{\{0,1\}}(x_i), \quad p \in \Omega = \{.2, .8\}.$$

In order to define the most powerful level-$\alpha$ test of the hypothesis $H_0\colon p = .2$ versus the hypothesis $H_a\colon p = .8$, use the Neyman–Pearson lemma and examine

$$C_r(k) = \{\mathbf{x} : f(\mathbf{x}; .20)/f(\mathbf{x}; .80) \le k\} = \{\mathbf{x} : (.25)^{\sum_{i=1}^{20} x_i}(4)^{20-\sum_{i=1}^{20} x_i} \le k\}$$

$$= \left\{\mathbf{x} : \left(\sum_{i=1}^{20} x_i\right)\ln(.25) + \left(20 - \sum_{i=1}^{20} x_i\right)\ln(4) \le \ln k\right\}$$

$$= \left\{\mathbf{x} : \sum_{i=1}^{20} x_i \ge 10 - .36067 \ln k\right\} \quad \text{(to five decimal places).}$$

Since $Z = \sum_{i=1}^{20} X_i$ has a binomial distribution with parameters $p$ and $n = 20$, the relationship between the choice of $k$ and the size of the test implied by $C_r(k)$ is given by

$$\alpha(k) = P(\mathbf{x} \in C_r(k); p = .20) = \sum_{z \ge 10 - .36067 \ln(k)} \binom{20}{z}(.20)^z(.80)^{20-z} I_{\{0,1,\ldots,20\}}(z).$$

Some possible choices of test size are given as follows:

to depend not only on the outcomes of **X** but also on auxiliary random variables that are independent of **X**, so as to allow any level of test size to be achieved. However, the fact that the test outcome can depend on random variables that are independent of the experiment under investigation has discouraged its use in practice. For an introduction to the ideas involved, see M. Kendall, and A. Stuart (1979), *The Advanced Theory of Statistics*, Vol. 2, 4th ed. New York: MacMillan, pp. 180–181. Also, see Problem 9.8.

| ln($k$) | 10 − .36067 ln($k$) | $\alpha(k)$ |
|---|---|---|
| 24.95356 | 1 | .9885 |
| 22.18094 | 2 | .9308 |
| 19.40832 | 3 | .7939 |
| 16.63571 | 4 | .5886 |
| 13.86309 | 5 | .3704 |
| 11.09047 | 6 | .1958 |
| 8.31785 | 7 | .0867 |
| 5.54524 | 8 | .0321 |
| 2.77261 | 9 | .0100 |

Note that there are no other choices of $\alpha$ within the range [.01, .9885] other than the ones displayed above (why?). The most powerful statistical test of $H_0: p = .20$ versus $H_a: p = .80$ having level .01, say, is thus given by

$$\mathbf{x} \in C_r \Rightarrow \text{reject } H_0: p = .20 \quad \text{and} \quad \mathbf{x} \notin C_r \Rightarrow \text{accept } H_0: p = .20,$$

where we use $k = exp(2.77261) = 16$ to define

$$C_r = \{\mathbf{x}: f(\mathbf{x}; .20)/f(\mathbf{x}; .80) \leq 16\},$$

The critical region can be represented more simply using the test statistic $t(\mathbf{X}) = \sum_{i=1}^{n} X_i$ as

$$C_r = \left\{\mathbf{x}: \sum_{i=1}^{20} x_i \geq 9\right\}.$$

Suppose upon observing the successes and failures in the 20 treatments, the value of the test statistic was calculated to be $\sum_{i=1}^{20} x_i = 5$. It follows that $\mathbf{x} \notin C_r$, and the hypothesis $H_0: p = .2$ is *not* rejected. The government concludes that your competitor's claim has merit and instructs your company to refrain from claiming that your hair-restoring chemical is 80 percent effective. (Point to ponder: Would it be better if the *null* hypothesis were $H_0: p = .8$ and the *alternative* hypothesis were $H_a: p = .2$ in performing the statistical test? Why or why not?) □

The reader may have noticed that in Ex. 9.10, the choice of $k$ used in the Neyman–Pearson lemma to define a UMP level-$\alpha$ test of $H_0: \Theta = \Theta_0$ versus $H_a: \Theta = \Theta_a$ was not unique. For example, any choice of $k$ such that $2.77261 \leq \ln(k) < 5.54524$, would correspond to a critical region $C_r^* = \{\mathbf{x}: \sum_{i=1}^{20} x_i \geq k_*\}$ for $k_* \in (8, 9]$ that would have defined a UMP critical region of level .01. This is due to the fact that none of the values of $\sum_{i=1}^{20} x_i \in (8, 9)$, or equivalently, none of the points in $\{\mathbf{x}: 2.77261 < \ln(f(\mathbf{x}; \Theta_0)/f(\mathbf{x}; \Theta_a)) < 5.54524\}$ are in the support of $f(\mathbf{x}; \Theta_0)$, i.e., the event $\sum_{i=1}^{20} x_i \in (8, 9)$ has probability zero under $f(\mathbf{x}; \Theta_0)$. Henceforth, it will be tacitly understood that the value of $k$ is chosen in the Neyman–Pearson lemma so that $\mathbf{x}$ values in the event $\{\mathbf{x}: f(\mathbf{x}; \Theta_0) = kf(\mathbf{x}; \Theta_a)\}$ are in the support of $f(\mathbf{x}; \Theta_0)$. Nothing of any practical consequence is lost by the assumption, and it allows one to avoid an irrelevant ambiguity in the definition of UMP tests.

**Example 9.11** Your company is about to ship 10 LCD screens to a notebook computer manufacturing firm. The manufacturer requested that the screens have an expected operating life of 10,000 hours. Your company manufactures two types of screens—one having a 10,000-hour mean life, and one having a 50,000-hour mean life. The label on the batch of 10 screens you are about to send to the manufacturer is missing, and it is not known which type of screen you are about to send (although they were all taken from one of the screen production lines, so it is known they are all of the same type). The screens with a mean life of 50,000 hours cost considerably more to produce, and so you would rather not send the batch of 10 screens if they were the more expensive variety. The lifetimes of the screens can be nondestructively and inexpensively determined by a test you perform. Having observed the lifetimes of the screens, you wish to test the hypothesis that the mean lifetime of the screens is 10,000 hours versus the hypothesis that their mean lifetime is 50,000 hours. The lifetime distribution of each type of screen belongs to the exponential family of distributions as

$$X_i \sim \theta^{-1} e^{-x_i/\theta} I_{[0,\infty)}(x_i), \quad \theta \in \Omega = \{1, 5\},$$

where $x_i$ is measured in 10,000-hour units.

Let $(X_1, \ldots, X_{10})$ be the random sample of size 10 from the appropriate (unknown) exponential population distribution, so that

$$f(\mathbf{x}; \theta) = \theta^{-10} \exp\left(-\sum_{i=1}^{10} x_i/\theta\right) \prod_{i=1}^{10} I_{[0,\infty)}(x_i), \quad \theta \in \Omega = \{1, 5\}.$$

In order to define the most powerful level-$\alpha$ test of the hypothesis $H_0: \theta = 1$ versus the hypothesis $H_a: \theta = 5$, use the Neyman–Pearson lemma and examine

$$\begin{aligned} C_r(k) &= \{\mathbf{x}: f(\mathbf{x}; 1)/f(\mathbf{x}; 5) \leq k\} \\ &= \left\{\mathbf{x}: (.2)^{-10} \exp\left(-.8 \sum_{i=1}^{10} x_i\right) \leq k\right\} \\ &= \left\{\mathbf{x}: -10 \ln(.2) - .8 \sum_{i=1}^{10} x_i \leq \ln(k)\right\} \\ &= \left\{\mathbf{x}: \sum_{i=1}^{10} x_i \geq 20.11797 - 1.25 \ln(k)\right\} \quad \text{(to five decimal places).} \end{aligned}$$

Since $Z = \sum_{i=1}^{10} X_i$ has a Gamma distribution with parameters $\alpha = 10$ and $\beta = \theta$, the relationship between the choice of $k$ and the size of the test implied by $C_r(k)$ is given by

$$\alpha(k) = P(\mathbf{x} \in C_r(k); \theta = 1) = \int_{20.11797 - 1.25 \ln(k)}^{\infty} \frac{1}{\Gamma(10)} z^9 \exp(-z) dz,$$

integrating the Gamma distribution with $\alpha = 10$ and $\beta = 1$. Note that in this case, any choice for the size of the test, $\alpha$, within the range $(0, 1)$ is possible (why?). For example, to define the most powerful statistical test of $H_0: \theta = 1$ versus $H_a: \theta = 5$ having level .05, the value of $k$ would be found by solving the integral equation $\alpha(k) = .05$ for $k$ (recall the definition of $\alpha(k)$ above). Through the use of numerical integration on a computer, it can be shown that

$$\int_{15.70522}^{\infty} \frac{1}{\Gamma(10)} z^9 \exp(-z) dz = .05 \quad \text{(to five decimal places).}$$

Therefore, $k = 34.13079$ (to five decimal places), and the most powerful statistical test of level .05 is given by

$$\mathbf{x} \in C_r \Rightarrow \text{reject } H_0: \theta = 1 \quad \text{and} \quad \mathbf{x} \notin C_r \Rightarrow \text{accept } H_0: \theta = 1,$$

where

$$C_r = \{\mathbf{x}: f(\mathbf{x}; 1)/f(\mathbf{x}; 5) \leq 34.13079\}.$$

The critical region can be represented more simply using the test statistic $t(\mathbf{X}) = \sum_{i=1}^{n} X_i$ as

$$C_r = \{\mathbf{x}: \sum_{i=1}^{10} x_i \geq 15.70522\}.$$

Suppose the sum of the lifetimes of the 10 screens ultimately had an outcome equal to $\sum_{i=1}^{10} x_i = 54.4$. It follows that $\mathbf{x} \in C_r$, and the hypothesis $H_0: \theta = 1$ is *rejected* at the .05 level. You conclude that the batch of 10 screens are the type having a mean life of 50,000 hours. (Point to ponder: Would it be better if the *null* hypothesis were $H_0: \theta = 5$ and the *alternative* hypothesis were $H_a: \theta = 1$ in performing the statistical test? Why or why not?) □

The density ratio implied in the statement of the Neyman–Pearson lemma can itself be viewed as a test statistic, i.e., $t(\mathbf{X}) = f(\mathbf{X}; \Theta_a)/f(\mathbf{X}; \Theta_0)$ is a test statistic for given (hypothesized) values of $\Theta_0$ and $\Theta_a$. The test implied by the lemma can then be conducted by determining whether $t(\mathbf{x}) \geq c = k^{-1}$. However, it is often the case in practice that the density function of this statistic is difficult to identify and/or work with, and thus determining the appropriate choice of $k$ for an $\alpha$-size test, and determining the power function of the test, may be difficult or intractable with respect to the test statistic in this form.

We also note that the Neyman–Pearson lemma can be (and often is) stated in terms of likelihood functions instead of probability density functions. The alternative statement of the lemma then utilizes

$$C_r(k) = \{\mathbf{x}: L(\Theta_0; \mathbf{x}) \leq kL(\Theta_a; \mathbf{x})\}, \quad \text{for } k > 0$$

in the statement of Theorem 9.1. Of course, the proof of the restated lemma would be identical with that of Theorem 9.1 with $f(\mathbf{x}; \Theta_0)$ and $f(\mathbf{x}; \Theta_a)$ replaced by $L(\Theta_0; \mathbf{x})$ and $L(\Theta_a; \mathbf{x})$. When the critical region is expressed in terms of *likelihood ratio values* as $C_r = \{\mathbf{x}: L(\Theta_a; \mathbf{x})/L(\Theta_0; \mathbf{x}) \geq k^{-1}\}$, the test implied by the lemma is then referred to as a **likelihood ratio test**.

A most powerful test of $H_0: \Theta = \Theta_0$ versus $H_a: \Theta = \Theta_a$ defined via the Neyman–Pearson lemma is also an unbiased test, as we state formally next.

**Theorem 9.2** (Unbiasedness of Most Powerful Test of $H_0: \Theta = \Theta_0$ versus $H_a: \Theta = \Theta_a$) *Let $C_r$ represent a most powerful level-$\alpha$ critical region defined by the Neyman–Pearson lemma for testing $H_0: \Theta = \Theta_0$ versus $H_a: \Theta = \Theta_a$. Then the test implied by $C_r$ is unbiased.*

**Proof** Let $Z$ represent a Bernoulli random variable that is independent of the random sample $\mathbf{X}$ and for which $p(z = 1) = \alpha$ and $p(z = 0) = 1 - \alpha$. Suppose that regardless (i.e., independent) of the outcome of $\mathbf{X}$, the hypothesis $H_0$ is subsequently rejected or accepted based on whether the outcome of $Z$ is in the critical region $C_r^* = \{1\}$. Let $E_{\mathbf{X}}$ and $E_Z$ represent expectations taken with respect to the probability distributions of $\mathbf{X}$ and $Z$, respectively. It follows by the independence of $\mathbf{X}$ and $Z$ that $E_Z(I_{\{1\}}(Z) \mid \mathbf{x}) = \alpha$ is the probability of rejecting $H_0$ conditional on any given outcome of $\mathbf{X}$, *regardless* of the value of $\Theta \in \{\Theta_0, \Theta_a\}$. Then the (unconditional) probability of rejecting $H_0$ is given by an application of the double expectation theorem as

$$\pi_{C_r^*}(\Theta) = P_*(\text{reject } H_0; \Theta) = \mathrm{E}_{\mathbf{X}}\mathrm{E}_Z(I_{\{1\}}(Z) \mid \mathbf{X}) = \alpha, \quad \Theta \in \{\Theta_0, \Theta_a\}.$$

Now let $C_r$ represent a most powerful level-$\alpha$ test of $H_0: \Theta = \Theta_0$ versus $H_a: \Theta = \Theta_a$ defined via the Neyman–Pearson lemma. Using an argument analogous to that used in the sufficiency proof of the lemma (replacing $I_{C_r^*}(\mathbf{x})$ with $I_{\{1\}}(z)$,

$$[I_{C_r}(\mathbf{x}) - I_{\{1\}}(z)] f(\mathbf{x}; \Theta_a) h(z) \geq k^{-1} [I_{C_r}(\mathbf{x}) - I_{\{1\}}(z)] f(\mathbf{x}; \Theta_0) h(z), \quad \forall\, \mathbf{x}, z,$$

where $h(z)$ represents the Bernoulli density function for $Z$. Integrating both sides of the inequality over $\mathbf{x} \in R^n$ if $\mathbf{X}$ is continuous or else summing over all $\mathbf{x}$ values for which $f(\mathbf{x}; \Theta_0) > 0$ or $f(\mathbf{x}; \Theta_a) > 0$ if $\mathbf{X}$ is discrete, and then summing over the range of $Z$ obtains

$$P(\mathbf{x} \in C_r; \Theta_a) - \alpha \geq k^{-1} [P(\mathbf{x} \in C_r; \Theta_0) - \alpha] = 0,$$

where the right-hand side of the inequality equals 0 because $C_r$ is an $\alpha$-size critical region. Then $P(\mathbf{x} \in C_r; \Theta_a) \geq \alpha$, so that the test implied by $C_r$ is unbiased. ■

The significance of Theorem 9.2 for applications is that once a most powerful level-$\alpha$ test of $H_0: \Theta = \Theta_0$ versus $H_a: \Theta = \Theta_a$ has been derived via the Neyman–Pearson lemma, there is no need to check whether the test is unbiased, since such tests are *always* unbiased. As an illustration, the tests in Ex. 9.10 and Ex. 9.11 are unbiased since they were most powerful Neyman–Pearson tests. We note, however, that unbiased tests are not necessarily most powerful.

*Composite Hypotheses* In some cases, the Neyman–Pearson lemma can also be used to identify UMP tests when the alternative hypothesis is composite. The basic idea is to show via the Neyman–Pearson lemma that the critical region, $C_r$, of the most powerful test of $H_0: \Theta = \Theta_0$ versus $H_a: \Theta = \Theta_a$ is the *same* $\forall\, \Theta_a \in \Omega_a$, where $\Omega_a$ is the set of alternative hypothesis values. It would then follow that the critical region defines a *uniformly* most powerful test of

$H_0\colon \Theta = \Theta_0$ versus $H_a\colon \Theta \in \Omega_a$, since the *same* critical region would be most powerful for each and every $\Theta \in \Omega_a$, i.e., uniformly in $\Theta \in \Omega_a$. We formalize the approach in the following theorem.

**Theorem 9.3** (UMP Test of $H_0\colon \Theta = \Theta_0$ Versus $H_a\colon \Theta \in \Omega_a$ Using Neyman–Pearson lemma) *The given critical region, $C_r$, of a statistical test of $H_0\colon \Theta = \Theta_0$ versus $H_a\colon \Theta \in \Omega_a$ defines a UMP level-$\alpha$ test if $P(\mathbf{x} \in C_r; \Theta_0) = \alpha$ and $\forall\, \Theta_a \in \Omega_a$, $\exists\, k(\Theta_a) \geq 0$ such that the given critical region can be defined as*

$$C_r = \{\mathbf{x}\colon f(\mathbf{x};\Theta_0) \leq k(\Theta_a) f(\mathbf{x};\Theta_a)\}.$$

*Furthermore, $C_r$ is then the unique (with probability 1) UMP critical region of size $\alpha$.*

**Proof** Since $C_r$ is defined via the Neyman–Pearson lemma, it represents the most powerful level-$\alpha$ test of $H_0\colon \Theta = \Theta_0$ versus $H_a\colon \Theta = \Theta_a$. Since the *same* critical region, $C_r$, applies for any $\Theta_a \in \Omega_a$ the test defined by $C_r$ is also most powerful for *any* choice of $\Theta_a \in \Omega_a$ and is thus a UMP level-$\alpha$ test of $H_0\colon \Theta = \Theta_0$ versus $H_a\colon \Theta \in \Omega_a$. Finally, $C_r$ is the unique size $\alpha$ critical region (with probability 1) since the most powerful critical region of size $\alpha$ is unique (with probability 1) by the Neyman–Pearson lemma $\forall\, \Theta_a$. ■

**Example 9.12** Recall Ex. 9.10 in which hypotheses regarding the effectiveness of a hair-restoring chemical were being analyzed. Examine the problem of testing the simple hypothesis $H_0\colon p = .2$ versus the composite alternative hypothesis $H_a\colon p > .2$ using a UMP test having level .01. Let $p_a$ represent any choice of $p \in (.2, 1]$, and define the critical region of the test of $H_0\colon p = .2$ versus $H_a\colon p = p_a$ using the Neyman–Pearson lemma:

$$\begin{aligned} C_r &= \{\mathbf{x}\colon f(\mathbf{x}; .2)/f(\mathbf{x}; p_a) \leq k(p_a)\} \\ &= \left\{\mathbf{x}\colon (.2/p_a)^{\sum_{i=1}^{20} x_i}(.8/(1-p_a))^{20-\sum_{i=1}^{20} x_i} \leq k(p_a)\right\} \\ &= \left\{\mathbf{x}\colon \sum_{i=1}^{20} x_i \geq \frac{\ln(k(p_a)) - 20\ln(.8/(1-p_a))}{\ln(.25(1-p_a)/p_a)}\right\}. \end{aligned}$$

Using reasoning identical to that used in Ex. 9.10, $C_r$ will define a size .01 test iff

$$C_r = \{\mathbf{x}\colon \sum_{i=1}^{20} x_i \geq 9\}$$

*regardless* of the value of $p_a$, since under $H_0$ the statistic $\sum_{i=1}^{20} X_i$ has a binomial distribution with $n = 20$ and $p = .2$. The nonnegative values of $k(p_a)\, \forall\, p_a \in (.2, 1]$ can be solved for accordingly, and thus by Theorem 9.3, $C_r = \{\mathbf{x}\colon \sum_{i=1}^{20} x_i \geq 9\}$ defines a UMP level-.01 test of $H_0\colon p = .2$ versus $H_a\colon p > .2$. □

**Example 9.13** Recall Ex. 9.11 in which hypotheses regarding the operating lives of screens for notebook computers were being analyzed. Examine the problem of testing the simple hypothesis $H_0: \theta = 1$ versus the composite hypothesis $H_a: \theta > 1$ using a UMP test having level .05. Let $\theta_a$ represent any choice of $\theta > 1$, and define the critical region of the test of $H_0: \theta = 1$ versus $H_a: \theta = \theta_a$ using the Neyman–Pearson lemma:

$$\begin{aligned} C_r &= \{\mathbf{x}: f(\mathbf{x}; 1)/f(\mathbf{x}; \theta_a) \leq k(\theta_a)\} \\ &= \left\{\mathbf{x}: \theta_a^{10} \exp\left(-\sum_{i=1}^{10} x_i \left[1 - \theta_a^{-1}\right]\right) \leq k(\theta_a)\right\} \\ &= \left\{\mathbf{x}: \sum_{i=1}^{10} x_i \geq \frac{10 \ln(\theta_a) - \ln(k(\theta_a))}{(1 - \theta_a^{-1})}\right\}. \end{aligned}$$

Using reasoning identical to that used in Ex. 9.11, $C_r$ will define a size .05 test iff

$$C_r = \left\{\mathbf{x}: \sum_{i=1}^{10} x_i \geq 15.70522\right\}$$

*regardless* of the value of $\theta_a$, since under $H_0$ the statistic $\sum_{i=1}^{10} X_i$ has a Gamma distribution with parameters $\alpha = 10$ and $\beta = 1$. The nonnegative values of $k(\theta_a)\ \forall\ \theta_a \in (1, \infty)$ can be solved for accordingly, and thus by Theorem 9.3, $C_r = \{\mathbf{x}: \sum_{i=1}^{10} x_i \geq 15.70522\}$ defines a UMP level-.05 test of $H_0: \theta = 1$ versus $H_a: \theta > 1$. □

The uniqueness result of Theorem 9.3 can be used to demonstrate that a UMP test of $H_0: \Theta = \Theta_0$ versus $H_a: \Theta \in \Omega_a$ having size $\alpha$ does *not* exist, as in the following example.

**Example 9.14** Nonexistence of a UMP Test

An untrusting gambler wishes to use a size .10 test that is also UMP level .10 to test the hypothesis that a roulette wheel being used for betting purposes in an Atlantic City casino is fair. The gambler suggests that the wheel be spun 100 times and the random-sample outcome of red and black outcomes be used to test the "fairness" hypothesis. The joint density of the random sample in this case is given by

$$f(\mathbf{x}; p) = p^{\sum_{i=1}^{100} x_i}(1 - p)^{100 - \sum_{i=1}^{100} x_i} \prod_{i=1}^{100} I_{\{0,1\}}(x_i)$$

(assuming the red/black outcomes are iid Bernoulli trials), where $x_i = 1$ denotes red and $x_i = 0$ denotes black. The null hypothesis to be tested is $H_0: p = .5$ versus the alternative hypothesis that $H_a: p \neq .5$. In the notation of Theorem

9.3, $\Omega_a = [0,1] - .5$. Following Theorem 9.3, define a critical region as

$$C_r = \left\{\mathbf{x} : \frac{f(\mathbf{x}; .5)}{f(\mathbf{x}; p_a)} \le k(p_a)\right\}$$

$$= \left\{\mathbf{x} : (.5)^{100} / \left[p_a^{\sum_{i=1}^{100} x_i}(1-p_a)^{100-\sum_{i=1}^{100} x_i}\right] \le k(p_a)\right\}$$

$$= \left\{\mathbf{x} : \ln(p_a/(1-p_a))\sum_{i=1}^{100} x_i \ge \gamma(p_a)\right\},$$

where $\gamma(p_a) = 100\ln(.5/(1-p_a)) - \ln(k(p_a))$. The critical region depends on $p_a$ (i.e., $C_r$ is *not* the same $\forall\, p_a \in \Omega_a$), and thus there is *no* UMP level-.10 statistical test of the "fairness" hypothesis that has size .10. To show how $C_r$ depends on $p_a$, note that for $p_a > .5$, $\ln(p_a/(1-p_a)) > 0$, while for $p_a < .5$, $\ln(p_a/(1-p_a)) < 0$. It follows that

$$C_r = \left\{\mathbf{x} : \sum_{i=1}^{100} x_i \ge \eta(p_a)\right\} \quad \text{if } p_a > .5$$

or

$$C_r = \left\{\mathbf{x} : \sum_{i=1}^{100} x_i \le \eta(p_a)\right\} \quad \text{if } p_a < .5,$$

where $\eta(p_a) = \gamma(p_a)/\ln(p_a/(1-p_a))$. Given that the size of the test is to be .10, the reader can show by using the binomial distribution of $\sum_{i=1}^{n} X_i$ with parameters $n = 100$ and $p = .5$ that $P\left(\sum_{i=1}^{100} x_i \ge 57; p = .5\right) = P\left(\sum_{i=1}^{100} x_i \le 43; p = .5\right) = .10$ (to two decimal places). Thus, for $P(\mathbf{x} \in C_r; .5) = .10$, *two different* critical regions are defined as

$$C_r = \left\{\mathbf{x} : \sum_{i=1}^{100} x_i \ge 57\right\} \quad \text{for } p_a > .5$$

or

$$C_r = \left\{\mathbf{x} : \sum_{i=1}^{100} x_i \le 43\right\} \quad \text{for } p_a < .5.$$

For values of $p_a > .5$, the first critical region is UMP of level .10 by Theorem 9.3, while for values of $p_a < .5$, the second critical region is UMP of level .10 by Theorem 9.3. Since the UMP level-.10 critical regions for $p_a > .5$ and $p_a < .5$ are not the same, no UMP level-.10 critical region exists for testing the fairness hypothesis $H_0: p = .5$ versus $H_a: p \neq .5$. □

The previous three examples illustrate the notion of *one-sided* and *two-sided alternative hypotheses*. In the current context, a **one-sided alternative hypothesis** is one for which *either* $(\Theta_a - \Theta_0) > 0\ \forall\ \Theta_a \in \Omega_a$, *or* $(\Theta_a - \Theta_0) < 0\ \forall\ \Theta_a \in \Omega_a$ (i.e., all values of $\Theta_a \in \Omega_a$ are larger than $\Theta_0$, or else they are all smaller than $\Theta_0$, so that the values of $\Theta_a$ are all on "one side" of $\Theta_0$). A **two-sided alternative hypothesis** is such that $\exists\ \Theta_a \in \Omega_a$ for which $(\Theta_a - \Theta_0) > 0$ *and* $\exists\ \Theta_a \in \Omega_a$ for which $(\Theta_a - \Theta_0) < 0$ (i.e., some values of $\Theta_a \in \Omega_a$ are larger than $\Theta_0$ *and* some values of $\Theta_a \in \Omega_a$ are smaller than $\Theta_0$, so there are values of $\Theta_a$ on "both sides" of $\Theta_0$). In practice, UMP level-$\alpha$ tests of simple hypotheses versus *one-sided alternative hypotheses* often exist when $\Theta$ is a scalar, but such is not the case when the alternative hypothesis is two-sided. In the latter case, one must generally resort to seeking a UMP test within a smaller class of tests, such as the class of unbiased tests. We will examine this case in more detail later.

A UMP level-$\alpha$ test of $H_0$: $\Theta = \Theta_0$ versus $H_a$: $\Theta \in \Omega_a$ defined using Theorem 9.3 is also an unbiased test, as indicated in the following extension of Theorem 9.2.

**Theorem 9.4** (Unbiasedness of Uniformly Most Powerful Test of $H_0$: $\Theta = \Theta_0$ versus $H_a$: $\Theta \in \Omega_a$) *Let $C_r$ represent a UMP level-$\alpha$ critical region for testing $H_0$: $\Theta = \Theta_0$ versus $H_a$: $\Theta \in \Omega_a$ defined using the Neyman–Pearson lemma as indicated in Theorem 9.3. Then the test implied by $C_r$ is unbiased.*

**Proof** Let $\Theta_a$ be any choice of $\Theta \in \Omega_a$. Using an argument analogous to the proof of Theorem 9.2, it can be shown that

$$P(\mathbf{x} \in C_r; \Theta_a) - \alpha \geq k(\Theta_a)^{-1}[P(\mathbf{x} \in C_r; \Theta_0) - \alpha] = 0,$$

where, again, the right-hand side of the inequality is 0 because $C_r$ is a size $\alpha$ critical region. Since this holds $\forall\ \Theta_a \in \Omega_a$, it follows that $P(\mathbf{x} \in C_r; \Theta_a) \geq \alpha$ $\forall\ \Theta_a \in \Omega_a$, so that the test implied by $C_r$ is unbiased. ■

The theorem implies that once a UMP level-$\alpha$ test of $H_0$: $\Theta = \Theta_0$ versus $H_a$: $\Theta \in \Omega_a$ has been found via the Neyman–Pearson approach, one need not check to see whether the test is unbiased, since such tests are *always* unbiased. Examples 9.12 and 9.13 illustrate this fact, where both UMP level-$\alpha$ tests are also unbiased tests. However, it should be noted that an unbiased test is not necessarily a UMP test.

### Monotone Likelihood Ratio Approach

In this subsection we will establish results that will be useful for constructing UMP level-$\alpha$ tests of the *composite* null hypothesis $H_0$: $\Theta \leq \Theta_0$ (or $H_0$: $\Theta \geq \Theta_0$) versus the composite *one-sided alternative* hypothesis $H_a$: $\Theta > \Theta_0$ (or $H_a$: $\Theta < \Theta_0$). The results in this section will apply equally well to the case where $H_0$: $\Theta = \Theta_0$, and thus this section can be interpreted as also providing additional results for testing $H_0$: $\Theta = \Theta_0$ versus the composite *one-sided alternative* $H_a$: $\Theta > \Theta_0$ (or $H_a$: $< \Theta_0$).

The procedure for defining UMP level-$\alpha$ tests that we will present relies on the concept of a *monotone likelihood ratio in the statistic* $T = t(\mathbf{X})$.

**Definition 9.13** *Monotone likelihood ratio in statistic $T = t(X)$*

Let $f(\mathbf{x}; \Theta)$, $\Theta \in \Omega$, be a family of probability density functions indexed by a scalar parameter $\Theta$. The family of density functions is said to have a **monotone likelihood ratio in the statistic $T = t(\mathbf{X})$** iff $\forall\ \Theta_1, \Theta_2 \in \Omega$ for which $\Theta_1 > \Theta_2$, the likelihood ratio $L(\Theta_1; \mathbf{x})/L(\Theta_2; \mathbf{x})$ can be expressed as a nondecreasing function of $t(\mathbf{x})$ $\forall\ \mathbf{x} \in \{\mathbf{x}: f(\mathbf{x}; \Theta_1) > 0 \text{ and/or } f(\mathbf{x}; \Theta_2) > 0\}$.

Verification of whether a family of density functions has a monotone likelihood ratio in some statistic $t(\mathbf{X})$ generally requires some degree of ingenuity. However, if the family of density functions belongs to the exponential class of densities, the verification process is often simplified by the following result.

**Theorem 9.5** (Monotone Likelihood Ratio and the Exponential Class of Densities) *Let $f(\mathbf{x};\Theta)$, $\Theta \in \Omega$, be a one-parameter density family belonging to the exponential class of densities, as*

$$f(\mathbf{x}; \Theta) = \exp(c(\Theta)g(\mathbf{x}) + d(\Theta) + z(\mathbf{x}))I_A(\mathbf{x}), \quad \Theta \in \Omega.$$

*If $c(\Theta)$ is a nondecreasing function of $\Theta$, then $f(\mathbf{x};\Theta)$, $\Theta \in \Omega$, has a monotone likelihood ratio in the statistic $g(\mathbf{X})$.*

**Proof** Let $\Theta_1 > \Theta_2$, and $\forall\ \mathbf{x} \in A$ examine the likelihood ratio

$$L(\Theta_1; \mathbf{x})/L(\Theta_2; \mathbf{x}) = \exp\left([c(\Theta_1) - c(\Theta_2)]\, g(\mathbf{x}) + d(\Theta_1) - d(\Theta_2)\right) = \eta(g(\mathbf{x})) = e^{c_* g(\mathbf{x}) + d_*},$$

where $c_* = c(\Theta_1) - c(\Theta_2)$ and $d_* = d(\Theta_1) - d(\Theta_2)$. Since $c(\Theta)$ is a nondecreasing function of $\Theta$, $c_* \geq 0$ so that the likelihood ratio can be expressed as a nondecreasing function of $g(\mathbf{x})$, $\forall\ \mathbf{x} \in A$. ■

The following examples illustrate the use of Theorem 9.5 for verifying the monotone likelihood ratio property.

**Example 9.15** Let $(X_1, \ldots, X_n)$ be a random sample from a Bernoulli distribution representing the population of television viewers in a certain region who can $(x_i = 1)$ or cannot $(x_i = 0)$ recall seeing a certain television commercial. Then

$$\begin{aligned} f(\mathbf{x}; p) &= p^{\sum_{i=1}^n x_i}(1 - p)^{n - \sum_{i=1}^n x_i} \prod_{i=1}^n I_{\{0,1\}}(x_i) \\ &= \exp(c(p)g(\mathbf{x}) + d(p) + z(\mathbf{x}))I_A(\mathbf{x}), \quad p \in (0, 1), \end{aligned}$$

where $c(p) = \ln(p/(1 - p))$, $g(\mathbf{x}) = \sum_{i=1}^n x_i$, $d(p) = n \ln(1 - p)$, $z(\mathbf{x}) = 0$, and $A = \times_{i=1}^n \{0, 1\}$. Then since $dc(p)/dp = [p(1 - p)]^{-1} > 0$, $c(p)$ is strictly increasing in $p$, and $f(\mathbf{x}; p)$, $p \in (0, 1)$, has a monotone likelihood ratio in the statistic $g(\mathbf{X}) = \sum_{i=1}^n X_i$. □

**Example 9.16** Let $(X_1, \ldots, X_n)$ be a random sample from a Gamma population distribution with $\beta = 2$ representing the survival time of cancer patients treated with a new form of chemotherapy. Then

$$f(\mathbf{x};\alpha) = \frac{1}{2^{n\alpha}[\Gamma(\alpha)]^n}\prod_{i=1}^{n} x_i^{\alpha-1}e^{-\sum_{i=1}^{n} x_i/2}\prod_{i=1}^{n} I_{(0,\infty)}(x_i)$$

$$= \exp(c(\alpha)g(\mathbf{x}) + d(\alpha) + z(\mathbf{x}))I_A(\mathbf{x}), \quad \alpha > 0,$$

where $c(\alpha) = \alpha - 1$, $g(\mathbf{x}) = \sum_{i=1}^{n} \ln x_i$, $d(\alpha) = -n\ln(2^{\alpha}\Gamma(\alpha))$, $z(\mathbf{x}) = -(1/2)\sum_{i=1}^{n} x_i$, and $A = \times_{i=1}^{n}(0,\infty)$. Then since $dc(\alpha)/d\alpha = 1 > 0$, $c(\alpha)$ is strictly increasing in $\alpha$ and $f(\mathbf{x};\alpha)$, $\alpha > 0$, has a monotone likelihood ratio in the statistic $g(\mathbf{X}) = \sum_{i=1}^{n} \ln X_i$. □

In the next example, we illustrate verification of the monotone likelihood ratio property for a family of density functions that does *not* belong to the exponential class of densities.

**Example 9.17** Let $X$ have a hypergeometric probability density function in which the parameters $n$ and $M$ have known values $n_0$ and $M_0$, respectively, with $X$ representing the number of defectives found in a random sample, without replacement, of size $n_0$ from a shipment of $M_0$ compact disks. The density function for $X$ can then be represented as

$$f(x;K) = \frac{\binom{K}{x}\binom{M_0-K}{n_0-x}}{\binom{M_0}{n_0}} I_{\{0,1,2,\ldots,K\}}(x)$$

for $K \in \{0, 1, 2, \ldots, M_0\}$. Examine the likelihood ratio

$$\frac{L(K;x)}{L(K-1;x)} = \left[\frac{K}{(K-x)}\right]\left[\frac{(M_0 - K - n_0 + 1 + x)}{(M_0 - K + 1)}\right]\left[\frac{I_{\{0,1,\ldots,K\}}(x)}{I_{\{0,1,\ldots,K-1\}}(x)}\right]$$

(recall Ex. 8.8), and note that the ratio is a nondecreasing function of $x \in \{0, 1, \ldots, K\} = \{x : f(x;K) > 0 \text{ and/or } f(x;K-1) > 0\}$, since the ratio is a product of three (in brackets) nondecreasing functions of $x$. Now let $K_1 > K_2$, and note that the likelihood ratio

$$\frac{L(K_1;x)}{L(K_2;x)} = \prod_{i=0}^{K_1-K_2-1}\left[\frac{L(K_1 - i;x)}{L(K_1 - (i+1);x)}\right]$$

$$= \prod_{i=0}^{K_1-K_2-1}\left(\left[\frac{K_1 - i}{K_1 - i - x}\right]\left[\frac{(M_0 - (K_1 - i) - n_0 + 1 + x)}{(M_0 - (K_1 - i) + 1)}\right]\right) \quad \text{for } x \in \{0, 1, \ldots, K_2\}$$

$$= \infty \quad \text{for } x \in \{K_2 + 1, \ldots, K_1\}$$

is a nondecreasing function of $x \in \{0, 1, \ldots, K_1\} = \{x : f(x;K_1) > 0 \text{ and/or } f(x;K_2) > 0\}$, since it is strictly increasing for $x \in \{0, 1, \ldots, K_2\}$, and equals $\infty$ (and hence nondecreasing) for $x \in \{K_2 + 1, \ldots, K_1\}$. Then the hypergeometric family of densities $f(x;K)$, $K \in \{0, 1, \ldots, M_0\}$, has a monotone likelihood ratio in the statistic $t(X) = X$. □

If the joint density function of the random sample can be shown to have a monotone likelihood ratio in some statistic, then UMP level-$\alpha$ tests of $H_0: \Theta \leq \Theta_0$ versus $H_a: \Theta > \Theta_0$, or $H_0: \Theta \geq \Theta_0$ versus $H_a: \Theta < \Theta_0$, or $H_0: \Theta = \Theta_0$ versus one of *either* $H_a: \Theta > \Theta_0$ *or* $H_a: \Theta < \Theta_0$ will exist. In fact, ahead it will be seen that a monotone likelihood ratio is effectively a sufficient condition for the Neyman–Pearson approach of Theorem 9.3 to be applicable to the testing problem.

**Theorem 9.6** (Monotone Likelihood Ratios and UMP Level-$\alpha$ Tests) *Let $f(\mathbf{x};\Theta), \Theta \in \Omega$, be a family of density functions having a monotone likelihood ratio in the statistic $t(\mathbf{X})$. Let $C_r = \{\mathbf{x}: f(\mathbf{x};\Theta_0) \leq kf(\mathbf{x};\Theta_a)\}$ be a size $\alpha$ critical region for testing $H_0: \Theta = \Theta_0$ versus $H_a: \Theta = \Theta_a$*

1. *If $\Theta_0 < \Theta_a$, then $C_r$ is UMP level $\alpha$ for testing either*
   - **a.** *$H_0: \Theta = \Theta_0$ versus $H_a: \Theta > \Theta_0$, or*
   - **b.** *$H_0: \Theta \leq \Theta_0$ versus $H_a: \Theta > \Theta_0$.*
2. *If $\Theta_0 > \Theta_a$, then $C_r$ is UMP level $\alpha$ for testing either*
   - **a.** *$H_0: \Theta = \Theta_0$ versus $H_a: \Theta < \Theta_0$, or*
   - **b.** *$H_a: \Theta \geq \Theta_0$ versus $H_a: \Theta < \Theta_0$.*

**Proof** We prove part (1). The proof of part (2) is analogous with appropriate inequality reversals and is left to the reader.

Since $f(\mathbf{x};\Theta_a)/f(\mathbf{x};\Theta_0)$ has a monotone likelihood ratio in $t(\mathbf{x})$,

$f(\mathbf{x};\Theta_a)/f(\mathbf{x};\Theta_0) \geq k^{-1}$ iff $g(t(\mathbf{x})) \geq k^{-1}$

for some nondecreasing function $g(\cdot)$, and then $g(t(\mathbf{x})) \geq k^{-1}$ iff $t(\mathbf{x}) \geq c$, where $c$ is chosen to satisfy[15] $\min_c g(c) \geq k^{-1}$. A Neyman–Pearson most powerful level-$\alpha$ test of $H_0: \Theta = \Theta_0$ versus $H_a: \Theta = \Theta_a$, for $\Theta_a > \Theta_0$, is defined by choosing $c$ so that $P(t(\mathbf{x}) \geq c; \Theta_0) = \alpha$. Note that this probability depends only on $\Theta_0$, so that the choice of $c$ does *not* depend on $\Theta_a$ and thus the *same* critical region $C_r = \{\mathbf{x}: t(\mathbf{x}) \geq c\}$ defines a Neyman–Pearson level-$\alpha$ test of $H_0: \Theta = \Theta_0$ versus $H_a: \Theta = \Theta_a$, *regardless* of the value of $\Theta_a > \Theta_0$. Thus, by Theorem 9.3, $C_r$ as defined in the current theorem is UMP level-$\alpha$ for testing $H_0: \Theta = \Theta_0$ versus $H_a: \Theta > \Theta_0$.

Now examine $\Theta_* < \Theta_0$, and let $\alpha_* = P(\mathbf{x} \in C_r; \Theta_*)$. From the preceding argument, it is known that $C_r$ represents a UMP level-$\alpha_*$ test of $H_0: \Theta = \Theta_*$ versus $H_a: \Theta > \Theta_*$, and since the test is then unbiased (Theorem 9.4), we know that $P(\mathbf{x} \in C_r; \Theta_*) \leq P(\mathbf{x} \in C_r; \Theta_0) = \alpha$, which holds $\forall\ \Theta_* < \Theta_0$. Thus, $C_r$ represents a level-$\alpha$ test of $H_0: \Theta \leq \Theta_0$ versus $H_a: \Theta > \Theta_0$.

Finally, let $C_r^*$ represent any other level-$\alpha$ test for testing $H_0: \Theta \leq \Theta_0$ versus $H_a: \Theta > \Theta_0$. Then by definition $\alpha_0 = P(x \in C_r^*; \Theta_0) \leq \sup_{\Theta \leq \Theta_0} P(\mathbf{x} \in C_r^*; \Theta) \leq \alpha$, so that $C_r^*$ has size $\alpha_0 \leq \alpha$ for testing $H: \Theta = \Theta_0$ versus $H_a: \Theta > \Theta_0$. Then

[15] If the likelihood ratio is strictly increasing in $t(\mathbf{x})$, then $c$ can be chosen to satisfy $g(c) = k^{-1}$.

since $C_r$ is UMP level-$\alpha$ for testing this hypothesis, $P(\mathbf{x} \in C_r; \Theta) \geq P(\mathbf{x} \in C_r^*; \Theta)$ $\forall\, \Theta > \Theta_0$, and thus $C_r$ is also a UMP level $\alpha$ test of $H_0\colon \Theta \leq \Theta_0$ versus $H_a\colon \Theta > \Theta_0$. ■

The UMP level-$\alpha$ critical region defined by Theorem 9.6 can be defined alternatively in terms of the statistic $t(\mathbf{X})$ for which the likelihood ratio is monotone.

**Corollary 9.1** (UMP Level-$\alpha$ Tests in Terms of the $t(\mathbf{X})$ of a Monotone Likelihood Ratio) *Let $f(\mathbf{x};\Theta), \Theta \in \Omega$, be a family of density functions having a monotone likelihood ratio in $t(\mathbf{X})$. Then*

1. $C_r = \{\mathbf{x}\colon\ t(\mathbf{x}) \geq c\}$, *for the choice of c such that* $P(t(\mathbf{x}) \geq c; \Theta_0) = \alpha$, *is a UMP level-$\alpha$ critical region for testing* $H_0\colon \Theta = \Theta_0$ *versus* $H_a\colon \Theta > \Theta_0$, *or* $H_0\colon \Theta \leq \Theta_0$ *versus* $H_a\colon \Theta > \Theta_0$.
2. $C_r = \{\mathbf{x}\colon\ t(\mathbf{x}) \leq c\}$, *for choice of c such that* $P(t(\mathbf{x}) \leq c; \Theta_0) = \alpha$, *is a UMP level-$\alpha$ critical region for testing* $H_0\colon \Theta = \Theta_0$ *versus* $H_a\colon \Theta < \Theta_0$, *or* $H_0\colon \Theta \geq \Theta_0$ *versus* $H_a\colon \Theta < \Theta_0$.

**Proof** The $C_r$'s are equivalent representations of the Neyman–Pearson critical regions referred to in Theorem 9.6. ■

It is also true that the UMP level-$\alpha$ tests based on the monotone likelihood ratio procedure are unbiased tests. We state this fact as a second corollary to that theorem.

**Corollary 9.2** (Unbiasedness of UMP Tests Based on Monotone Likelihood Ratios) *The critical region, $C_r$, defined in Theorem 9.6 and Corollary 9.1, defines an unbiased size $\alpha$ test of the respective hypotheses stated in the theorem.*

**Proof** This follows immediately from the characteristics of the power function of the test established in the proof of Theorem 9.6. ■

**Example 9.18** Your company manufactures personal computers and is in the process of evaluating the purchase of floppy disk controllers from various input suppliers. Among other considerations, a necessary condition for a disk controller to be used in the manufacture of your PCs is that the controller have a minimum expected life of more than 5,000 operating hours. The lifetimes of all of the various brands of disk controllers are characterized by exponential densities, as

$$X_i \sim f(x_i;\theta) = \theta^{-1}\exp\left[-x_i/\theta\right] I_{(0,\infty)}(x_i), \quad \theta > 0,$$

where $x_i$ is measured in thousand's of operating hours. Define a UMP level-.01 test of $H_0\colon \theta \leq 5$ versus $H_a\colon \theta > 5$ based on a random sample of 100 lifetimes of a given brand of disk controller.

**Answer**: The joint density of the random sample is given by $f(\mathbf{x};\theta) = \theta^{-100}\exp[-\sum_{i=1}^{100} x_i/\theta] \prod_{i=1}^{100} I_{(0,\infty)}(x_i), \theta > 0$. The density is in the exponential class with $c(\theta) = -\theta^{-1}$, which is a nondecreasing function of $\theta$, so that $f(\mathbf{x};\theta)$ has a monotone likelihood ratio by Theorem 9.6. In particular, the likelihood ratio $L(\theta_1; \mathbf{x})/L(\theta_2; \mathbf{x}) \propto \exp[-\sum_{i=1}^{100} x_i(\theta_1^{-1} - \theta_2^{-1})]$ is a *strictly increasing* function of $t(\mathbf{x}) = \sum_{i=1}^{100} x_i$ for $\theta_1 > \theta_2 \; \forall \, \mathbf{x} \in \times_{i=1}^{100}(0, \infty)$. Then Theorem 9.6 and its corollaries, are applicable, so that the UMP critical region of level .01 is given by

$$C_r = \{\mathbf{x} : \sum_{i=1}^{100} x_i \geq c\} \quad \text{for } c \text{ chosen so that } P(\mathbf{x} \in C_r; \theta = 5) = .01.$$

To calculate the appropriate value of $c$, first note that $Z = \sum_{i=1}^{100} X_i$ has a Gamma distribution, with parameters $\alpha = 100$ and $\beta = \theta$.[16] Then $c$ is the solution to the integral equation

$$\int_c^\infty \frac{1}{\theta^{100}\Gamma(100)} z^{99} \exp(-z/\theta) dz = .01$$

with $\theta = 5$, which can be found with the aid of a computer to be $c = 623.61$. The UMP level-.01 and unbiased test of $H_0: \theta \leq 5$ versus $H_a: \theta > 5$ is then defined by the critical region for the test statistic $T = t(\mathbf{X}) = \sum_{i=1}^n X_i$ given by $C_r^T = [623.61, \infty)$. An alternative test statistic for performing this UMP level-.01 and unbiased test is the sample mean so that the test can also be conducted using the critical region for the test statistic $\bar{X}$ given by $C_r^{\bar{x}} = [6.2361, \infty)$. □

**Example 9.19** A shipment of 300 blank video cassettes arrives at your video store. You intend to randomly sample, *without replacement*, 25 video cassettes and observe whether or not they are defective. On the basis of the sample outcome, you wish to test the hypothesis that the shipment of video cassettes contains no more than 5 percent defectives. Define a UMP and unbiased test of the hypothesis having level .05 and a size as close to .05 as possible.[17]

**Answer**: The density function associated with the number of defectives observed in a random sample, without replacement, of size 25 from the population of 300 video cassettes can be defined as

$$f(x; K) = \frac{\binom{K}{x}\binom{300-K}{25-x}}{\binom{300}{25}} I_{\{0,1,2,\ldots,25\}}(x),$$

i.e., hypergeometric with $M = 300$, $n = 25$, and $K \in \{0, 1, \ldots, 300\}$. As we have shown in Ex. 9.17, this hypergeometric density has a monotone likelihood ratio in the statistic $t(X) = X$. Note that the likelihood ratio $L(K_1; x)/L(K_2; x)$, $K_1 > K_2$, is in fact *strictly increasing* in $t(\mathbf{x}) = x$ for all values of $x$ in the support of $f(x; K_2)$, as was shown in Ex. 9.17. Theorem 9.6 and its corollaries are applicable so that the UMP critical region of size $\alpha$ for testing $H_0: K \leq 15$ versus $H_a: K > 15$

[16] This can be shown via the MGF approach, since the MGF of $\sum_{i=1}^{100} X_i = \prod_{i=1}^{100} M_{X_i}(t) = (1 - \theta t)^{-1} = (1 - \theta t)^{-100}$ for $t < \theta^{-1}$, which is of the gamma form with $\beta = \theta$, $\alpha = 100$.

[17] Again, it is possible to use a *randomized test* to achieve a size of .05 exactly, but the test can depend on the outcome of a random variable that has nothing to do with the experiment being analyzed. See Problem 9.8 for an example of this approach. Randomized tests are not often used in practice.

is represented by $C_r = \{x\colon x \geq c\}$ for $c$ chosen so that $P(x \in C_r; K = 15) = \alpha$. Computing the values of the hypergeometric CDF on the computer reveals that $\alpha = .027$, when $c = 4$. This is the choice of $c$ that generates a value of $\alpha$ closest to .05 without exceeding it. Thus, a UMP level-.05 and unbiased size .027 test of $H_0\colon K \leq 15$ versus $H_a\colon K > 15$ in a random sample, without replacement, of size 25 from the population of 300 video cassettes is defined by the critical region $C_r = [4, 25]$ for $x$. □

There does *not* generally exist a UMP level-$\alpha$ test of the hypothesis $H_0 : \Theta = \Theta_0$ versus the *two-sided* alternative $H_a\colon \Theta \neq \Theta_0$ in the case of monotone likelihood ratios. This follows from the fact that the UMP level-$\alpha$ critical region of the test of such a hypothesis is different, depending on "which side" of $\Theta_0$ the alternative value $\Theta$ is on. We demonstrate this phenomenon for the case where the monotone likelihood ratio statistic is a continuous random variable.

**Theorem 9.7** (Nonexistence of UMP Level-$\alpha$ Test of $H_0\colon \Theta = \Theta_0$ versus $H_a\colon \Theta \neq \Theta_0$ for monotone likelihood ratios) *Let $f(\mathbf{x};\Theta), \Theta \in \Omega$, be a family of PDFs having a monotone likelihood ratio in $t(\mathbf{x})$, where $t(\mathbf{X})$ is a continuous random variable. Then there does not exist a UMP critical region with level $\alpha \in (0,1)$ for testing $H_0\colon \Theta = \Theta_0$ versus $H_a\colon \Theta \neq \Theta_0$.*

**Proof** From Theorem 9.6 and Corollary 9.1, the UMP size-$\alpha$ critical region for testing $H_0\colon \Theta = \Theta_0$ versus $H_a\colon \Theta > \Theta_0$ is of the form $C_r = \{\mathbf{x}\colon t(\mathbf{x}) \geq c\}$ with $P(t(\mathbf{x}) \geq c) = \alpha$ while the UMP size $\alpha$ critical region for testing $H_0\colon \Theta = \Theta_0$ versus $H_a\colon \Theta < \Theta_0$ is of the form $C_r = \{\mathbf{x}\colon t(\mathbf{x}) \leq c\}$ with $P(t(\mathbf{x}) \leq c) = \alpha$. Appropriate choices of $c$ exist $\forall\ \alpha \in (0, 1)$ since $t(\mathbf{X})$ is a continuous random variable. It follows that there is no critical region, $C_r^*$, that is UMP level for *both* $\Theta < \Theta_0$ and $\Theta > \Theta_0$. ■

An implication of Theorem 9.7 is that we must consider alternative criteria than the UMP property when defining statistical tests of $H_0\colon \Theta = \Theta_0$ versus $H_0\colon \Theta \neq \Theta_0$ and the sampling density has a monotone likelihood ratio. A similar, albeit somewhat more complicated, argument can also be made when $t(\mathbf{X})$ is a discrete random variable. In the next subsection we will examine situations in which UMP level-$\alpha$ tests of $H_0 : \Theta = \Theta_0$ versus $H_a : \Theta \neq \Theta_0$ exist *within the unbiased class of tests.*[18] A UMP test within the class of unbiased tests will be referred to as a **uniformly most powerful unbiased (UMPU) test.**

### Exponential Class of Densities

For cases where the joint density of the random sample belongs to the exponential class of densities with a scalar parameter $\Theta$, our preceding discussion

[18] To this point, we have established UMP tests in the class of *all* tests of a certain level $\alpha$. The reader should note that in all cases examined heretofore, we have shown that UMP level-$\alpha$ tests were also unbiased. This is clearly different than examining *only unbiased* tests of a certain level $\alpha$ and within this restricted set of tests, attempting to find one that is UMP.

of monotone likelihood ratios can be used to motivate the general existence of UMP level-$\alpha$ tests of

$$H_0\colon \Theta = \Theta_0 \text{ versus } H_a\colon \Theta > \Theta_0,$$

$$H_0\colon \Theta = \Theta_0 \text{ versus } H_a\colon \Theta < \Theta_0,$$

$$H_0\colon \Theta \leq \Theta_0 \text{ versus } H_a\colon \Theta > \Theta_0,$$

$$H_0\colon \Theta \geq \Theta_0 \text{ versus } H_a\colon \Theta < \Theta_0,$$

and the general *nonexistence* of a UMP level-$\alpha$ test of

$$H_0\colon \Theta = \Theta_0 \text{ versus } H_a\colon \Theta \neq \Theta_0.$$

We now reconsider the problem of defining a test of the latter hypothesis. We will find that if the joint density of the random sample belongs to the exponential class with a scalar parameter $\Theta$, then it will generally be possible to find a critical region that is UMP level $\alpha$ within the class of *unbiased* critical regions for testing $H_0\colon \Theta = \Theta_0$ versus $H_a\colon \Theta \neq \Theta_0$. We will also establish UMPU critical regions for testing $H_0\colon \Theta_1 \leq \Theta \leq \Theta_2$ versus $H_a\colon \Theta \notin [\Theta_1, \Theta_2]$.[19] Our discussion will be facilitated by the following result concerning the differentiability of the expectation of a function taken with respect to an exponential class density.

**Lemma 9.1**
***Differentiability of $\gamma(\Theta) = \mathrm{E}_\Theta\phi(X)$ in the Case of an Exponential Class Density***

Let $f(\mathbf{x};\Theta) = \exp(c(\Theta)g(\mathbf{x}) + d(\Theta) + z(\mathbf{x}))I_A(\mathbf{x}),\ \Theta \in \Omega$, be an exponential class density with $\Omega$ defined as an open interval contained in $R$, and with $c(\Theta)$ and $d(\Theta)$ being differentiable functions of $\Theta \in \Omega$. Let $\mathrm{E}_\Theta$ denote an expectation taken with respect to $f(\mathbf{x};\Theta)$.

If the function $\phi(\mathbf{x})$ is such that $\gamma(\Theta) = E_\Theta\phi(\mathbf{X})$ exists $\forall\ \Theta \in \Omega$, then $\gamma(\Theta) = \mathrm{E}_\Theta\phi(\mathbf{X})$ is a differentiable function of $\Theta \in \Omega$. Furthermore,

$$(\textit{continuous}) \qquad \frac{\partial^r}{\partial\Theta^r}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\phi(\mathbf{x})f(\mathbf{x};\Theta)d\mathbf{x} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\phi(\mathbf{x})\frac{\partial^r f(\mathbf{x};\Theta)}{\partial\Theta^r}d\mathbf{x},$$

$$(\textit{discrete}) \qquad \frac{\partial^r}{\partial\Theta^r}\sum_{\mathbf{x}\in R(\mathbf{X})}\cdots\sum\phi(\mathbf{x})f(\mathbf{x};\Theta) = \sum_{\mathbf{x}\in R(\mathbf{X})}\cdots\sum\phi(\mathbf{x})\frac{\partial^r f(\mathbf{x};\Theta)}{\partial\Theta^r},$$

i.e., differentiation can occur under the integral or summation sign.

**Proof**: See E. L. Lehmann (1986), *Testing Statistical Hypotheses*, 2nd ed., New York: John Wiley, pp. 59–60, or D. V. Widder (1946), *The Laplace Transform*, Princeton. pp. 240–241.

[19] Results are available for a more general class of densities referred to as Polya distributions, which subsumes the exponential class densities as a special case. However, the mathematics involved in analyzing the more general distributions is beyond the scope of our study. Interested readers can consult the work of S. Karlin (1957), "Polya type distributions II." *Ann. Math. Stat.*, 28, pp. 281–308.

The lemma allows one to establish the differentiability (and thus also continuity) of the power function associated with the critical region of a hypothesis test when the joint density of the random sample is an exponential class distribution. In particular, by defining $\phi(\mathbf{x}) = I_{C_r}(\mathbf{x})$, Lemma 9.1 becomes a result concerning the differentiability of power functions.

**Theorem 9.8** (Differentiability (and Continuity) of Power Function in the Case of an Exponential Class Density) *Let the joint density function of a random sample be a member of the exponential class of densities*

$$f(\mathbf{x};\Theta) = \exp(c(\Theta)g(\mathbf{x}) + d(\Theta) + z(\mathbf{x}))I_A(\mathbf{x}), \quad \Theta \in \Omega,$$

*with $\Omega$ being an open interval contained in $R$, and with $c(\Theta)$ and $d(\Theta)$ being differentiable functions of $\Theta \in \Omega$. Let $C_r$ be any critical region for testing some hypothesis $H_0: \Theta \in \Omega_0$ versus $H_a: \Theta \in \Omega_a$.*[20] *Then the power function $\pi_{C_r}(\Theta) = P(\mathbf{x} \in C_r;\Theta)$ is differentiable (and hence continuous) with respect to $\Theta \in \Omega$.*

**Proof** Let $\phi(\mathbf{x}) \equiv I_{C_r}(\mathbf{x})$ in Lemma 9.1. Then since $\mathrm{E}_\Theta\phi(\mathbf{X}) = \mathrm{E}_\Theta I_{C_r}(\mathbf{X}) = \pi_{C_r}(\Theta) = P(\mathbf{x} \in C_r;\Theta)\ \forall\,\Theta \in \Omega$, the power function $\pi_{C_r}(\Theta)$ is differentiable $\forall\,\Theta \in \Omega$. Continuity of $\pi_{C_r}(\Theta)$ follows from differentiability of $\pi_{C_r}(\Theta)$. ■

We now turn to the main result concerning the definition of UMPU level-$\alpha$ two-sided tests when the random sample has an exponential class density.

**Theorem 9.9** (UMPU Level-$\alpha$ Two-Sided Tests for Exponential Class Densities) *Let the joint density of the random sample be given by $f(\mathbf{x};\Theta) = \exp(c(\Theta)g(\mathbf{x}) + d(\Theta) + z(\mathbf{x}))I_A(\mathbf{x}), \Theta \in \Omega$, with $\Omega$ being an open interval contained in $R$, $c(\Theta)$ and $d(\Theta)$ being differentiable functions of $\Theta \in \Omega$, and $c(\Theta)$ being strictly monotonic (either increasing or decreasing) in $\Theta$. Define a critical region as*

$$C_r = \{\mathbf{x}: g(\mathbf{x}) \le c_1 \text{ or } g(\mathbf{x}) \ge c_2\},$$

*where $c_1 < c_2$.*

1. *A size $\alpha$ $C_r$ defines a UMPU level-$\alpha$ test of $H_0: \Theta = \Theta_0$ versus $H_a: \Theta \neq \Theta_0$ iff $\Theta_0 = \arg\min_{\Theta \in H_0 \cup H_a}(\pi_{C_r}(\Theta)) = \arg[d\pi_{C_r}(\Theta)/d\Theta = 0]$, i.e., the power of the test is minimized at $\Theta_0$.*
2. *A size $\alpha$ $C_r$ defines a UMPU level-$\alpha$ test of $H_0: \Theta_1 \le \Theta \le \Theta_2$ versus $H_a: \Theta \notin [\Theta_1,\Theta_2]$, $\Theta_1 < \Theta_2$, iff $\pi_{C_r}(\Theta_1) = \pi_{C_r}(\Theta_2) = \alpha$.*

**Proof** Necessity: For result (1), if the power function is not minimized at $\Theta_0$, then $C_r$ is not an $\alpha$-size unbiased test since then $\exists\Theta \in H_a$ such that $\pi_{C_r}(\Theta) < \pi_{C_r}(\Theta_0) = \alpha$. The power function is differentiable by Theorem 9.8, and can be shown (see reference below) to be strictly convex, so that the minimum occurs at $\Theta_0$ iff $d\pi_{C_r}(\Theta_0)/d\Theta = 0$.

[20] We are assuming that $C_r$ is such that a power function is defined, i.e., $C_r$ can be assigned probability by $f(\mathbf{x};\Theta), \Theta \in \Omega$.

For result (2), since the power function is continuous by Theorem 9.8, it is necessary that $\pi_{C_r}(\Theta_1) = \pi_{C_r}(\Theta_2) = \alpha$, for consider the contrary, where $\pi_{C_r}(\Theta_1) < \alpha$ and/or $\pi_{C_r}(\Theta_2) < \alpha$ (note that $\pi_{C_r}(\Theta_i) > \alpha$ for $i = 1, 2$ is ruled out since the test has size $\alpha$). Then there exists a value of $\Theta \in H_a$ close to $\Theta_1$ or $\Theta_2$ such that $\pi_{C_r}(\Theta) < \alpha$ by the continuity of $\pi_{C_r}(\Theta)$, contradicting unbiasedness.

Sufficiency: E. Lehmann, *Testing Statistical Hypotheses, 2nd ed.*, pp. 135–137. ■

The application of Theorem 9.9 to find UMPU level-$\alpha$ tests of $H_0\colon \Theta = \Theta_0$ versus $H_a\colon \Theta \neq \Theta_a$ or of $H_0\colon \Theta_1 \leq \Theta \leq \Theta_2$ versus $H_a\colon \Theta \notin [\Theta_1, \Theta_2]$ is conceptually straightforward although a computer will often be needed as an aid in making the necessary calculations. Essentially, one searches for the appropriate $c_1$ and $c_2$ values in $C_r = \{\mathbf{x}\colon g(\mathbf{x}) \leq c_1 \text{ or } g(\mathbf{x}) \geq c_2\}$ that produce an $\alpha$-size critical region and that also satisfy either $d\pi_{C_r}(\Theta_0)/d\Theta = 0$ or $\pi_{C_r}(\Theta_1) = \pi_{C_r}(\Theta_2)$, respectively. The following examples illustrate the process.

**Example 9.20** Bernoulli Population Distribution—$H_0: p = p_0$ versus $H_a: p \neq p_0$

Recall Ex. 9.14 in which an untrusting gambler desired to test whether a roulette wheel used by an Atlantic City casino was fair. A random sample of 100 spins of the wheel was to be used for the test. The joint density of the random sample was in the form of a product of Bernoulli densities, which can be represented as a member of the exponential class

$$f(\mathbf{x}; p) = \exp[c(p)g(\mathbf{x}) + d(p) + z(\mathbf{x})]I_A(\mathbf{x}),$$

where $c(p) = \ln(p/(1-p))$, $g(\mathbf{x}) = \sum_{i=1}^{100} x_i$, $d(p) = 100\ln(1-p)$, $z(\mathbf{x}) = 0$, and $A = \times_{i=1}^{100}\{0, 1\}$, with $p \in \Omega = (0, 1)$. We demonstrated in Ex. 9.14 that there does *not* exist a UMP level-.10 test of the "fairness" hypothesis $H_0\colon p = .5$ versus $H_a\colon p \neq .5$ having size .10. We now use Theorem 9.9 to show that a UMPU level and size $\alpha$ test of the fairness hypothesis does exist. The critical region we seek has the form

$$C_r = \left\{\mathbf{x}\colon \sum_{i=1}^{100} x_i \leq c_1 \text{ or } \sum_{i=1}^{100} x_i \geq c_2\right\}$$

for choices of $c_1$ and $c_2$ that produce an $\alpha$-size critical region satisfying $d\pi_{C_r}(.5)/dp = 0$. Since $Z = \sum_{i=1}^{100} X_i$ has a binomial distribution, $C_r$ will have size $\alpha$ when

$$1 - \pi_{C_r}(.5) = P\left(\mathbf{x} \in \overline{C_r}; p = .5\right) = \sum_{z=c_1+1}^{c_2-1} \binom{100}{z}(.5)^z(.5)^{100-z} I_{\{0,1,\ldots,100\}}(z) = 1 - \alpha. \quad (1)$$

The derivative of the power function can be represented as

$$\frac{d\pi_{C_r}(p)}{dp} = -\sum_{z=c_1+1}^{c_2-1} \binom{100}{z} p^{z-1}(1-p)^{99-z}[z - 100p]I_{\{0,1,\ldots,100\}}(z),$$

which when evaluated at $p = .5$ and then set equal to 0 results in the appropriate condition on the choices of $c_1$ and $c_2$, as

$$\frac{d\pi_{C_r}(.5)}{dp} = -\sum_{z=c_1+1}^{c_2-1} \binom{100}{z} (.5)^{98}[z - 50] = 0. \tag{2}$$

Equation (2) implies that $c_1$ and $c_2$ must be chosen so that $c_1 < z < c_2$ is a symmetric interval around the value 50, which is a direct consequence of the binomial density being a symmetric density around the point $z = 50$ when $p = .5$. Thus, the possible choices of $c_1$ and $c_2$ are given by the finite set of two-tuples $A = \{(c_1, c_2): c_1 = 50 - i, c_2 = 50 + i, \ i = 1, \ldots, 50\}$, which restricts the admissible choices of test sizes to be

$$B = \{\alpha : \alpha = P(z \leq c_1 \text{ or } z \geq c_2; p = .5), \ (c_1, c_2) \in A\}.$$

Some specific possibilities for the size of the test are given as follows (calculated using a computer):

| $c_1, c_2$ | $\alpha = P(x \in C_r; p = .5)$ |
|---|---|
| 37, 63 | .012 |
| 38, 62 | .021 |
| 39, 61 | .035 |
| 40, 60 | .057 |
| 41, 59 | .089 |
| 42, 58 | .133 |

Suppose a test size of $\alpha = .057$ is acceptable to the concerned gambler. Then the critical region of the UMPU level-.057 test of the fairness hypothesis is given by

$$C_r = \left\{ \mathbf{x} : \sum_{i=1}^{100} x_i \leq 40 \quad \text{or} \ \sum_{i=1}^{100} x_i \geq 60 \right\}$$

and the hypothesis that the roulette wheel was fair would be rejected if red $(x_i = 1)$ occurred 40 or fewer times, or 60 or more times. Note that the values of the power function (see Figure 9.9) for $p \in (0, 1)$ are given by

$$\pi_{C_r}(p) = 1 - \sum_{z=41}^{59} \binom{100}{z} p^z (1 - p)^{100-z} I_{\{0,1,\ldots,100\}}(z).$$

Some selected values of the power function are given below.

| $p$ | $\pi_{C_r}(p)$ |
|---|---|
| .5 | .057 |
| .45,.55 | .185 |
| .4, .6 | .543 |
| .3, .7 | .988 |
| .2, .8 | $\cong$ 1.000 |

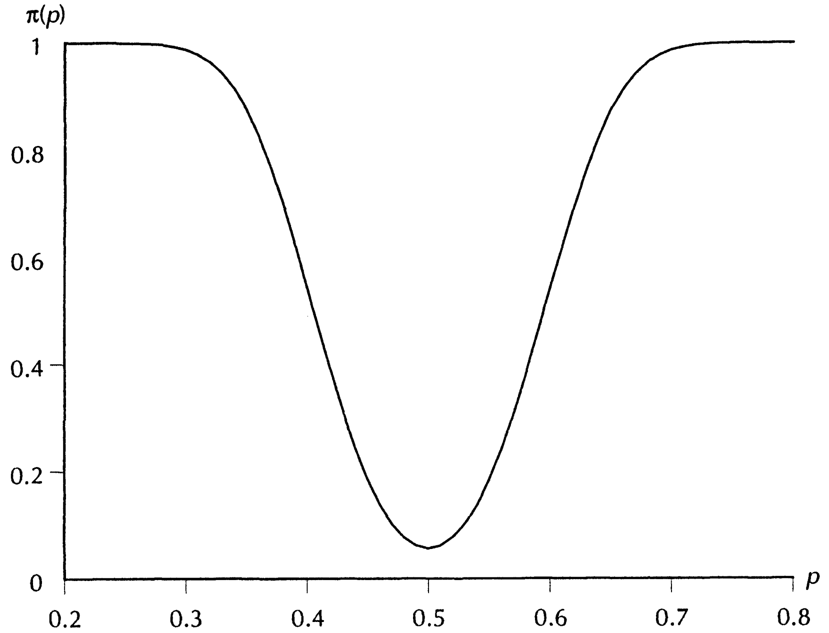


**Figure 9-9** Power function for the UMPU level-.057 test of $H_0: p = .5$ versus $H_a: p \neq .5$.

(Point to ponder: Do you think the power function given in Figure 9.9 should be acceptable to the gambler? Why or why not? If not, what could you do to alter the power function to make it more acceptable to the gambler?) □

**Example 9.21** Normal Population Distribution—$H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$

The manufacturer of a certain battery-powered notebook computer claims that the average operating time of the computer between full battery charges is 2 hours. It can be assumed that the operating time, in hours, obtained from a full battery charge is a random variable having a normal distribution with standard deviation equal to .2, i.e., $X_i \sim N(\mu, .04)$. A random sample of the operating times of 200 of the notebook computers is to be used to test the hypothesis that the average operating time is indeed 2 hours, i.e., a test of $H_0: \mu = 2$ versus $H_a: \mu \neq 2$ is to be conducted. Define a test that is UMPU level $\alpha$ for testing $H_0$ versus $H_a$.[21]

**Answer**: The joint density of the random sample can be represented in the form

$$f(\mathbf{x}; \mu) = \frac{1}{(2\pi)^{100}(.2)^{200}} \exp\left[-12.5 \sum_{i=1}^{200}(x_i - \mu)^2\right].$$

The density can be written in the exponential class form

$$f(\mathbf{x}; \mu) = \exp[c(\mu)g(\mathbf{x}) + d(\mu) + z(\mathbf{x})]I_A(\mathbf{x}),$$

[21] One can show using the monotone likelihood ratio approach that a UMP level-$\alpha$ test of $H_0$ versus $H_a$ does *not* exist. Recall Theorem 9.7.

with

$$c(\mu) = 25\mu, \qquad g(\mathbf{x}) = \sum_{i=1}^{200} x_i, \qquad d(\mu) = -2500\mu^2,$$

$$z(\mathbf{x}) = -\ln\left((2\pi)^{100}(.2)^{200}\right) - 12.5\sum_{i=1}^{200} x_i^2, \quad \text{and } A = R^n.$$

From Theorem 9.9, we seek a critical region of the form

$$C_r = \left\{\mathbf{x} : \sum_{i=1}^{200} x_i \leq c_1 \text{ or } \sum_{i=1}^{200} x_i \geq c_2\right\}$$

for choices of $c_1$ and $c_2$ that produce an $\alpha$-size critical region satisfying $d\pi_{C_r}(2)/d\mu = 0$. Using results for linear combinations of normally distributed random variables we know that $Z = \sum_{i=1}^{200} X_i \sim N(200\mu, 8)$, and thus $C_r$ will have size $\alpha$ when

$$1 - \pi_{C_r}(2) = P(\mathbf{x} \in \bar{C}_r; \mu = 2) = \int_{c_1}^{c_2} \frac{1}{\sqrt{2\pi}\sqrt{8}} \exp\left[-\frac{1}{16}(z - 400)^2\right] dz = 1 - \alpha. \quad (1)$$

The derivative of the power function can be represented as

$$\frac{d\pi_{C_r}(\mu)}{d\mu} = -\int_{c_1}^{c_2} \frac{25[z - 200\mu]}{\sqrt{2\pi}\sqrt{8}} \exp\left[-\frac{1}{16}(z - 200\mu)^2\right] dz,$$

which when evaluated at $\mu = 2$ and then set equal to 0 results in the condition

$$\frac{d\pi_{C_r}(2)}{d\mu} = -25\int_{c_1}^{c_2} (z - 400)N(z; 400, 8)dz = 0. \quad (2)$$

Equation (2) implies that $c_1$ and $c_2$ must be chosen so that $c_1 < z < c_2$ is a symmetric interval around the value 400, which is a direct consequence of the symmetry of the normal distribution around the point $z = 400$. Thus, the possible choices of $c_1$ and $c_2$ are given by the uncountably infinite set of two-tuples $A = \{(c_1, c_2) : c_1 = 400 - c, c_2 = 400 + c, c \in (0, \infty)\}$, which then implies that the choice of test size can be any value $\in (0, 1)$, since

$$B = \{\alpha : \alpha = P(z \leq c_1 \text{ or } z \geq c_2; \mu = 2), (c_1, c_2) \in A\} = (0, 1).$$

Suppose $\alpha = .10$ is chosen as the size of the test. The appropriate values of $c_1$ and $c_2$ can be found by solving for $c$ in the integral equation

$$\int_{400-c}^{400+c} N(z; 400, 8)dz = \int_{-c/\sqrt{8}}^{c/\sqrt{8}} N(z; 0, 1)dz = .90$$

and then setting $c_1 = 400 - c$ and $c_2 = 400 + c$. Using the computer, or the table of the standard normal distribution, we find that $c/\sqrt{8} = 1.645$, or $c = 4.653$. Then $c_1 = 395.347$ and $c_2 = 404.653$, and the critical region of the UMPU size

.10 test is given by

$$C_r = \left\{ \mathbf{x} : \sum_{i=1}^{200} x_i \leq 395.347 \text{ or } \sum_{i=1}^{200} x_i \geq 404.653 \right\}.$$

The critical region can be equivalently expressed using the sample mean as a test statistic, in which case

$$C_r = \{\mathbf{x} : \bar{x} \leq 1.977 \text{ or } \bar{x} \geq 2.023\}.$$

Thus, the hypothesis is rejected if the average operating time of the 200 sampled notebook computers $\leq 1.977$ hours or $\geq 2.023$ hours.

Note that the values of the power function (see Figure 9.10) for $\mu \in (-\infty, \infty)$ are given by

$$\pi_{C_r}(\mu) = 1 - \int_{395.347}^{404.653} N(z; 200\mu, 8)dz = 1 - \int_{(395.347-200\mu)/\sqrt{8}}^{(404.653-200\mu)/\sqrt{8}} N(z; 0, 1)dz.$$

Some selected values of the power function are given below.

| $\mu$ | $\pi_{C_r}(\mu)$ |
|---|---|
| 2 | .100 |
| 1.99,2.01 | .183 |
| 1.98,2.02 | .410 |
| 1.97,2.03 | .683 |
| 1.96,2.04 | .882 |
| 1.95,2.05 | .971 |
| 1.94,2.06 | .995 |

□

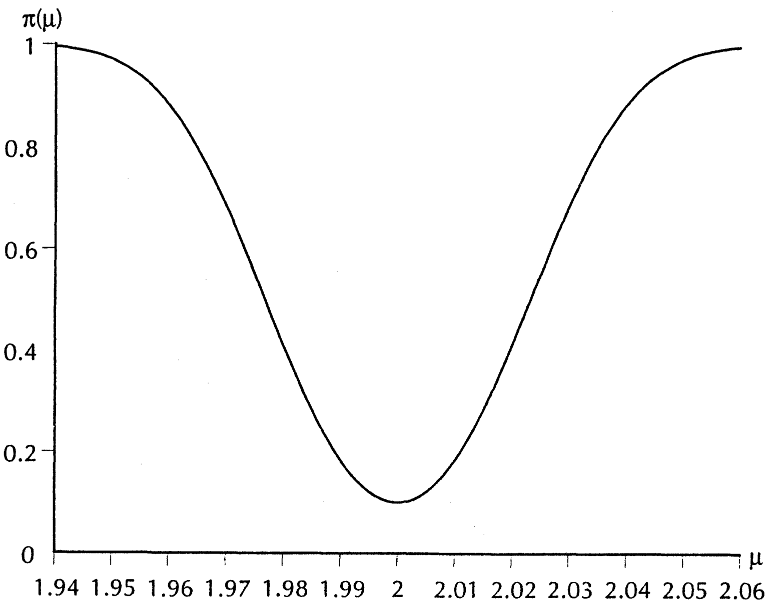


**Figure 9-10** Power function of the UMPU level-.10 test of $H_0: \mu = 2$ versus $H_a: \mu \neq 2$.

The reader may have noticed that in the preceding two examples, the characterization of admissible values for $c_1$ and $c_2$ was simplified by the fact that $g(\mathbf{X})$, which was a test statistic in both examples, had a distribution that was symmetric about its mean under the assumption $H_0$ is true. In cases where this symmetry in the distribution of $g(\mathbf{X})$ does not occur, the characterization of admissible $(c_1, c_2)$ values is somewhat more involved.

**Example 9.22** Exponential Population Distribution—$H_0: \theta = \theta_0$ versus $H_a: \theta \neq \theta_0$

The operating life of a memory chip manufactured by the Elephant Computer Chip Co. can be viewed as a random variable with a density function belonging to the exponential family of densities. A random sample of the lifetimes of 500 memory chips is to be used to test the hypothesis that the mean life of the chip is 5,000 hours. With the $X_i$'s measured in thousands of hours, the joint density of the random sample is given by

$$f(\mathbf{x};\theta) = \theta^{-500} \exp\left(-\sum_{i=1}^{500} x_i/\theta\right) \prod_{i=1}^{500} I_{(0,\infty)}(x_i)$$

for $\theta \in \Omega = (0, \infty)$. Define a UMPU level-.05 test of $H_0: \theta = 5$ versus $H_a: \theta \neq 5$.
**Answer**: The joint density can be written in the exponential class form

$$f(\mathbf{x};\theta) = \exp[c(\theta)g(\mathbf{x}) + d(\theta) + z(\mathbf{x})]I_A(\mathbf{x}),$$

with

$$c(\theta) = -\theta^{-1}, \qquad g(\mathbf{x}) = \sum_{i=1}^{500} x_i, \qquad d(\theta) = \ln(\theta^{-500}), \qquad z(\mathbf{x}) = 0, \qquad \text{and } A = \times_{i=1}^{500}(0,\infty).$$

From Theorem 9.9, we seek a critical region of the form

$$C_r = \left\{\mathbf{x}: \sum_{i=1}^{500} x_i \leq c_1 \text{ or } \sum_{i=1}^{500} x_i \geq c_2\right\}$$

for choices of $c_1$ and $c_2$ that produce an $\alpha$-size critical region satisfying $d\pi_{C_r}(5)/d\theta = 0$. Using the MGF approach, it can be established that $Z = \sum_{i=1}^{500} X_i$ has a Gamma$(z; 500, \theta)$ distribution, i.e.,

$$Z \sim (1/\theta^{500}\Gamma(500))z^{499} \exp(-z/\theta)I_{(0,\infty)}(z)$$

(recall Ex. 9.18).

Then $C_r$ will have size $\alpha$ when

$$1 - \pi_{C_r}(5) = \int_{c_1}^{c_2} \frac{1}{5^{500}\Gamma(500)} z^{499} \exp(-z/5)I_{(0,\infty)}(z) = 1 - \alpha. \tag{1}$$

The derivative of the power function can be represented as

$$d\pi_{C_r}(\theta)/d\theta = -\int_{c_1}^{c_2} [(z/\theta^2) - 500\theta^{-1}] \quad \text{Gamma}\,(z; 500, \theta)dz,$$

which when evaluated as $\theta = 5$, then set equal to 0, results in the condition

$$d\pi_{C_r}(5)/d\theta = -\int_{c_1}^{c_2} [(z/25) - 100] \quad \text{Gamma}\,(z; 500, 5)dz = 0. \tag{2}$$

Unlike the previous two examples, the distribution of $Z$ is not symmetric about its mean, and (2) does *not* imply that $(c_1, c_2)$ is a symmetric interval around $\mathrm{E}Z = 2,500$. Nonetheless, (2) defines an implicit functional relationship between $c_2$ and $c_1$, say $c_2 = \gamma(c_1)$, that determines admissible values of $c_1$ and $c_2$. Using the computer, the simultaneous equations (1) and (2) can be numerically solved for $c_1$ and $c_2$, given a choice of the test size $\alpha \in (0, 1)$. This procedure was followed, yielding the following $c_1$ and $c_2$ values for selected test sizes:

| $\alpha$ | $c_1$ | $c_2$ |
|---|---|---|
| .01 | 2222.920 | 2799.201 |
| .05 | 2287.190 | 2725.618 |
| .10 | 2320.552 | 2688.469 |
| .20 | 2359.419 | 2646.057 |

To define a UMPU level-.05 test of $H_0: \theta = 5$ versus $H_a: \theta \neq 5$, the critical region would be defined as

$$C_r = \left\{\mathbf{x}: \sum_{i=1}^{500} x_i \leq 2287.190 \text{ or } \sum_{i=1}^{500} x_i \geq 2725.618\right\}.$$

One could also use the sample mean as a test statistic, in which case the critical region could be defined alternatively as

$$C_r = \{\mathbf{x}: \bar{x} \leq 4.574 \text{ or } \bar{x} \geq 5.451\}.$$

Thus, the hypothesis $H_0: \theta = 5$ will be rejected if the average life of the 500 sampled chips $\leq 4,574$ hours or $\geq 5,451$ hours.

Power function values for $\theta > 0$ are given by (see Figure 9.11)

$$\pi_{C_r}(\theta) = 1 - \int_{2287.190}^{2725.618} \frac{1}{\theta^{500}\Gamma(500)} z^{499} \exp(-z/\theta) dz.$$

Some selected values of the power function are given below.

| $\theta$ | $\pi_{C_r}(\theta)$ | $\theta$ | $\pi_{C_r}(\theta)$ |
|---|---|---|---|
| 4 | .999 | 5.10 | .073 |
| 4.25 | .954 | 5.25 | .196 |
| 4.50 | .649 | 5.50 | .573 |
| 4.75 | .206 | 5.75 | .878 |
| 4.90 | .073 | 6.00 | .981 |
| 5.00 | .050 | 6.25 | .999 |

Unlike the previous two examples, the power function is *not* symmetric about the point $\theta_0 = 5$. However, it is "nearly" symmetric because of the fact that

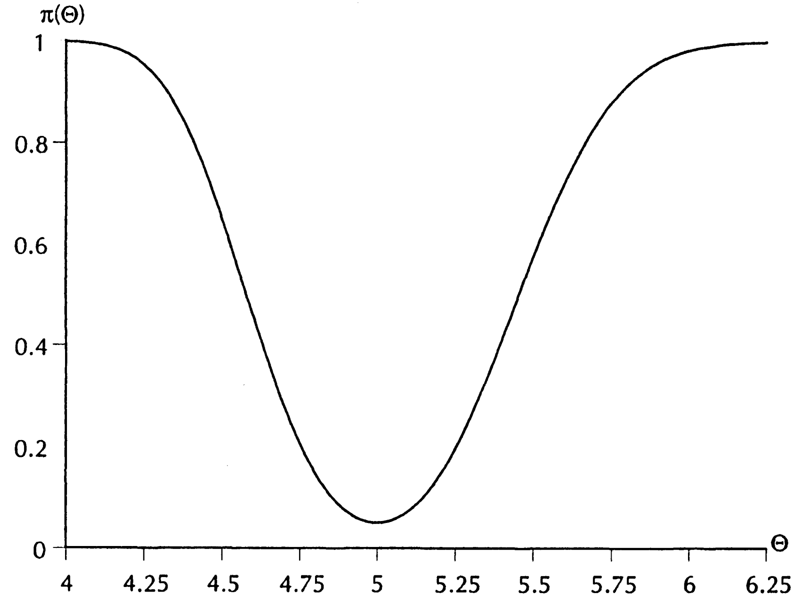


**Figure 9-11** Power function of UMPU level-.05 test of $H_0 : \theta = 5$ versus $H_a : \theta \neq 5$.

$Z_n = \sum_{i=1}^{n} X_i$ has an asymptotic normal distribution, and a sample size of $n = 500$ is sufficient for the asymptotic distribution to provide a good approximation to $Z_n$'s Gamma distribution. □

**Example 9.23** Normal Population Distribution—$H_0 : \mu \in [\mu_1, \mu_2]$ versus $H_a : \mu \notin [\mu_1, \mu_2]$

Recall Ex. 9.21 regarding the operating times of notebook computers between full battery charges. Suppose instead of testing the hypothesis $H_0 : \mu = 2$, a *range* of values for the mean operating time of the computer is to be tested. In particular, suppose we wish to test the hypothesis $H_0 : 1.75 \leq \mu \leq 2.25$ versus $H_a : \mu < 1.75$ or $\mu > 2.25$. We know from Ex. 9.21 that $f(\mathbf{x}; \mu)$ belongs to the exponential class of densities and satisfies the other conditions of Theorem 9.9. Using Theorem 9.9, we seek an $\alpha$-size critical region defined by

$$C_r = \{\mathbf{x} : \sum_{i=1}^{200} x_i \leq c_1 \text{ or } \sum_{i=1}^{200} x_i \geq c_2\}$$

for choices of $c_1$ and $c_2$ that satisfy

$$1 - \pi_{C_r}(350) = \int_{c_1}^{c_2} N(z; 350, 8)dz = \int_{c_1}^{c_2} N(z; 450, 8)dz = 1 - \pi_{C_r}(450) = 1 - \alpha.$$

Choosing a specific value of $\alpha$, the preceding condition becomes a system of *two* integral equations which can be solved simultaneously (using the computer) for the two unknowns $c_1$ and $c_2$.

Suppose the size of the test is chosen to be $\alpha = .05$. Using a nonlinear, simultaneous equations solver,[22] the solution to the preceding two-equation system was found to be $c_1 = 345.348$ and $c_2 = 454.652$. Thus, the critical region of the UMPU level-.05 test defined by Theorem 9.9 is given by

$$C_r = \left\{\mathbf{x}\colon \sum_{i=1}^{200} x_i \leq 345.348 \text{ or } \sum_{i=1}^{200} x_i \geq 454.652\right\}.$$

One could also use the sample mean for the test statistic, in which case $C_r$ could be defined alternatively as

$$C_r = \{\mathbf{x}\colon \bar{x} \leq 1.727 \text{ or } \bar{x} \geq 2.273\}.$$

Power function values are given by (see Figure 9.12)

$$\pi_{C_r}(\mu) = 1 - \int_{345.348}^{454.652} N(z; 200\mu, 8)dz$$

Some selected values of the power function are given below.

| $\mu$ | $\pi_{C_r}(\mu)$ | $\mu$ | $\pi_{C_r}(\mu)$ |
|---|---|---|---|
| 2 | $\approx 0$ | 1.72, 2.28 | .68 |
| 1.76,2.24 | .01 | 1.71,2.29 | .88 |
| 1.75,2.25 | .05 | 1.70,2.30 | .97 |
| 1.74,2.26 | .17 | 1.69,2.31 | $\approx 1.00$ |
| 1.73,2.27 | .41 | | |

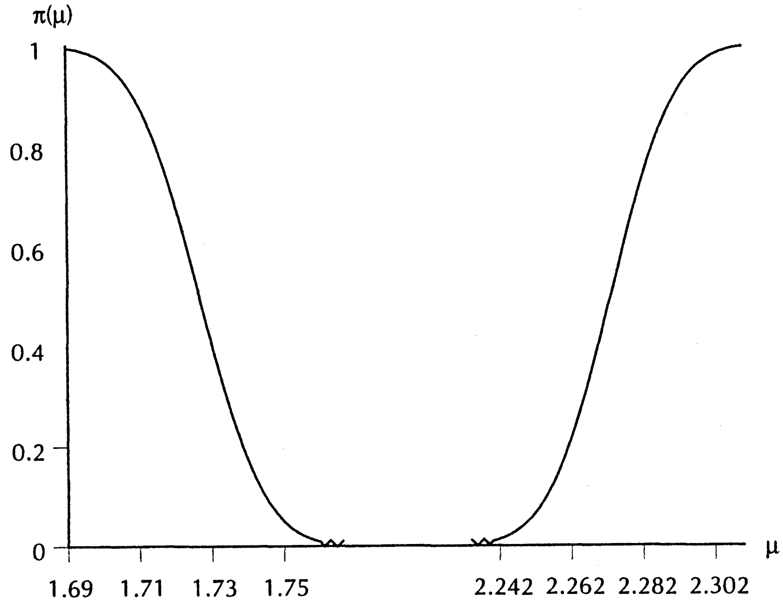


**Figure 9-12**
Power function of UMP unbiased level .05 test of $H_0: 1.75 \leq \theta \leq 2.25$ versus $H_a: \theta < 1.75$ or $\theta > 2.25$

[22] The algorithm actually used was the NLSYS procedure in the GAUSS Matrix language.

The power function verifies that the test defined by $C_r$ is unbiased and has size .05. The hypothesis $H_0$: $1.75 \leq \mu \leq 2.25$ is then rejected by the UMPU level-.05 test if the average operating time of the 200 sampled notebook computers $\leq 1.727$ hours or $\geq 2.273$ hours. □

**Example 9.24** Exponential Population Distribution—$H_0: \theta \in [\theta_1, \theta_2]$ versus $H_a: \theta \notin [\theta_1, \theta_2]$

Recall Ex. 9.22 regarding the operating life of memory chips. Suppose instead of testing the hypothesis $H_0$: $\theta = 5$, a range of values for the mean life of the memory chips is to be tested. In particular, suppose we wish to test the hypothesis $H_0$: $4.9 \leq \theta \leq 5.1$ versus $H_a$: $\theta < 4.9$ or $\theta > 5.1$. We know from Ex. 9.19 that $f(\mathbf{x}; \theta)$ belongs to the exponential class of densities and adheres to the other conditions of Theorem 9.9. Using Theorem 9.9, we seek a size $\alpha$ critical region of the form

$$C_r = \left\{\mathbf{x}: \sum_{i=1}^{500} x_i \leq c_1 \text{ or } \sum_{i=1}^{500} x_i \geq c_2\right\}$$

for choices of $c_1$ and $c_2$ that satisfy

$$1 - \pi_{C_r}(4.9) = \int_{c_1}^{c_2} \text{Gamma}(z; 500, 4.9)dz = \int_{c_1}^{c_2} \text{Gamma}(z; 500, 5.1)dz1 - \pi_{C_r}(5.1) = 1 - \alpha.$$

Choosing a specific value of test size $\alpha$, the preceding condition becomes a system of *two* integral equations, which can be solved simultaneously (using the computer) for the two unknowns $c_1$ and $c_2$.

Suppose the size of the test is chosen to be $\alpha = .05$. Using a nonlinear equation solver,[23] the solution to the two-equation system was found to be $c_1 = 2268.031$ and $c_2 = 2746.875$. Thus, the critical region of the UMPU level-.05 test defined by Theorem 9.9 is given by

$$C_r = \left\{\mathbf{x}: \sum_{i=1}^{500} x_i \leq 2268.031 \text{ or } \sum_{i=1}^{500} x_i \geq 2746.875\right\}.$$

One could also use the sample mean for the test statistic, in which case $C_r$ could be defined alternatively as

$$C_r = \{\mathbf{x}: \bar{x} \leq 4.536 \text{ or } \bar{x} \geq 5.494\}.$$

Power function values are given by (see Figure 9.13)

$$\pi_{C_r}(\Theta) = 1 - \int_{2268.031}^{2746.875} \text{Gamma}(z; 500, \theta)dz$$

Some selected values of the power function are given below.

[23] The algorithm actually used was the NLSYS procedure in the GAUSS matrix language.

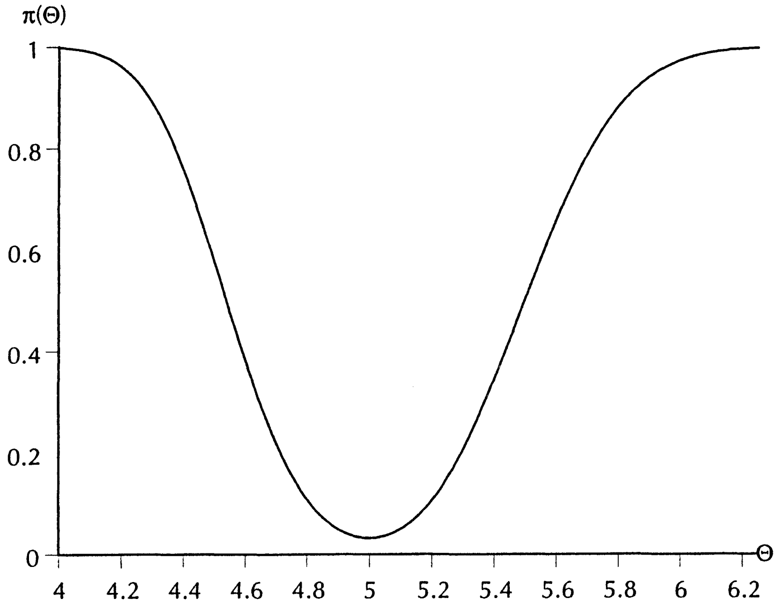


**Figure 9-13** Power function of UMPU level-.05 test of $H_0: 4.9 \leq \theta \leq 5.1$ versus $H_a: \theta < 4.9$ or $\theta > 5.1$.

| $\theta$ | $\pi_{C_r}(\theta)$ | $\theta$ | $\pi_{C_r}(\theta)$ |
|---|---|---|---|
| 4.00 | .998 | 6.00 | .973 |
| 4.25 | .932 | 5.75 | .841 |
| 4.50 | .576 | 5.50 | .504 |
| 4.75 | .157 | 5.25 | .151 |
| 4.90 | .050 | 5.10 | .050 |
| 5.00 | .032 | | |

The power function verifies that the test defined by $C_r$ is a size .05 unbiased test. The hypothesis $H_0$: $4.9 \leq \theta \leq 5.1$ is then rejected by the UMPU level .05 test if the average life of the 500 sampled memory chips $\leq 4,536$ hours or $\geq 5,494$ hours. □

We observed in previous examples that the definition of the UMPU level-$\alpha$ critical region for testing $H_0$: $\Theta = \Theta_0$ versus $H_a$: $\Theta \neq \Theta_0$ was substantially simplified if the test statistic had a density that was symmetric about its mean. This is a general property of the construction of UMPU tests, based on Theorem 9.9, that we now formalize.

**Theorem 9.10** (Two-Sided UMPU Tests of $H_0$: $\Theta = \Theta_0$ Versus $H_a$: $\Theta \neq \Theta_0$ When Test Statistics Have Symmetric PDFs) *Assume the conditions of Theorem 9.9, and let the density of the test statistic* $g(\mathbf{X})$ *be symmetric about its mean* $\gamma = \mathrm{E}g(\mathbf{X})$ *when* $\Theta = \Theta_0$. *If* $c_1$ *and* $c_2$ *are such that* $P(g(\mathbf{x}) \geq c_2; \Theta_0) = \alpha/2$ *and* $c_1 = 2\gamma - c_2$, *then* $C_r = \{\mathbf{x}: g(\mathbf{x}) \leq c_1 \text{ or } g(\mathbf{x}) \geq c_2\}$ *defines a UMPU level-*$\alpha$ *test of* $H_0$: $\Theta = \Theta_0$ *versus* $H_a$: $\Theta \neq \Theta_0$.

**Proof** The random variable $Z = g(\mathbf{X}) - \gamma$ is symmetric about zero, and thus

$$\begin{aligned} P(g(\mathbf{x}) \geq c_2; \Theta_0) = P(z \geq c_2 - \gamma; \Theta_0) &= P(z \leq \gamma - c_2; \Theta_0) \\ &= P(g(\mathbf{x}) \leq 2\gamma - c_2; \Theta_0) \\ &= P(g(\mathbf{x}) \leq c_1; \Theta_0) = \alpha/2. \end{aligned}$$

It follows that $C_r$ defines an $\alpha$-size test.

To see that $d\pi_{C_r}(\Theta_0)/d\Theta = 0$ is satisfied, first note that since $\pi_{C_r}(\Theta) = \mathrm{E}_\Theta I_{C_r}(\mathbf{X})$, it follows from the exponential class representation of the density of $\mathbf{X}$ and Lemma 9.1 that (continuous case—discrete case is analogous)

$$\begin{aligned} d\pi_{C_r}(\Theta)/d\Theta &= \frac{d\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} I_{C_r}(\mathbf{x}) \exp[c(\Theta)g(\mathbf{x}) + d(\Theta) + z(\mathbf{x})] I_A(\mathbf{x})}{d\Theta} \\ &= \mathrm{E}_\Theta I_{C_r}(\mathbf{X}) g(\mathbf{X}) \frac{dc(\Theta)}{d\Theta} + \frac{dd(\Theta)}{d\Theta} P(\mathbf{x} \in C_r; \Theta). \end{aligned}$$

Therefore,

$$d\pi_{C_r}(\Theta_0)/d\Theta = 0 \text{ iff } \mathrm{E}_{\Theta_0} I_{C_r}(\mathbf{X}) g(\mathbf{X}) \frac{dc(\Theta_0)}{d\Theta} = -\alpha \frac{dd(\Theta_o)}{d\Theta}.$$

The right-hand side of the iff statement does in fact hold under the current assumptions, since

$$\begin{aligned} E_{\Theta_0} I_{C_r}(\mathbf{X}) g(\mathbf{X}) \frac{dc(\Theta_0)}{d\Theta} &= \mathrm{E}_{\Theta_0}\left[(g(\mathbf{X}) - \gamma) I_{C_r}(\mathbf{X}) \frac{dc(\Theta_0)}{d\Theta}\right] + \mathrm{E}_{\Theta_0} \gamma I_{C_r}(\mathbf{X}) \frac{dc(\Theta_0)}{d\Theta} \\ &= \alpha\gamma \frac{dc(\Theta_0)}{d\Theta} = -\alpha \frac{dd(\Theta_0)}{d\Theta}, \end{aligned}$$

where the last equality follows from the fact that $\mathrm{E}_{\Theta_0} g(\mathbf{X})(dc(\Theta_0)/d\Theta) = -(dd(\Theta_0)/d\Theta)$, since

$$0 = \frac{d\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp[c(\Theta)g(\mathbf{x}) + d(\Theta) + z(\mathbf{x})] I_A(\mathbf{x}) d\mathbf{x}}{d\Theta} = \mathrm{E}_\Theta g(\mathbf{X}) \frac{dc(\Theta)}{d\Theta} + \frac{dd(\Theta)}{d\Theta}. \blacksquare$$

**Example 9.25** Revisit Ex. 9.20 and Ex. 9.21. In the first example, since the binomial density of $Z = \sum_{i=1}^{100} X_i$ is symmetric about $\mathrm{E}Z = 50$, we know from Theorem 9.10 that $c_1$ and $c_2$ such that $P(z \geq c_2; .5) = \alpha/2$ and $c_1 = 2(50) - c_2$ will define a UMPU level-$\alpha$ critical region $C_r = \{\mathbf{x}\colon \sum_{i=1}^{100} x_i \leq c_1 \text{ or } \sum_{i=1}^{100} x_i \geq c_2\}$. In particular, $P(z \geq 60; .5) = .0285$, $c_1 = 100 - 60 = 40$, and then $C_r = \{\mathbf{x}\colon \sum_{i=1}^{100} x_i \leq 40 \text{ or } \sum_{i=1}^{100} x_i \geq 60\}$ defines the appropriate size .057 critical region.

In the second example, since the normal distribution of $Z = \sum_{i=1}^{100} X_i$ is symmetric about $\mathrm{E}Z = 400$, we know from Theorem 9.10 that $c_1$ and $c_2$ such that $P(z \geq c_2; 400) = \alpha/2$ and $c_1 = 2(400) - c_2$ will define a UMPU level-$\alpha$ critical region $C_r = \{\mathbf{x}\colon \sum_{i=1}^{200} x_i \leq c_1 \text{ or } \sum_{i=1}^{200} x_i \geq c_2\}$. In particular, $P(z \geq 404.653) = .05$, $c_1 = 800 - 404.653 = 395.347$, and then $C_r = \{\mathbf{x}\colon \sum_{i=1}^{200} x_i \leq 395.347 \text{ or } \sum_{i=1}^{200} x_i \geq 404.653\}$ defines the appropriate size .10 critical region. □

It should also be noted that in the discrete case, it may not be possible to define an unbiased test, let alone a UMPU test, without resorting to so-called randomized tests (see Problem 9.8). The reader should revisit Ex. 9.20 and examine alternative choices of $p$ to explore the difficulty. The problem does not arise in the continuous case.

### *Conditioning in the Multiple Parameter Case

In the multiple parameter case, the random sample, $\mathbf{X}$, has a joint distribution, $f(\mathbf{x}; \boldsymbol{\Theta})$, that belongs to a family of distributions indexed by a $(k \times 1)$ parameter vector $\boldsymbol{\Theta} \in \Omega$, where $k > 1$. We will examine hypotheses of the form $H_0: \Theta_i \in \Omega_0^i$ versus $H_a: \Theta_i \in \Omega_a^i$, where $\Omega^i = \Omega_0^i \cup \Omega_a^i$ represents the admissible values cf the $i$th parameter $\Theta_i$. Note that for both $H_0$ and $H_a$, the values of $\Theta_j$, $j \neq i$, are left unspecified, and so it is tacitly understood that $\Theta_1, \Theta_2, \ldots, \Theta_{i-1}, \Theta_{i+1}, \ldots, \Theta_k$ can assume any admissible values for which $\Theta \in \Omega$. In this hypothesis-testing context, the parameters $\Theta_j$, $j \neq i$, are often referred to as **nuisance parameters**, since we are not interested in them from the standpoint of our stated hypotheses, but they must nonetheless be dealt with in defining a statistical test of $H_0$ versus $H_a$. We will also examine explicitly tests of the more general hypotheses $H_0: \gamma(\boldsymbol{\Theta}) \in \Omega_0^{\gamma}$ versus $H_a: \gamma(\boldsymbol{\Theta}) \in \Omega_a^{\gamma}$, where $\gamma(\cdot)$ is a scalar function of the parameter vector $\boldsymbol{\Theta}$.

The approach we will use for defining tests in the multiparameter case will allow us to use the previous results we have obtained concerning tests for the single parameter case. In particular, we will seek to transform the problem into a single parameter situation by conditioning on sufficient statistics and then apply previous results for the single parameter case to the transformed problem. In a significant number of cases of practical interest, the approach can be used to define UMP level-$\alpha$ tests, or at least UMP level-$\alpha$ tests within the unbiased class of tests. We will examine a number of other useful and more versatile test procedures for the multiparameter case in Chapter 10, although the procedures will not necessarily be UMP or UMPU tests.

There is a substantial literature on UMP and UMPU tests of statistical hypotheses in multiparameter situations that is more advanced than what we present here and that also applies to a wider array of problems. For additional reading, and a multitude of references, see E. Lehmann (1986), *Testing Statistical Hypotheses*, 2nd ed. New York: John Wiley, pp. 134–281.

*UMPU Level-$\alpha$ Tests via Conditioning* The basic idea of conditioning in order to transform a multiparameter problem into a problem involving only a single parameter, or a scalar function of parameters, is as follows. Let $S_1, \ldots, S_r$ be a set of sufficient statistics for $f(\mathbf{x}; \boldsymbol{\Theta})$. Suppose that the elements in the parameter vector $\boldsymbol{\Theta}$ have been ordered so that $\Theta_1$ is the parameter of interest in the hypothesis test, and $\Theta_2, \ldots, \Theta_k$ are nuisance parameters. Suppose further that, for each *fixed* value of $\Theta_1^0 \in \Omega^1$, $S_j, \ldots, S_r$ are sufficient statistics for the $k-1$ parameter density $f(\mathbf{x}; \Theta_1^0, \Theta_2, \ldots, \Theta_k)$, $(\Theta_1^0, \Theta_2, \ldots, \Theta_k) \in \Omega$. It then follows from the definition of sufficiency (recall Def. 7.18) that

$$f_*(\mathbf{x}; \Theta_1^0) = f(\mathbf{x}; \Theta_1^0, \Theta_2, \ldots, \Theta_k \mid s_j, \ldots, s_r), \quad \Theta_1^0 \in \Omega^1,$$

i.e., the *conditional* distribution of $\mathbf{X}$, given $s_j, \ldots, s_r$, does *not* depend on the nuisance parameters $\Theta_2, \ldots, \Theta_k$. In effect, by conditioning on $s_j, \ldots, s_r$ the problem will have been converted into one involving only the parameter $\Theta_1$. As long as we are concerned only with results obtained in the context of the conditional problem, single parameter UMP or UMPU hypothesis-testing results would apply.

An analogous argument can be applied to a *reparameterized* version of $f(\mathbf{x}; \Theta)$. Suppose the joint density function of $\mathbf{X}$ was reexpressed in terms of the parameter vector $\mathbf{c}$, where $c_i = c_i(\Theta)$ for $i = 1, \ldots, k$, so that $f_*(\mathbf{x}; \mathbf{c})$, for $\mathbf{c} \in \Omega_c = \{\mathbf{c}\colon c_i = c_i(\Theta), i = 1, \ldots, k, \Theta \in \Omega\}$, is an alternative representation of the parametric family of densities $f(\mathbf{x}; \Theta), \Theta \in \Omega$. Then if, for each fixed $c_1^0 \in \Omega_c^1$, $S_j^*, \ldots, S_r^*$ is a set of sufficient statistics for $f_*(\mathbf{x}; c_1^0, c_2, \ldots, c_k)$, $(c_1^0, c_2, \ldots, c_k) \in \Omega_c$, it would follow that

$$f_{**}(\mathbf{x}; c_1^0) = f_*(\mathbf{x}; c_1^0, c_2, \ldots, c_k \mid s_j^*, \ldots, s_r^*), \quad c_1^0 \in \Omega,$$

i.e., the conditional distribution of $\mathbf{X}$, given $s_j^*, \ldots, s_r^*$, would not depend on $c_2, \ldots, c_k$. Then single parameter UMP and UMPU results could be applied to testing hypotheses about $c_1(\Theta)$, *conditional* on $s_j^*, \ldots, s_r^*$.

For the conditioning procedure to have practical value, the optimal properties possessed by the critical region, $C_r$, in the *conditional* problem need to carry over to the context of the original *unconditional* problem. In particular, we will focus on determining when a $C_r$ that is UMPU level $\alpha$ for the conditional problem also possesses this property in the unconditional problem. The exponential class of densities represents an important set of cases in which the conditioning procedure works well. We will have use for the following generalizations of Lemma 9.1 and Theorem 9.8:

**Lemma 9.2**
***Differentiability of $\gamma(\Theta) = \mathrm{E}_\Theta\phi(X)$ in the Case of a Multiparameter Exponential Class Density***

Let $f(\mathbf{x}; \Theta) = \exp(\sum_{i=1}^k c_i(\Theta) g_i(\mathbf{x}) + d(\Theta) + z(\mathbf{x})) I_A(\mathbf{x})$, $\Theta \in \Omega$, be an exponential class density with $\Omega$ defined as an open rectangle contained in $R^k$ and with $c_1(\Theta), \ldots, c_k(\Theta)$ and $d(\Theta)$ being differentiable functions of $\Theta \in \Omega$. If the function $\phi(\mathbf{x})$ is such that $\gamma(\Theta) = \mathrm{E}_\Theta\phi(\mathbf{X})$ exists $\forall\ \Theta \in \Omega$, then $\gamma(\Theta) = \mathrm{E}_\Theta(\mathbf{X})$ is a differentiable function of $\Theta \in \Omega$. Furthermore,

$$\text{(continuous)} \quad \frac{\partial^{r+s}}{\partial\Theta_i^r \partial\Theta_j^s} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \phi(\mathbf{x}) f(\mathbf{x}; \Theta) d\mathbf{x} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \phi(\mathbf{x}) \frac{\partial^{r+s} f(\mathbf{x}; \Theta)}{\partial\Theta_i^r \partial\Theta_j^s} d\mathbf{x}$$

$$\text{(discrete)} \quad \frac{\partial^{r+s}}{\partial\Theta_i^r \partial\Theta_j^s} \sum_{\mathbf{x} \in R(\mathbf{X})} \cdots \sum \phi(\mathbf{x}) f(\mathbf{x}; \Theta)$$

$$= \sum_{\mathbf{x}\in R(\mathbf{X})} \cdots \sum \phi(\mathbf{x}) \frac{\partial^{r+s} f(\mathbf{x};\Theta)}{\partial\Theta_i^r \partial\Theta_j^s}$$

for $i$ and $j = 1, \ldots, k$, i.e., differentiation can occur under the integral or summation sign.

**Proof**: See references listed for Lemma 9.1.

**Theorem 9.11** (Differentiability (and Continuity) of Power Function in the Case of a Multiparameter Exponential Class Density) *Let the joint density of a random sample be a member of the exponential class of densities*

$$f(\mathbf{x};\Theta) = \exp\left(\sum_{i=1}^{k} c_i(\Theta) g_i(\mathbf{x}) + d(\Theta) + z(\mathbf{x})\right) I_A(\mathbf{x}),$$

$\Theta \in \Omega$, *with* $\Omega$ *defined as an open rectangle contained in* $R^k$ *and with* $c_1(\Theta), \ldots, c_k(\Theta)$, $d(\Theta)$ *being differentiable functions of* $\Theta \in \Omega$. *Let* $C_r$ *be any critical region for testing some hypothesis* $H_0: \Theta \in \Omega_0$ *versus* $H_a: \Theta \in \Omega_a$.[24] *Then the power function* $\pi_{C_r}(\Theta) = P(\mathbf{x} \in C_r;\Theta)$ *is differentiable (and hence continuous) with respect to* $\Theta \in \Omega$.

**Proof** This follows from Lemma 9.2 with a proof analogous to the proof of Theorem 9.8. ■

We will state our main results in terms of hypotheses concerning the value of the $c_1(\Theta)$ function in the definition of the exponential class density.[25] There is no loss of generality in stating our results this way—indeed, there is a gain in generality from doing so. First note that the ordering of the functions $c_1(\Theta), \ldots, c_k(\Theta)$ in the exponential class definition is arbitrary, and so the hypothesis-testing results we will present can be interpreted as referring to any of the originally specified $c_i(\Theta)$ functions given an appropriate reordering and relabeling of the functions $c_1(\Theta), \ldots, c_k(\Theta)$. Furthermore, since the parameterization of a density function is not unique, the $c_1(\Theta)$ function can be defined to represent various functions of $\Theta$ that might be of interest to the researcher, depending on the chosen parameterization of the density function. In particular, a parameterization of the density of $\mathbf{X}$ that results in $c_1(\Theta) = \Theta_i$ places hypotheses about $\Theta_i$ under consideration. We will also be able to extend our results to the case of

[24] We are assuming that $C_r$ is such that a power function is defined, i.e., $C_r$ can be assigned probability by $f(\mathbf{x};\Theta), \Theta \in \Omega$.

[25] Note that $\mathbf{c} = (c_1, \ldots, c_k)'$ could be viewed as an alternative parameterization of the exponential class of densities, where

$$f_*(\mathbf{x};\mathbf{c}) = \exp\left(\sum_{i=1}^{k} c_i g_i(\mathbf{x}) + d_*(\mathbf{c}) + z(\mathbf{x})\right) I_A(\mathbf{x}), \quad \mathbf{c} \in \Omega_c,$$

with $d_*(\mathbf{c}) = \ln\left(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(\sum_{i=1}^{k} c_i g_i(\mathbf{x}) + z(\mathbf{x})\right) I_A(\mathbf{x}) d\mathbf{x}\right)^{-1}$ (Use summation in discrete case). This parameterization is referred to as the **natural parameterization of the exponential class** of densities. Note that the definition of $d_*(\mathbf{c})$ is a direct result of the fact that the density must integrate (or sum) to 1.

testing hypotheses concerning linear combinations of the functions $c_1(\Theta), \ldots, c_k(\Theta)$. Thus, types of hypotheses that can be tested using the following theorem are much more general than they might at first appear.

**Theorem 9.12** (Hypothesis Testing in the Multiparameter Exponential Class) *Let the density of the random sample* $\mathbf{X}$ *be given by* $f(\mathbf{x};\Theta) = \exp(\sum_{i=1}^{k} c_i(\Theta) g_i(\mathbf{x}) + d(\Theta) + z(\mathbf{x})) I_A(\mathbf{x}), \Theta \in \Omega \subset R^k$, *where* $\Omega$ *is an open rectangle,* $s_i(\mathbf{X}) = g_i(\mathbf{X}), i = 1, \ldots, k$, *represents a set of complete sufficient statistics for* $f(\mathbf{x};\Theta)$, *and where* $c(\Theta) = (c_1(\Theta), \ldots, c_k(\Theta))$ *and* $d(\Theta)$ *are differentiable functions of* $\Theta \in \Omega$. *If the size-*$\alpha$ *critical region* $C_r$ *is unbiased, then the following relationships between* $H_0$, $H_a$, *and* $C_r$ *hold:*

| **Case** | $H_0$ | $H_a$ | $C_r$ **for UMPU Level-**$\alpha$ **Test** |
|---|---|---|---|
| (1) | $c_1(\Theta) = c_1^0$<br>$c_1(\Theta) \leq c_1^0$ | $c_1(\Theta) > c_1^0$<br>$c_1(\Theta) > c_1^0$ | $C_r = \{\mathbf{x}: s_1(\mathbf{x}) \geq h(s_2, \ldots, s_k)\}$<br>such that $P(\mathbf{x} \in C_r; c_1^0 \mid s_2, \ldots, s_k) = \alpha$,<br>$\forall (s_2, \ldots, s_k)^{26}$ |
| (2) | $c_1(\Theta) = c_1^0$<br>$c_1(\Theta) \geq c_1^0$ | $c_1(\Theta) < c_1^0$<br>$c_1(\Theta) < c_1^0$ | $C_r = \{\mathbf{x}: s_1(\mathbf{x}) \leq h(s_2, \ldots, s_k)\}$<br>such that $P(\mathbf{x} \in C_r; c_1^0 \mid s_2, \ldots, s_k) = \alpha$,<br>$\forall (s_2, \ldots, s_k)^{26}$ |
| (3) | $c_1(\Theta) = c_1^0$ | $c_1(\Theta) \neq c_1^0$ | $C_r = \{\mathbf{x}: s_1(\mathbf{x}) \leq h_1(s_2, \ldots, s_k)$,<br>or $s_1(\mathbf{x}) \geq h_2(s_2, \ldots, s_k)\}$<br>such that $P(x \in C_r; c_1^0 \mid s_2, \ldots, s_k) = \alpha$,<br>$\forall (s_2, \ldots, s_k)^{26}$ |
| (4) | $c_1(\Theta) \in [c_1^\ell, c_1^h]$ | $c_1(\Theta) \notin [c_1^\ell, c_1^h]$ | $C_r = \{\mathbf{x}: s_1(\mathbf{x}) \leq h_1(s_2, \ldots, s_k)$<br>or $s_1(\mathbf{x}) \geq h_2(s_2, \ldots, s_k)\}$<br>such that $P(\mathbf{x} \in C_r; c_1^\ell \mid s_2, \ldots, s_k) = \alpha$,<br>$P(\mathbf{x} \in C_r; c_2^h \mid s_2, \ldots, s_k) = \alpha$,<br>$\forall (s_2, \ldots, s_k)^{26}$ |

***Sketch of the Proof** We focus on the case where the density function of the random sample is *discrete*. The proofs in the continuous case can be constructed using a similar approach, except the conditional densities involved would be degenerate, so that line integrals would be needed to assign conditional probabilities to events (recall the footnote to our definition of sufficient statistics Def. 7.18, regarding the degenerate conditional distribution $f(\mathbf{x} \mid s)$.)

Given the definition of $f(\mathbf{x};\Theta)$, examine the distribution of $\mathbf{X}$, conditional on $s_2, \ldots, s_k$:

$$f(\mathbf{x}; \Theta \mid s_2, \ldots, s_k) = \frac{P(\{\mathbf{x}\} \cap \{\mathbf{x}: s_i(\mathbf{x}) = s_i, i = 2, \ldots, k\})}{P(s_2, \ldots, s_k)}$$

[26] Except, perhaps, on a set having probability zero.

$$= \frac{\exp\left(c_1(\Theta)s_1(\mathbf{x}) + \sum_{i=2}^{k} c_i(\Theta)s_i + d(\Theta) + z(\mathbf{x})\right) I_{A_*}(\mathbf{x})}{\exp\left(\sum_{i=2}^{k} c_i(\Theta)s_i + d(\Theta)\right) \sum \cdots \sum_{\{\mathbf{x}:\ s_i(\mathbf{x})=s_i, i=2,\ldots,k\}} \exp(c_1(\Theta)s_1(\mathbf{x}) + z(\mathbf{x}))I_A(\mathbf{x})}$$

$$= \exp(c_1(\Theta)s_1(\mathbf{x}) + d_*(c_1) + z(\mathbf{x}))I_{A_*}(\mathbf{x}),$$

where $A_* = A \cap \{\mathbf{x}:\ s_i(\mathbf{x}) = s_i,\ i = 2, \ldots, k\}$, and

$d_*(c_1) = \ln[\sum \cdots \sum_{x \in A_*} \exp(c_1(\Theta)s_1(\mathbf{x}) + z(\mathbf{x}))]^{-1}$.

Thus, $f_*(\mathbf{x}; c_1(\Theta) \mid s_2, \ldots, s_k) = f(\mathbf{x}; \Theta \mid s_2, \ldots, s_k)$ is a *one-parameter exponential class density* with parameter $c_1(\Theta)$ and parameter space $\Omega_{c_1} = \{c_1 : c_1 = c_1(\Theta), \Theta \in \Omega\}$. Note that $c_1$ is differentiable with respect to $c_1$, and it is also true that $d_*(c_1)$ is differentiable. The latter result follows from the chain rule of differentiation, since the natural logarithmic function is differentiable for all positive values of its argument, and the positive-valued bracketed term in the definition of $d_*(c_1)$ is differentiable by Lemma 9.1 with $\phi(\mathbf{x}) = 1$.

Since the assumptions of Theorem 9.11 are met under the stated conditions, all power functions for testing $H_0$ versus $H_a$ are continuous and differentiable. It follows that, *conditional* on $(s_2, \ldots, s_k)$, all of the previous results regarding UMP and UMPU tests for the case of a single parameter exponential class density apply. In particular, the definitions of the $C_r$'s in cases (1)–(4) are all *conditionally* UMPU level-$\alpha$, based on Theorems 9.5 and 9.6 and Corollaries 9.1 and 9.2 for results (1) and (2), and on Theorem 9.9 for results (3) and (4).

Examine result (1). Let $C_r^*$ be any other *unbiased* level-$\alpha$ test, so that $\pi_{C_r^*}(c_1^0) = \alpha_* \leq \alpha$. Then it must be the case that $\pi_{C_r^*}(c_1^0 \mid S_2, \ldots, S_k) \leq \alpha$ with probability 1. To see this, note that by the double expectation theorem,

$$\pi_{C_r^*}(c_1^0) - \alpha_* = \mathrm{E}_{c_1^0} I_{C_r^*}(\mathbf{X}) - \alpha_* = \mathrm{E}\left[\mathrm{E}_{c_1^0}\left(I_{C_r^*}(\mathbf{X}) \mid S_2, \ldots, S_k\right) - \alpha_*\right] = 0,$$

and since the bracketed expression is a function of the *complete* sufficient statistics $(S_2, \ldots, S_k)$ it follows by definition that the bracketed expression must equal 0 and thus $\mathrm{E}_{c_1^0}(I_{C_r^*}(\mathbf{X}) \mid S_2, \ldots, S_k) = \alpha_* \leq \alpha$ with probability 1 (recall Def. 7.20). Thus, $C_r^*$ must be *conditionally* unbiased of level-$\alpha$.

Now since $C_r$ has maximum power for $c_1 > c_1^0$ among all unbiased level-$\alpha$ tests conditional on $(s_2, \ldots, s_k)$, it follows from the double expectation theorem that for $c_1 > c_1^0$,

$$\pi_{C_r}(c_1) = \mathrm{E}_c\left(\mathrm{E}_{c_1}(I_{C_r}(\mathbf{X}) \mid S_2, \ldots, S_k)\right) \geq \mathrm{E}_c\left(\mathrm{E}_{c_1}(I_{C_r^*}(\mathbf{X}) \mid S_2, \ldots, S_k)\right) = \pi_{C_r^*}(c_1)$$

so that $C_r$ defines a UMPU level-$\alpha$ test of the hypotheses in case (1).

The proof of case (2) follows directly from the proof of case (1) via appropriate inequality reversals. Moderate extensions of the preceding argument can be used to prove results (3) and (4) (see T. Ferguson (1967), *Mathematical Statistics*, New York: Academic Press, pp. 230–232). ■

We now study examples illustrating how Theorem 9.12 can be used in practice to define UMPU level-$\alpha$ tests of statistical hypotheses. Some of the examples will illustrate the use of *reparameterization* for transforming a problem into a form that allows Theorem 9.12 to be used for defining a statistical test. Where appropriate, we point out general hypothesis-testing contexts to which the results of the examples can be applied.

**Example 9.26** Testing Hypotheses About the Variance of a Normal Population Distribution

The number of miles that a certain brand of automobile tire can be driven before the tread-wear indicators are visible and the tires need to be replaced is a random variable having a normal distribution with unknown mean *and* variance, i.e., $X_i \sim N(\mu, \sigma^2)$. A random sample of 28 tires is used to obtain information about the mean and variance of the population distribution of mileage obtainable from this brand of tire. We wish to test the hypotheses that $H_0: \sigma^2 \leq 4$ versus $H_a: \sigma^2 > 4$, where the mileage measurement, $x_i$, is measured in thousands of miles. Define a UMPU level-.05 test of $H_0$ versus $H_a$. Given that $\bar{x} = 45.175$ and $s^2 = 2.652$, test the hypothesis.

**Answer**: First note that the joint density of the random sample, $f(\mathbf{x}; \mu, \sigma^2) = (1/(2\pi)^{n/2}\sigma^n)\exp(-(1/2)\sum_{i=1}^n (x_i - \mu)^2/\sigma^2)$, for $n = 28$, can be written in exponential class form as

$$f(\mathbf{x}; \mu, \sigma^2) = \exp\left[\sum_{i=1}^{2} c_i(\mu, \sigma^2) g_i(\mathbf{x}) + d(\mu, \sigma^2) + z(\mathbf{x})\right],$$

where

$$c_1(\mu, \sigma^2) = -(2\sigma^2)^{-1}, \qquad g_1(\mathbf{x}) = \sum_{i=1}^{n} x_i^2,$$

$$c_2(\mu, \sigma^2) = \left(\frac{\mu}{\sigma^2}\right), \qquad g_2(\mathbf{x}) = \sum_{i=1}^{n} x_i,$$

$$d(\mu, \sigma^2) = -\frac{n}{2}\left(\frac{\mu^2}{\sigma^2}\right) - \ln((2\pi)^{n/2}\sigma^n), \quad \text{and} \quad z(\mathbf{x}) = 0.$$

Letting $s_i(\mathbf{x}) = g_i(\mathbf{x})$ for $i = 1, 2$, $s_1(\mathbf{X})$ and $s_2(\mathbf{X})$ are complete sufficient statistics for $f(\mathbf{x}; \mu, \sigma^2)$. Examining the parameterization of the exponential class density, it is evident that hypotheses concerning $\sigma^2$ can be framed in terms of hypotheses about $c_1$. In particular, $H_0: \sigma^2 \leq 4$ versus $H_a: \sigma^2 > 4$ can be alternatively expressed as $H_0: c_1 \leq -.125$ versus $H_a: c_1 > -.125$.

Given result (1) of Theorem 9.12, we seek $h(s_2)$ such that for $c_1^0 = -.125$,

$$\pi_{C_r}(c_1^0) = P\left(s_1(\mathbf{x}) \geq h(s_2); c_1^0 \mid s_2\right) = P\left(\sum_{i=1}^{n} x_i^2 \geq h\left(\sum_{i=1}^{n} x_i\right); c_1^0 \mid \sum_{i=1}^{n} x_i\right) = \alpha = .05$$

$\forall s_2 = \sum_{i=1}^n x_i$ (except, perhaps, on a set of $s_2$ values having probability zero). We can simplify this problem substantially by first recalling (Theorem 6.12) that the two random variables $\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n^{-1}\left(\sum_{i=1}^n X_i\right)^2$ and $\sum_{i=1}^n X_i$ are *independent* random variables, given that the $X_i$'s are iid $N(\mu, \sigma^2)$.

Then the preceding probability equality can be written alternatively as

$$P\left(\sum_{i=1}^{n}(x_i-\bar{x})^2 \geq h_*\left(\sum_{i=1}^{n}x_i\right); c_1^0\right)=\alpha=.05,$$

where $h_* = h - n^{-1}\left(\sum_{i=1}^{n} x_i\right)^2$. Note that we have eliminated the conditional probability notation since the probability distribution of the random variable $\sum_{i=1}^{n}(X_i-\bar{X})^2$ is *unaffected* by the value of $\sum_{i=1}^{n} x_i$ given the aforementioned independence property.

We can simplify the problem still further by first recalling that $\sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2 \sim \chi^2_{n-1}$ when the $X_i$'s are iid $N(\mu, \sigma^2)$ (recall Theorem 6.12). Now rewrite the preceding probability equality as

$$P\left(\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{\sigma_0^2} \geq h_*\left(\sum_{i=1}^{n}x_i\right)/\sigma_0^2; c_1^0\right)=\alpha=.05 \qquad (*)$$

where $\sigma_0^2 = 4$. Note in (*) that the probability is being calculated using the parameter value $c_1^0 = -.125$, which coincides with $\sigma_0^2 = 4$. In other words, the probability is being calculated assuming that $X_i \sim N(\mu, 4)\ \forall\, i$, so that $\sum_{i=1}^{n}(X_i - X)^2/\sigma_0^2 = \sum_{i=1}^{n}(X_i - X)^2/4 \sim \chi^2_{n-1}$. It follows that the value of $h_*(\sum_{i=1}^{n} x_i)/\sigma_0^2$ that solves (*) can be obtained from the table of the $\chi^2$ distribution, with $n-1=27$ degrees of freedom in this case. The tabled value of 40.1 corresponding to $\alpha = .05$ implies that

$$\begin{aligned} h_*\left(\sum_{i=1}^{n}x_i\right)/\sigma_0^2 &= \left[h\left(\sum_{i=1}^{n}x_i\right)-n^{-1}\left(\sum_{i=1}^{n}x_i\right)^2\right]/\sigma_0^2 \\ &= \left[h\left(\sum_{i=1}^{28}x_i\right)-\frac{1}{28}\left(\sum_{i=1}^{28}x_i\right)^2\right]/4 = 40.1, \end{aligned}$$

so that the solution for $h(\sum_{i=1}^{n} x_i)$ is

$$h\left(\sum_{i=1}^{28}x_i\right)=160.4+\left(\sum_{i=1}^{28}x_i\right)^2/28.$$

We are given that $\sum_{i=1}^{28} x_i = 28(45.175) = 1264.9$, so that $h(1264.9) = 57,302.258$. Then by Theorem 9.12, the critical region becomes

$$C_r = \left\{\mathbf{x}: \sum_{i=1}^{28}x_i^2 \geq 57,302.258\right\}.$$

Equivalently, using (*) with $\sigma_0^2 = 4$ and $n = 28$,

$$C_r = \{\mathbf{x}: 28s^2/4 \geq 40.1\}.$$

Then since $28s^2/4 = 28(2.652)/4 = 18.564 \not\geq 40.1$, the null hypothesis $\sigma^2 \leq 4$ is *not* rejected.

The power function of the test can be defined in terms of the parameter $\sigma^2$ as

$$\pi_{C_r}(\sigma^2) = P\left(\frac{28s^2}{\sigma^2}\left(\frac{\sigma^2}{4}\right) \geq 40.1; \sigma^2\right)$$

$$= P\left(\frac{28s^2}{\sigma^2} \geq \frac{160.4}{\sigma^2}; \sigma^2\right).$$

The power function could also have been expressed as a function of $c_1$, but power in terms of $\sigma^2$ is more natural given the original statement of the hypothesis. Since $28S^2/\sigma^2 \sim \chi^2_{27}$ when the variance of the $X_i$'s equals $\sigma^2$, we can use the CDF of the $\chi^2_{27}$ distribution to plot a graph of the power function for various potential values of $\sigma^2$:

| $\sigma^2$ | $P(28s^2/\sigma^2 \geq 160.4/\sigma^2; \sigma^2) = P(\chi^2_{27} \geq 160.4/\sigma^2; \sigma^2)$ |
|---|---|
| 4 | .050 |
| 3.9, 4.1 | .040, .062 |
| 3.75, 4.25 | .028, .082 |
| 3.50, 4.50 | .013, .123 |
| 3.00, 5.00 | .002, .229 |
| 6.00 | .478 |
| 7.00 | .700 |
| 8.00 | .829 |
| 10.00 | .952 |
| 13.00 | .993 |

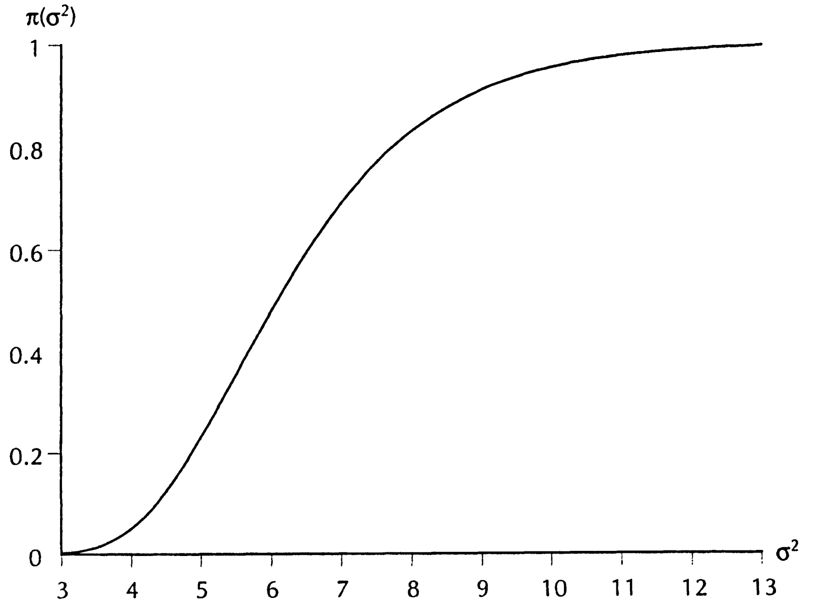


**Figure 9-14**
Power function of UMPU level-.05 unbiased test of $H_0 : \sigma^2 \leq 4$ versus $H_a : \sigma^2 > 4$.

Note that even though $C_r$ defines a *UMPU level-.05 test* (by Theorem 9.12), the test is not very powerful for values of $\sigma^2 \in (4, 7)$, say. To increase the power of the test in this range, one would either need to increase the size of the random sample on which the test is based, or else use a test size larger than .05 (although the latter approach would, of course, increase the type I error probability associated with the test). □

We can generalize the results of the previous example as follows.

**Definition 9.14** ***UMPU level-α statistical test of $H_0: \sigma^2 \le \sigma_0^2$ versus $H_a: \sigma^2 > \sigma_0^2$ when sampling from a normal population distribution***

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. The critical region

$$C_r = \left\{\mathbf{x}: \frac{ns^2}{\sigma_0^2} \ge \chi^2_{n-1;\alpha}\right\},$$

where $\chi^2_{n-1;\alpha}$ is a value solving $\int_{\chi^2_{n-1;\alpha}}^{\infty} f(z; n-1)dz = \alpha$, and $f(z; n-1)$ is a $\chi^2$ density function having $n-1$ degrees of freedom, defines a UMPU level-$\alpha$ statistical test of $H_0: \sigma^2 \le \sigma_0^2$ versus $H_0: \sigma^2 > \sigma_0^2$.

The next example illustrates the fact that a great deal of ingenuity may be required, concomitant with a rather high degree of mathematical sophistication, in order to implement the results of Theorem 9.12.

***Example 9.27** (Testing Hypotheses About the Mean of a Normal Population Distribution) Recall Ex. 9.26 regarding mileage obtained with automobile tires. Define UMPU level-.05 tests of $H_0: \mu = 40$ versus $H_a: \mu \neq 40$, and of $H_0: \mu \le 40$ versus $H_a: \mu > 40$.

The parameterization used in Ex. 9.26 is not directly useful for implementing the results of Theorem 9.12 to define statistical tests of the hypotheses under consideration. In considering alternative parameterizations, the following lemma is useful.

**Lemma 9.3** ***Reparameterizing Exponential Class Densities via Linear Transformations***

Let an exponential class density be given by

$$f(\mathbf{x}; \mathbf{c}) = \exp\left[\sum_{i=1}^{k} c_i g_i(\mathbf{x}) + d(\mathbf{c}) + z(\mathbf{x})\right] I_A(\mathbf{x}).$$

Let $c_1^* = \sum_{i=1}^{k} a_i c_i$, with $a_1 \neq 0$, and let $\mathbf{c}^* = [c_1^*, c_2, \ldots, c_k]'$. Then $f(\mathbf{x}; \mathbf{c})$ can be alternatively represented as

$$f(\mathbf{x}; \mathbf{c}^*) = \exp\left[c_1^* g_1^*(\mathbf{x}) + \sum_{i=2}^{k} c_i g_i^*(\mathbf{x}) + d^*(\mathbf{c}^*) + z(\mathbf{x})\right] I_A(\mathbf{x}),$$

where

$$g_1^*(\mathbf{x}) = \frac{g_1(\mathbf{x})}{a_1}; \qquad g_i^*(\mathbf{x}) = g_i(\mathbf{x}) - \left(\frac{a_i}{a_1}\right) g_1(\mathbf{x}), \quad i = 2, \ldots, k;$$

and

$$d^*(\mathbf{c}^*) = \ln\left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[c_1^* g_1^*(\mathbf{x}) + \sum_{i=2}^{k} c_i g_i^*(\mathbf{x}) + z(\mathbf{x})\right] I_A(\mathbf{x}) d\mathbf{x}\right]^{-1}.$$

**Proof**: The proof is immediate upon substituting the definitions of $c_1^*$ and $g_i^*(\mathbf{x}), i = 1, \ldots, k$, into the expression for $f(\mathbf{x}; \mathbf{c}^*)$, and defining $d(\mathbf{c}^*)$ so that the density integrates to 1. Replace integration with summation in the discrete case.

The point of the lemma is that any linear combination of the $c_i(\cdot)$ functions in the definition of an exponential class density can be used as a parameter in a reparameterization of the density (note that the $c_i(\cdot)$ functions can always be reordered and relabeled so that the condition $a_1 \neq 0$ is not a restriction in practice). Then Theorem 9.12 can be used to test hypotheses about $c_1^*$ or, equivalently, about $\sum_{i=1}^k a_i c_i$.

In the case at hand, *reorder* the $c_i(\cdot)$ functions and $g_i(\mathbf{x})$ functions so that

$$c_1^* = \sum_{i=1}^{2} a_i c_i(\mu, \sigma^2) = a_1\left(\frac{\mu}{\sigma^2}\right) + a_2\left(-\frac{1}{2\sigma^2}\right).$$

Without knowledge of the value of $\sigma^2$, it is clear that for $c_1^* = c_1^0$ to imply a unique value of $\mu$, the condition $c_1^0 = 0$ is required, in which case $\mu = a_2/(2a_1)$. Then letting $c_1^0 = 0$ and $a_2/(2a_1) = \mu_0 = 40$, it follows that $c_1^* = 0$ iff $\mu = 40$, and $c_1^* \leq 0$ iff $\mu_0 \leq 40$, assuming $a_1 > 0$. This establishes an equivalence between the hypotheses concerning $\mu$ and hypotheses concerning $c_1^*$ as follows:

$$\left\{\begin{array}{ccc} H_0\colon \mu = 40 & & H_0\colon c_1^* = 0 \\ \text{versus} & \Longleftrightarrow & \text{versus} \\ H_a\colon \mu \neq 40 & & H_a\colon c_1^* \neq 0 \end{array}\right\} \quad \text{and} \quad \left\{\begin{array}{ccc} H_0\colon \mu \leq 40 & & H_0\colon c_1^* \leq 0 \\ \text{versus} & \Longleftrightarrow & \text{versus} \\ H_a\colon \mu > 40 & & H_a\colon c_1^* > 0 \end{array}\right\}$$

Test of $H_0\colon \mu \leq 40$ versus $H_a\colon \mu > 40$: The reparameterized exponential class density can be written as

$$f(\mathbf{x}; \mathbf{c}^*) = \exp\left[c_1^*\left(\sum_{i=1}^{n} \frac{x_i}{a_1}\right) + c_2\left(\sum_{i=1}^{n} x_i^2 - \frac{a_2}{a_1}\sum_{i=1}^{n} x_i\right) + d^*(\mathbf{c}^*)\right] I_A(\mathbf{x})$$

(recall Lemma 9.3, and the fact that $z(\mathbf{x}) = 0$ in this case). Our approach will be simplified somewhat if we rewrite $f(\mathbf{x}; \mathbf{c}^*)$ as

$$\begin{aligned} f(\mathbf{x}; \mathbf{c}^*) &= \exp\left[c_1^* \sum_{i=1}^{n} \frac{(x_i - \mu_0)}{a_1} + c_2 \sum_{i=1}^{n} (x_i - \mu_0)^2 + d^0(\mathbf{c}^*)\right] I_A(\mathbf{x}) \\ &= \exp\left[c_1^* s_1^* + c_2 s_2^* + d^0(\mathbf{c}^*)\right] I_A(\mathbf{x}) \end{aligned}$$

where $d^0(\mathbf{c}^*) = d^*(\mathbf{c}^*) - n\mu_0^2 c_2 + n\mu_0 c_1^*/a_1, \mu_0 = 40$, and we have used the fact that $-a_2/a_1 = -2\mu_0$.

Result (1) of Theorem 9.12 indicates that a UMPU level-$\alpha$ test of $H_0: c_1^* \leq 0$ versus $H_a: c_1^* > 0$ can be defined by first finding a function $h(s_2^*)$ such that

$$P\left(s_1^*(\mathbf{x}) \geq h(s_2^*); c_1^0 \mid s_2^*\right) = P\left(\sum_{i=1}^{n} \frac{(x_i - \mu_0)}{a_1} \geq h\left(\sum_{i=1}^{n}(x_i - \mu_0)^2\right); c_1^0 \mid s_2^*\right) = \alpha = .05 \quad (*)$$

$\forall\, s_2^* = \sum_{i=1}^{n}(x_i - \mu_0)^2$ (except, perhaps, on a set of $s_2^*$ values having probability zero), with $c_1^0 = 0$. Letting $\mathbf{c}^0 = [c_1^0, c_2]' = [0, c_2]'$, it is useful to recognize that

$$f(\mathbf{x}; \mathbf{c}^0) = \exp\left[c_2 s_2^*(\mathbf{x}) + d^0(\mathbf{c}^0)\right] I_A(\mathbf{x}),$$

which indicates that the value of $f(\mathbf{x}; \mathbf{c}^0)$ changes *only* when the value of $s_2^*(\mathbf{x})$ changes. This observation then clarifies the nature of the conditional distribution of $\mathbf{X}$ *given* $s_2^*$, namely, the conditional distribution of $\mathbf{X}$ is a *uniform* distribution on the range $R(\mathbf{X} \mid s_2^*) = \{\mathbf{x}: \sum_{i=1}^{n}(x_i - \mu_0)^2 = s_2^*\}$, where $R(\mathbf{X} \mid s_2^*)$ defines the boundary of an $n$-dimensional hypersphere having radius $(s_2^*)^{1/2}$. Specifically,[27]

$$f(\mathbf{x}; \mathbf{c}^0 \mid s_2^*) = \left[\frac{n(s_2^*)^{(n-1)/2}\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}\right]^{-1} I_{R(\mathbf{X}|s_2^*)}(\mathbf{x}).$$

It follows from the definition of the support of the density of $\mathbf{X}$ that $f(\mathbf{x}; \mathbf{c}^0 \mid s_2^*)$ is a *degenerate* density function, since all of the probability mass for the $(n \times 1)$ vector $\mathbf{X}$ is concentrated on the surface $R(\mathbf{X} \mid s_2^*)$, which has dimension $n - 1$.

We can now define the function $h(s_2^*)$ that satisfies $(*)$. Let $h(s_2^*) = \gamma_\alpha (s_2^*)^{1/2}$, where $\gamma_\alpha$ is a positive constant. Multiplying both sides of the inequality in $(*)$ by $a_1$ then defines

$$P\left(\sum_{i=1}^{n}(x_i - \mu_0) \geq \gamma_\alpha^*(s_2^*)^{1/2}; c_1^0 \mid s_2^*\right) = \alpha = .05, \quad (**)$$

where $\gamma_\alpha^* = a_1\gamma_\alpha$ is to be chosen so that $(**)$ is true. The choice of $\gamma_\alpha^*$ will be motivated by examining the case where $n = 2$, although an analogous argument applies to arbitrary finite $n$. Examine a circle with center $(\mu_0, \mu_0)$ and radius $r = (s_2^*)^{1/2}$, as in Figure 9.15. Recall from basic trigonometry that with $(\mu_0, \mu_0)$ as the origin, the point $(x_1^0 - \mu_0, x_2^0 - \mu_0)$ can be represented in terms of polar coordinates as $(r\cos\Theta, r\sin\Theta)$. The function of $\Theta$ defined by $\sum_{i=1}^{2}(x_i - \mu_0) = r[\cos\Theta + \sin\Theta]$ is strictly concave for $\Theta \in (-135, 225)$, attains a maximum at $\Theta = 45$ (the maximized value being $1.4142r$ to four decimals), and is symmetric around $\Theta = 45$. It follows that an $\alpha$ proportion of the points $(x_1, x_2)$ on the circumference of the circle can be defined by appropriately choosing $\gamma_\alpha^* < 1.4142$

[27] The surface area of an $n$-dimensional hypersphere is given by $A = (nr^{n-1}\pi^{n/2})/\Gamma((n/2)+1)$, where $r$ is the radius of the hypersphere. See R. G. Bartle (1976), *The Elements of Real Analysis*, 2nd ed. New York: John Wiley, pp. 454–455, and note that the surface area can be defined by differentiating the volume of the hypersphere with respect to $r$. For $n = 2$, the bracketed expression simply becomes $2\pi(s_2^*)^{1/2}$, which is the familiar $2\pi r$.

as

$$D_\alpha = \{(x_1, x_2): (x_1 - \mu_1) = r\cos\Theta, (x_2 - \mu_2) = r\sin\Theta, r(\cos\Theta + \sin\Theta) \geq \gamma_\alpha^* r\}$$

$$= \left\{(x_1, x_2): \sum_{i=1}^{2}(x_i - \mu_o) \geq \gamma_\alpha^*(s_2^*)^{1/2}\right\},$$

and since $(x_1, x_2)$ is distributed uniformly on the circumference, $P(D_\alpha \mid s_2^*) = \alpha$.

Now note that the preceding argument applies for *arbitrary* $r = (s_2^*)^{1/2} > 0$, so that $\gamma_\alpha^*$ is the *same* value $\forall\ r > 0$. An analogous argument leading to the constancy of $\gamma_\alpha^*$ can be applied to the case of an $n$-dimensional hypersphere.

Thus the critical region for the test will have the form

$$C_r = \left\{\mathbf{x}: \sum_{i=1}^{n}(x_i - \mu_0) \geq \gamma_\alpha^*(s_2^*)^{1/2}\right\},$$

regardless of the value of $s_2^*$. The remaining task is then to find the appropriate value of $\gamma_\alpha^*$ that defines a test of size $\alpha$.

It turns out that there is an alternative representation of the critical region in terms of the *t-statistic*, which is the form in which the test is usually defined in practice. To derive this alternative representation, first note that[28]

$$s_2^* = \sum_{i=1}^{n}(x_i - \mu_0)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2\left[1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right].$$

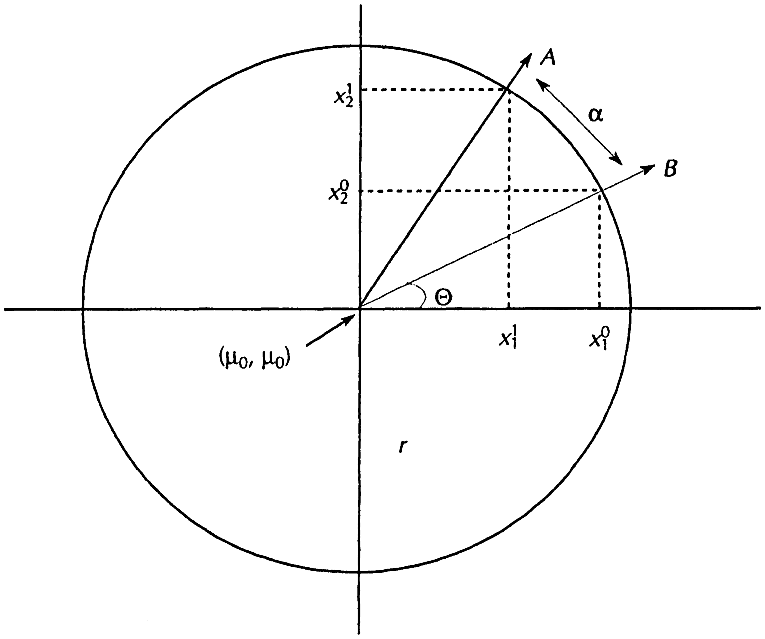


**Figure 9-15**
Graph of $R(X \mid r) = \left\{\mathbf{x}: \sum_{i=1}^{2}(x_i - \mu_0)^2 \leq r^2\right\}$.

[28] Expanding $\sum_{i=1}^{n}[(x_i - \bar{x}) + (\bar{x} - \mu_0)]^2$ leads to the result.

Then the critical region can be represented as

$$C_r = \left\{ \mathbf{x} : \frac{n(\bar{x} - \mu_0)}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^{1/2}} \geq \gamma_\alpha^* \left[1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]^{1/2} \right\}.$$

Multiplying both sides of the set-defining inequality by $[(n-1)/n]^{1/2}$ and defining $t = n^{1/2}(\bar{x} - \mu_0)/[\sum_{i=1}^n (x_i - \bar{x})^2/(n-1)]^{1/2}$, we can equivalently represent $C_r$ as

$$C_r = \left\{ \mathbf{x} : \frac{t}{(t^2 + n - 1)^{1/2}} \geq \gamma_\alpha^0 \right\},$$

where $\gamma_\alpha^0 = \gamma_\alpha^*/n^{1/2}$. Since $t/(t^2 + n - 1)^{1/2} = (1 + (n-1)/t^2)^{-1/2}$ is *strictly monotonically increasing* in $t$ (for $n > 1$), there exists a value $t_\alpha$ such that

$$\frac{t}{(t^2 + n - 1)^{1/2}} \geq \gamma_\alpha^0 \iff t \geq t_{n-1;\alpha},$$

so that $C_r$ can be substantially simplified to

$$C_r = \{\mathbf{x} : t \geq t_{n-1;\alpha}\}.$$

Assuming $\mu = \mu_0$ to be true, it would follow that $T$ has the student $t$-distribution with $n - 1$ degrees of freedom so that the value of $t_{n-1;\alpha}$ can be found as the value that makes $P(t \geq t_{n-1;\alpha}; \mu_0) = \alpha = .05$ true. From the table of the student $t$-distribution, with degrees of freedom equal to $n - 1 = 27$, it is found that $t_{27;.05} = 1.703$. Given that $\bar{x} = 45.175$ and $s^2 = 2.652$, it follows that for $\mu_0 = 40$, $t = (28)^{1/2}(45.175 - 40)/(2.75)^{1/2} = 16.513$. Since $t > 1.703$, we *reject* $H_0 : \mu \leq 40$ in favor of $H_a : \mu > 40$.

In order to graph the power function of the test, we must calculate values of $\pi(\mu) = P(t \geq 1.703; \mu)$, which in turn requires that we identify the distribution of $T$ when $\mu \neq \mu_0$. To see what is involved, note that we can represent $T$ as

$$T = \frac{(\bar{X} - \mu) + (\mu - \mu_0)}{\sigma/\sqrt{n}} \Big/ \left[\frac{nS^2}{\sigma^2}/(n-1)\right]^{1/2},$$

so that if $\mu \neq \mu_0$, we have

$$T = \left[Z + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right] \Big/ \left[\frac{Y}{(n-1)}\right]^{1/2},$$

where $Z \sim N(0,1)$, $Y \sim \chi_{n-1}^2$, and $Z$ and $Y$ are independent. Letting $\lambda = (\mu - \mu_0)/(\sigma/\sqrt{n})$, the random variable $T$ has a **noncentral *t*-distribution**, with **noncentrality parameter $\lambda$**, and degrees of freedom $n - 1$, denoted as $t_{n-1}(\lambda)$. When $\lambda = 0$, the ordinary *central* student $t$-distribution is defined. We discuss properties of the noncentral $t$-distribution in Section 9.6 of this chapter. For now, note that $\lambda$ depends not only on $\mu$, but also on $\sigma$, which effectively implies that we must either plot the power of the test in two dimensions (in terms of values of $\mu$ and $\sigma$), or else in a single dimension as a function of the noncentrality parameter. The latter is the approach most often followed in practice, and we follow it here as well (see Figure 9.16). Integrals of the noncentral $t$-distribution can be evaluated by many modern statistical software packages,

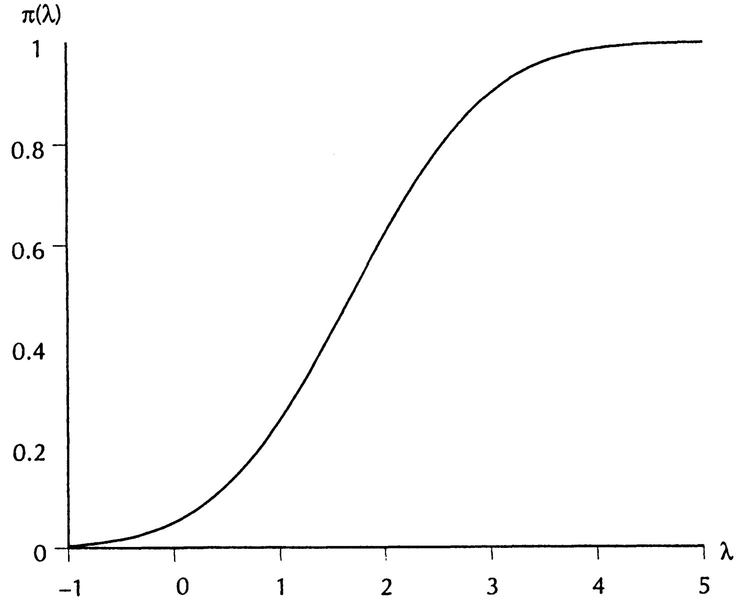


**Figure 9-16** Power function of UMPU level-.05 test of $H_0: \mu \leq 40$ versus $H_a: \mu > 40$, where $\lambda = (\mu - 40)/(\sigma/\sqrt{28})$.

such as SAS and GAUSS.

| $\lambda$ | $\pi(\lambda) = P(t \geq 1.703; \lambda)$ |
|---|---|
| 0 | .050 |
| −.1, .1 | .041, .061 |
| −.5, .5 | .016, .124 |
| −1, 1 | .004, .251 |
| 1.5 | .428 |
| 2 | .620 |
| 3 | .900 |
| 4 | .988 |
| 5 | .999 |

Test of $H_0: \mu = 40$ versus $H_a: \mu \neq 40$: In order to define a test of $H_0: \mu = 40$ versus $H_a: \mu \neq 40$, result (3) of Theorem 9.12 suggests that the critical region can be defined as

$$C_r = \{\mathbf{x}: s_1^*(\mathbf{x}) \leq h_1(s_2^*) \text{ or } s_1^*(\mathbf{x}) \geq h_2(s_2^*)\}$$

for $h_1$ and $h_2$ chosen such that $P(\mathbf{x} \in C_r; c_1^0 \mid s_2^*) = \alpha = .05 \; \forall \; s_2^*$ (except, perhaps, for a set of $s_2^*$ values having probability zero). Using an argument along the lines of the previous discussion (see T. Ferguson, *Mathematical Statistics*, p. 232), it can be shown that

$$C_r = \{\mathbf{x}: t \leq t_1 \text{ or } t \geq t_2\},$$

where $t$ is defined as before, and $t_1$ and $t_2$ are constants chosen so that when $\mu = \mu_0 = 40$, $P(\mathbf{x} \in C_r; \mu_0) = .05$. The additional condition in result (3) of

Theorem 9.12 is met by choosing $t_1 = -t_2$ since $T$ has the symmetric student $t$-distribution when $\mu = \mu_0$ and Theorem 9.10 applies. In particular, $C_r$ is defined by

$$C_r = \{\mathbf{x}:\ t \leq -t_{n-1;\alpha/2} \text{ or } t \geq t_{n-1;\alpha/2}\},$$

where $t_{n-1;\alpha/2}$ solves $\int_{t_{n-1;\alpha/2}}^{\infty} f(t; n-1)dt = \alpha/2$ with $f(t; n-1)$ being a student $t$-distribution with $n-1$ degrees of freedom. For a $t$-distribution with 27 degrees of freedom, $t_{27;.025} = 2.052$. Since $t = 16.513 > 2.052$, $H_0: \mu = 40$ is rejected in favor of $H_a: \mu \neq 40$.

The power function of the test is defined by values of $P(\mathbf{x} \in C_r; \mu)$. As before, $T$ has a noncentral $t$-distribution (see Section 9.6) with noncentrality parameter $\lambda = (\mu - 40)/(\sigma/\sqrt{28})$. Defining the power function in terms of $\lambda$ yields the following results.

| $\lambda$ | $\pi(\lambda) = P(t \leq -2.052$ or $t \geq 2.052; \lambda)$ |
|---|---|
| 0 | .050 |
| −.1, .1 | .051 |
| −.5, .5 | .077 |
| −1, 1 | .161 |
| −2, 2 | .488 |
| −3, 3 | .824 |
| −4, 4 | .971 |
| −5, 5 | .998 |

The power function is graphed in Figure 9.17. □

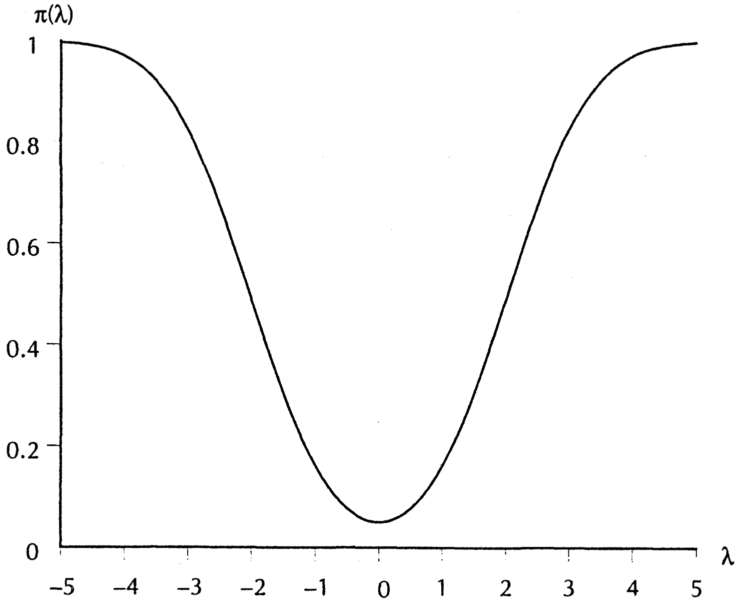


**Figure 9-17**
Power function of UMPU level-.05 test of $H_0: \mu = 40$ versus $H_a: \mu \neq 40$, where $\lambda = (\mu - 40)/(\sigma/\sqrt{28})$.

We can generalize the results of the previous example as follows.

**Definition 9.15** ***UMPU level-$\alpha$ statistical test of $H_0: \mu \leq \mu_0$ versus $H_a: \mu > \mu_0$, and of $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$, when sampling from a normal population distribution***

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, let $f(t; v)$ denote the student $t$-distribution with $v$ degrees of freedom, and define $T = (n-1)^{1/2}(\bar{X} - \mu_0)/S$.

1. $C_r = \{\mathbf{x}: t \geq t_{n-1;\alpha}\}$ defines a UMPU level-$\alpha$ statistical test of $H_0: \mu \leq \mu_0$ versus $H_a: \mu > \mu_0$, where $\int_{t_{n-1;\alpha}}^{\infty} f(t; n-1)dt = \alpha$.
2. $C_r = \{\mathbf{x}: t \leq -t_{n-1;\alpha/2}$ or $t \geq t_{n-1;\alpha/2}\}$ defines a UMPU level-$\alpha$ statistical test of $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$, where $\int_{t_{n-1;\alpha/2}}^{\infty} f(t; n-1)dt = \alpha/2$.

### Concluding Remarks

In this section we examined some theory and procedures related to the definition of statistical tests that were either UMP or UMPU level-$\alpha$ for various types of hypotheses. As is evident from the proofs of the theorems and the illustrative examples, attempting to define UMP or UMPU level-$\alpha$ tests can sometimes be a substantial challenge. While the results provided in this section provide direction to the search for a statistical test, application of the results still requires considerable ingenuity, and the procedures are not always tractable or even applicable.

In the next chapter we will examine well-defined procedures that will lead to specific statistical tests with less effort than the procedures introduced so far. While leading to the definition of statistical tests that *generally work well in practice*, the procedures do not explicitly incorporate the objective of finding a UMP or UMPU level-$\alpha$ test. As a result, they generally cannot be relied upon to produce tests with these properties. However, the power functions of such tests are nonetheless often quite adequate for hypothesis-testing applications, and the tests frequently do possess some optimal properties.

## 9.6 Noncentral *t*-Distribution

### Family Name: Noncentral *t*-Distribution

*Parameterization*

$v$ (degrees of freedom) and $\lambda$ (noncentrality parameter)

$(v, \lambda) \in \Omega = \{(v, \lambda): v \text{ is a positive integer}, \lambda \in (-\infty, \infty)\}$

*Density Definition*

$$f(t; v, \lambda) = \frac{v^{v/2}}{\pi^{1/2}\Gamma(v/2)2^{(v-1)/2}}(t^2 + v)^{-(v+1)/2} \exp\left[-\frac{\lambda^2 v}{2(t^2+v)}\right] \int_0^{\infty} \exp\left[-\frac{1}{2}\left(x - \frac{\lambda t}{(t^2+v)^{1/2}}\right)^2\right] x^v \, dx$$

*Moments*

$\mu = a_1\lambda$ for $v > 1$, $\sigma^2 = a_2\lambda^2 + a_3$ for $v > 2$, $\mu_3 = a_4\lambda^3 + a_5\lambda$ for $v > 3$, where

$$a_1 = (v/2)^{1/2}\Gamma\left(\frac{v-1}{2}\right)/\Gamma\left(\frac{v}{2}\right), \quad a_2 = \frac{v}{v-2} - a_1^2, \quad a_3 = \frac{v}{v-2},$$

$$a_4 = a_1\left[\frac{v(7-2v)}{(v-2)(v-3)} + 2a_1^2\right], \quad \text{and } a_5 = \left[\frac{3v}{((v-2)(v-3))}\right]a_1.$$

*MGF* Does not exist

The noncentral *t*-distribution is the distribution of the random variable

$$T = \frac{Z}{\sqrt{Y/v}},$$

where $Z \sim N(\lambda, 1)$, $Y \sim \chi^2_v$, and $Z$ and $Y$ are independent. Note that when $\lambda = 0$, $T$ will have the ordinary (*central*) student *t*-distribution derived in Chapter 6. The density definition can be found by the change-of-variables approach, similar to the derivation of the *t*-density in Section 6.7.

When $\lambda = 0$, the *t*-distribution is symmetric around zero. When $\lambda > 0$ ($\lambda < 0$), the mean of the distribution moves to the right (left) of zero, and the distribution becomes skewed to the right (left).

Our principal use of the noncentral *t*-distribution is in analyzing the power function of statistical tests based on the *t-statistic*. We have already encountered a case where the statistic $T = (n-1)^{1/2}(\bar{X} - \mu_0)/S$ was used to test hypotheses about the mean of a normal distribution, such as $H_0\colon \mu = \mu_0$ versus $H_a\colon \mu \neq \mu_0$, or $H_0\colon \mu \leq \mu_0$ versus $H_a\colon \mu > \mu_0$ (recall Ex. 9.27 and Def. 9.15). In particular, letting

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \Big/ \left(\frac{nS^2}{\sigma^2}\Big/(n-1)\right)^{1/2} = \frac{Z}{(Y/(n-1))^{1/2}},$$

where $Z \sim N((\mu - \mu_0)/(\sigma/\sqrt{n}), 1)$, $Y \sim \chi^2_{n-1}$, and $Z$ and $Y$ are independent (as in the context of Ex. 9.27), it follows from the definition of the noncentral *t*-density that $T \sim f(t; n-1, \lambda)$, with $\lambda = (\mu - \mu_0)/(\sigma/\sqrt{n})$. Then the power function of any test defined by $C_r = \{\mathbf{x}\colon t \in C_r^t\}$ can be expressed as a function of $\lambda$ as

$$\pi(\lambda) = \int_{t \in C_r^t} f(t; n-1, \lambda)\, dt.$$

The noncentral *t*-distribution can be integrated numerically, or the integral can be approximated via series expansions. Integration of the noncentral *t*-distribution adequate for performing power calculations is readily available in software packages such as GAUSS, SAS, and Shazam. Regarding power function calculations, it is useful to note that

$$\int_c^\infty f(t; v, \lambda)\, dt \text{ increases as } \lambda \text{ increases,}$$

$\int_{-\infty}^{c} f(t;\, v, \lambda)dt$ increases as $\lambda$ decreases,and

$\int_{-c}^{c} f(t;\, v, \lambda)dt$ decreases as $|\lambda|$ increases.

For further information, see the excellent survey article by D. B. Owen (1968), "A survey of properties and applications of the noncentral $t$-distribution." *Technometrics*, pp. 445–478.

## Key Words, Phrases, and Symbols

statistical hypothesis
simple hypothesis
composite hypothesis
$H_0$, $H_a$
$H$: set-defining conditions
range of $\mathbf{X}$ over $H$, $R(\mathbf{X} \mid H)$
test of a statistical hypothesis
rejection (or critical) region, $C_r$
acceptance region, $C_a$
statistical test defined by $C_r$
type I and type II errors
ideal statistical test
controlling type I and type II errors
type I/type II error tradeoff
protection against type I error
$H$ is at least approximately true
null hypothesis
alternative hypothesis
parametric hypothesis testing
maintained hypothesis
power function of a statistical test
power of the test at $\Theta$
operating characteristic function of a statistical test
ideal power function
size of test
significance level of the test
$H$ is (not) rejected at the $\alpha$-level of significance
$H$ is (not) rejected using a size-$\alpha$ test
unbiasedness of a test rule
$\pi_{C_r}(\Theta)$
uniformly most powerful level-$\alpha$ test
more powerful level-$\alpha$ test
consistent test sequence of level $\alpha$
$P$ value
Neyman–Pearson lemma
test statistic
likelihood ratio test
likelihood ratio values
sufficient statistic representation of likelihood ratio test
unbiasedness of most powerful test of $H_0$: $\Theta = \Theta_0$ versus $H_a$: $\Theta = \Theta_a$
UMP test of $H_0$: $\Theta = \Theta_0$ versus $H_a$: $\Theta \in \Omega_a$, Neyman–Pearson approach
one-sided and two-sided alternative hypotheses
monotone likelihood ratio (MLR) in the statistic $t(\mathbf{X})$
monotone likelihood ratio and the exponential class of densities
monotone likelihood ratios and UMP level-$\alpha$ tests
UMP level-$\alpha$ test of $H_0$: $\Theta = \Theta_0$ versus $H_a$: $\Theta \neq \Theta_0$ nonexistence in case of MLR
UMPU level-$\alpha$ test of $H_0$: $\Theta = \Theta_0$ versus $H_a$: $\Theta \neq \Theta_0$ in case of exponential class density
UMPU level-$\alpha$ test of $H_0$: $\Theta_1 \leq \Theta \leq \Theta_2$ versus $H_a$: $\Theta < \Theta_1$ or $\Theta > \Theta_2$ in case of exponential class density
nuisance parameters
conditioning
multiparameter exponential class and hypothesis testing
UMPU level-$\alpha$ test of $H_0$: $\sigma^2 \leq \sigma_0^2$ versus $H_a$: $\sigma^2 > \sigma_0^2$ when sampling from normal distribution
reparameterizing exponential class densities via linear transformation
natural parameterization of the exponential class
noncentral $t$-distribution
noncentrality parameter
UMPU level-$\alpha$ test of $H_0$: $\mu \leq \mu_0$ versus $H_a$: $\mu > \mu_0$, and of $H_0$: $\mu = \mu_0$ versus $H_a$: $\mu \neq \mu_0$, normal population

## Problems

**1.** A shipment of 20 projection screen television sets is at the receiving dock of a large department store. The department store has a policy of not accepting shipments that contain more than 10 percent defective merchandise. The receiving clerk is instructed to have a quality inspection done on two sets that are randomly drawn, without replacement, from the shipment. Letting $k$ represent the unknown number of defective sets in the shipment of 20 sets, the null hypothesis $H_0: k \leq 2$ will be rejected iff both sets that are inspected are found to be defective.

a. Calculate the probabilities of committing type I errors when $k = 0$, 1, or 2.

b. Calculate the probabilities of committing type II errors when $k \geq 3$.

c. Plot the power function of this testing procedure. Interpret the implications of the power function from the standpoint of both the department store and the television manufacturer.

**2.** A pharmaceutical company is analyzing the effectiveness of a new drug that it claims can stimulate hair growth in balding men. For the purposes of an advertising campaign, the marketing department would like to be able to claim that the drug will be effective for at least 50 percent of the balding men who use it. To test the claim, a random sample of 25 balding men are given the drug treatment, and it is found that 10 applications were effective in stimulating hair growth. The population of balding men is sufficiently large that you may treat this as a problem of random sampling *with* replacement.

a. Test the null hypothesis $H_0: p \geq .50$ using as close to a .10-size test as you can.

b. Plot the power function for this test. Interpret the power function from the standpoint of both the pharmaceutical company and the consuming public.

c. Is the test you used in (a) a UMP test? Is it a UMPU test? Is it an unbiased test? Is it a consistent test?

d. Calculate and interpret the $p$ value for the test.

**3.** In each case below, identify whether the null and the alternative hypotheses are simple or composite hypotheses.

a. You are random sampling from a Gamma population distribution, and you are testing $H_0: \alpha \leq 2$ versus $H_a: \alpha > 2$.

b. You are random sampling from a geometric population distribution, and you are testing $H_0: p = .01$ versus $H_a: p > .01$.

c. The joint density of the random sample $Y = x\beta + \varepsilon$ is $N(x\beta, \sigma^2 I)$, and you are testing whether $\beta = [\mathbf{0}]$.

d. You are random sampling from a Poisson population distribution, and you are testing $H_0: \lambda = 2$ versus $H_a: \lambda = 3$.

**4.** A large metropolitan branch of a savings and loan is examining staffing issues and wants to test the hypothesis that the expected number of customers requiring the services of bank personnel during the midweek (Tuesday–Thursday) noon hour $\leq 50$. The bank has obtained the outcome of a random sample consisting of 100 observations on the number of noon hour customers requiring service from bank personnel. It was observed that $\bar{x} = 54$. You may assume that the population distribution is Poisson in this case.

a. Design a UMP level-.05 test of the null hypothesis having size as close to .05 as possible. Test the hypothesis.

b. Plot the power function for the test. Interpret the power function both from the standpoint of management's desire for staff reductions and the need to provide quality customer service to bank customers.

**5.** The annual proportion of new restaurants that survive in business for at least one year in a U.S. city with population $\geq 500,000$ people is assumed to be the outcome of some Beta population distribution. Part of the maintained hypothesis is that $b = 1$ in the Beta distribution, so that the population distribution is assumed to be Beta$(a, 1)$. A random sample of size 50 from the beta population distribution results in the geometric mean $\bar{x}_g = .84$.

a. Define a UMP level-.05 test of the hypothesis that less than three-quarters of new restaurants are expected to survive at least one year in business in U.S. cities of size $\geq 500,000$. Test the hypothesis.

b. Plot the power function for the test. Interpret the power function both from the standpoint of a potential investor in a restaurant and from the perspective of the managing director of a chamber of commerce.

c. Calculate and interpret the $p$ value for the test.

**6.** A complaint has been lodged against a major domestic manufacturer of potato chips stating that its 16-oz. bags of chips are being underfilled. The manufacturer claims that its filling process produces fill weights that are normally distributed with a mean of 16.1 ounces and a standard deviation of $\leq .05$, so that over 97 percent of its product has a weight of $\geq 16$ oz. It suggests that its product be randomly sampled and its claims be tested for accuracy. Two *independent* random samples of observations on fill weights, each of size 250, resulted in the following summary statistics: $\bar{x}_1 = 16.05$, $\bar{x}_2 = 16.11$, $s_1^2 = .0016$, and $s_2^2 = .0036$.

a. Define a UMPU level-.05 test of $H_0: \mu = 16.1$ versus $H_a: \mu \neq 16.1$ based on a random sample of size 250. Test the hypothesis using the statistics associated with the first random-sample outcome. Plot and interpret the power function of this test.

b. Define a UMPU level-.05 test of $H_0: \sigma \leq .05$ versus $H_a: \sigma > .05$ based on a random sample of size 250. Test the hypothesis using the statistics associated with the second random-sample outcome. Plot and interpret the power function of this test.

c. Calculate and interpret the $p$ values of the tests in (a) and (b).

d. Treating the hypotheses in (a) and (b) as a joint hypothesis on the parameter vector of the normal population distribution, what is the probability of type I error for the joint hypothesis $H_0: \mu = 16.1$ and $\sigma \leq .05$ when using the outcome of the two test statistics above to determine acceptance or rejection of the joint null hypothesis? Does the complaint against the company appear to be valid?

*e. Repeat (a)–(c) using a pooled sample of 500 observations. (Hint: In (c), it might be useful to consider Bonferroni's inequality for placing an upper bound on the probability of type I error.)

**7.** In a random sample of size 10 from a Bernoulli population distribution, how many (nonrandomized) critical regions can you define that have size $\leq .10$ and that are also unbiased for testing the null hypothesis $H_0: p = .4$ versus $H_a: p \neq .4$?

**8. Randomized Test** It was demonstrated in Ex. 9.10 that the choices of size for most powerful tests of the hypothesis $H_0: p = .2$ versus $H_a: p = .8$ were quite limited. Suppose that a .05 level test of the null hypothesis was desired and that you were willing to utilize a *randomized* test. In particular, examine the following randomized test rule:

$x \geq 8 \Rightarrow$ reject $H_0$,

$x = 7 \Rightarrow$ reject $H_0$ with probability $\tau$,

accept $H_0$ with probability $(1 - \tau)$,

$x \leq 6 \Rightarrow$ accept $H_0$.

To implement the rule when $x = 7$ occurs, a uniform random number $z$ with range $(0,1)$ could be drawn, and if $z \leq \tau$, $H_0$ would be rejected, and if $z > \tau$, $H_0$ would be accepted.

a. Find a value of $\tau$ that defines a size-.05 test of the null hypothesis.

b. Is it possible that two analysts, using exactly the same random sample outcome and exactly the same test rule could come to different conclusions regarding the validity of the null hypothesis? Explain. (This feature of randomized tests has discouraged their use.)

c. Is the test you defined in (a) an unbiased size .05 test of the null hypothesis?

*d. Is the test you defined in (a) a most powerful size .05 test of the null hypothesis?

**9.** The number of work-related injuries per week that occur at the manufacturing plant of the Excelsior Corporation is a Poisson-distributed random variable with mean $\lambda \geq 3$, according to company analysts. In an attempt to lower insurance costs, the corporation institutes a program of intensive safety instruction for all employees. Upon completion of the program, a 12-week period produced an average of 2 accidents per week.

a. Design a uniformly most powerful level-.10 test of the null hypothesis $H_0: \lambda \geq 3$ versus the alternative hypothesis $H_a: \lambda < 3$ having size as close as possible to, without exceeding, .10.

b. Test the null hypothesis using the test you defined in part (a). What can you say about the effectiveness of the safety program?

**10.** Being both quality and cost conscious, a major foreign manufacturer of compact disc players is contemplating its warranty policy. The standard warranty for compact disc players sold by competing producers is one year. The manufacturer is considering a two-year warranty. The operating life until failure of the manufacturer's compact disc player has the density

$f(x; \beta) = (x/\beta^2)\exp(-x/\beta)I_{(0,\infty)}(x)$

for some value of $\beta > 0$, where $x$ is measured in years.

a. The manufacturer wants its exposure to warranty claims to be, on the average, no more than 5 percent of the units sold. Find the values of $\beta$ for which $P(x \leq 2; \beta) = \int_0^2 f(x; \beta)dx \leq .05$.

b. Based on a random sample of size 50, design a uniformly most powerful level-.10 test of the null hypothesis that $\beta$ will be in the set of values you identified in (a).

c. A nondestructive test of the disc players that determines their operating life until failure is applied to 50 players that are randomly chosen from the assembly line. The measurements resulted in $\bar{x} = 4.27$. Test the null hypothesis in (b).

d. Plot the power curve of this test, and interpret its meaning to the management of the manufacturing firm.

**11.** Referring to Def. 9.14, state the form of the UMPU level-$\alpha$ test of the null hypothesis $H_0: \sigma^2 \geq \sigma_0^2$ versus $H_a: \sigma^2 < \sigma_0^2$. Justify your answer.

**12.** Referring to Def. 9.15, state the form of the UMPU level-$\alpha$ test of the null hypothesis $H_0: \mu \geq \mu_0$ versus $H_0: \mu < \mu_0$. Justify your answer.

**13.** Your company supplies an electronic component that is critical to the navigational systems of large jet aircraft. The operating life of the component has an exponential distribution with some mean value $\nu$, where operating life is measured in 100,000-hour units. You are seeking a contract to supply these components to a major aircraft manufacturer on the West Coast. The contract calls for a minimum mean operating life of 750,000 hours for the component, and you must provide evidence on the reliability of your product. You have a random sample of observations on tests of 300 of your components that provide measurements on the components' operating lives. According to the tests performed, the mean operating life of the components was 783,824 hours.

a. In designing a UMP level-$\alpha$ test in this situation, should the null hypothesis be defined as $\theta \geq 7.5$ or $\theta \leq 7.5$? Base your discussion on the characteristics of the power function of each of the tests.

b. Design a UMP size $\alpha$ test of whichever null hypothesis you feel is appropriate based on your discussion in (a). Choose whatever size test you feel is appropriate, and discuss your choice of size.

c. Conduct the hypothesis test. Should your company get the contract? Why or why not?

***14.** The *Pareto* distribution

$f(\mathbf{x}; \theta, c) = c^\theta \theta x^{-(1+\theta)} I_{(c,\infty)}(\mathbf{x})$,

for $\theta > 0$ and $c > 0$, has been used to model the distribution of incomes in a given population of individuals, where $c$ represents the minimum level of income in the population. In a certain large state on the East Coast, the office of fiscal management is investigating a claim that professors at four-year universities have average annual salaries that exceed the state average annual salary for middle management, white collar workers, which is known to be $62,471. A random sample of 250 professors' salaries was obtained, and the *geometric* mean of the observations was found to be $(\prod_{i=1}^n x_i)^{1/n} = 61.147$, where the $x_i$'s are measured in thousands of dollars. As part of the maintained hypothesis, the value of $c$ is taken to be 30.

a. Express the mean level of income as a function of the parameter $\theta$.

b. Define a test statistic on which you can base a UMP level-$\alpha$ test of the null hypothesis $H_0: \mu \leq 62.471$ versus $H_a: \mu > 62.471$.

c. Define a UMP size .05 test of the null hypothesis $H_0: \mu \leq 62.471$ versus $H_a: \mu > 62.471$. (Hint: Use a test statistic for which you can apply an asymptotic normal distribution, and use the normal approximation.)

d. Test the hypothesis. Are university professors paid more than white collar, middle management workers in this state?

**15.** Suppose that a random sample of size $n$ is drawn from a normal population distribution for which $\sigma^2$ is assumed to be *known* and equal to the given value $\sigma_*^2$. Define UMPU level-$\alpha$ tests of the following null hypotheses:

a. $H_0: \mu \leq \mu_0$ versus $H_a: \mu > \mu_0$

b. $H_0: \mu \geq \mu_0$ versus $H_a: \mu < \mu_0$

c. $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$

**16.** Suppose that a random sample of size $n$ is drawn from a normal population distribution for which $\mu$ is assumed to be *known* and equal to the given value $\mu_*$. Define UMPU level-$\alpha$ tests of the following null hypotheses:

a. $H_0: \sigma^2 \leq \sigma_0^2$ versus $H_a: \sigma^2 > \sigma_0^2$

b. $H_0: \sigma^2 \geq \sigma_0^2$ versus $H_a: \sigma^2 < \sigma_0^2$

c. $H_0: \sigma^2 = \sigma_0^2$ versus $H_a: \sigma^2 \neq \sigma_0^2$

**17. Control Charting** A large mail-order house has initiated a quality-control program. It randomly samples 100 of each day's orders and monitors whether or not the mail order-taking process is "under control" in the sense that errors in order taking are at minimum levels. The number of orders per day is sufficiently large that one can assume that the sampling is done with replacement. The daily error proportion prior to the initiation of this program has been 3.2 percent.

a. Define a UMPU level-$\alpha$ test of the hypothesis $H_0: p = .032$ versus $H_a: p \neq .032$. You may use the asymptotic normal distribution of the test statistic in defining the critical region.

b. Conduct a size .05 UMPU test of the null hypothesis on a day where $\bar{x} = 3.4$. Is the order process under control?

c. After quality-control training and closer monitoring of clerical workers, a daily random sample resulted in $\bar{x} = 2.6$. Is there evidence that quality has increased over what it has been in the past? Why or why not?

(The mail-order company of Alden's Inc. was one of the earliest companies to use control charting techniques for monitoring clerical work. See John Neter (1952), "Some applications of statistics for auditing." *J. American Statistical Assoc.*, March, pp. 6–24.)

**18.** Revisit Ex. 9.27 and the power function graph in Figure 9.16 and consider the implications of the power function graph in the two-dimensional parameter space $(\mu, \sigma)$.

a. Plot the power surface in three dimensions, with the axes referring to power, the value of $\mu$, and the value of $\sigma$. (This is probably best done with the aid of a computer!)

b. Plot the *isopower contour* in the $(\mu, \sigma)$-plane for a power level of .90. (An *isopower contour* is the set of $(\mu, \sigma)$ points that result in the same level of power, which in the case at hand is equivalent to the set of $(\mu, \sigma)$ points that result in the same value of the noncentrality parameter $\lambda$.) Interpret the isopower contour with respect to the ability of the test to detect deviations from the null hypothesis.

**19.** The Gibralter Insurance Co. is reevaluating the premiums it charges on car insurance and is analyzing classifications of cars into high risk, average risk, and low risk on the basis of frequency of accidents. It is currently examining an imported mid-size four-door sedan and wishes to examine whether its frequency of claims history is significantly different than a range of values considered to be consistent with the expected frequency of claims of the average risk class, the range being between 2 percent and 4 percent of the vehicles insured. A random sample with replacement of 400 insured vehicles of the type in question resulted in a claims percentage of 7 percent.

a. Design a UMPU level-$\alpha$ test of the null hypothesis $H_0: \mu \in [2,4]$ versus $H_a: \mu \notin [2,4]$. You can base your test on the asymptotic normal distribution of the test statistic.

b. Test the null hypothesis using a size .05 UMPU test. Does the outcome contradict classifying the vehicle in the average risk class? Why or why not?

c. Supposing you rejected the hypothesis, what would be your conclusion? Is further statistical analysis warranted?

***20.** A certain business uses national telephone solicitation to sell its product. Its sales staff have individual weekly sales quotas of 10 sales that they must meet or else their job performance is considered to be unsatisfactory and they receive only base pay and no sales commission. In hiring sales staff, the company has claimed that the proportion of customers solicited that will ultimately buy the company's product is .05, so that on average, 200 phone calls per week should produce the required 10 sales. The company requires that a salesperson keep a record of how many phone solicitations were made, and when the tenth sale is made, the salesperson must indicate the number of phone calls that were made to obtain the 10 sales. The data on the last 200 weekly quotas that were met by various salespersons indicated that 289 phone calls were needed, on average, to meet the quota. A disgruntled employee claims that the company has overstated the market for the product, and she wants the quota lowered.

a. Define a UMP level-$\alpha$ test of the hypothesis $H_0: p = .05$ versus $H_a: p < .05$. You may use an asymptotic normal distribution for the test statistic, if it has one.

b. Test the hypothesis with a UMP size .10 test. Does the disgruntled employee have a legitimate concern?

c. Examine the asymptotic power function of the test (i.e., construct a power function based on the asymptotic normal distribution of the test statistic). Interpret the implications of the power function for the test you performed, both from the perspective of the company and from the perspective of the employee. If you were primarily interested in workers' rights, might you design the test differently and/or would you consider testing a different null hypothesis? Explain.

d. Suppose there was a substantial difference in the abilities of salespersons to persuade consumers to purchase the company's product. Would this have an impact on your statistical analysis above? Explain.

# 10 Hypothesis-Testing Methods

## 10.1 Introduction

In this chapter we examine general methods that can be used to define explicit rules for testing statistical hypotheses. In particular, the likelihood ratio, Wald, and Lagrange multiplier methods for constructing statistical tests are widely used in empirical work, and they provide well-defined procedures for defining test statistics and critical regions in given hypothesis-testing contexts. In addition, it is possible to find useful test statistics based entirely on a heuristic principle of test construction. None of these four methods is guaranteed to produce a statistical test with optimal properties in all cases. In fact, no method of defining statistical tests can provide such a guarantee. The virtues of these methods are that they are relatively straightforward to apply (in comparison to direct implementation of many of the theorems in Section 9.5), they are applicable to a wide class of problems that are relevant in applications, they generally have excellent asymptotic properties, they often have good power in finite samples, they are sometimes unbiased and/or UMP, and they have intuitive appeal.

These four methods of defining test rules, together with the results on UMP and UMPU testing presented in Section 9.5, by no means exhaust the ways in which a statistical test can be defined. (See E. Lehmann, *Testing Statistical Hypotheses*, for further reading.) However, the methods we present cover a large portion of the approaches to defining test rules that are used in practice. Ultimately, regardless of the genesis of a test rule, whether a statistical test is useful in an application depends on its properties. We will often refer to the concepts presented in Chapter 9 to assess the usefulness of the tests that we examine in this chapter.

This chapter will also provide a brief introduction to nonparametric tests that have been designed to assess assumptions relating to the functional form of the joint density of a random sample as well as to assess the ubiquitous iid assumption. Finally, we will introduce the concepts of confidence intervals and confidence regions as they relate to statistical tests of statistical hypotheses.

## 10.2 Heuristic Approach

In the heuristic approach to defining statistical tests, one attempts to implement the following **heuristic principle of test construction**: "Discover a test statistic whose probabilistic behavior is different under $H_0$ and $H_a$, and exploit the difference in defining a critical region for a statistical test." For the choice of an appropriate test statistic $T = t(\mathbf{X})$, one might examine a good estimator of $\Theta$ or $\mathbf{q}(\Theta)$, such as a maximum likelihood, least-squares, or method-of-moments estimator. Alternatively, a (minimal and/or complete) sufficient statistic for $\Theta$ or $\mathbf{q}(\Theta)$ might be useful in defining a test statistic.

If the range of $T$ over $H_0 \cup H_a$ can be partitioned as $C_r^T \cup C_a^T$ so that $P(t \in C_r^T; \Theta) \leq \alpha \; \forall \; \Theta \in H_0$, a level-$\alpha$ test of $H_0$ will have been defined. If it is also true that $P(t \in C_r^T; \Theta) \geq \alpha \; \forall \; \Theta \in H_a$, the test will be unbiased. If the power function $\pi_{C_r^T}(\Theta) = P(t \in C_r^T; \Theta)$ is acceptable to the analyst (and the reviewers of her analysis), the test represents at least a useful, if not optimal, statistical test of $H_0$ versus $H_a$. The heuristic approach can be a substantially less complicated method of defining a test statistic than the UMP and UMPU approaches, which we discussed in Section 9.5. The following examples illustrate the heuristic method.

**Example 10.1** At the Union Bank a debate is raging over the expected rate at which customers arrive at the teller windows. If the expected rate $\leq 2$ per minute, management feels that staffing can be reduced. A random sample of 100 observations from the (assumed) Poisson $(\lambda)$ population of customer arrivals/minute was obtained, and the outcome of the complete sufficient statistic, and MLE estimator of $\lambda$, $\bar{X}$, was 2.48. Test $H_0: \lambda \leq 2$ versus $H_a: \lambda > 2$, where

$$f(\mathbf{x}; \lambda) = \left[ e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i} \Big/ \prod_{i=1}^n x_i! \right] \prod_{i=1}^n I_{\{0,1,2,\ldots\}}(x_i),$$

with $n = 100$.

**Answer**: Note that the statistic $n\bar{X} = \sum_{i=1}^n X_i \sim \text{Poisson}(n\lambda)$ (straightforwardly shown via the MGF approach), and thus the event $100\bar{x} \geq c$ will be more probable for higher values of $\lambda$ than for lower values. Thus heuristically, it seems reasonable to reject $H_0: \lambda \leq 2$ for large values of $100\bar{x}$ and not to reject for small values. Furthermore, the probability of rejection, $P(100\bar{x} \geq c; 100\lambda)$ *increases* as $\lambda$ increases, and so to identify a value of $c$ that defines a level-$\alpha$ test of $H_0$, it suffices to choose $c$ so that $P(100\bar{x} \geq c; 200) = \alpha$, since then $P(100\bar{x} \geq c; 100\lambda) < \alpha \; \forall \; \lambda < 2$. The preceding observation also implies that a critical region defined by $100\bar{x} \geq c$ defines an *unbiased* test.

With the aid of the computer, we identify the following potential test sizes:

| $c$ | $P(100\bar{x} \geq c; 200) = 1 - \sum_{j=0}^{c-1} e^{-200}200^j/j!$ |
|---|---|
| 219 | .097 |
| 220 | .086 |
| 221 | .075 |
| 222 | .066 |
| 223 | .058 |
| 224 | .050 |
| 225 | .044 |

Assuming the labor union and bank management agree that a level-.05 test is acceptable (should they agree on this level?), the critical region expressed in terms of the test statistic $\bar{X}$ is $C_r^{\bar{X}} = \{\bar{x}: \bar{x} \geq 2.24\}$. Since the outcome $\bar{x} = 2.48 \in C_r^{\bar{X}}$, $H_0$ is *rejected* at the .05 level. Using the results of Section 9.5, it can be shown that $C_r^{\bar{X}}$ defines the UMP level-.05 test of $H_0: \lambda \leq 2$ versus $H_a: \lambda > 2$. Note that the $p$ value for this test is $P(\bar{x} \geq 2.48; 200) = P(n\bar{x} \geq 248; 200) < .001$, which indicates there is strong evidence against $H_0$. □

The next example deals with a complicated case involving nuisance parameters, and yet the heuristic principle of test construction leads rather straightforwardly to a test with accepted power characteristics. In fact, the test is a UMPU level-$\alpha$ test.

**Example 10.2** Revisit Ex. 8.3 relating to the estimation of a linear relationship between the number of new homes sold and hypothesized determinants of home sales. Assume that the disturbance term of the linear model is approximately normally distributed (why *must* this be only an approximation?), so that (approximately) $\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{x}'\mathbf{x})^{-1})$. The realtor wishes to test whether mortgage interest rates actually impact home sales, and so she specifies the hypothesis $H_0: \beta_2 = 0$ versus $H_a: \beta_2 \neq 0$, where $\beta_2$ is the GLM parameter associated with the mortgage interest rate variable. Test the hypothesis at the .05 level, and also calculate the $p$ value of the test.

**Answer**: Under the prevailing assumptions, $\hat{\beta}_2 \sim N(\beta_2, \text{var}(\hat{\beta}_2))$, where $\text{var}(\hat{\beta}_2)$ is the (2,2) entry in the covariance matrix $\sigma^2(\mathbf{x}'\mathbf{x})^{-1}$. Consider the statistic $T = \hat{\beta}_2/S_{\hat{\beta}_2}$ for defining a statistical test of the hypothesis, where $S^2_{\hat{\beta}_2}$ is the (2,2) entry of the covariance matrix estimator $\hat{S}^2(\mathbf{x}'\mathbf{x})^{-1}$. We know from Section 8.2 that $\hat{\beta}_2$ and $\hat{S}^2$ are independent random variables, and in the case at hand $(\hat{\beta}_2 -$

$\beta_2)/[\mathrm{var}(\hat{\beta}_2)]^{1/2} \sim N(0,1)$ and $(n-k)\hat{S}^2/\sigma^2 \sim \chi^2_{n-k}$. It follows that

$$\frac{\left(\hat{\beta}_2-\beta_2\right)/\left[\mathrm{var}(\hat{\beta}_2)\right]^{1/2}}{\left[\hat{S}^2/\sigma^2\right]^{1/2}} = \frac{\hat{\beta}_2-\beta_2}{S_{\hat{\beta}_2}}$$

has a (central) $t$-distribution with $n-k$ degrees of freedom. Then under $H_0: \beta_2 = 0$, $T = \hat{\beta}_2/S_{\hat{\beta}_2}$ has a central $t$-distribution with $n-k$ degrees of freedom, whereas under $H_a$: $\beta_2 \neq 0$, $T$ has a noncentral $t$-distribution with $n-k$ degrees of freedom and noncentrality parameter $\lambda = \beta_2/[\mathrm{var}(\hat{\beta}_2)]^{1/2}$ (recall Section 9.6). We then know that $P(|t| \geq c; \lambda)$ increases monotonically as $|\lambda| \to \infty$, so that the event $|t| \geq c$ becomes ever more probable as $|\beta_2| \to \infty$. Thus heuristically, it seems reasonable to reject $H_0: \beta_2 = 0$ for large values of $|t|$ and not to reject for small values. Therefore, the critical region for the test statistic $T$ will be $C_r^T = [-\infty, -c] \cup [c, \infty)$. Given the aforementioned behavior of $P(|t| \geq c; \lambda)$ as a function of $\lambda$, it follows that $C_r^T$ will define an unbiased test of $H_0$.

Referring to the table of the (central) $t$-distribution, we find that $P(t \geq c) = .025$ when $c = 2.571$ and the degrees of freedom are $n-k = 10-5 = 5$. Then, by the symmetry of the (central) $t$-distribution, it follows that $P(t \in C_r^T) = .05$ where $C_r^T = (-\infty, -2.571] \cup [2.571, \infty)$. Since the outcome of $T$ is $t = -3.33 \in C_r^T$, we *reject* $H_0$ and conclude that mortgage interest rates do impact home sales. The $p$ value of the test is given by $P(|t| \geq 3.33) = .021$, which suggests that the evidence against $H_0$ is substantial.

The power function of the test can be calculated with the aid of a computer (we use the GAUSS procedure CDFTNC here). Selected values of the power function are displayed in the following table, and a graph of the power function as a function of $\lambda = \beta_2/[\mathrm{var}(\hat{\beta}_2)]^{1/2}$ is given in Figure 10.1.

| $\lambda = \beta_2/[\mathrm{var}(\hat{\beta}_2)]^{1/2}$ | $\pi(\lambda) = P(t \in C_r^T; \lambda)$ |
|---|---|
| 0 | .05 |
| .5, −.5 | .07 |
| 1, −1 | .13 |
| 1.5, −1.5 | .23 |
| 2, −2 | .37 |
| 3, −3 | .67 |
| 4, −4 | .89 |
| 5, −5 | .98 |

Note that $\lambda$ effectively measures deviations of $\beta_2$ from 0 in terms of standard deviation units (i.e., $\beta_2$ is relative to the standard deviation $[\mathrm{var}(\hat{\beta}_2)]^{1/2}$). The power function suggests that $\beta_2$ must be a number of standard deviations away from zero before the test has appreciable power in rejecting $H_0$. This is typical for power functions associated with tests of the significance of parameters estimated by least squares, and it suggests that to be able to detect small departures of $\beta_2$ from zero, $\beta_2$ must be estimated quite accurately (i.e., $\mathrm{var}(\hat{\beta}_2)$ must be small).

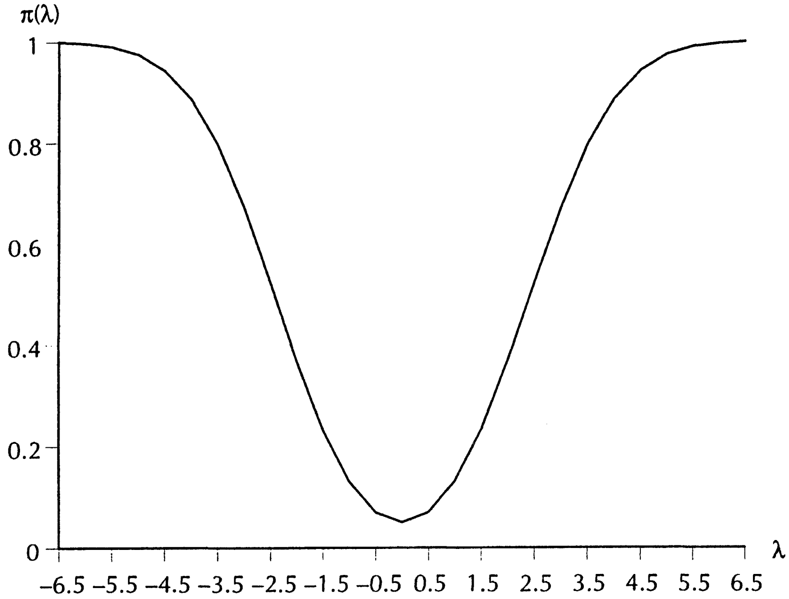


**Figure 10-1** Power function of UMPU size .05 test of $H_0: \beta_2 = 0$ versus $H_a: \beta_2 \neq 0$.

In this application, it might be argued, based on economic theory considerations, that $\beta_2$ must be nonpositively valued. If this is adopted as part of the maintained hypothesis, then a one-sided alternative hypothesis might be considered as $H_a: \beta_2 < 0$. The reader is invited to reexamine the problem of testing $H_0: \beta_2 = 0$ using the one-sided alternative hypothesis and a .05-level test. In such a case, the power of the test for $\beta_2 \in H_a$ is increased. □

In the next example we revisit Ex. 10.1 and illustrate how the asymptotic distribution of a test statistic can be used to define an asymptotic test of $H_0$. In the case at hand, this will allow the test to be conducted using standard normal tables and without the computer. It also allows one to circumvent the limited choice of test sizes in this discrete case, albeit in an approximate sense.

**Example 10.3** Revisit Ex. 10.1 and consider using the asymptotic normal distribution of the test statistic $\bar{X}$ to conduct the test of $H_0$: $\lambda \leq 2$ versus $H_a$: $\lambda > 2$. In this case, where we are random sampling from the Poisson population distribution, we know that $\bar{X} \overset{a}{\sim} N(\lambda, n^{-1}\lambda)$, and thus $Z = n^{1/2}(\bar{X} - \lambda)/\lambda^{1/2} \overset{a}{\sim} N(0,1)$. Using heuristic reasoning analogous to that used in Ex. 10.1, we should reject $H_0$ for large values of $\bar{x}$, and thus for large values of $z$. This suggests a critical region of the form $C_r^Z = [c, \infty)$ for outcomes of $Z$. Consulting the standard normal table under the assumption that a .05-level test is desired, we choose $c = 1.645$, so that $P(z \geq 1.645) = .05$ (as an asymptotic approximation).

The reader may have noticed a serious practical flaw in the use of $Z$ as a test statistic—namely, $\lambda$ is unknown and so $Z$ is *not* a statistic! However, this problem is overcome by assigning $\lambda$ a numerical value. Setting $\lambda = 2$ in the definition of $Z$ is the logical choice under the current circumstances. This follows because if $\bar{X} \sim N(\lambda, n^{-1}\lambda)$, then $Z = n^{1/2}(\bar{X} - 2)/2^{1/2} \sim N\left((n/2)^{1/2}(\lambda - 2), \lambda/2\right)$, and thus (asymptotically)

$$\pi_{C_r}(\lambda) = P(z \geq 1.645) \begin{bmatrix} \leq \\ > \end{bmatrix} .05, \quad \forall\, \lambda \begin{bmatrix} \leq 2 \\ > 2 \end{bmatrix},$$

so that $C_r^Z$ then defines a .05-level unbiased test of $H_0: \lambda \leq 2$ versus $H_a: \lambda > 2$. Restating the test in terms of a critical region for the outcome of $\bar{X}$ itself, and recalling that $n = 100$ in this case, we have that $z \geq 1.645$ iff $\bar{x} \geq 2.23$, so that $C_r^{\bar{X}} = [2.23, \infty)$, which is virtually the same critical region as in Ex. 10.1 based on the Poisson population distribution. The near equivalence of the two critical regions is due to the accuracy of the asymptotic normal distribution of $\bar{X}$. Since $\bar{x} = 2.48 \in C_r^{\bar{X}}$, we reject $H_0$, as before. An asymptotic $p$ value for the test can be calculated based on the observed value of $z = 10(2.48 - 2)/2^{1/2} = 3.394$, so that $p$ value $= \int_{3.394}^{\infty} N(z; 0, 1)dz < .001$, providing strong evidence for the rejection of $H_0$. Also, under the assumption of normality for $\bar{X}$, it could be argued, using the results of Section 9.5, that the preceding test is a UMP level-.05 test. In this case, we then state that the test is *asymptotically* UMP level .05. The reader might consider plotting the power function of the test. □

In each of the preceding examples, a specific random-sample size is involved. One might also consider whether sequences of level-.05 unbiased tests defined analogously to the preceding tests have the property of consistency as sample sizes increase without bound. Using asymptotic theory applied to the respective test statistics and the definitions of critical regions, one can argue that the respective test sequences are indeed consistent.

**Example 10.4** Revisit Ex. 10.1 to Ex. 10.3, and consider whether the respective sequences of .05-level unbiased tests are *consistent*. Regarding Ex. 10.1, the $n$th element in the appropriate sequence of critical regions will be of the form $C_{rn}^{\bar{X}} = [c_n, \infty)$ such that $P(\bar{x}_n \geq c_n) \leq .05 \;\forall\, n$, where $\lambda = 2$ is assumed when defining the probability of the critical region for $\bar{X}_n$, and $c_n$ is chosen as small as possible in order to maximize power for $\lambda \in H_a$. The preceding sequence of probability inequalities can be written alternatively as

$$P(n^{1/2}(\bar{x}_n - \lambda)/\lambda^{1/2} \geq n^{1/2}(c_n - \lambda)/\lambda^{1/2}) \leq .05, \quad \forall\, n.$$

Also, since $Z_n = n^{1/2}(\bar{X}_n - \lambda)/\lambda^{1/2} \xrightarrow{d} N(0, 1)$ by the LLCLT, we can infer that $n^{1/2}(c_n - \lambda)/\lambda^{1/2} \rightarrow 1.645$, and thus $c_n \rightarrow \lambda$. Since $\lambda = 2$ is used when defining the sequence of critical regions, then $c_n \rightarrow 2$. By the consistency of $\bar{X}_n$ for $\lambda$, we also know that $\bar{X}_n \xrightarrow{p} \lambda$. If $\lambda > 2$, so that $H_a$ is true, then $\bar{X}_n - c_n \xrightarrow{p} \lambda - 2 > 0$, so that the probability that $H_0$ is rejected when $H_a$ is true converges to 1. Thus the sequence of tests is consistent, and a similar argument can be applied to demonstrate consistency in the case of Ex. 10.3.

Regarding Ex. 10.2, note that the critical region for the $T$-statistic will ultimately converge to $C_{rn}^T \rightarrow C_r^T = (-\infty, -1.96] \cup [1.96, \infty)$ since the (central) $T_n$ random variable is such that $T_n \xrightarrow{d} Z \sim N(0, 1)$ and $P(z \in C_r^T) = .05$. Assuming $\sigma^2(\mathbf{x}'\mathbf{x})^{-1} \to [0]$, so that $\hat{\beta} \xrightarrow{p} \beta$, and assuming that $\hat{S}^2 \xrightarrow{p} \sigma^2$ (e.g., assume that the error terms in the linear model are iid), it follows that the noncentrality parameter of the $t$-distribution associated with $T_n = \hat{\beta}_2/S_{\hat{\beta}_2}$, given by $\lambda = \beta_2/[\text{var}(\hat{\beta}_2)]^{1/2}$, diverges to infinity as $n \to \infty$ whenever $\beta_2 \neq 0$. Then $P(t_n \in C^T_{rn}; \lambda) \to P(t_n \in C^T_r; \lambda) \to 1$, as $n \to \infty$, when $H_a$ is true, and thus the test is consistent. □

## 10.3 Generalized Likelihood Ratio Tests

As its name implies, a generalized likelihood ratio (GLR) test of a statistical hypothesis is based on a test rule that is defined in terms of a ratio of likelihood function values. The adjective *generalized* is used here to distinguish the GLR test from the simple ratio of two likelihood function values presented in our discussion of the Neyman–Pearson lemma (Theorem 9.1). In the latter context, the likelihood function was evaluated at two distinct values of the parameter vector, $\Theta$. In the current context, the likelihood ratio is "generalized" by forming the ratio of likelihood function suprema (or maxima if maximums exist), where the two suprema are taken with respect to two different feasible sets—the set of null hypothesis values for $\Theta$, and the set of values of $\Theta$ represented by the union of the null and alternative hypotheses. The GLR test is a natural procedure to use for testing hypotheses about $\Theta$ or functions of $\Theta$ when maximum likelihood estimation is being used in a statistical analysis.

**Definition 10.1**
***Generalized likelihood ratio (GLR) test of size $\alpha$***

Let the random sample $(X_1, \ldots, X_n)$ have the joint probability density function $f(x_1, \ldots, x_n; \Theta)$ and associated likelihood function $L(\Theta; x_1, \ldots, x_n)$. The **generalized likelihood ratio (GLR)** is defined as

$$\lambda(x_1, \ldots, x_n) = \frac{\sup_{\Theta \in H_0} L(\Theta; x_1, \ldots, x_n)}{\sup_{\Theta \in H_0 \cup H_a} L(\Theta; x_1, \ldots, x_n)},$$

and a generalized likelihood ratio test for testing $H_0$ versus $H_a$ is given by the following test rule:

reject $H_0$ iff $\lambda(x_1, \ldots, x_n) \leq c$,

or equivalently,

reject $H_0$ iff $I_{[0,c]}(\lambda) = 1$.

For a size $\alpha$ test, the constant $c$ is chosen to satisfy

$$\sup_{\Theta \in H_0} \pi(\Theta) = \sup_{\Theta \in H_0} P(\lambda(x_1, \ldots, x_n) \leq c; \Theta) = \alpha.$$

In order to provide some intuitive rationale for the GLR test, first note that the numerator of the GLR is essentially the largest likelihood value that can be associated with the sample outcome $(x_1, \ldots, x_n)$ when we are able to choose only among probability distributions that are contained in the null hypothesis, $H_0$. The numerator can then be interpreted as the likelihood function evaluated at the *constrained* maximum likelihood estimate of $\Theta$, the constraint being $\Theta \in H_0$. The denominator of the GLR is the largest likelihood value that can be associated with $(x_1, \ldots, x_n)$ when we can choose any probability distribution contained in $H_0$ and/or $H_a$. In most applications where $H_0 \cup H_a$ is the entire parameter space, $\Omega$, the denominator is then the likelihood function evaluated at the maximum likelihood estimate of $\Theta$. Since likelihood functions are nonnegatively valued, and since the feasible space for the numerator supremum problem is contained in the feasible space for the denominator supremum problem, we can infer that $\lambda \in [0, 1]$.

To see why the critical region of the test statistic $\lambda(x_1, \ldots, x_n)$ is defined in terms of the *lower tail*, $[0, c]$, of the range of $\lambda$, note that the smaller the value of $\lambda$, the larger the maximum likelihood for values of $\Theta \in H_0 \cup H_a$ is relative to the maximum likelihood for $\Theta \in H_0$. Intuitively, this means that when $\lambda$ is small, there is a value of $\Theta \in H_a$ that is notably "more likely" to have characterized the true density $f(\mathbf{x}; \Theta)$ associated with the outcome $x$ than any other value of the parameter $\Theta \in H_0$. When $f(\mathbf{x}; \Theta)$ is a discrete density function, we could also infer that there is a value of $\Theta \in H_a$ that implies a notably higher *probability* of observing $\mathbf{x}$ than does any value of $\Theta \in H_0$. Thus, for small values of $\lambda$, say $c$ or less, it appears *reasonable* to reject $H_0$ as containing the true probability distribution of $\mathbf{x}$ and to conclude instead that a better representation of the true probability distribution of $\mathbf{x}$ resides in the set $H_a$.

### Test Properties: Finite Sample

Intuition aside, whether the GLR test represents a good statistical test for a given statistical hypothesis ultimately depends on the properties of the test in the given application. There is no guarantee that a GLR test is UMP, or even unbiased, in finite samples. Finite sample properties of the GLR test must be established on a case-by-case basis and will depend on both the characteristics of $f(\mathbf{x}; \Theta)$ and the definition of the sets underlying $H_0$ and $H_a$. Nonetheless, before we proceed to asymptotic properties for which general results do exist, we point out some parallels with the results presented in Chapter 9.

*Simple Hypotheses* In the case where both $H_0$ and $H_a$ are simple hypotheses, the size $\alpha$ GLR test and the most powerful level-$\alpha$ test based on the Neyman–Pearson lemma (Theorem 9.1) will be equivalent.

**Theorem 10.1** (Equivalence of GLR Test and Neyman–Pearson Most Powerful Test When $H_0$ and $H_a$ Are Simple Hypotheses) *Suppose a size $\alpha$ GLR test of $H_0: \Theta = \Theta_0$ versus $H_a: \Theta = \Theta_a$ exists with critical region*

$$C_r^{\mathrm{GLR}} = \{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}, \quad \textit{where } P(\mathbf{x} \in C_r^{\mathrm{GLR}}; \Theta_0) = \alpha \in (0, 1).$$

*Furthermore, suppose a Neyman–Pearson most powerful level-$\alpha$ test also exists with critical region*

$$C_r = \{\mathbf{x}\colon\ L(\Theta_0;\mathbf{x}) \le kL(\Theta_a;\mathbf{x})\},\quad \textit{where } P(\mathbf{x}\in C_r;\Theta_0)=\alpha.$$

*Then the GLR test and the Neyman–Pearson most powerful test are equivalent.*

**Proof** Since $\lambda(\mathbf{x}) \in [0,1]$, and given that $\alpha \in (0,1)$, it follows that $P(\lambda(\mathbf{x}) \le c;\Theta_0) = \alpha$ only if $c < 1$. Now let $\hat{\theta} = \hat{\Theta}(\mathbf{x}) = \arg\max_{\Theta\in\{\Theta_0,\Theta_a\}}[L(\Theta;\mathbf{x})]$,[1] so that the GLR can be represented as $\lambda(\mathbf{x}) = L(\Theta_0;\mathbf{x})/L(\hat{\theta};\mathbf{x})$. Note the following relationship relating to a dichotomous partition of the range of $\mathbf{X}$:

$$\mathbf{x}\in A = \{\mathbf{x}\colon\ \hat{\Theta}(\mathbf{x}) = \Theta_a\} \Rightarrow \frac{L(\Theta_0;\mathbf{x})}{L(\Theta_a;\mathbf{x})} = \frac{L(\Theta_0;\mathbf{x})}{L(\hat{\theta};\mathbf{x})} \le 1,$$

$$\mathbf{x}\in B = \{\mathbf{x}\colon\ \hat{\Theta}(\mathbf{x}) = \Theta_0\} \Rightarrow \frac{L(\Theta_0;\mathbf{x})}{L(\Theta_a;\mathbf{x})} \ge \frac{L(\Theta_0;\mathbf{x})}{L(\hat{\theta};\mathbf{x})} = 1.$$

It follows that $L(\Theta_0;\mathbf{x})/L(\Theta_a;\mathbf{x}) \le c < 1$ only if $\lambda(\mathbf{x}) = L(\Theta_0;\mathbf{x})/L(\hat{\theta};\mathbf{x}) \le c < 1$. When both preceding inequalities hold, $\hat{\theta} = \Theta_a$ and $L(\Theta_0;\mathbf{x})/L(\Theta_a;\mathbf{x}) = L(\Theta_0;\mathbf{x})/L(\hat{\theta};\mathbf{x})$. Thus, for $c < 1$ and $\alpha \in (0,1)$, if $P(L(\Theta_0;\mathbf{x})/L(\hat{\theta};\mathbf{x}) \le c;\Theta_0) = \alpha$ and $P(L(\Theta_0;\mathbf{x})/L(\Theta_a;\mathbf{x}) \le c;\Theta_0) = \alpha$, then $C_r = C_r^{\mathrm{GLR}}$ and the GLR test is equivalent to the Neyman–Pearson most powerful test of size $\alpha$. ■

Since the GLR test is equivalent to the Neyman–Pearson most powerful test when $H_0$ and $H_a$ are simple, we also know immediately by Theorem 9.2 that the GLR test is also *unbiased*. We revisit a previous example, in which a most powerful test was found via the Neyman–Pearson Lemma to illustrate the GLR approach to the problem.

**Example 10.5** Recall Ex. 9.11, in which a decision is to be made regarding the population mean life of a type of LCD notebook computer screen. The likelihood function for the parameter $\theta$ (the mean life of the screens) given the 10 observations on screen lifetimes represented by $(x_1,\ldots,x_{10})$ is given by (suppressing the indicator functions)

$$L(\theta;\mathbf{x}) = \theta^{-10}\exp\left(-\sum_{i=1}^{10} x_i/\theta\right).$$

The null and alternative hypotheses under consideration are $H_0\colon\ \theta = 1$ and $H_a\colon\ \theta = 5$, respectively.

[1] Recall that "arg max$_{\Theta\in A}[\cdot]$" represents the argument value of $\Theta$ contained in the set $A$ that maximizes the function of $\Theta$ contained in brackets.

The GLR for this problem is given by

$$\lambda(\mathbf{x}) = \frac{L(1;\mathbf{x})}{\max\limits_{\theta\in\{1,5\}} L(\Theta;\mathbf{x})} = \min\left\{1, \frac{\exp\left(-\sum_{i=1}^{10} x_i\right)}{(5)^{-10}\exp\left(-\sum_{i=1}^{10} x_i/5\right)}\right\}$$

$$= \min\left\{1, (.2)^{-10}\exp\left(-.8\sum_{i=1}^{10} x_i\right)\right\}.$$

It follows that, for $c < 1$, the probability that $\lambda(\mathbf{x}) \le c$ is given by

$$P(\lambda(\mathbf{x}) \le c) = P\left((.2)^{-10}\exp\left(-.8\sum_{i=1}^{10} x_i\right) \le c\right)$$

$$= P\left(\sum_{i=1}^{10} x_i \ge 20.11797 - 1.25\ln(c)\right) \quad \text{(to five decimal places).}$$

A size $\alpha$ GLR test with critical region $[0, c]$ is defined by choosing $c$ so that $P(\lambda(\mathbf{x}) \le c) = \alpha$, where the probability value can be determined by utilizing the fact that $\sum_{i=1}^{10} X_i \sim \text{Gamma}(10, \theta)$ and $\theta = 1$ under $H_0$. Comparing this result to the definition of the most powerful critical region given in Ex. 9.11, it is evident that the two critical regions are identical, and so the GLR test is both unbiased and the most powerful test of $H_0\colon \theta = 1$ versus $H_a\colon \theta = 5$. For $\alpha = .05$, the critical region of the GLR test would be given by $[0, 34.13079]$, which can be transformed into a critical region stated in terms of the test statistic $\sum_{i=1}^{10} X_i$ as $[15.70522, \infty)$. □

*Composite Hypotheses* Similar to the Neyman–Pearson lemma's extension to the case of defining UMP tests for testing simple null versus composite alternative hypotheses (Theorem 9.3), the result of Theorem 10.1 can be extended to the case of testing simple null versus composite alternative hypotheses as follows:

**Theorem 10.2** (UMP Level-$\alpha$ GLR Test of $H_0\colon \Theta = \Theta_0$ Versus $H_a\colon \Theta \in \Omega_a$ When $C_r$ is Invariant) *Suppose the given critical region, $C_r^{\text{GLR}} = \{\mathbf{x}\colon \lambda(\mathbf{x}) \le c\}$, of the GLR test of $H_0\colon \Theta = \Theta_0$ versus $H_a\colon \Theta \in \Omega_a$ defines a size $\alpha \in (0,1)$ test and $\forall\, \Theta_a \in H_a\; \exists c_{\Theta_a} \ge 0$ such that*

$$C_r^{\text{GLR}} = \{\mathbf{x}\colon \lambda_{\Theta_a}(\mathbf{x}) \le c_{\Theta_a}\},$$

*where*

$$\lambda_{\Theta_a}(\mathbf{x}) = \frac{L(\Theta_0;\mathbf{x})}{\max_{\Theta\in\{\Theta_0,\Theta_a\}} L(\Theta;\mathbf{x})} \quad \text{and} \quad P(\lambda_{\Theta_a}(\mathbf{x}) \le c_{\Theta_a};\Theta_0) = \alpha.$$

*Furthermore, suppose a Neyman-Pearson UMP test of $H_0$ versus $H_a$ having size $\alpha$ exists. Then $C_r^{\text{GLR}}$ defines a UMP level-$\alpha$ test of $H_0$ versus $H_a$.*

**Proof** Given that a Neyman–Pearson UMP test having size $\alpha$ exists, it is the Neyman–Pearson most powerful size $\alpha$ test for every pair $(\Theta_0, \Theta_a)$, by Theorem 9.3, where $C_r$ is invariant to the choice of $\Theta_a \in H_a$. Because $\forall$ $(\Theta_0, \Theta_a)$ pair, both the Neyman–Pearson and GLR size $\alpha \in (0, 1)$ tests exist, the tests are equivalent $\forall$ $(\Theta_0, \Theta_a)$ by Theorem 10.1. Then $C_r^{\text{GLR}}$ defines a UMP level-$\alpha$ test of $H_0$ versus $H_a$. ■

The theorem implies that if the critical region of a size $\alpha$ GLR test of $H_0: \Theta = \Theta_0$ versus $H_a: \Theta = \Theta_a$ is the *same* $\forall$ $\Theta_a \in \Omega_a$, then if a Neyman–Pearson UMP test of $H_0: \Theta = \Theta_0$ versus $H_a: \Theta \in \Omega_a$ having size $\alpha$ exists, it is given by the GLR test. The UMP test would also be *unbiased* (Theorem 9.4).

**Example 10.6** Recall Ex. 9.11, 9.13, and 10.1 regarding the operating lives of computer screens. Examine the problem of defining a size .05 GLR test of $H_0: \theta = 1$ versus $H_a: \theta \in \Omega_a$, where $\Omega_a = (1, \infty)$. The GLR for this problem is given by

$$\lambda(\mathbf{x}) = \frac{L(1; \mathbf{x})}{\sup_{\theta \in [1,\infty)} L(\theta; \mathbf{x})} = \frac{\exp\left(-\sum_{i=1}^{10} x_i\right)}{\sup_{\theta \in [1,\infty)} \theta^{-10} \exp\left(-\sum_{i=1}^{10} x_i/\theta\right)}.$$

The maximum of $L(\theta; \mathbf{x})$ for $\theta \in [1, \infty)$ can be defined by first solving the first-order condition

$$\frac{d \ln L(\theta; \mathbf{x})}{d\theta} = \frac{-10}{\theta} + \frac{\sum_{i=1}^{10} x_i}{\theta^2} = 0,$$

which yields $\theta = \sum_{i=1}^{10} x_i/10$ as the choice of $\theta$ that maximizes $\ln L(\theta; \mathbf{x})$, and hence maximizes $L(\theta; \mathbf{x})$, when there are *no constraints* on $\theta$. Then, if the constraint $\theta \geq 1$ is recognized, $\lambda(\mathbf{x})$ can be defined ultimately as[2]

$$\lambda(\mathbf{x}) = \begin{cases} \left(\sum_{i=1}^{10} x_i/10\right)^{10} \exp\left(10 - \sum_{i=1}^{10} x_i\right), & \text{for } \sum_{i=1}^{10} x_i > 10, \\ 1 & \text{otherwise.} \end{cases}$$

It follows that, for $c < 1$, the probability that $\lambda(\mathbf{x}) \leq c$ is given by

$$\begin{aligned} P(\lambda(\mathbf{x}) \leq c) &= P\left(\left(\sum_{i=1}^{10} x_i/10\right)^{10} \exp\left(10 - \sum_{i=1}^{10} x_i\right) \leq c\right) \\ &= P\left(\sum_{i=1}^{10} x_i - 10 \ln\left(\sum_{i=1}^{10} x_i\right) \geq -13.0259 - \ln(c)\right]\right). \end{aligned} \qquad (*)$$

Recall that $\sum_{i=1}^{10} X_i \sim \text{Gamma}(10, 1)$ when $\theta = 1$. Also, from Ex. 9.11 we know that $P(\sum_{i=1}^{10} x_i \geq 15.70522) = .05$. Note that $\sum_{i=1}^{10} x_i - 10 \ln(\sum_{i=1}^{10} x_i)$ is strictly monotonically increasing in the value of $\sum_{i=1}^{10} x_i$ for values of $\sum_{i=1}^{10} x_i > 10$, so that there will exist a value of $c$ in $(*)$ that defines the event $\sum_{i=1}^{10} x_i \geq 15.70522$.

[2] A more elegant solution procedure for this inequality-constrained maximization problem could be formulated in terms of Kuhn–Tucker conditions.

In particular, the appropriate value is $c = .30387$ (to five decimal places). Thus, the critical region for the GLR is $[0, .30387]$, and the associated critical region for $\mathbf{x}$ is $C_r^{\mathrm{GLR}} = \{\mathbf{x} : \sum_{i=1}^{10} x_i \geq 15.70522\}$, which we know (from Ex. 9.13) defines the UMP and unbiased test of $H_0 : \theta = 1$ versus $H_a : \theta > 1$.

To see that the GLR test satisfies Theorem 10.2—so that we could declare the test to be UMP and unbiased independently of knowing the result of Ex. 9.13—the reader can retrace the steps followed in Ex. 10.1, replacing $H_a : \theta = 5$ with $H_a : \theta = \theta_a$, where $\theta_a$ is an arbitrary choice of $\theta \in (1, \infty)$. The critical region will invariably have the form $C_r^{\mathrm{GLR}} = \{\mathbf{x} : \sum_{i=1}^{10} x_i \geq 15.70522\}$ and will agree with the Neyman–Pearson most powerful critical region. □

The GLR test can also lead to UMP and unbiased tests of the simple or *composite* null hypothesis $H_0 : \Theta \in \Omega_0$ versus the composite alternative hypothesis $H_a : \Theta \in \Omega_a$ when $\Theta$ is a scalar, the alternative hypothesis is one-sided, and the problem is characterized by a monotone likelihood ratio.

**Theorem 10.3** (UMP and Unbiased GLR Level-$\alpha$ Test of $H_0$: $\Theta \in \Omega_0$ Versus One-Sided $H_a$: $\Theta \in \Omega_a$ in Case of Monotone Likelihood Ratio) *Let the sampling density of* $\mathbf{X}$, *given by* $f(\mathbf{x};\Theta)$ *for scalar* $\Theta \in \Omega$, *be a family of density functions having a monotone likelihood ratio in the statistic* $T = t(\mathbf{X})$. *Then the GLR test of* $H_0 : \Theta \in \Omega_0$ *versus one-sided* $H_a : \Theta \in \Omega_a$ *is a UMP and unbiased level-*$\alpha$ *test if* $P(\lambda(\mathbf{x}) \leq c) = \alpha$ *and either of the following conditions holds*:

1. $H_0 : \Theta = \Theta_0$ *or* $\Theta \leq \Theta_0$, $H_a : \Theta > \Theta_0$, *and* $\lambda(\mathbf{x}) \leq c$ *iff* $t(\mathbf{x}) \geq c^*$; *or*
2. $H_0 : \Theta = \Theta_0$ *or* $\Theta \geq \Theta_0$, $H_a : \Theta < \Theta_0$, *and* $\lambda(\mathbf{x}) \leq c$ *iff* $t(\mathbf{x}) \leq c^*$.

**Proof** The proof follows immediately from Corollaries 9.1 and 9.2 upon recognition that $\lambda(\mathbf{x}) \leq c$ is simply an alternative representation of the UMP and unbiased level-$\alpha$ critical region for the respective hypothesis tests based on the properties of monotone likelihood ratios. ■

**Example 10.7** A personal computer manufacturer claims that $\geq 80$ percent of its new computers are shipped to its customers without any defects whatsoever. A random sample, with replacement, of 20 purchases of the company's products resulted in 6 reports of initial defects upon delivery of computers. The responses of the purchasers are viewed as outcomes of iid Bernoulli random variables as

$$X_i \sim p^{x_i}(1-p)^{1-x_i} I_{\{0,1\}}(x_i),$$

where $x_i = 1 \Rightarrow$ defect reported and $x_i = 0 \Rightarrow$ no defect reported. We will construct a size .01 GLR test of the hypothesis that $H_0 : p \leq .20$ versus $H_a : p > .20$, i.e., we are testing the null hypothesis that the proportion of computers shipped that are defective is less than or equal to .20 versus the alternative that the proportion is greater than .20.

The GLR for this problem is

$$\lambda(\mathbf{x}) = \frac{\sup_{p \leq .20} p^{\sum_{i=1}^{20} x_i}(1-p)^{20-\sum_{i=1}^{20} x_i}}{\sup_{p \in [0,1]} p^{\sum_{i=1}^{20} x_i}(1-p)^{20-\sum_{i=1}^{20} x_i}}.$$

The value of $p$ that solves the denominator supremum problem is simply the MLE outcome for $p$, $\hat{p} = \sum_{i=1}^{20} x_i/20$. The value of $p$ that solves the numerator supremum problem is defined as

$$\hat{p}_0 = \begin{bmatrix} \sum_{i=1}^{20} x_i/20 & \text{if } \sum_{i=1}^{20} x_i/20 \le .20 \\ .20 & \text{otherwise} \end{bmatrix}.$$

Therefore, $\lambda(\mathbf{x})$ can be represented as

$$\lambda(\mathbf{x}) = \begin{cases} \dfrac{(.20)^z(.80)^{20-z}}{(z/20)^z(1-z/20)^{20-z}} & \text{if } \sum_{i=1}^{20} x_i > 4, \\ 1 & \text{otherwise} \end{cases}$$

where $Z = \sum_{i=1}^{20} X_i \sim \text{Binomial}(20, p)$.

Note that $\lambda(\mathbf{x})$ is a strictly decreasing function of $z$ for $z > 4$. Since $P(z \ge 9) = .01$, it follows that

$$P(\lambda(\mathbf{x}) \le .04172) = P(z \ge 9) = .01,$$

and so the critical region for the GLR test is

$$C_r^{\text{GLR}} = \{\mathbf{x}: \lambda(\mathbf{x}) \le .04172\} = \left\{\mathbf{x}: \sum_{i=1}^{20} x_i \ge 9\right\}.$$

The reader can verify that the statistic $t(\mathbf{X})$ of the monotone likelihood ratio can be specified as $t(\mathbf{X}) = \sum_{i=1}^{20} X_i$. From Theorem 10.3, it follows that the GLR test is UMP and unbiased with level .01. Given that $\sum_{i=1}^{20} x_i = 6$, the hypothesis $H_0: p \le .20$ *cannot* be rejected at the .01 level. □

At this point we focus the reader's attention on a common procedure that was used in the preceding examples to assign probabilities to events for the GLR of the form $\lambda(\mathbf{x}) \le c$. Namely, in each case we were able to find a strictly monotonically increasing or decreasing function of $\lambda(\mathbf{x})$, say $h(\lambda(\mathbf{x}))$, whose PDF had a known tractable form. Then probability was assigned using the PDF of $h(\lambda(\mathbf{X}))$ as $P(\lambda(\mathbf{x}) \le c) = P(h \le k)$ or $P(h \ge k)$ for $h(\cdot)$ monotonically increasing or decreasing, respectively. It is often the case that the probability density of the GLR, $\lambda(\mathbf{X})$, is difficult to define or intractable to work with for defining a size $\alpha$ critical region. In applications of GLR tests, one must often seek a *test statistic*, $h(\lambda(\mathbf{X}))$, having a tractable probability density in order to be able to both define a size $\alpha$ critical region and investigate the finite sample properties of the GLR test. Unfortunately, in practice, it is not always possible to define such a test statistic.

We will refrain from attempting to provide additional results concerning finite sample properties of GLR tests, mainly because few additional generalizations can be made. Typically, a GLR test is constructed in a given problem context, and then an attempt to assess its properties is made. It is sometimes the case that little can be definitively established regarding the finite sample properties of a GLR test. Fortunately, it is typically the case that the large sample properties of GLR tests are very good, and the results apply quite generally to simple or composite hypotheses involving scalar or multidimensional parameters for problems with or without nuisance parameters. We examine this topic next.

### Test Properties: Asymptotics

The GLR test is generally a consistent test, and in cases where $H_0$ is defined by functional restrictions on the parameter space, the asymptotic distribution of $-2\ln\lambda(\mathbf{X})$ is generally a $\chi^2$-distribution. In many cases, the verification of consistency and the identification of the asymptotic distribution of the GLR test can best be accomplished by analyzing either the properties of the random variable $\lambda(\mathbf{X})$ directly or else the properties of a test statistic $h(\lambda(\mathbf{X}))$ that is a function of the GLR statistic. In other cases, there exist regularity conditions that ensure the consistency and asymptotic distribution of the GLR test. We present some results for the asymptotic properties of the GLR test ahead. Additional results can be found in S. Wilks (1962), *Mathematical Statistics*. New York: John Wiley, p. 419; and R. J. Serfling (1980), *Approximation Theorems of Mathematical Statistics*. New York: John Wiley, pp. 151–160.

We begin with the property of consistency and introduce a sufficient condition that applies quite generally and is often not difficult to establish.

**Theorem 10.4** (Consistency of the GLR Test) *Assume the conditions for consistency of the maximum likelihood estimator (MLE) given by any of Theorems 8.14–8.16. Let* $[0,c_n]$, *for* $n = 1, 2, 3, \ldots$, *represent level-*$\alpha$ *critical regions of the GLR statistic for testing* $H_0: \Theta \in \Omega_0$ *versus* $H_a: \Theta \in \Omega_a$ *based on increasing sample size n. Then* $\lim_{n\to\infty} P(\lambda(\mathbf{x}) \le c_n;\Theta) = 1\ \forall\ \Theta \in H_a$, *so that the sequence of GLR tests is consistent, if either of the following conditions holds*:

**a.** *the GLR statistic is bounded below 1 with probability* $\to 1$ *as* $n \to \infty$ $\forall\ \Theta \in H_a$, *i.e.,* $\lim_{n\to\infty} P(\lambda(\mathbf{x}) \le \tau;\Theta) = 1\ \forall\ \Theta \in H_a$, *where* $\tau < 1$; *or*

**b.** $\operatorname{plim}\lambda(\mathbf{X}) = \delta(\Theta) \le \tau < 1\ \forall\ \Theta \in H_a$.

**Proof** **a.** Assuming that the true $\Theta_0 \in H_0$, it follows that $\operatorname{plim}\lambda(X_1,\ldots,X_n) = 1$. To see this, let $\hat\Theta_0$ and $\hat\Theta$ represent the MLEs for $\Theta \in H_0$ and $\Theta \in H_0 \cup H_a$, respectively, and expand the logarithm of the likelihood function $L(\hat\Theta_0;\mathbf{X})$ in a Taylor series around the point $\hat\Theta$ to obtain $\ln\lambda(\mathbf{X}) = \ln L(\hat\Theta_0;\mathbf{X}) - \ln L(\hat\Theta;\mathbf{X}) = (\partial \ln L(\Theta_*;\mathbf{X})/\partial\Theta)'(\hat\Theta_0 - \hat\Theta)$, where $\Theta_* = \eta\hat\Theta_0 + (1-\eta)\hat\Theta$ and $\eta \in [0,1]$.[3] Since both $\hat\Theta_0 \xrightarrow{p} \Theta_0$ and $\hat\Theta \xrightarrow{p} \Theta_0$ (recall $\Theta_0 \in H_0$), then $\Theta_* \xrightarrow{p} \Theta_0$, and it follows that both $\partial \ln L(\Theta_*;\mathbf{X})/\partial\Theta$ and $(\hat\Theta_0 - \hat\Theta) \xrightarrow{p} [0]$. Therefore, $\ln\lambda(\mathbf{X}) \xrightarrow{p} 0$, so that $\lambda(\mathbf{X}) \xrightarrow{p} 1$, and if the GLR test is to be of size $\le \alpha\ \forall\ n$, so that $\forall\ n$ and $\forall\ \Theta \in H_0$, $P(\lambda(\mathbf{x}) \le c_n;\Theta) \le \alpha$, it follows

[3] R. G. Bartle, *Real Analysis*, p. 371.

that $c_n \to 1$ as $n \to \infty$. Finally, if $\lim_{n\to\infty} P(\lambda(\mathbf{x}) \leq \tau; \Theta) = 1$, with $\tau < 1$ $\forall\, \Theta \in H_a$, then $\lim_{n\to\infty} P(\lambda(\mathbf{x}) \leq c_n; \Theta) = 1\ \forall\, \Theta \in H_a$.

**b.** $\forall\, \Theta \in H_a$, $\operatorname{plim} \lambda(\mathbf{X}) = \delta(\Theta) \leq \tau < 1 \Rightarrow \lim_{n\to\infty} P(\lambda(\mathbf{x}) \leq \tau; \Theta) = 1$ which leads to case (a). ■

**Example 10.8** Revisit Ex. 10.6 regarding the operating lives of computer screens. Assume that the true $\theta_0 \in H_a = (1, \infty)$, and examine the behavior of the GLR $\lambda(\mathbf{X})$. Letting $Z_n = \bar{X}_n \exp(n(1 - \bar{X}_n))$, the GLR can be represented as $\lambda(\mathbf{X}) = I_{[0,1]}(\bar{X}_n) + Z_n I_{(1,\infty)}(\bar{X}_n)$; if $\theta_0 \in H_a$, then $I_{[0,1]}(\bar{X}_n) \xrightarrow{\text{p}} 0$ and $I_{(1,\infty)}(\bar{X}_n) \xrightarrow{\text{p}} 1$. Then $\operatorname{plim} \lambda(\mathbf{X}) = \operatorname{plim}(Z_n)$ if the latter probability limit exists. Note that $n^{-1} \ln Z_n = n^{-1} \ln \bar{X}_n + (1 - \bar{X}_n) \xrightarrow{\text{p}} 1 - \theta_0 = \xi < 0$ since $\bar{X}_n \xrightarrow{\text{p}} \theta_0 > 1$. It follows from the definition of convergence in probability that, $\forall\, \varepsilon > 0$,

$$P(n^{-1} \ln z_n < \xi + \varepsilon) \geq P(\xi - \varepsilon < n^{-1} \ln z_n < \xi + \varepsilon) \to 1.$$

Then choosing $\varepsilon > 0$ small enough so that $\xi + \varepsilon < 0$,

$$P\left(n^{-1} \ln z_n < \xi + \varepsilon\right) = P\left(\ln z_n < n(\xi + \varepsilon)\right) = P\left(z_n < \exp\left(n(\xi + \varepsilon)\right)\right) \to 1.$$

Since $\exp(n(\xi + \varepsilon)) \to 0$ and $z_n \geq 0$, $\lim_{n\to\infty} P(z_n \in [0, \tau)) = 1\ \forall\, \tau > 0$, so that $\operatorname{plim} Z_n = 0$, and thus $\operatorname{plim} \lambda(\mathbf{X}) = 0$. Then the sequence of GLR tests of $H_0: \theta = 1$ versus $H_a: \theta > 1$ is consistent by Theorem 10.4.b. □

**Example 10.9** Recall Ex. 10.7 regarding the claim that $\geq 80$ percent of new computers are shipped free of defects. Assume that the true $p_0 \in H_a = (.20, 1]$, and examine the behavior of the GLR $\lambda(\mathbf{X})$. Letting $w_n = (.20)^{n\bar{x}}(.80)^{n(1-\bar{x})}/[\bar{x}^{n\bar{x}}(1 - \bar{x})^{n(1-\bar{x})}]$, the GLR can be represented as $\lambda(\mathbf{X}) = I_{[0,.20]}(\bar{X}_n) + W_n I_{(.20,\infty)}(\bar{X}_n)$, and if $p_0 \in H_a$, then $I_{[0,.20]}(\bar{X}_n) \xrightarrow{\text{p}} 0$ and $I_{(.20,\infty)}(\bar{X}_n) \xrightarrow{\text{p}} 1$. Then $\operatorname{plim} \lambda(\mathbf{X}) = \operatorname{plim} W_n$ if the latter probability limit exists. Note that

$$(W_n)^{1/n} = \frac{(.20)^{\bar{X}}(.80)^{(1-\bar{X})}}{\left[\bar{X}^{\bar{X}}(1 - \bar{X})^{(1-\bar{X})}\right]} \xrightarrow{\text{p}} \frac{(.20)^{p}(.80)^{1-p}}{p^{p}(1 - p)^{1-p}}$$

$$\xrightarrow{\text{p}} \left(\frac{.20}{p}\right)^{p} \left(\frac{.80}{(1 - p)}\right)^{1-p} = \xi < 1 \quad \forall\, p \in (.2, 1].$$

Following reasoning similar to that in the previous example, one can establish that $P(w_n < (\xi + \varepsilon)^n) \to 1$ for $\varepsilon > 0$ small enough such that $\xi + \varepsilon$ is positive and less than 1. It follows from $(\xi + \varepsilon)^n \to 0$ and $w_n \geq 0$ that

$$\lim_{n\to\infty} P(w_n \in [0, \tau)) = 1 \quad \forall\, \tau > 0,$$

so that $W_n \xrightarrow{\text{p}} 0$, and thus $\lambda(\mathbf{X}) \xrightarrow{\text{p}} 0$. Then the sequence of GLR tests of $H_0: p \leq .20$ versus $H_a: p > .20$ is consistent by Theorem 10.4.b. □

We have shown in the preceding examples that sequences of statistical tests are consistent, so that for large enough sample sizes, one is essentially certain to reject the null hypothesis if it is false. It is useful to note that consistent tests are effectively **asymptotically unbiased tests** in the sense that for large enough $n$, the rejection probability will eventually exceed whatever level $\alpha < 1$ is associated with the consistent test sequence since the probability of rejecting a false $H_0 \to 1$. Thus, the tests in the sequence eventually become unbiased (if they weren't unbiased to begin with) as $n \to \infty$.

In the examples of statistical tests presented heretofore, the probability density of either the GLR statistic or a function of the GLR statistic has been readily identifiable and tractable to work with. In cases where the identification and/or tractability of the probability distributions of GLR test statistics is problematic, there are asymptotic results relating to the GLR test that can be helpful as long as sample sizes are not too small. We will examine one such result relating to the asymptotic distribution of the natural logarithm of the GLR statistic for a special but important and prevalent form of null hypothesis. A discussion of additional results, and further readings, can be found in D. R. Cox and D. V. Hinkley (1979), *Theoretical Statistics*. London: Chapman and Hall, pp. 311–342.

When considering the asymptotic distribution of the GLR, it will prove useful to focus attention on the function $-2\ln(\lambda(\mathbf{x}))$ rather than on $\lambda(\mathbf{x})$ itself. A general result regarding the asymptotic distribution of $-2\ln(\lambda(\mathbf{X}))$ can be obtained when $H_0$ is defined via functional restrictions on the parameter space. In particular, we will be examining null hypotheses of the form $H_0 = \{\Theta : \mathbf{R}(\Theta) = \mathbf{r}, \Theta \in \Omega\}$, where $\mathbf{R}(\Theta)$ is a $(q \times 1)$ differentiable vector function and $\mathbf{R}(\Theta) = \mathbf{r}$ places linear and/or nonlinear constraints on the elements of the parameter vector $\Theta$. It will be assumed that *none* of the $q$ coordinate functions in $\mathbf{R}(\Theta)$ is *redundant*. In this case, we can show that when the null hypothesis is true, $-2\ln(\lambda(\mathbf{X})) \overset{a}{\sim} \chi^2_q$ so that an asymptotically valid size $\alpha$ GLR test of $H_0: \mathbf{R}(\Theta) = \mathbf{r}$ versus $H_a: \mathbf{R}(\Theta) \neq \mathbf{r}$ can be conducted as

$$-2\ln(\lambda(\mathbf{x})) \geq \chi^2_{q;\alpha} \Rightarrow \text{reject } H_0$$

or in terms of the GLR statistic itself,

$$\lambda(\mathbf{x}) \leq \exp\left(-\frac{1}{2}\chi^2_{q;\alpha}\right) \Rightarrow \text{reject } H_0,$$

where $\chi^2_{q;\alpha}$ is the value of a $\chi^2$-random variable with $q$ degrees of freedom such that the event $(\chi^2_{q;\alpha}, \infty)$ is assigned probability $\alpha$. Furthermore, $-2\ln\lambda(\mathbf{X})$ will have a *noncentral* $\chi^2$ asymptotic distribution (see Section 10.9) when $H_0$ is false (and when we examine *local* alternative hypotheses, to be discussed shortly), allowing asymptotically valid power functions to be constructed. The formal result on the asymptotic distribution of the GLR test when $H_0$ is true is given next.

**Theorem 10.5** (Asymptotic Distribution of GLR Test of $H_0: \mathbf{R}(\Theta) = \mathbf{r}$ Versus $H_a: \mathbf{R}(\Theta) \neq \mathbf{r}$ When $H_0$ is True) *Assume the conditions for the consistency, asymptotic normality, and asymptotic efficiency of the MLE of the* $(k \times 1)$ *vector* $\Theta$ *as given in Theorem 8.18. Let* $\lambda(\mathbf{x}) = \sup_{\Theta\in H_0} L(\Theta;\mathbf{x})/\sup_{\Theta\in H_0 \cup H_a} L(\Theta;\mathbf{x})$ *be the GLR statistic for testing* $H_0: \mathbf{R}(\Theta) = \mathbf{r}$ *versus* $H_a: \mathbf{R}(\Theta) \neq \mathbf{r}$*, where* $\mathbf{R}(\Theta)$

*is a $(q \times 1)$ continuously differentiable vector function having nonredundant coordinate functions and $q \leq k$. Then $-2\ln\lambda(\mathbf{X}) \xrightarrow{d} \chi^2_q$ when $H_0$ is true.*

***Proof** The proof is somewhat challenging. For a related proof sketch, see T. Amemiya (1985), *Advanced Econometrics*. Cambridge, MA: Harvard Univ. Press, pp. 142–144. The original work of A. Wald (1943), "Tests of statistical hypotheses concerning several parameters when the number of observations is large." *Transactions of the American Mathematical Society*. 54, pp. 426–482, provides an alternative, albeit more difficult and restrictive, proof of the asymptotic distribution result.

Expand $\ln L(\Theta_0; \mathbf{X})$ in a second-order Taylor series around the ML estimator $\hat{\Theta}$ as[4]

$$\ln L(\hat{\Theta}; \mathbf{X}) - \ln L(\Theta_0; \mathbf{X}) = -\frac{1}{2}(\hat{\Theta} - \Theta_0)' \frac{\partial^2 \ln L(\Theta_*; \mathbf{X})}{\partial\Theta\partial\Theta'}(\hat{\Theta} - \Theta_0), \tag{1}$$

where $\Theta_* = \tau\hat{\Theta} + (1-\tau)\Theta_0$ for $\tau \in [0, 1]$, and the first-order term is dropped since $\partial L(\hat{\Theta}; \mathbf{X})/\partial\Theta = [0]$ by the first-order conditions defining the MLE. Let $\mathbf{R}(\Theta) = \mathbf{r}$ be used to define $q$ of the entries in $\Theta$ as functions of the remaining entries; without loss of generality, presume that the entries in $\Theta$ have been ordered so that the first $q\ \Theta_i$'s are functions of the remaining $k - q\ \Theta_i$'s, say as $\Theta^a = \mathbf{g}(\Theta^b)$, where $\Theta^a = [\Theta_1, \ldots, \Theta_q]'$ and $\Theta^b = [\Theta_{q+1}, \ldots, \Theta_k]'$. Then under the constraint $\mathbf{R}(\Theta) = \mathbf{r}$, the feasible $\Theta$ vectors can be characterized as

$$\Theta = \begin{bmatrix} \mathbf{g}(\Theta^b) \\ \hline \Theta^b \end{bmatrix} = \mathbf{h}(\Theta^b).$$

Making this substitution in the likelihood function defines the *restricted* likelihood function, $L_r(\Theta^b; \mathbf{x}) \equiv L(\mathbf{h}(\Theta^b); \mathbf{x})$. To simplify notation, define $\eta \equiv \Theta^b$. Assuming $H_0$: $\mathbf{R}(\Theta) = \mathbf{r}$ to be true, then $\Theta_0 = \mathbf{h}(\eta_0)$, and we can expand $\ln L_r(\eta_0; \mathbf{x})$ in a second-order Taylor series around the ML estimator $\hat{\eta}$, analogous to the preceding expansion, as

$$\ln L_r(\hat{\eta}; \mathbf{X}) - \ln L_r(\eta_0; \mathbf{X}) = -\frac{1}{2}(\hat{\eta} - \eta_0)' \frac{\partial^2 \ln L_r(\eta_*; \mathbf{X})}{\partial\eta\partial\eta'}(\hat{\eta} - \hat{\eta}_0). \tag{2}$$

Now note that since $\hat{\Theta} \xrightarrow{p} \Theta_0$ and $\hat{\eta} \to \eta_0$, then $\Theta_* \xrightarrow{p} \Theta_0$ and $\eta_* \to \eta_0$. Then rewrite the right-hand side of (1) as

$$-\frac{1}{2}(\hat{\Theta} - \Theta_0)' n^{1/2} \left[n^{-1} \frac{\partial^2 \ln L(\Theta_*; \mathbf{x})}{\partial\Theta\partial\Theta'}\right] n^{1/2}(\hat{\Theta} - \Theta_0),$$

and note that the bracketed expression converges in probability to $\mathbf{M}(\Theta_0)$, a positive definite, symmetric matrix. Then $\ln L(\hat{\Theta}; \mathbf{X}) - \ln L(\Theta_0; \mathbf{X})$ and $-1/2(\hat{\Theta} - \Theta_0)' n^{1/2}\mathbf{M}(\Theta_0) n^{1/2}(\hat{\Theta} - \Theta_0)$ will share the same limiting distribution, by Theorem 5.7. An analogous argument applied to (2) indicates that $\ln L_r(\hat{\eta}; \mathbf{X}) -$

[4] R. G. Bartle, *Real Analysis*, p. 371.

$\ln L_r(\eta_0; \mathbf{x})$ and $-1/2(\hat{\eta} - \eta_0)'n^{1/2}\mathbf{M}_r(\eta_0)n^{1/2}(\hat{\eta} - \eta_0)$ share the same limiting distribution. Letting $\overset{d}{=}$ indicate equivalence in terms of limiting distributions, it follows from $L(\Theta_0; \mathbf{x}) = L_r(\eta_0; \mathbf{x})$ that

$$-2\ln\lambda(\mathbf{X}) \overset{d}{=} (\hat{\Theta} - \Theta_0)'n^{1/2}\mathbf{M}(\Theta_0)n^{1/2}(\hat{\Theta} - \Theta_0) - (\hat{\eta} - \eta_0)'n^{1/2}\mathbf{M}_r(\eta_0)n^{1/2}(\hat{\eta} - \eta_0). \tag{3}$$

We now relate the limiting distributions of $n^{1/2}(\hat{\Theta} - \Theta_0)$ and $n^{1/2}(\hat{\eta} - \eta_0)$. It follows from the proof of Theorem 8.17 that

$$n^{1/2}(\hat{\Theta} - \Theta_0) \overset{d}{=} \mathbf{M}(\Theta_0)^{-1}n^{-1/2}\frac{\partial \ln L(\Theta_0; \mathbf{X})}{\partial \Theta} \xrightarrow{d} N\left([\mathbf{0}], \mathbf{M}(\Theta_0)^{-1}\right)$$

and

$$n^{1/2}(\hat{\eta} - \eta_0) \overset{d}{=} \mathbf{M}_r(\eta_0)^{-1}n^{-1/2}\frac{\partial \ln L_r(\eta_0; \mathbf{X})}{\partial \eta} \xrightarrow{d} N\left([\mathbf{0}], \mathbf{M}_r(\eta_0)^{-1}\right).$$

Since

$$\frac{\partial \ln L_r(\eta_0; \mathbf{X})}{\partial \eta} = \frac{\partial h(\eta_0)}{\partial \eta}\frac{\partial \ln L(\Theta_0; \mathbf{x})}{\partial \Theta},$$

it follows that

$$n^{1/2}(\hat{\eta} - \eta_0) \overset{d}{=} \mathbf{M}_r(\eta_0)^{-1}n^{-1/2}\frac{\partial h(\eta_0)}{\partial \eta}n^{1/2}\left[n^{-1/2}\frac{\partial \ln L(\Theta_0; \mathbf{X})}{\partial \Theta}\right].$$

Substituting for $n^{1/2}(\hat{\Theta} - \Theta_0)$ and $n^{1/2}(\hat{\eta} - \eta_0)$ on the right-hand side of (3), and letting $\mathbf{Z}_n = n^{-1/2}(\partial \ln L(\Theta_0; \mathbf{X})/\partial \Theta)$ to simplify notation,

$$-2\ln\lambda(\mathbf{X}) \overset{d}{=} \mathbf{Z}_n'\left[\mathbf{M}(\Theta_0)^{-1} - \mathbf{G}'\mathbf{M}_r(\eta_0)^{-1}\mathbf{G}\right]\mathbf{Z}_n, \tag{4}$$

where $\mathbf{G} = \partial h(\eta_0)/\partial \eta$.

Define $\mathbf{V}_n = \mathbf{M}(\Theta_0)^{-1/2}\mathbf{Z}_n$, so that $\mathbf{V}_n \xrightarrow{d} \mathbf{V} \sim N([\mathbf{0}], \mathbf{I})$, and rewrite (4) as

$$-2\ln\lambda(\mathbf{X}) \overset{d}{=} \mathbf{V}_n'\left[\mathbf{I} - \mathbf{M}(\Theta_0)^{1/2}\mathbf{G}'\mathbf{M}_r(\eta_0)^{-1}\mathbf{G}\mathbf{M}(\Theta_0)^{1/2}\right]\mathbf{V}_n. \tag{5}$$

The bracketed matrix is idempotent, which can be demonstrated by multiplying the bracketed matrix by itself and using the condition $\mathbf{G}\mathbf{M}(\Theta_0)\mathbf{G}' = \mathbf{M}_r(\eta_0)$. Furthermore, the trace of the idempotent matrix is seen to be $k - (k - q) = q$. Then, representing the quadratic form in (5) in terms of the characteristic roots and vectors of the idempotent matrix as $\mathbf{V}_n'\mathbf{P}\mathbf{\Lambda}\mathbf{P}'\mathbf{V}_n$, analogous to the proof of Theorem 6.12b, we finally obtain

$$-2\ln\lambda(\mathbf{X}) \xrightarrow{d} \mathbf{V}'\mathbf{P}\mathbf{\Lambda}\mathbf{P}'\mathbf{V} = \sum_{i=1}^{q} W_i^2 \sim \chi_q^2,$$

where $\mathbf{W} = \mathbf{P}'\mathbf{V} \sim N([\mathbf{0}], \mathbf{I})$ and $\mathbf{\Lambda}$ is a diagonal matrix having $q$ 1's and $(k - q)$ 0's along the diagonal. ■

In Ex. 10.5 and Ex. 10.6, which were based on random sampling from an exponential population distribution, we knew from our study of the MLE in

Chapter 8 that the MLE adheres to the conditions of Theorem 8.18 and is consistent, asymptotically normal, and asymptotically efficient. It follows by Theorem 10.5 that the GLR statistic for testing the simple null hypothesis $H_0: \theta = 1$ is such that $-2\ln\lambda(\mathbf{X}) \overset{a}{\sim} \chi^2_1$. The asymptotic result would also apply to any *simple* hypotheses tested in the context of Ex. 10.7, where sampling was from a Bernoulli population distribution. While neither of these previous cases presented significant difficulties in determining critical regions of the test in terms of the distribution of the GLR test statistic, it is useful to note that the asymptotic result for the GLR test provides an approximate method for circumventing the inherently limited choices of test sizes in the discrete case (recall Ex. 9.10 and Problem 9.7). In particular, since the $\chi^2$-distribution is continuous, *any* size test can be defined in terms of the *asymptotic* distribution of $-2\ln\lambda(\mathbf{X})$, whether or not $f(\mathbf{x}; \Theta)$ is continuous, although the size will be an *approximation* that improves as the sample size increases.

In order to establish an asymptotically valid method of investigating the power of the GLR test, we now consider the asymptotic distribution of the GLR when the *alternative* hypothesis is true. We will analyze **local alternatives** to the null hypothesis $H_0: \mathbf{R}(\Theta) = \mathbf{r}$. In particular, we will focus on alternatives of the form $\mathbf{R}(\Theta) = \mathbf{r} + n^{-1/2}\boldsymbol{\phi}$. In this context, the vector $\boldsymbol{\phi}$ specifies in what direction alternative hypotheses will be examined, and since $n^{-1/2}\boldsymbol{\phi} \to [\mathbf{0}]$ as $n \to \infty$, we are ultimately analyzing alternatives that are close, or *local,* to $\mathbf{R}(\Theta) = \mathbf{r}$ for large enough $n$. This type of local alternatives is also referred to as a **Pittman drift**.

A primary reason for examining local alternatives as opposed to fixed alternative hypotheses is that in the latter case, power will always be 1 for consistent tests. Thus, no further information about the operating characteristics of the test is gained from asymptotic considerations other than what is already known about consistent tests, namely, that one is sure to reject a false null hypothesis when $n \to \infty$. There is an analog to degenerate limiting distributions, which are equally uninformative about the characteristics of random variables other than that they converge in probability to a constant. In the latter case, the random-variable sequence was centered and scaled to obtain a nondegenerate limiting distribution that was more informative about the random-variable characteristics of interest, e.g., variance. In the current context, the alternative hypotheses are scaled to establish "nondegenerate" power function behavior.

**Theorem 10.6** (Asymptotic Distribution of GLR Test of $H_0: \mathbf{R}(\Theta) = \mathbf{r}$ Versus $H_a: \mathbf{R}(\Theta) \neq \mathbf{r}$ for Local Alternatives when $H_0$ is False) *Consider the GLR test of* $H_0: \mathbf{R}(\Theta) = \mathbf{r}$ *versus* $H_a: \mathbf{R}(\Theta) \neq \mathbf{r}$ *under the conditions and notation of Theorem 10.5, and let a sequence of local alternatives be defined by* $H_{an}: \mathbf{R}(\Theta) = \mathbf{r} + n^{-1/2}\boldsymbol{\phi}$. *Assume further that* $\partial\mathbf{R}(\Theta_0)/\partial\Theta'$ *has full row rank. Then the limiting distribution of* $-2\ln\lambda(\mathbf{X})$ *under the sequence of local alternatives is noncentral* $\chi^2$ *as*

$$-2\ln\lambda(\mathbf{X}) \xrightarrow{d} \chi^2_q(\lambda) \quad \textit{where } \lambda = \frac{1}{2}\boldsymbol{\phi}'\left[\frac{\partial\mathbf{R}(\Theta_o)'}{\partial\Theta}\mathbf{M}(\Theta_0)^{-1}\frac{\partial\mathbf{R}(\Theta_0)}{\partial\Theta}\right]^{-1}\boldsymbol{\phi}.$$

***Sketch of the Proof** Let $\hat{\Theta}_r$ and $\hat{\Theta}$ represent the MLEs for $\Theta \in H_0$ and $\Theta \in H_0 \cup H_a$, respectively, and expand $L(\hat{\Theta}_r; \mathbf{X})$ in a second-order Taylor series around the point $\hat{\Theta}$ to obtain

$$-2\left[\ln L(\hat{\Theta}_r; \mathbf{X}) - \ln L(\hat{\Theta}; \mathbf{X})\right] = -2\ln\lambda(\mathbf{X}) = (\hat{\Theta}_r - \hat{\Theta})'\left[-\frac{\partial^2 \ln L(\Theta_+; \mathbf{X})}{\partial\Theta\partial\Theta'}\right](\hat{\Theta}_r - \hat{\Theta}),$$

where we have used the fact that $\partial \ln L(\hat{\Theta}; \mathbf{X})/\partial\Theta = [\mathbf{0}]$ in the first-order term of the Taylor series and $\Theta_+ = \tau\hat{\Theta} + (1-\tau)\hat{\Theta}_r$ for some $\tau \in [0,1]$. We know from the proofs of Theorems 8.17 and 8.18 that

$$n^{1/2}(\hat{\Theta} - \Theta_0) \overset{\mathrm{d}}{=} \mathbf{M}(\Theta_0)^{-1}\left[n^{-1/2}\frac{\partial \ln L(\Theta_0; \mathbf{X})}{\partial\Theta}\right]$$

(recall that "$\overset{\mathrm{d}}{=}$" means equivalence in terms of limiting distributions).

Under the sequence of local alternatives $H_{an}: \mathbf{R}(\Theta) = \mathbf{r} + n^{-1/2}\boldsymbol{\phi}$, $\hat{\Theta}_r \xrightarrow{\mathrm{p}} \Theta_0$, and it can be shown that (see the subsequent proofs of Theorems 10.7 and 10.8)

$$\begin{aligned} n^{1/2}(\hat{\Theta}_r - \Theta_0) &\overset{\mathrm{d}}{=} \mathbf{M}(\Theta_0)^{-1}\left[n^{-1/2}\frac{\partial \ln L(\Theta_0; \mathbf{X})}{\partial\Theta}\right] \\ &\quad + \mathbf{M}(\Theta_0)^{-1}\mathbf{G}'\left[-\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\mathbf{G}'\right]^{-1}\left[\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\left(n^{-1/2}\frac{\partial \ln L(\Theta_0; \mathbf{X})}{\partial\Theta}\right) - \boldsymbol{\phi}\right], \end{aligned}$$

where $\mathbf{G} = \partial\mathbf{R}(\Theta_0)/\partial\Theta'$. Then

$$\begin{aligned} n^{1/2}(\hat{\Theta}_r - \hat{\Theta}) &= n^{1/2}(\hat{\Theta}_r - \Theta_0) - n^{1/2}(\hat{\Theta} - \Theta_0) \\ &\overset{\mathrm{d}}{=} \mathbf{M}(\Theta_0)^{-1}\mathbf{G}'\left[-\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\mathbf{G}'\right]^{-1}\left[\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\left(n^{-1/2}\frac{\partial \ln L(\Theta_0; \mathbf{X})}{\partial\Theta}\right) - \boldsymbol{\phi}\right], \\ &\overset{\mathrm{d}}{=} \mathbf{M}(\Theta_0)^{-1}\mathbf{G}'\mathbf{Z}_n, \end{aligned}$$

where $\mathbf{Z}_n \xrightarrow{\mathrm{d}} N([\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\mathbf{G}']^{-1}\boldsymbol{\phi}, [\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\mathbf{G}']^{-1})$.

Noting that $-2\ln\lambda(\mathbf{X}) \overset{\mathrm{d}}{=} n^{1/2}(\hat{\Theta}_r - \hat{\Theta})'\mathbf{M}(\Theta_0)(\hat{\Theta}_r - \hat{\Theta})n^{1/2}$ because

$$-n^{-1}\frac{\partial^2 \ln \mathrm{L}(\Theta_+; \mathbf{X})}{\partial\Theta\partial\Theta'} \xrightarrow{\mathrm{p}} \mathbf{M}(\Theta_0),$$

it follows that

$$-2\ln\lambda(\mathbf{X}) \overset{\mathrm{d}}{=} \mathbf{Z}_n'\left[\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\mathbf{G}'\right]\mathbf{Z}_n \xrightarrow{\mathrm{d}} \chi_q^2(\lambda)$$

with $\lambda = (1/2)\boldsymbol{\phi}'\left[\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\mathbf{G}'\right]^{-1}\boldsymbol{\phi}$ since

$$\left[\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\mathbf{G}'\right]^{1/2}\mathbf{Z}_n \xrightarrow{\mathrm{d}} N([\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\mathbf{G}']^{-1/2}\boldsymbol{\phi}, \mathbf{I})$$

(see Section 10.9 on properties of the noncentral $\chi^2$-distribution). ■

The GLR statistic and the LM and Wald statistics discussed in subsequent sections all share the same limiting distribution under sequences of local alternatives of the type identified here and are thus referred to as **asymptotically**

**equivalent tests**. This asymptotic equivalence will be established when we examine the alternative testing procedures. For related readings on the relationships between the triad of tests, see S. D. Silvey (1959), "The Lagrangian multiplier test." *Annals of Mathematical Statistics*, 30, pp. 389–407, R. Davidson and J. MacKinnon (1993), *Estimation and Inference in Econometrics*. New York: Oxford Univ. Press, pp. 445–449.

**Example 10.10** Asymptotic Power of GLR Test of $H_0: \theta = \theta_0$ in Exponential Population

Revisit Ex. 10.6, and consider the asymptotic power of the GLR test of $H_0: \theta = \theta_0$ versus $H_a: \theta > \theta_0$ for the sequence of local alternatives $\theta_n = \theta_0 + n^{-1/2}\delta$, where in this application $\theta_0 = 1$. For this case we know that $-2\ln\lambda(\mathbf{X}) \overset{a}{\sim} \chi_1^2(\lambda)$, with $\lambda = (1/2)\delta'\delta = \delta^2/2$, and $n = 10$ according to Ex. 10.6 (which is a bit small for the asymptotic calculations to be accurate). The asymptotic power function can be plotted in terms of the noncentrality parameter, or in this single parameter case, in terms of $\delta$. It is more conventional to graph power in terms of noncentrality, and we do this for a size .05 test in Figure 10.2, where $\pi(\lambda) = \int_{\chi^2_{1;.05}}^{\infty} f(w; 1, \lambda)dw = \int_{3.841}^{\infty} f(w; 1, \lambda)dw$ and $f(w; 1, \lambda)$ is a noncentral $\chi^2$-density with 1 degree of freedom and noncentrality parameter $\lambda$ (the GAUSS matrix language, procedure CDFCHINC, was used to calculate the integrals). As is typical of other tests we have examined, the closer $\delta$ is to zero, and thus the closer $H_a: \theta = \theta_0 + n^{-1/2}\delta$ is to $H_0: \theta = \theta_0$, the less power there is for detecting a false $H_0$. □

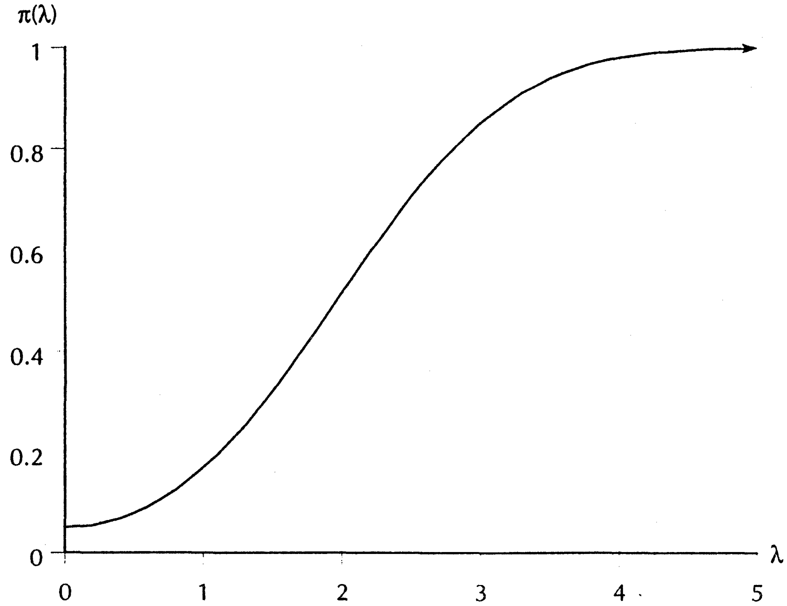


**Figure 10-2** Asymptotic power of GLR test for $H_0: \theta = \theta_0$ versus $H_a: \theta > \theta_0$, local alternatives $\theta_n = \theta_0 + n^{-1/2}\delta, \lambda = \delta^2/2$.

## 10.4 Lagrange Multiplier Tests

In comparison to the GLR approach, the Lagrange multiplier (LM) test of a statistical hypothesis utilizes an alternative measure of the discrepancy between restricted (by $H_0$) and unrestricted estimates of the parameters of the joint density of a random sample in order to design a test rule. In particular, instead of analyzing the ratio of restricted and unrestricted likelihood function values as in the GLR approach, the LM approach assesses the significance of the magnitude of the Lagrange multipliers appearing in the restricted ML optimization problem. The intuition here is that the Lagrange multipliers in such a problem indicate the marginal changes to the optimized likelihood function value from constraint relaxation. Large values of Lagrange multipliers indicate that large increases in likelihood function values are possible from constraint relaxation. If the LM values are significantly different from zero, the indication is that the likelihood function can be significantly increased by examining parameter values contained in $H_a$, suggesting that $H_0$ is false and should be rejected. Like the GLR test, the LM test is a natural testing procedure to use for testing hypotheses about $\Theta$ or functions of $\Theta$ when (restricted) maximum likelihood estimation is being used in a statistical analysis. We will focus on the asymptotic properties of the test.[5]

In order to establish the form of the test rule, examine the problem of maximizing $\ln L(\Theta; \mathbf{x})$ subject to the constraint $\mathbf{R}(\Theta) = \mathbf{r}$ specified in $\mathrm{H}_0$, as

$$\max_{\Theta,\lambda} \ln L(\Theta; \mathbf{x}) - \lambda'[\mathbf{R}(\Theta) - \mathbf{r}],$$

where $\lambda$ is $(q \times 1)$ vector of LMs. The first-order conditions for this problem are

$$\frac{\partial \ln L(\Theta; \mathbf{x})}{\partial \Theta} - \frac{\partial \mathbf{R}(\Theta)}{\partial \Theta}\lambda = [0] \quad \text{and} \quad \mathbf{R}(\Theta) - \mathbf{r} = [0].$$

Letting $\hat{\Theta}_r$ represent the restricted MLE that solves the first-order conditions and $\Lambda_r$ represent the corresponding value of the LM, it follows that

$$\frac{\partial \ln L(\hat{\Theta}_r; \mathbf{X})}{\partial \Theta} - \frac{\partial \mathbf{R}(\hat{\Theta}_r)}{\partial \Theta}\Lambda_r = [0] \quad \text{and} \quad \mathbf{R}(\hat{\Theta}_r) - \mathbf{r} = [0].$$

We now establish the LM test and its asymptotic distribution under $H_0$.

**Theorem 10.7** (The LM Test of $H_0: \mathbf{R}(\Theta_0) = \mathbf{r}$ Versus $H_a: \mathbf{R}(\Theta_0) \neq \mathbf{r}$) *Assume the conditions and notation of Theorem 8.18 ensuring the consistency, asymptotic normality, and asymptotic efficiency of the MLE, $\hat{\Theta}$, of $\Theta$. Let $\hat{\Theta}_r$ and $\Lambda_r$ be the restricted MLE and value of the Lagrangian multiplier that satisfy*

$$\max_{\Theta,\lambda} L(\Theta; \mathbf{x}) - \lambda'(\mathbf{R}(\Theta) - \mathbf{r}),$$

[5] Excellent references for additional details include L. G. Godfrey (1988), *Misspecification Tests in Econometrics*. New York: Cambridge Univ. Press, pp. 5–20; and R. F. Engle (1984), "Wald, likelihood ratio and Lagrange multiplier tests." in *Handbook of Econometrics*, vol. 2, Z. Giliches and M. Intriligator, eds. Amsterdam: North-Holland, pp. 775–826.

*where* $\mathbf{R}(\Theta)$ *is a continuously differentiable* $(q \times 1)$ *vector function that contains no redundant coordinate functions. If* $\mathbf{G} = \partial \mathbf{R}(\Theta_0)/\partial \Theta'$ *has full row rank, then under* $H_0$ *it follows that:*[6]

1.

$$\mathbf{W} = \Lambda_r' \frac{\partial \mathbf{R}(\hat{\Theta}_r)'}{\partial \Theta}\left[-\frac{\partial^2 \ln L(\hat{\Theta}_r;\mathbf{X})}{\partial \Theta \partial \Theta'}\right]^{-1} \frac{\partial \mathbf{R}(\hat{\Theta}_r)}{\partial \Theta} \Lambda_r \xrightarrow{d} \chi^2_q,$$

2. *an asymptotic size* $\alpha$ *and consistent test of* $H_0$: $\mathbf{R}(\Theta_0) = \mathbf{r}$ *versus* $H_a$: $\mathbf{R}(\Theta_0) \neq \mathbf{r}$ *is given by*

$$w \begin{bmatrix} \geq \\ < \end{bmatrix} \chi^2_{q;\alpha} \Rightarrow \begin{bmatrix} \textit{reject } H_0 \\ \textit{accept } H_0 \end{bmatrix},$$

*and*

3. *an alternative and equivalent representation of* $\mathbf{W}$ *is given by*

$$\mathbf{W} = \frac{\partial \ln L(\hat{\Theta}_r;\mathbf{X})'}{\partial \Theta}\left[-\frac{\partial^2 \ln L(\hat{\Theta}_r;\mathbf{X})}{\partial \Theta \partial \Theta'}\right]^{-1} \frac{\partial \ln L(\hat{\Theta}_r;\mathbf{X})}{\partial \Theta}.$$

*(The test based on this alternative form of* $\mathbf{W}$ *was called the* **scoring test** *by C. R. Rao (1948), "Large sample tests of statistical hypotheses." Proceedings of the Cambridge Philosophical Society, 44, pp. 50–57.)*

***Proof** Expanding both $\partial \ln L(\hat{\Theta}_r;\mathbf{X})/\partial \Theta$ and $\mathbf{R}(\hat{\Theta}_r) - \mathbf{r} = [\mathbf{0}]$ in a first-order Taylor series around the true $\Theta_0$ allows the first-order conditions of the ML problem to be written as

$$\frac{\partial \ln L(\Theta_0;\mathbf{X})}{\partial \Theta} + \frac{\partial^2 \ln L(\Theta_*;\mathbf{X})}{\partial \Theta \partial \Theta'}(\hat{\Theta}_r - \Theta_0) - \frac{\partial \mathbf{R}(\hat{\Theta}_r)}{\partial \Theta}\Lambda_r = [\mathbf{0}],$$

$$\frac{\partial \mathbf{R}(\Theta_+)'}{\partial \Theta}(\hat{\Theta}_r - \Theta_0) = [\mathbf{0}],$$

where $\Theta_*$ and $\Theta_+$ each lie between $\Theta_0$ and $\hat{\Theta}_r$.[7] Note the second equation incorporates the fact that the first term in the Taylor series, $\mathbf{R}(\Theta_0) - \mathbf{r}$, is zero

[6] $\Lambda_r$ is the random vector whose outcome is $\lambda_r$.

[7] We must alert the reader to a technical point that we suppress notationally regarding the use of Taylor series representations of *vector* functions. Specifically, such a representation is actually a collection of Taylor series representations, one for each entry in the vector function, say $\mathbf{f}(\mathbf{z})$. As such, the point of evaluation of the final derivative terms in each Taylor series can differ for each coordinate function. For example, if $\mathbf{f}(\mathbf{z})$ is $(j \times 1)$, then in the Taylor series representation

$$\mathbf{f}(\mathbf{z}) = \mathbf{f}(\mathbf{z}_0) + \frac{\partial \mathbf{f}(\mathbf{z}_*)'}{\partial \mathbf{z}}(\mathbf{z} - \mathbf{z}_0),$$

it can be that each *row* of $\partial \mathbf{f}(\mathbf{z}_*)\partial \mathbf{z}'$ must be evaluated at a *different* $\mathbf{z}_* = \tau \mathbf{z} + (1 - \tau)\mathbf{z}_0$, $\tau \in [0, 1]$. Having alerted the reader to this situation, we will tacitly assume henceforth that this is understood. What is most important for our purposes is that $\mathbf{z}_* \to \mathbf{z}_0$ as $\mathbf{z} \to \mathbf{z}_0$, so in this case, in the limit, all rows of $\partial \mathbf{f}(\mathbf{z}_*)/\partial \mathbf{z}'$ will be evaluated at the same point.

under $H_0$, which we now assume is true. Premultiplying the first of the preceding equations by $n^{-1/2}$ and the second by $n^{1/2}$ leads to the matrix equation

$$\left[\begin{array}{c|c} -n^{-1}\dfrac{\partial^2 \ln L(\Theta_*;\mathbf{X})}{\partial\Theta\partial\Theta'} & \dfrac{\partial\mathbf{R}(\hat{\Theta}_r)}{\partial\Theta} \\ \hline \dfrac{\partial\mathbf{R}(\Theta_+)'}{\partial\Theta} & [0] \end{array}\right] \left[\begin{array}{c} n^{1/2}(\hat{\Theta}_r - \Theta_0) \\ \hline n^{-1/2}\Lambda_r \end{array}\right] = \left[\begin{array}{c} n^{-1/2}\dfrac{\partial \ln L(\Theta_0;\mathbf{X})}{\partial\Theta} \\ \hline [0] \end{array}\right] \qquad (*)$$

Observe that the right-hand side of the matrix equation $(*)$ converges in distribution to the vector $\left[\begin{array}{c}\mathbf{Z}\\ \hline [0]\end{array}\right]$, where $\mathbf{Z} \sim N([0], \mathbf{M}(\Theta_0))$.

In examining the *asymptotic* behavior of $n^{1/2}(\hat{\Theta}_r - \Theta_0)$ and $n^{-1/2}\Lambda_r$, it will be permissible by Theorem 5.9 to write the first matrix of $(^*)$ as

$$\left[\begin{array}{c|c} \mathbf{M}(\Theta_0) & \dfrac{\partial\mathbf{R}(\Theta_0)}{\partial\Theta} \\ \hline \dfrac{\partial\mathbf{R}(\Theta_0)'}{\partial\Theta} & [0] \end{array}\right]$$

since $\hat{\Theta}_r$, and thus $\Theta_*$ and $\Theta_+$, $\xrightarrow{p} \Theta_0$, $\partial\mathbf{R}(\Theta)/\partial\Theta$ is continuous, and

$$n^{-1}\frac{\partial^2 \ln L(\Theta_*;\mathbf{X})}{\partial\Theta\partial\Theta'} \xrightarrow{p} -\mathbf{M}(\Theta_0).$$

Using partitioned inversion,[8] $n^{-1/2}\Lambda_r$ can be solved for, yielding

$$n^{-1/2}\Lambda_r = \left[\frac{\partial\mathbf{R}(\Theta_0)'}{\partial\Theta}\mathbf{M}(\Theta_0)^{-1}\frac{\partial\mathbf{R}(\Theta_0)}{\partial\Theta}\right]^{-1}\frac{\partial\mathbf{R}(\Theta_0)'}{\partial\Theta}\mathbf{M}(\Theta_0)^{-1}\left[n^{-1/2}\frac{\partial \ln L(\Theta_0;\mathbf{X})}{\partial\Theta}\right].$$

If $\mathbf{G} = \partial\mathbf{R}(\Theta_0)/\partial\Theta'$ has full row rank, it follows by Slutsky's theorems that

$$n^{-1/2}\Lambda_r \xrightarrow{d} N([0], [\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\mathbf{G}']^{-1}),$$

and

$$[\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\mathbf{G}']^{1/2}n^{-1/2}\Lambda_r \xrightarrow{d} N([0], \mathbf{I}).$$

Since the limiting distribution is unaffected by replacing $\mathbf{G}$ with $\partial\mathbf{R}(\hat{\Theta}_r)/\partial\Theta'$ and $\mathbf{M}(\Theta_0)$ by $[-n^{-1}(\partial^2 \ln L(\hat{\Theta}_r;\mathbf{X})/\partial\Theta\partial\Theta')]$, we finally have (recall $\overset{d}{=}$ means

[8] Note since the lower right block of the matrix is [0], Lemma 8.3 cannot be used. One can use the following result for symmetric matrices (H. Theil (1971), *Principles of Econometrics*. New York: John Wiley, p. 18), proven by direct multiplication:

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{C} \\ \hline \mathbf{C}' & \mathbf{B} \end{array}\right]^{-1} = \left[\begin{array}{c|c} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{C}(\mathbf{B}-\mathbf{C}'\mathbf{A}^{-1}\mathbf{C})^{-1}\mathbf{C}'\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{C}(\mathbf{B}-\mathbf{C}'\mathbf{A}^{-1}\mathbf{C})^{-1} \\ \hline -(\mathbf{B}-\mathbf{C}'\mathbf{A}^{-1}\mathbf{C})^{-1}\mathbf{C}'\mathbf{A}^{-1} & (\mathbf{B}-\mathbf{C}'\mathbf{A}^{-1}\mathbf{C})^{-1} \end{array}\right].$$

equivalent in limiting distribution)

$$\mathbf{W} = n^{-1}\mathbf{\Lambda}_r'[\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\mathbf{G}']\mathbf{\Lambda}_r \overset{d}{=} \mathbf{\Lambda}_r' \frac{\partial \mathbf{R}(\hat{\Theta}_r)'}{\partial \Theta}\left[-\frac{\partial^2 \ln L(\hat{\Theta}_r;\mathbf{X})}{\partial\Theta\partial\Theta'}\right]^{-1}\frac{\partial \mathbf{R}(\hat{\Theta}_r)}{\partial\Theta}\mathbf{\Lambda}_r,$$

$$\xrightarrow{d} \chi_q^2 \quad \text{under } H_0.$$

The size $\alpha$ test indicated in part (2) of the theorem follows immediately from the limiting distribution of $\mathbf{W}$. Consistency can be demonstrated by retracing the previous argument beginning with the initial Taylor series expansion of the first-order conditions, but with $\mathbf{R}(\Theta_0) - \mathbf{r} = \boldsymbol{\phi} \neq [\mathbf{0}]$ so that $[\mathbf{0}]$ is replaced by $-n^{1/2}\boldsymbol{\phi}$ in the vector on the right-hand side of the equality in $(*)$. The net result is that $p(w > \tau) \to 1$ as $n \to \infty\ \forall\ \tau > 0$, and so $H_0$ is rejected for any critical region of size $\in (0, 1)$ of the form $[\chi^2_{q;\alpha}, \infty)$. See S. D. Silvey, *The Lagrangian Multiplier Test*, pp. 389–407, for further discussion.

The alternative representation of $\mathbf{W}$ in part (3) of the theorem follows immediately from the first-order conditions $\partial \ln L(\hat{\Theta}_r;\mathbf{X})/\partial\Theta = (\partial\mathbf{R}(\hat{\Theta}_r)/\partial\Theta)\mathbf{\Lambda}_r$. ■

In certain applications, the LM test can have a computational advantage over the GLR test since the latter involves *both* the restricted and unrestricted estimates of $\Theta$, whereas the former requires *only* the restricted estimates of $\Theta$. In the following example, the relative convenience of the LM test is illustrated in the case of testing a hypothesis relating to the Gamma distribution. Note the GLR approach in this case would be complicated by the fact that the unrestricted MLE cannot be obtained in closed form. Obtaining the restricted MLE in the following case is relatively straightforward.

**Example 10.11** LM Test of $H_0 : \alpha = 1$ (Exponential Family) Versus $H_a : \alpha \neq 1$ in Gamma Population Distribution

The operating life of a new CD player produced by a major manufacturer is considered to be Gamma distributed as

$$z \sim \frac{1}{\beta^\alpha \Gamma(\alpha)} z^{\alpha-1} e^{-z/\beta} I_{(0,\infty)}(z).$$

The marketing department is contemplating a warranty policy on the CD player and wants to test the hypothesis that the CD player is "as good as new while operating," i.e., does the CD player's operating life adhere to an exponential population distribution? To test the null hypothesis $H_0 : \alpha = 1$ versus $H_a : \alpha \neq 1$, consider performing a size .05 LM test using a sample of 50 observations on the operating lives of the new CD player.

The likelihood function is (suppressing the indicator function)

$$L(\alpha, \beta; \mathbf{x}) = \frac{1}{\beta^{n\alpha}\Gamma^n(\alpha)} \prod_{i=1}^{n} x_i^{\alpha-1} e^{-\sum_{i=1}^n x_i/\beta},$$

with $n = 50$. We know from Ex. 8.15 that the conditions of Theorem 8.18 apply, and so Theorem 10.7 is applicable. Under the constraint $\alpha = 1$, the restricted

ML estimate is the value of $\hat{\beta}$ such that

$$\hat{\beta} = \arg\max_{\beta} \frac{1}{\beta^{50}} e^{-\sum_{i=1}^{50} x_i/\beta},$$

which we know to be the sample mean, $\bar{x}$.

To implement the LM test, consider calculating the LM statistic as indicated in Theorem 10.7.3. The second order derivatives of $\ln L(\alpha, \beta; \mathbf{x})$ are given in Ex. 8.15. In order to evaluate these derivatives at $\alpha = 1$ and $\beta = \bar{x}$, note that (to five decimal places)

$$\frac{d\Gamma(\alpha)}{d\alpha}\Big|_{\alpha=1} = -.57722 \quad \text{and} \quad \frac{d^2\Gamma(\alpha)}{d\alpha^2}\Big|_{\alpha=1} = 1.64493$$

(see M. Abramowitz and I. Stegun (1970), *Handbook of Mathematical Functions*, New York: Dover Publications, pp. 258–260, or else calculate numerically on a computer). Then letting $\Theta = [\alpha, \beta]'$,

$$\frac{\partial^2 \ln L(\Theta; \mathbf{x})}{\partial\Theta\partial\Theta'}\Big|_{\alpha=1,\beta=\bar{x}} = n\left[\begin{array}{c|c} -1.31175 & -\bar{x}^{-1} \\ \hline -\bar{x}^{-1} & -\bar{x}^{-2} \end{array}\right],$$

where in this application, $n = 50$.

The first derivatives of $\ln L(\alpha, \beta; \mathbf{x})$ are given in Ex. 8.10, and when evaluated at $\alpha = 1$ and $\beta = \bar{x}$, they yield

$$\frac{\partial \ln L(\alpha, \beta; \mathbf{x})}{\partial\alpha}\Big|_{\substack{\alpha=1\\ \beta=\bar{x}}} = -n\ln\bar{x} + .57722n + \sum_{i=1}^{n} \ln x_i,$$

$$\frac{\partial \ln L(\alpha; \beta; \mathbf{x})}{d\beta}\Big|_{\substack{\alpha=1\\ \beta=\bar{x}}} = -n\bar{x}^{-1} + n\bar{x}^{-1} = 0.$$

The LM test statistic then takes the form (note that $\hat{\theta}_r = [1, \bar{x}]'$

$$w = \frac{\partial \ln L(\hat{\theta}_r; \mathbf{x})'}{\partial\Theta}\left[-\frac{\partial^2 \ln L(\hat{\theta}_r; \mathbf{x})}{\partial\Theta\partial\Theta'}\right]^{-1}\frac{\partial \ln L(\hat{\theta}_r; \mathbf{x})}{\partial\Theta}$$

$$= \frac{n(\ln\bar{x}_G - \ln\bar{x} + .57722)}{(.31175)},$$

where $\bar{x}_G = (\prod_{i=1}^{n} x_i)^{1/n}$ is the geometric mean of the $x_i$'s, and $n = 50$ in this application. The test rule is then

$$w = 160.38492\left(\ln\left(\frac{\bar{x}_G}{\bar{x}}\right) + .57722\right)\begin{bmatrix}\geq\\<\end{bmatrix} 3.84146 \Rightarrow \begin{bmatrix}\text{reject } H_0\\ \text{accept } H_0\end{bmatrix}$$

where $\chi^2_{1;.05} = 3.84146$.

Suppose the 50 observations on CD player lifetimes yielded $\bar{x} = 3.00395$ and $\bar{x}_G = 2.68992$. The value of the LM test statistic is $w = 74.86822$, and since $w \geq 3.84146$, $H_0$ is rejected. We conclude that the population distribution is *not* exponential at significance level .05. (In actuality, $\bar{x}$ and $\bar{x}_G$ were simulated from a random sample of 50 observations from a Gamma(2, 10) population distribution, and so the test made the *correct* decision in this case). □

Analogous to the limiting distribution of $-2ln\lambda(\mathbf{X})$ under local alternatives in the case of the GLR test, the limiting distribution of the LM statistic is noncentral $\chi^2$ under general conditions.

**Theorem 10.8** (Asymptotic Distribution of LM Test of $H_0$: $\mathbf{R}(\Theta) = \mathbf{r}$ Versus $H_a$: $\mathbf{R}(\Theta) \neq \mathbf{r}$ For Local Alternatives When $H_0$ Is False) *Consider the LM test of* $H_0$: $\mathbf{R}(\Theta) = \mathbf{r}$ *versus* $H_a$: $\mathbf{R}(\Theta) \neq \mathbf{r}$ *under the conditions and notation of Theorem 10.7, and let a sequence of local alternatives be defined by* $H_{an}$: $\mathbf{R}(\Theta) = \mathbf{r}+n^{-1/2}\phi$. *Then the limiting distribution of the LM statistic is noncentral* $\chi^2$ *as*

$$\mathbf{W} = \Lambda_r' \frac{\partial \mathbf{R}(\hat{\Theta}_r)'}{\partial \Theta} \left[ -\frac{\partial^2 \ln L(\hat{\Theta}_r; \mathbf{X})}{\partial \Theta \partial \Theta'} \right]^{-1} \frac{\partial \mathbf{R}(\hat{\Theta}_r)}{\partial \Theta} \Lambda_r \xrightarrow{d} \chi_q^2(\lambda),$$

*where the noncentrality parameter is as follows*:

$$\lambda = \frac{1}{2}\phi' \left[ \frac{\partial \mathbf{R}(\Theta_0)'}{\partial \Theta} \mathbf{M}(\Theta_0)^{-1} \frac{\partial \mathbf{R}(\Theta_0)}{\partial \Theta} \right]^{-1} \phi.$$

**Sketch of the Proof** Assuming $\mathbf{R}(\Theta) - \mathbf{r} = n^{-1/2}\phi \to [0]$, as implied by the sequence of local alternatives, it follows that the sequence of restricted MLEs of $\Theta_0$ is such that $\hat{\Theta}_r \xrightarrow{p} \Theta_0$. The matrix equation (∗) in the proof of Theorem 10.7 applies here, except that [0] in the vector on the right-hand side of the equality changes to $-\phi$ (since the Taylor series expansion of $\mathbf{R}(\hat{\Theta}_r) - \mathbf{r} = [0]$ is now $n^{-1/2}\phi + \partial \mathbf{R}(\Theta_+)/\partial \Theta'(\hat{\Theta}_r - \Theta_0) = [0]$). Following the proof of Theorem 10.7 in applying partitioned inversion and the convergence result of Theorem 5.9, the revised matrix equation (∗) yields

$$n^{-1/2}\Lambda_r \overset{d}{=} \left[ \frac{\partial \mathbf{R}(\Theta_0)'}{\partial \Theta} \mathbf{M}(\Theta_0)^{-1} \frac{\partial \mathbf{R}(\Theta_0)}{\partial \Theta} \right]^{-1} \left[ \frac{\partial \mathbf{R}(\Theta_0)'}{\partial \Theta} \mathbf{M}(\Theta_0)^{-1} \left( n^{-1/2} \frac{\partial \ln L(\Theta_0; \mathbf{X})}{\partial \Theta} \right) - \phi \right].$$

By Slutsky's theorem, the rightmost bracketed term converges in distribution to

$$N\left( -\phi, \frac{\partial \mathbf{R}(\Theta_0)'}{\partial \Theta} \mathbf{M}(\Theta_0)^{-1} \frac{\partial \mathbf{R}(\Theta_0)}{\partial \Theta} \right),$$

and letting $\mathbf{G} = \partial \mathbf{R}(\Theta_0)/\partial \Theta'$, it follows that

$$[\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\mathbf{G}']^{1/2} n^{-1/2} \Lambda_r \xrightarrow{d} N\left( -[\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\mathbf{G}']^{-1/2}\phi, \mathbf{I} \right).$$

Finally, since $\partial \mathbf{R}(\hat{\Theta}_r)/\partial \Theta' \xrightarrow{p} \mathbf{G}$ and $n^{-1}(\partial^2 \ln L(\hat{\Theta}_r; \mathbf{X})/\partial \Theta \partial \Theta') \xrightarrow{p} -\mathbf{M}(\Theta_0)$, then

$$\left[ \frac{\partial \mathbf{R}(\hat{\Theta}_r)'}{\partial \Theta} \left[ -n^{-1} \frac{\partial^2 \ln L(\hat{\Theta}_r; \mathbf{X})}{\partial \Theta \partial \Theta'} \right]^{-1} \frac{\partial R(\hat{\Theta}_r)}{\partial \Theta} \right]^{1/2} \xrightarrow{p} [\mathbf{G}\mathbf{M}(\Theta_0)^{-1}\mathbf{G}']^{1/2}$$

so that by Slutsky's theorem,

$$w = \Lambda_r' \frac{\partial \mathbf{R}(\hat{\Theta}_r)'}{\partial \Theta} \left[ -\frac{\partial^2 \ln L(\hat{\Theta}_r; \mathbf{X})}{\partial \Theta \partial \Theta'} \right]^{-1} \frac{\partial \mathbf{R}(\hat{\Theta}_r)}{\partial \Theta} \Lambda_r \xrightarrow{\mathrm{d}} \chi_q^2(\lambda),$$

with $\lambda = (1/2)\phi'[\mathbf{GM}(\Theta_0)^{-1}\mathbf{G}']^{-1}\phi$ (see Section 10.9 for properties of the noncentral $\chi^2$-distribution). ■

Given Theorem 10.8, the asymptotic power of the LM test as a function of λ can be calculated and graphed. The asymptotic power function is monotonically increasing in the noncentrality parameter λ. The LM test is also *asymptotically unbiased* for testing $H_0$: $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ versus $H_a$: $\mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}$ since the test is consistent. Being that the LM statistic has the same limiting distribution under local alternatives as the statistic $-2[\ln L(\hat{\Theta}_r; X) - \ln L(\hat{\Theta}; x)]$ used in the GLR test, the two procedures cannot be distinguished on the basis of asymptotic power considerations. In applications the choice of whether to use a GLR or LM test is often made on the basis of convenience. Comparisons of the finite sample properties of the tests continue to be researched and are generally a case-by-case affair.

Historically, the LM test procedure is the most recent of the GLR, Wald, and LM triad of tests, listed in order of discovery. Recently, substantial progress has been made in applying the LM approach to a wide range of testing contexts in the econometrics literature. Interested readers can consult the article by R. F. Engle, *Lagrange Multiplier Tests*, and the book by L. G. Godfrey, *Misspecification Tests*, to begin further reading.

## 10.5 Wald Tests

The Wald test for testing statistical hypotheses utilizes a third alternative measure of the discrepancy between restricted (by $H_0$) and unrestricted estimates of the parameters of the joint density of a random sample in order to define a test rule. In particular, the Wald test assesses the significance of the difference between the unrestricted estimate of $\mathbf{R}(\Theta)$ and the value of $\mathbf{R}(\Theta)$ that is specified in $H_0$, $\mathbf{r}$. Significantly large values of $\mathbf{R}(\hat{\Theta}) - \mathbf{r}$ indicate significant discrepancies between the hypothesized value of $\mathbf{R}(\Theta)$ and the unrestricted estimate of $\mathbf{R}(\Theta)$ provided by the data, suggesting that $H_0$ is false. It is not necessary that the estimator $\hat{\Theta}$ be a MLE for the Wald test to be applicable. In fact, one need not specify a likelihood function at all to perform the test, and in this respect the Wald test is more general than either the GLR or LM test procedure, as presented heretofore.

**Theorem 10.9** (Wald Test of Asymptotic Size $\alpha$ When $\hat{\Theta} \overset{a}{\sim} N(\Theta_0, n^{-1}\Sigma)$) *Let the random sample* $(X_1, \ldots, X_n)$ *have the joint probability density function* $f(x_1, \ldots, x_n; \Theta_0)$, *let* $\hat{\Theta}$ *be a consistent estimator for* $\Theta_0$ *such that* $n^{1/2}(\hat{\Theta} - \Theta_0) \xrightarrow{\mathrm{d}} N([\mathbf{0}], \Sigma)$, *and let* $n\hat{\Sigma}_n$ *be a consistent estimator of the positive definite* $\Sigma$. *Furthermore, let the* *null and alternative hypotheses be of the form $H_0$: $\mathbf{R}(\Theta) = \mathbf{r}$ and $H_a$: $\mathbf{R}(\Theta) \neq \mathbf{r}$, where $\mathbf{R}(\Theta)$ is a $(q \times 1)$ continuously differentiable vector function of $\Theta$ for which $q \leq k$ and $\mathbf{R}(\Theta)$ contains no redundant coordinate functions. Finally, let $\partial\mathbf{R}(\Theta_0)/\partial\Theta'$ have full row rank. Then a consistent and asymptotic size $\alpha$ test of $H_0$ versus $H_a$ is given by*

$$w\begin{bmatrix}\geq\\<\end{bmatrix}\chi^2_{q;\alpha} \Rightarrow \begin{bmatrix}\textit{reject } H_0\\ \textit{accept } H_0\end{bmatrix},$$

*where*

$$W = \left[\mathbf{R}(\hat{\Theta}) - \mathbf{r}\right]' \left[\frac{\partial\mathbf{R}(\hat{\Theta})'}{\partial\Theta}\hat{\boldsymbol{\Sigma}}_n\frac{\partial\mathbf{R}(\hat{\Theta})}{\partial\Theta}\right]^{-1}\left[\mathbf{R}(\hat{\Theta}) - \mathbf{r}\right].$$

**Proof** Under the stated assumptions, Theorem 5.40 is applicable, so that

$$n^{1/2}\left[\mathbf{R}(\hat{\Theta}) - \mathbf{r}\right] \xrightarrow{d} N\left([0], \frac{\partial\mathbf{R}(\Theta_0)'}{\partial\Theta}\boldsymbol{\Sigma}\frac{\partial\mathbf{R}(\Theta_0)}{\partial\Theta}\right)$$

under $H_0$. Letting $\boldsymbol{\Sigma}_r$ represent the covariance matrix of this limiting distribution, it follows that

$$n^{1/2}\boldsymbol{\Sigma}_r^{-1/2}\left[\mathbf{R}(\hat{\Theta}) - \mathbf{r}\right] \xrightarrow{d} N([0], \mathbf{I}_q),$$

so that by Theorem 5.3

$$n\left[\mathbf{R}(\hat{\Theta}) - \mathbf{r}\right]'\boldsymbol{\Sigma}_r^{-1}\left[\mathbf{R}(\hat{\Theta}) - \mathbf{r}\right] \xrightarrow{d} \chi^2_q.$$

Now note that

$$\left[\frac{\partial\mathbf{R}(\hat{\Theta})'}{\partial\Theta}(n\hat{\boldsymbol{\Sigma}}_n)\frac{\partial\mathbf{R}(\hat{\Theta})}{\partial\Theta}\right]^{-1} \xrightarrow{p} \boldsymbol{\Sigma}_r^{-1}$$

by the continuity of $\partial\mathbf{R}(\Theta)/\partial\Theta$ and the continuity of the inverse matrix function, in which case it follows from Theorem 5.9 that $W \xrightarrow{d} \chi^2_q$, so that the test as defined above has size $\alpha$ asymptotically.

Regarding consistency, assume $\alpha > 0$ and thus $c\chi^2_{q;\alpha} < \infty$. Since $\hat{\Theta} \xrightarrow{p} \Theta_0$ and $\mathbf{R}(\Theta)$ is continuous, $\mathbf{R}(\hat{\Theta}) - \mathbf{r} \xrightarrow{p} \phi \neq [0]$, assuming $H_a$: $\mathbf{R}(\Theta) \neq \mathbf{r}$ is true. Then $n^{-1}W \xrightarrow{p} \phi'\boldsymbol{\Sigma}_r^{-1}\phi > 0$, so that $\lim_{n\to\infty} P(w \geq c) = 1$ for any $c < \infty$, and the test is consistent. ■

**Example 10.12** Revisit Ex. 10.5, and suppose that in addition to a hypothesis regarding the mean life of the computer screen, a hypothesis regarding the variance of the operating life is to be investigated. In particular, suppose $H_0$: $\theta^2 = 1$ versus $H_a$: $\theta^2 \neq 1$ is under consideration. We know in this case that

$$n^{1/2}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2), \quad \text{and} \quad n\hat{\boldsymbol{\Sigma}}_n = n\left(\frac{S^2}{n}\right) \xrightarrow{p} \sigma^2.$$

Thus, Theorem 10.9 is applicable, with $\mathbf{R}(\theta) = \theta^2$. Since $d\mathbf{R}(\theta)/d\theta = 2\theta$, it follows that a Wald test of asymptotic size .05 can be defined as

$$w \begin{bmatrix} \geq \\ < \end{bmatrix} 3.84 \Rightarrow \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix}$$

where $P(x \geq 3.84) = .05$ when $X \sim \chi^2_1$, and letting $n = 10$,

$$w = (\hat{\theta}^2 - 1)' \left[ \frac{4\hat{\theta}^2 s^2}{10} \right]^{-1} (\hat{\theta}^2 - 1) = \frac{10(\bar{x}^2 - 1)^2}{(4\bar{x}^2 s^2)}.$$

Supposing that $\bar{x} = 1.4$ and $s^2 = 2.1$, then $w = .56 < 3.84$, and thus $H_0$ would *not* be rejected using the Wald test. □

The asymptotic distribution of the Wald statistic when $H_0: \mathbf{R}(\Theta) = \mathbf{r}$ is false can be established for the case of *local alternatives* and leads to a noncentral $\chi^2$-distribution, as in the case of the LM and GLR procedures.

**Theorem 10.10** (Asymptotic Distribution of Wald Statistic Under $H_a: \mathbf{R}(\Theta) \neq \mathbf{r}$) *Let the conditions and notation of Theorem 10.9 hold for testing* $H_0: \mathbf{R}(\Theta) = \mathbf{r}$ *versus* $H_a: \mathbf{R}(\Theta) \neq \mathbf{r}$. *Then for a sequence of local alternatives* $H_{an}: \mathbf{R}(\Theta) = \mathbf{r} + n^{-1/2}\phi$, *it follows that* $W \xrightarrow{d} \chi^2_q(\lambda)$ *with noncentrality parameter*

$$\lambda = \frac{1}{2}\phi' \left[ \frac{\partial \mathbf{R}(\Theta_0)'}{\partial \Theta} \Sigma \frac{\partial \mathbf{R}(\Theta_0)}{\partial \Theta} \right]^{-1} \phi.$$

**Proof** It follows from Theorem 5.40 that $n^{1/2}(\mathbf{R}(\hat{\Theta}) - \mathbf{R}(\Theta_0)) \xrightarrow{d} N([\mathbf{0}], \mathbf{G}\Sigma\mathbf{G}')$, where $\mathbf{G} = \partial\mathbf{R}(\Theta_0)/\partial\Theta'$. Under the sequence of local alternatives, $\mathbf{R}(\Theta) = \mathbf{r} + n^{-1/2}\phi \to \mathbf{r}$ and $n^{1/2}(\mathbf{R}(\hat{\Theta}) - (\mathbf{r} + n^{-1/2}\phi)) \xrightarrow{d} N([\mathbf{0}], \mathbf{G}\Sigma\mathbf{G}')$, in which case $n^{1/2}(\mathbf{R}(\hat{\Theta}) - \mathbf{r}) \xrightarrow{d} N(\phi, \mathbf{G}\Sigma\mathbf{G}')$ and $[\mathbf{G}\Sigma\mathbf{G}']^{-1/2} n^{1/2}(\mathbf{R}(\hat{\Theta}) - \mathbf{r}) \xrightarrow{d} N([\mathbf{G}\Sigma\mathbf{G}']^{-1/2}\phi, \mathbf{I})$. Since $\partial\mathbf{R}(\hat{\Theta})/\partial\Theta' \xrightarrow{p} \mathbf{G}$ and $n\hat{\Sigma}_n \xrightarrow{p} \Sigma$, it follows from Theorem 5.9 and properties of the noncentral $\chi^2$-distribution that

$$W = n^{1/2}(\mathbf{R}(\hat{\Theta}) - \mathbf{r})' \left[ \frac{\partial \mathbf{R}(\hat{\Theta})'}{\partial \Theta} (n\hat{\Sigma}_n) \frac{\partial \mathbf{R}(\hat{\Theta})}{\partial \Theta} \right]^{-1} (\mathbf{R}(\hat{\Theta}) - \mathbf{r}) n^{1/2}$$

$$\overset{d}{=} n^{1/2}(\mathbf{R}(\hat{\Theta}) - \mathbf{r})'[\mathbf{G}\Sigma\mathbf{G}']^{-1}(\mathbf{R}(\hat{\Theta}) - \mathbf{r}) n^{1/2} \xrightarrow{d} \chi^2_q(\lambda),$$

with $\lambda = (1/2)\phi'(\mathbf{G}\Sigma\mathbf{G}')^{-1}\phi$. ■

**Example 10.13** Wald Test Based on a GMM Estimator

Revisit Ex. 8.28 concerning GMM estimation of $\beta$ in the demand function $\mathbf{Y} = \mathbf{X}\beta + \mathbf{V}$, when $\mathbf{X}$ is random and $\mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V} \neq [\mathbf{0}]$. The GMM estimator $\hat{\beta}_G = (\mathbf{z}'\mathbf{X})^{-1}\mathbf{z}'\mathbf{Y}$ as defined in the example is a consistent and asymptotically normal estimator of $\beta$ as $n^{1/2}(\hat{\beta}_G - \beta_0) \xrightarrow{d} N([\mathbf{0}], \sigma^2(\mathbf{A}'_{zx}\mathbf{A}^{-1}_{zz}\mathbf{A}_{zx})^{-1})$. A consistent estimator of the covariance matrix of the limiting distribution is given by $n\hat{\sigma}^2(\mathbf{X}'\mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{X})^{-1}$, where $\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_G)'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_G)/n$. Let $H_0\colon \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ be any hypothesis concerning linear combinations of the parameters of the demand function, where $\mathbf{R}$ has full row rank. Then a Wald test of $H_0$ having asymptotic size $\alpha$ can be based on outcomes of

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}}_G - \mathbf{r})'\left[\hat{\sigma}^2\mathbf{R}(\mathbf{X}'\mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_G - \mathbf{r}),$$

where the critical region is defined by $C_r^w = [\chi^2_{q;\alpha}, \infty)$. How would the test change if the null hypothesis were *nonlinear* in $\boldsymbol{\beta}$, i.e., $H_0\colon \mathbf{R}(\boldsymbol{\beta}) = \mathbf{r}$? □

The asymptotic power function of the Wald test is monotonically increasing in the noncentrality parameter $\lambda$ since $P(w \geq c; \lambda)$ for $c < \infty$ is monotonically increasing in $\lambda$ if $W \sim \chi^2_q(\lambda)$. The Wald test is also asymptotically unbiased for testing $H_0\colon \mathbf{R}(\Theta_0) = \mathbf{r}$ versus $H_0\colon \mathbf{R}(\Theta_0) \neq \mathbf{r}$ based on asymptotic power considerations since the test is consistent.

As we had noted at the end of our LM discussion, each of the tests in the GLR–LM–Wald triad has the same limiting distribution under sequences of local alternatives when all of the tests are based on the unrestricted and/or restricted MLEs of $\Theta$. As such, the tests cannot be distinguished on the basis of asymptotic power comparisons. If convenience dictates the choice of test procedure, then the Wald procedure will be most useful when the unrestricted ML estimate is relatively easy to obtain. Consideration of small sample properties requires a case-by-case analysis. The Wald test has flexibility in being quite generally applicable to a wide array of estimation contexts other than ML estimation. We note, however, that this flexibility is not exclusive to the Wald test. In particular, the LM approach has been gradually expanding into contexts more general than maximum likelihood estimation (see Engle, *Lagrange Multiplier Tests*, Godfrey, *Misspecification Tests*, and the references contained therein).

## 10.6 Tests in the GLM

In this section we examine some specific statistical tests that have been applied in the context of the GLM. The collection of tests that are presented here will allow testing of statistical hypotheses relating to values of the $\boldsymbol{\beta}$ and $\sigma^2$ parameters. We begin in the context of multivariate normality of the disturbance term of $\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and later examine asymptotically valid procedures that do not depend on the normality of $\boldsymbol{\varepsilon}$.

### Tests When $\varepsilon$ Is Multivariate Normal

In this subsection we begin by assuming that the classical assumptions of the GLM hold and in addition that $\boldsymbol{\varepsilon} \sim N([0], \sigma^2\mathbf{I})$ in $\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. We concentrate here on testing hypotheses relating to the value of *linear* combinations of $\boldsymbol{\beta}$, i.e., $\mathbf{R}\boldsymbol{\beta}$, and the value of $\sigma^2$. We will examine hypotheses concerning nonlinear functions of $\boldsymbol{\beta}$ when we examine the more general nonnormal case.

*Testing* $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ *When* $\mathbf{R}$ *Is* $q \times k$*: F-Tests* Under the prevailing assumptions, $\mathbf{Y} \sim N(\mathbf{x}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{x}'\mathbf{x})^{-1})$. Examine a Wald-type test of the null hypothesis $H_0$: $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ versus $H_a$: $\mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}$, where $\mathbf{R}$ is $(q \times k)$ with full row rank. We know that $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N([0], \sigma^2(\mathbf{n}^{-1}\mathbf{x}'\mathbf{x})^{-1})$, and thus $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N([0], \sigma^2\mathbf{Q}^{-1})$, assuming $n^{-1}\mathbf{x}'\mathbf{x} \to \mathbf{Q}$, a positive definite matrix. Also, $\hat{S}^2(n^{-1}\mathbf{x}'\mathbf{x})^{-1}$ is a consistent estimator of the covariance matrix of the limiting distribution for $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Then Theorem 10.9 suggests that outcomes of the Wald statistic

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'\left[\hat{S}^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}'\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})$$

can be used to test $H_0$, where the critical region $C_r^w = [\chi^2_{q;\alpha}\infty)$ defines an asymptotic size $\alpha$, asymptotically unbiased, and consistent test.

In this case, we have sufficient information to establish the small sample distribution of $W$, and so we will also be able to define an exact size $\alpha$ test of $H_0$. In particular, rewrite $W$ in the following fractional form:

$$W = \frac{\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)'\left[\sigma^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}'\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{\left[(n-k)\hat{S}^2/\sigma^2(n-k)\right]} = \frac{Y_q}{(Z_{n-k}/(n-k))}.$$

Note that the numerator random variable is the sum of the squares of $q$ independent standard normal variates since $\mathbf{V} = [\sigma^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \sim N([0], \mathbf{I}_q)$ under $H_0$, and the numerator can be represented as $\mathbf{V}'\mathbf{V}$. We know from Section 8.2 that $(n-k)\hat{S}^2/\sigma^2 \sim \chi^2_{n-k}$, and thus the denominator random variable is a $\chi^2_{n-k}$ random variable divided by its degrees of freedom. Finally, recall that $\hat{\boldsymbol{\beta}}$ and $\hat{S}^2$ are independent random variables (Section 8.2) so that $Y_q$ and $Z_{n-k}$ are independent. It follows from Section 6.7 that $W/q \sim F(q, n-k)$ under $H_0$,[9] so that $W$ itself is distributed as an $F$ random variable multiplied by $q$, where the $F$-distribution has $q$ numerator and $(n-k)$ denominator degrees of freedom.

An exact size $\alpha$ test of $H_0$: $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ versus $H_a$: $\mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}$ is defined by

$$w\begin{bmatrix}\geq \\ <\end{bmatrix} qF_\alpha(q, n-k) \Rightarrow \begin{bmatrix}\text{reject } H_0 \\ \text{accept } H_0\end{bmatrix},$$

where the value $F_\alpha(q, n-k)$ is such that $P(f \geq F_\alpha(q, n-k)) = \alpha$ when $F \sim F(q, n-k)$. This test is *unbiased* and is also *consistent* if $(\mathbf{x}'\mathbf{x})^{-1} \to [0]$ as $n \to \infty$. To see unbiasedness, note that if $\mathbf{R}\boldsymbol{\beta} - \mathbf{r} = \boldsymbol{\phi} \neq [0]$, so that $H_a$ is true, then $[\sigma^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r} - \boldsymbol{\phi}) \sim N([0], \mathbf{I})$, so that $[\sigma^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \sim N([\sigma^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{-1/2}\boldsymbol{\phi}, \mathbf{I})$. Thus, the numerator of the previous fractional representation of $W$ is a random variable having a *noncentral* $\chi^2$-distribution, i.e., $\mathbf{Y}_q \sim \chi^2_q(\lambda)$, with noncentrality parameter $\lambda = (1/2)\boldsymbol{\phi}'[\sigma^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{-1}\boldsymbol{\phi}$. It follows that $W/q$ is the ratio of a *noncentral* $\chi^2_q$ random variate to an independent

[9] Note that under $H_0$, $W/q$ is the ratio of two independent, central $\chi^2$-random variables, each divided by its respective degrees of freedom.

central $\chi^2_{n-k}$ variate, each divided by its respective degrees of freedom, and thus $W/q$ has a *noncentral* $F$-distribution, $F(q, n-k, \lambda)$ (see Section 10.9). Then since $P(w/q \geq c; \lambda)$ is monotonically increasing in $\lambda\ \forall\, c < \infty$, it follows that the power function of the $F$-test is strictly monotonically increasing in the noncentrality parameter, $\lambda$, so that the test is unbiased.

To motivate consistency, first note that if $(\mathbf{x}'\mathbf{x})^{-1} \to [\mathbf{0}]$, so that $\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}' \to [\mathbf{0}]$, then all of the characteristic roots, $\lambda_1, \ldots, \lambda_q$, of $\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}'$ limit to 0 as $n \to \infty$ since $\sum_{i=1}^{q} \lambda_i = \operatorname{tr}\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}' \to 0$ and $\lambda_i \geq 0\ \forall\, i$ by the positive semidefiniteness of $\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}'$. Since the characteristic roots of $[\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{-1}$ are the reciprocals of the characteristic roots of $\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}'$, all of the characteristic roots of $[\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{-1}$, say $\xi_1, \ldots, \xi_q$, will then increase without bound as $n \to \infty$. If $\mathbf{R}\boldsymbol{\beta} - \mathbf{r} = \boldsymbol{\phi} \neq [\mathbf{0}]$, it follows from results on the extrema of quadratic forms (see C. R. Rao, *Statistical Inference*, p. 62) that

$$\lambda = \frac{1}{2}\boldsymbol{\phi}'\left[\sigma^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}'\right]^{-1}\boldsymbol{\phi} \geq \frac{1}{2\sigma^2}\xi_s\boldsymbol{\phi}'\boldsymbol{\phi},$$

where $\xi_s$ is the smallest characteristic root of $[\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{-1}$, and thus if $\xi_s \to \infty$ and $\boldsymbol{\phi}'\boldsymbol{\phi} > 0$, then $\lambda \to \infty$. Finally, note that because $\hat{S}^2 \xrightarrow{p} \sigma^2$, it follows that $W \xrightarrow{d} \chi^2_q$ under $H_0$, and $qF_\alpha(q, n-k) \to \chi^2_{q;\alpha} = c < \infty$. Therefore, since $\lim_{\lambda\to\infty} P(w \geq c; \lambda) = 1$ when $w \sim \chi^2_q(\lambda)$, the test sequence is consistent.

**Example 10.14** Hypothesis Test Concerning Significance of Explanatory Variable Matrix in Representing E$Y$

In this example we examine a size $\alpha$ test of the null hypothesis that *none* of the explanatory variables, other than the intercept, is significant in explaining the expected value of the dependent variable in a GLM, $\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim N([\mathbf{0}], \sigma^2\mathbf{I})$. We suppose that the first column of $\mathbf{x}$ is a vector of 1's, so that $\beta_1$ represents the intercept. Then, from the foregoing discussion, a test of the hypothesis that none of the nonintercept explanatory variables is significant, i.e., $H_0: \beta_i = 0, i = 2, \ldots, k$, versus $H_a$: not $H_0$ can be performed in terms of the $F$-statistic $W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'(\hat{S}^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})$, where $\mathbf{r}_{(k-1)\times 1} = [\mathbf{0}]$ and $\mathbf{R}_{(k-1)\times k} = [[\mathbf{0}]|\mathbf{I}_{k-1}]$. In the current context, the test statistic can be represented alternatively as $W = \hat{\boldsymbol{\beta}}_*'(\widehat{\mathbf{Cov}}(\hat{\boldsymbol{\beta}}_*))^{-1}\hat{\boldsymbol{\beta}}_*$, where $\hat{\boldsymbol{\beta}}_* = (\hat{\beta}_2, \ldots, \hat{\beta}_k)'$ and $\widehat{\mathbf{Cov}}(\hat{\boldsymbol{\beta}}_*)$ is the sub-matrix of $\hat{S}^2(\mathbf{x}'\mathbf{x})^{-1}$ referring to the estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}_*$. The test is defined by

$$w\begin{bmatrix}\geq\\<\end{bmatrix}(k-1)F_\alpha(k-1, n-k) \Rightarrow \begin{bmatrix}\text{reject } H_0\\ \text{accept } H_0\end{bmatrix}.$$

As a numerical illustration, suppose $\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ represented a relationship that was hypothesized to explain quarterly average stock price levels for a certain composite of stocks sensitive to international market conditions. Let

$$\mathbf{b} = [1.2, 2.75, 3.13]' \quad \text{and} \quad \hat{s}^2(\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} .25 & .15 & .31\\ & .9 & .55\\ \multicolumn{2}{c}{\text{(symmetric)}} & 1.2\end{bmatrix}$$

represent least-squares–based estimates of $\boldsymbol{\beta}$ and $\sigma^2(\mathbf{x}'\mathbf{x})^{-1}$, where $n = 30$, $b_1 = 1.2$ represents the estimated intercept of the relationship, and $b_2 = 2.75$ and $b_3 = 3.13$ represent estimated coefficients on quarterly average composite exchange rates and western country aggregate real income. The value of the $F$-statistic for testing the significance of the explanatory variables other than the intercept is given by

$$w = \mathbf{b}_*'[s^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{-1}]\mathbf{b}_* = [2.75 \quad 3.13]\begin{bmatrix} .9 & .55 \\ .55 & 1.2 \end{bmatrix}^{-1}\begin{bmatrix} 2.75 \\ 3.13 \end{bmatrix} = 10.8347.$$

Then $H_0: \beta_2 = \beta_3 = 0$ is *rejected* using a size .05 unbiased and consistent test, since

$$w = 10.8347 \geq 3.35 = F_{.05}(2, 27).$$

The interpretation of the result of the statistical test is then that *at least one* of the two explanatory variables is significant in explaining the expected value of the stock price composite. □

It turns out that there does *not* exist a UMP or UMPU test of $H_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ versus $H_a: \mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}$ when $\mathbf{R}$ has *two* or more rows. Next we will see that for the case where $\mathbf{R}$ is a single row vector, a UMPU test does exist.

*Testing* $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, $\mathbf{R}\boldsymbol{\beta} \leq \mathbf{r}$, *Or* $\mathbf{R}\boldsymbol{\beta} \geq \mathbf{r}$ *When* $\mathbf{R}$ *Is* $(1 \times k)$*: T-Tests* In order to test $H_0: \mathbf{R}\boldsymbol{\beta} = r$ versus $H_a: \mathbf{R}\boldsymbol{\beta} \neq r$ when $\mathbf{R}$ is a *row* vector, the $F$-statistic in the preceding subsection could be used without modification. However, since $\mathbf{R}\boldsymbol{\beta}$ is a scalar, an alternative representation of $W$ is possible as

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - r)^2/[\hat{S}^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}'] \sim F(1, n-k) \quad \text{under } H_0.$$

An alternative test statistic based on the $t$-distribution whose square is equal to $W$ is given by

$$T = (\mathbf{R}\hat{\boldsymbol{\beta}} - r)/[\hat{S}^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{1/2} \sim T_{(n-k)} \quad \text{under } H_0.$$

Justification for the $t$-distribution of $T$ can be provided from first principles (recall Problem 3(a) of Chapter 8), or else from the fact that the square of a $T_{(n-k)}$ random variable is a random variable having a $F(1, n-k)$ distribution. In terms of the $t$-statistic, the test is given by

$$t\begin{bmatrix} \in \\ \notin \end{bmatrix}(-\infty, -t_{\alpha/2}(n-k)] \cup [t_{\alpha/2}(n-k), \infty) \Rightarrow \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix},$$

where $t_{\alpha/2}(n-k)$ denotes the critical value of a $t$-distribution with $n-k$ degrees of freedom for which $P(t \geq t_{\alpha/2}(n-k)) = \alpha/2$.

It can be shown using the principles in Section 9.5 that the $t$-test (and thus the $F$-test) is a UMPU level-$\alpha$ test of $H_0$: $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ versus $H_a$: $\mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}$. We will defer the demonstration of the UMPU property, as it is quite involved. Unbiasedness could alternatively be demonstrated from the equivalence of the $t$-test and $F$-test or from the power function of the test. In the latter context, when $\mathbf{R}\beta - r = \phi \neq 0$, the aforementioned $T$-statistic has a **noncentral $t$-distribution** with noncentrality parameter $\lambda = \phi/[\sigma^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{1/2}$. The power function, $\pi_{C_r}(\lambda) = 1 - \int_{-t_{\alpha/2}}^{t_{\alpha/2}} f(t; n-k, \lambda)dt$ achieves its minimum value when $\lambda = 0$ (i.e., $H_0$ is true) and is strictly monotonically increasing as $\lambda \to \infty$ or $\lambda \to -\infty$ (see Section 9.6). Thus the test is *unbiased*.

Consistency of the $t$-test sequence follows from consistency of the $F$-test sequence. As alternative motivation for consistency, first note that since $T_{n-k} \xrightarrow{d} N(0,1)$ under $H_0$, then $t_{\alpha/2}(n-k) \to z_{\alpha/2}$, where $\int_{z_{\alpha/2}}^{\infty} N(z;0,1)dz = \alpha/2$. Based on the $t$-test, the null hypothesis will be rejected iff $|(\mathbf{R}\hat{\beta} - r)/\hat{s}| \geq t_{\alpha/2}(n-k)[\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{1/2}$. Since the $\varepsilon_i$'s are iid, $\hat{S} \xrightarrow{p} \sigma$, and assuming $(\mathbf{x}'\mathbf{x})^{-1} \to [\mathbf{0}]$, then $\hat{\beta} \xrightarrow{p} \beta$. All of the preceding results imply collectively that the condition for rejecting $H_0$ becomes $|\phi/\sigma| > 0$ in the limit. Thus when $\phi \neq 0$, $H_0$ is rejected with probability $\to 1$ as $n \to \infty$, and the test sequence is consistent.

**Example 10.15** Revisit Ex. 10.14, and test the individual null hypotheses that each of the non-intercept explanatory variables is significant in explaining the daily composite stock value. That is, test $H_0: \beta_i = 0$ versus $H_a: \beta_i \neq 0$, for $i = 2, 3$. Use a size .05 test for each test.

**Answer**: To test the first null hypothesis, set $\mathbf{R} = [0 \;\; 1 \;\; 0]$ and $r = 0$, and calculate the $t$-statistic as

$$t = \frac{\mathbf{Rb} - r}{(\hat{s}^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}')^{1/2}} = \frac{2.75}{(.9)^{1/2}} = 2.8988.$$

Then since

$$t = 2.8988 \in C_r^T = (-\infty, -t_{.025}(27)] \cup [t_{.025}(27)) = (-\infty, -2.052] \cup [2.052, \infty)$$

we *reject* $H_0: \beta_2 = 0$.

To test the second null hypothesis, set $\mathbf{R} = [0 \;\; 0 \;\; 1]$ and $r = 0$, and calculate the $t$-statistic as

$$t = \frac{\mathbf{Rb} - r}{(\hat{s}^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}')^{1/2}} = \frac{3.13}{(1.2)^{1/2}} = 2.8573.$$

The critical region is the same as above, so that $t \in C_r^T$ and $H_0: \beta_3 = 0$ is *rejected*. It appears that *both* explanatory variables are significant in explaining the composite stock value; however, see the next subsection for the appropriate interpretation of this *joint* conclusion in terms of the probability of type I error. □

From the previous example it is clear that a test of the significance of any individual explanatory variable in the GLM will be based on a statistic whose value is equal to the explanatory variable's associated parameter estimate, $b_i$, divided by the estimated standard error of the estimate, $\hat{s}_{(\hat{\beta}_i)}$, as $t = b_i/\hat{s}_{(\hat{\beta}_i)}$. The statistic will be $t$-distributed with $n-k$ degrees of freedom, and the critical region will be of the form $(-\infty, -t_{\alpha/2}(n-k)] \cup [t_{\alpha/2}(n-k), \infty)$.

The $t$-statistic can also be used to define level-$\alpha$ UMPU tests of $H_0: \mathbf{R}\boldsymbol{\beta} \le r$ versus $H_a: \mathbf{R}\boldsymbol{\beta} > r$ or of $H_0: \mathbf{R}\boldsymbol{\beta} \ge r$ versus $H_a: \mathbf{R}\boldsymbol{\beta} < r$. The critical regions defined in terms of the $t$-statistic are as follows:

| **Case** | $H_0$ | $H_a$ | $C_r$ |
|---|---|---|---|
| (1) | $\mathbf{R}\boldsymbol{\beta} \le r$ | $\mathbf{R}\boldsymbol{\beta} > r$ | $t \in [t_\alpha(n-k), \infty)$ |
| (2) | $\mathbf{R}\boldsymbol{\beta} \ge r$ | $\mathbf{R}\boldsymbol{\beta} < r$ | $t \in (-\infty, -t_\alpha(n-k)]$ |

The proof of the UMPU property can be based on the results of Section 9.5, but it is quite involved and is deferred. Unbiasedness can be verified from an investigation of the power functions of the tests. Examine case (1), and note that in terms of the noncentrality parameter, $\lambda = (\mathbf{R}\boldsymbol{\beta} - r)/[\sigma^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{1/2}$, of the noncentral $t$-distribution, an equivalent representation of $H_0$ and $H_a$ is given by $H_0: \lambda \le 0$ versus $H_a: \lambda > 0$. Since the power function of the test

$$\pi_{C_r}(\lambda) = \int_{t_\alpha}^{\infty} f(t; n-k, \lambda)dt$$

is strictly monotonically increasing in $\lambda$, the test is an unbiased test. Unbiasedness of the test in case (2) follows from an analogous argument and is left to the reader.

Regarding consistency of the respective test sequences, again examine case (1) and recall that $t_\alpha(n-k) \to z_\alpha$, where $\int_{z_\alpha}^{\infty} N(z; 0, 1)dz = \alpha$. Based on the $t$-test, the null hypothesis will be rejected iff $(\mathbf{R}\mathbf{b} - r)/\hat{s} \ge t_\alpha(n-k)[\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{1/2}$. Given that $\hat{S} \xrightarrow{p} \sigma$, $(\mathbf{x}'\mathbf{x})^{-1} \to [\mathbf{0}]$, and thus $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$, the condition for rejecting $H_0$ becomes $\phi/\sigma > 0$ in the limit. Then if $H_a$ is true, so that $\phi > 0$, the null hypothesis will be rejected with probability $\to 1$ as $n \to \infty$, and the test sequence is consistent. Consistency for case (2) follows analogously.

**Example 10.16** A constant elasticity quarterly demand relationship is specified as $\ln \mathbf{Y} = \beta_1\iota + \beta_2 \ln \mathbf{p} + \beta_3 \ln \mathbf{p}_s + \beta_4 \ln \mathbf{m} + \boldsymbol{\varepsilon}$ where $\mathbf{Y}$ is quantity demanded of a commodity, $\mathbf{p}$ is its price, $\mathbf{p}_s$ is an index of substitute prices, $\mathbf{m}$ is disposable income, $\iota$ is a vector of 1's, and $\boldsymbol{\varepsilon} \sim N([\mathbf{0}], \sigma^2\mathbf{I})$. Forty quarterly observations were used to calculate the least-squares–based estimates

$$\mathbf{b} = [\,10.1 \quad -.77 \quad .41 \quad .56\,]'$$

and

$$\hat{s}^2(\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} 3.79 & .05 & .02 & .01 \\ & .04 & .9 \times 10^{-3} & .1 \times 10^{-3} \\ & & .03 & .1 \times 10^{-3} \\ \text{(symmetric)} & & & .02 \end{bmatrix}.$$

**a.** Using a size .05 test, test the significance of the matrix of explanatory variable values other than the intercept.

**b.** Test the individual significance of the effect of price, of the substitute price index, and of disposable income, each at the .05 level of significance.

**c.** Test whether the demand equation is homogeneous degree zero in prices and income.
**d.** Test whether the own price elasticity $\leq -1$.
**e.** Test whether the income elasticity is unitary elastic or larger.

**Answer**: For each of the hypotheses being tested, we list the specification of $\mathbf{R}$, $\mathbf{r}$, the relationship hypothesized between $\mathbf{R}\boldsymbol{\beta}$ and $\mathbf{r}$, whether the statistic used is an $F$-statistic ($w$) or a $t$-statistic ($t$), the value of the calculated statistic, the critical region of the test statistic with degrees of freedom and distribution indicated, and the outcome of the test.

| Test | R | $\mathbf{R}\boldsymbol{\beta}\begin{pmatrix}\leq\\=\\\geq\end{pmatrix}$ | r | w or t | $C_r$ | Outcome |
|---|---|---|---|---|---|---|
| (a) | $\begin{bmatrix}0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}$ | $=$ | $\begin{bmatrix}0\\0\\0\end{bmatrix}$ | $w = 36.6237$ | $w > 2.86$<br>$> F_{.05}(3, 36)$ | reject $H_0$ |
| (b) | $[0\ 1\ 0\ 0]$ | $=$ | 0 | $t = -3.85$ | $t \notin (-2.028, 2.028)$ | reject $H_0$ |
|  | $[0\ 0\ 1\ 0]$ | $=$ | 0 | $t = 2.3671$ | $\notin (-t_{.025}(36), t_{.025}(36))$ | reject $H_0$ |
|  | $[0\ 0\ 0\ 1]$ | $=$ | 0 | $t = 3.9598$ |  | reject $H_0$ |
| (c) | $[0\ 1\ 1\ 1]$ | $=$ | 0 | $t = .6587$ | $t \notin (-2.028, 2.028)$<br>$\notin (-t_{.025}(36), t_{.025}(36))$ | accept $H_0$ |
| (d) | $[0\ 1\ 0\ 0]$ | $\leq$ | $-1$ | $t = 1.15$ | $t \in [1.688, \infty)$<br>$\in [t_{.05}(36), \infty)$ | accept $H_0$ |
| (e) | $[0\ 0\ 0\ 1]$ | $\geq$ | 1 | $t = -3.1113$ | $t \in (-\infty, -1.688]$<br>$\in (-\infty, -t_{.05}(36)]$ | reject $H_0$ |

Test (a) indicates that at least one of the explanatory variables is significant at the .05 level; test (b) that each explanatory variable is individually significant at the .05 level (but see the next subsection on the interpretation of the collective outcome of the three tests); test (c) that homogeneity of degree zero is acceptable as is test (d)'s conjecture that price elasticity $\leq -1$; and test (e)'s conjecture that the income elasticity is unitary elastic or greater is rejected. Thus, one must conclude that the income response is inelastic. However, the reader should contemplate the power function of the tests used for the two hypotheses that were accepted in order to temper one's enthusiasm for *literal acceptance* of the respective hypotheses. □

*Bonferroni Joint Tests of* $\mathbf{R}_i\boldsymbol{\beta} = \mathbf{r}_i$, $\mathbf{R}_i\boldsymbol{\beta} \leq \mathbf{r}_i$, *or* $\mathbf{R}_i\boldsymbol{\beta} \geq \mathbf{r}_i$, $i = 1, \ldots, m$ Based on the Bonferroni probability inequality, it is possible to test *individually* a number of equality and/or inequality hypotheses concerning various linear combinations of $\boldsymbol{\beta}$ and still provide an upper bound to the probability of making a type I error with respect to the null hypotheses taken *collectively* or *simultaneously*. To see what is involved, suppose a collection of $m$ tests based on the aforementioned $t$-statistics are performed that relate to any combination of hypotheses of the form $H_{0i}: \mathbf{R}_i\boldsymbol{\beta} = r_i$ or $H_{0i}: \mathbf{R}_i\boldsymbol{\beta} \leq r_i$ or $H_{0i}: \mathbf{R}_i\boldsymbol{\beta} \geq r_i$ versus, respectively, $H_{ai}: \mathbf{R}_i\boldsymbol{\beta} \neq r_i$ or $H_{ai}: \mathbf{R}_i\boldsymbol{\beta} > r_i$ or $H_{ai}: \mathbf{R}_i\boldsymbol{\beta} < r_i$, where $\mathbf{R}_i$'s and $r_i$'s are $(1 \times k)$ vectors and scalars. Let $T_i$ and $C_r^i$, $i = 1, \ldots, m$, represent the associated $t$-statistics and critical regions, respectively, and suppose the sizes of the tests were $\alpha_i$, $i = 1, \ldots, m$. It follows by Bonferroni's inequality that if all of the null hypotheses are true,

$$P(t_i \notin C_r^i, i = 1, \ldots, m) \geq 1 - \sum_{i=1}^{m} P(t_i \in C_r^i) = 1 - \sum_{i=1}^{m} \alpha_i.$$

Thus the overall level of protection against type I error provided by the collection of tests is less than or equal to $\sum_{i=1}^{m} \alpha_i$.

For example, if four different tests are performed, each of size .025, then treating the four different null hypotheses collectively as a joint hypothesis regarding the parameter vector $\boldsymbol{\beta}$, the size of the joint test $\leq .10$. In performing the four tests, one knows the probability that one or more individual hypotheses will be rejected by mistake $\leq .10$. In practice, one generally sets the overall bound on type I error, say $\alpha$, that is desired for the joint test comprised of the $m$ individual tests, and then sets the size of each individual test to $\alpha/m$. Alternatively, any distribution of test sizes across the $m$ individual tests that sum to $\alpha$ will afford the same overall level of type I error protection. The Bonferroni approach can of course be applied in contexts other than the GLM.

While the individual tests in the collection of tests may have optimal properties, such as being UMPU, the properties do not transfer to the Bonferroni-type joint test in general. However, asymptotically the situation is favorable in the sense that if the individual test sequences are consistent, the joint test sequence will also be consistent. This follows from the fact that the joint test will reject the collection of $H_0$'s as being simultaneously true with probability

$$P\left(\cup_{i=1}^{m}\{\mathbf{y}: t_i(\mathbf{y}) \in C_r^i\}\right) \geq P\left(\{\mathbf{y}: t_i(\mathbf{y}) \in C_r^i\}\right) \quad \forall i,$$

and if $P(\{\mathbf{y}: t_i(\mathbf{y}) \in C_r^i\}) \to 1$ as $n \to \infty$ when $H_{ai}$ is true, then $P(\cup_{i=1}^{m}\{\mathbf{y}: t_i(\mathbf{y}) \in C_r^i\}) \to 1$ as $n \to \infty$. Thus the collection of null hypotheses $\cup_{i=1}^{m} H_{0i}$ is rejected with probability $\to 1$, and the joint test sequence is consistent.

A useful feature of the Bonferroni approach to joint testing of hypotheses is the ability to "look inside" a joint vector hypothesis $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ to analyze which linear combinations of the entries in $\boldsymbol{\beta}$ are causing a rejection of the vector hypothesis. Specifically, if the joint test $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ is rejected, say by a joint $F$-test, one can consider performing a series of tests of the individual null hypotheses $H_{0i}: \mathbf{R}_i\boldsymbol{\beta} = r_i$, $i = 1, \ldots, m$, to attempt to see *which* linear combination hypotheses are to be rejected and which are not. By choosing an overall level of type I error of $\alpha$, and then distributing $\alpha/m$ of this type I error probability to each of the individual tests, the collection of conclusions is protected at the level of type I error $\alpha$.

**Example 10.17** Revisit Ex. 10.14, Ex. 10.15, and Ex. 10.16. In each case, an $F$-test rejected a joint null hypothesis of the form $H_0: \beta_i = 0$, $i \in I$, and then individual $t$-tests were performed on each of the null hypotheses $H_{0i}: \boldsymbol{\beta}_i = 0$ separately. For Ex. 10.14 and Ex. 10.15, since each of the two tests had size .05, the overall conclusion that both explanatory variables are significant is protected against a type I error at level .10 by the Bonferroni approach. Likewise, the conclusion that each of the three explanatory variables is significant in Ex. 10.16 is protected against a type I error at level .15. □

*Tests for $\sigma^2$: $\chi^2$-Tests* In order to test hypotheses concerning the magnitude of $\sigma^2$, consider the GLR approach. The likelihood function for $(\boldsymbol{\beta}, \sigma^2)$ is given by

$$L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-1/2(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})/\sigma^2\right].$$

*$H_0$: $\sigma^2 = d$ Versus $H_a$: $\sigma^2 \neq d$* Let $H_0$: $\sigma^2 = d$ and $H_a$: $\sigma^2 \neq d$. In Section 9.3 we saw that the MLE for $\boldsymbol{\beta}$ and $\sigma^2$ is given by $\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ and $\hat{\sigma}^2(\mathbf{Y}) = (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\beta}})/n$, so that

$$\begin{aligned}\max_{\boldsymbol{\beta},\sigma^2} L(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) &= (2\pi\hat{\sigma}^2(\mathbf{y}))^{-n/2} \exp\left[-1/2(\mathbf{y} - \mathbf{x}\mathbf{b})'(\mathbf{y} - \mathbf{x}\mathbf{b})/\hat{\sigma}^2(\mathbf{y})\right] \\ &= (2\pi\hat{\sigma}^2(\mathbf{y}))^{-n/2} \exp[-n/2].\end{aligned}$$

Now consider the MLE subject to the restriction $\sigma^2 = d$. It is evident that $L(\boldsymbol{\beta}, d \mid y)$ is maximized when $(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})$ is minimized, so that $\hat{\boldsymbol{\beta}}$ remains the optimum choice of $\boldsymbol{\beta}$ under the restriction. The GLR statistic is then given by

$$\lambda(\mathbf{Y}) = \left(\frac{\hat{\sigma}^2(\mathbf{Y})}{d}\right)^{n/2} \exp\left[\frac{n}{2}\left(1 - \frac{\hat{\sigma}^2(\mathbf{Y})}{d}\right)\right].$$

In order to define a size $\alpha > 0$ test, we need to find a positive number $c < 1$ such that $P(\lambda \leq c) = \alpha$ when $H_0$ is true. The GLR statistic has a complicated probability distribution and is therefore not convenient for solving this problem. We seek an alternative test statistic for determining the critical region of the GLR test. Letting $z = \hat{\sigma}^2(\mathbf{y})/d$, note that $\ln\lambda = (n/2)\ln z + (n/2)(1 - z)$ is such that $d\ln\lambda/dz = (n/2)[(1/z) - 1]$ and $d^2\ln\lambda/dz^2 = -n/2z^2 < 0$ for $z > 0$, which indicates that $\ln\lambda$, and thus $\lambda$ itself, attains its maximum when $z = 1$ and is strictly monotonically decreasing as $z$ increases in value above 1 or decreases in value below 1 and moves toward zero. It follows that

$$\lambda \leq c \quad \text{iff} \quad \hat{\sigma}^2(\mathbf{y})/d \leq \tau_1 \text{ or } \geq \tau_2$$

for appropriate choices of $\tau_1 < \tau_2$.

We can simplify the problem further via multiplying $\hat{\sigma}^2(\mathbf{Y})/d$ by the constant $n$, resulting in $n\hat{\sigma}^2(\mathbf{Y})/d = (n - k)\hat{S}^2/d \sim \chi^2_{n-k}$ under $H_0$. We thus seek appropriate values of $\tau_1 < \tau_2$ that satisfy both the size $\alpha$ requirement, $1 - \int_{n\tau_1}^{n\tau_2} f(w; n-k)dw = \alpha$, where $f(w; n-k)$ is a $\chi^2_{n-k}$ density, and the requirement that the GLR $\lambda$ have the same value, $c$, at $\tau_1$ and $\tau_2$, implying that $\ln\tau_1 + (1 - \tau_1) = \ln\tau_2 + (1 - \tau_2)$, or $\ln(\tau_1/\tau_2) = \tau_1 - \tau_2$. The two simultaneous equations must be solved numerically to obtain the appropriate values of $\tau_1$ and $\tau_2$ that define the critical region $[0, n\tau_1] \cup [n\tau_2, \infty)$ for the $\chi^2$-test-statistic outcome $(n-k)\hat{s}^2/d$. Although obtaining the solution is currently highly feasible on personal computers, it is probably safe to say that most practitioners nevertheless approximate the generalized likelihood ratio test by using an "equal-tails" test that can be defined by consulting a table of the $\chi^2$ CDF. The equal-tails test would be

$$(n-k)\hat{s}^2/d \begin{bmatrix} \in \\ \notin \end{bmatrix} \left[0, \chi^2_{n-k;1-\alpha/2}\right] \cup \left[\chi^2_{n-k;\alpha/2}, \infty\right) \Rightarrow \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix}.$$

Note that the GLR test is a UMPU level-$\alpha$ test of $H_0$. The equal-tails test has level $\alpha$ but is not UMPU and in fact is not unbiased.

Both the GLR test and the equal-tails test are consistent. To motivate consistency in the case of the equal-tails test, note that $H_0$ is rejected iff $\hat{s}^2/d \notin ((n-k)^{-1}\chi^2_{n-k;1-\alpha/2}, (n-k)^{-1}\chi^2_{n-k;\alpha/2})$. Since $(n-k)^{-1}\chi^2_{n-k} \xrightarrow{p} 1$, in the limit the test amounts to deciding whether or not plim $\hat{S}^2/d = 1$, and since $\hat{S}^2 \xrightarrow{p} \sigma_0^2$, the test then amounts to whether $\sigma_0^2/d = 1$. Thus, if $d \neq \sigma_0^2$, so that $H_a$ is true, $H_0$ is rejected with probability → 1 as $n \to \infty$, and the test sequence is consistent.

The consistency of the GLR test sequence can be established by showing that Theorem 10.4 applies. To see this, recall that $\ln \lambda = (n/2)\ln(\hat{\sigma}^2(\mathbf{y})/d) + (n/2)(1 - \hat{\sigma}^2(\mathbf{y})/d)$ achieves its maximum value, equal to 0, when $\hat{\sigma}^2(\mathbf{y})/d = 1$. Now note that $\operatorname{plim} n^{-1}(2 \ln \lambda(\mathbf{Y})) = \ln(\sigma_0^2/d) + 1 - \sigma_0^2/d < 0 \ \forall\ \sigma_0^2 \neq d$, which implies that $\lambda(\mathbf{Y}) \xrightarrow{p} 0$ (recall Ex. 10.8). Then Theorem 10.4.b applies, and the GLR test sequence is consistent.

$H_0\colon \sigma^2 \leq d$ Versus $H_a\colon \sigma^2 > d$ or $H_0\colon \sigma^2 \geq d$ Versus $H_a\colon \sigma^2 < d$ In order to test the hypothesis $H_0\colon \sigma^2 \leq d$ versus $H_a\colon \sigma^2 > d$, the GLR approach of the previous subsection can be applied, where the constrained MLE is now found by maximizing $L(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y})$ subject to the inequality constraint $\sigma^2 \leq d$. The solution for $\boldsymbol{\beta}$ remains $\mathbf{b} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$, while the solution for $\sigma^2$ is $\hat{\sigma}_*^2(\mathbf{y}) = \hat{\sigma}^2(\mathbf{y}) = n^{-1}(\mathbf{y} - \mathbf{x}\mathbf{b})'(\mathbf{y} - \mathbf{x}\mathbf{b})$ if $\hat{\sigma}^2(\mathbf{y}) \leq d$, and $\hat{\sigma}_*^2(\mathbf{y}) = d$ otherwise. The resulting GLR statistic is therefore defined to be

$$\lambda(\mathbf{Y}) = \left(\frac{\hat{\sigma}^2(\mathbf{Y})}{\hat{\sigma}_*^2(\mathbf{Y})}\right)^{n/2} \exp\left[\frac{n}{2}\left(1 - \frac{\hat{\sigma}^2(\mathbf{Y})}{\hat{\sigma}_*^2(\mathbf{Y})}\right)\right].$$

Note that the definition of $\hat{\sigma}_*^2$ is such that $\hat{\sigma}^2/\hat{\sigma}_*^2 \geq 1$. Given our previous discussion of the behavior of $\lambda = (z)^{n/2}\exp[(n/2)(1-z)]$ indicating that $\lambda$ achieves its maximum value of 1 when $z = 1$ and $\lambda$ is strictly monotonically decreasing as $z$ increases in value above 1, it follows that

$$\lambda \leq c \quad \text{iff } \hat{\sigma}^2(\mathbf{y})/\hat{\sigma}_*^2(\mathbf{y}) \geq \tau$$

for some appropriate choice of $\tau > 1$ for a size $\alpha > 0$ test. Furthermore, since $\hat{\sigma}^2(\mathbf{y})/\sigma_*^2(\mathbf{y}) > 1$ only if $\hat{\sigma}_*^2(\mathbf{y}) = d$, it follows that

$$\lambda \leq c \quad \text{iff } \hat{\sigma}^2(\mathbf{y})/d \geq \tau.$$

Finally, by multiplying $\hat{\sigma}^2(\mathbf{y})/d$ by $n$ as before, we have

$$\lambda \leq c \quad \text{iff } (n-k)\hat{s}^2/d \geq n\tau.$$

To define the critical region of the size $\alpha$ GLR test, we need to find the value of $n\tau$ that satisfies

$$\int_{n\tau}^{\infty} f(w; n-k)dw = \alpha,$$

where $f(w; n-k)$ is a $\chi^2$ density function with $n-k$ degrees of freedom. The appropriate value of $n\tau$ can be found in tables of the $\chi^2$ CDF for the typical choices of $\alpha = .01, .05$, or $.10$, or else they can be found through use of the computer. The GLR test is then

$$\frac{(n-k)\hat{s}^2}{d}\begin{bmatrix}\in\\ \notin\end{bmatrix}[\chi^2_{n-k;\alpha}, \infty) \Rightarrow \begin{bmatrix}\text{reject } H_0\\ \text{accept } H_0\end{bmatrix}.$$

It can be shown that the GLR test is a UMPU level-$\alpha$ test of $H_0: \sigma^2 \leq d$ versus $H_a: \sigma^2 > d$. Consistency of the GLR test sequence can be demonstrated by showing that Theorem 10.4 applies following a similar argument to the one used in the previous subsection. In order to test the hypothesis $H_0: \sigma^2 \geq d$ versus $H_a: \sigma^2 < d$, the GLR approach can be followed with appropriate inequality reversals; it is left to the reader. The GLR test of $H_0$ versus $H_a$ is

$$\frac{(n-k)\hat{s}^2}{d}\begin{bmatrix}\in\\ \notin\end{bmatrix}[0, \chi^2_{n-k;1-\alpha}] \Rightarrow \begin{bmatrix}\text{reject } H_0\\ \text{accept } H_0\end{bmatrix}.$$

The test is UMPU level $\alpha$ and consistent.

**Example 10.18** Suppose that in Ex. 10.14, $\hat{s}^2 = .13$. Test the null hypotheses that (a) $\sigma^2 = .10$, (b) $\sigma^2 \leq .10$, and (c) $\sigma^2 \geq .25$. Use a size .05 test in each case.
**Answer:** In each case, $n - k = 40 - 4 = 36$. We use the *equal-tails* test for (a), and the GLR tests for (b) and (c). Noting that $\chi^2_{36;.975} = 21.335$, $\chi^2_{36;.025} = 54.437$, $\chi^2_{36;.05} = 50.998$, and $\chi^2_{36;.95} = 23.268$,

**a.** $(n-k)\hat{s}^2/d = 36(.13)/.10 = 46.8 \notin [0, 21.335] \cup [54.437, \infty) \Rightarrow$ accept $H_0$;
**b.** $(n-k)\hat{s}^2/d = 36(.13)/.10 = 46.8 \notin [50.998, \infty) \Rightarrow$ accept $H_0$;
**c.** $(n-k)\hat{s}^2/d = 36(.13)/.25 = 18.72 \in [0, 23.268] \Rightarrow$ reject $H_0$.

The reader may wish to investigate the power function of the tests used in (a) and (b) to gain perspective on the danger of interpreting "acceptance" in the *literal sense* in these cases. □

### Testing When $\varepsilon$ Is Not Multivariate Normal

In order to test hypotheses about the parameters of the GLM when $\varepsilon$ is not multivariate normally distributed, we resort to asymptotic distributions of test statistics to establish asymptotically valid size $\alpha$ and consistent tests. We assume that the classical GLM assumptions apply and that the $\varepsilon_i$'s are iid distributed according to some unknown, nonnormal PDF $f(\varepsilon_i)$. Furthermore, we assume that conditions exist so that $n^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} N([0], \sigma^2\mathbf{Q}^{-1})$ and $n^{1/2}(\hat{S}^2 - \sigma^2) \xrightarrow{d} N(0, \tau)$ (e.g., see Table 8.1).[10]

*Testing* $\mathbf{R}\beta = \mathbf{r}$ *or* $\mathbf{R}(\beta) = \mathbf{r}$ *When* $\mathbf{R}$ *Has q Rows: Asymptotic* $\chi^2$*-Tests* We first reexamine the use of the Wald statistic for testing $H_0: \mathbf{R}\beta = \mathbf{r}$ versus $H_a: \mathbf{R}\beta \neq \mathbf{r}$ where $\mathbf{R}$ is $(q \times k)$ with full row rank. Assume the conditions of case (5) hold in Table 8.1 so that $n^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} N([0], \sigma^2\mathbf{Q}^{-1})$, and a consistent estimator of $\sigma^2\mathbf{Q}^{-1}$ is given by $n\hat{S}^2(\mathbf{x}'\mathbf{x})^{-1}$. It follows directly from Theorem 10.9 that the Wald test of $H_0: \mathbf{R}\beta = \mathbf{r}$ versus $H_a: \mathbf{R}\beta \neq \mathbf{r}$ is given by

$$w = (\mathbf{Rb} - r)'\left[\hat{s}^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}'\right]^{-1}(\mathbf{Rb} - \mathbf{r})\begin{bmatrix} \geq \\ < \end{bmatrix}\chi^2_{q;\alpha} \Rightarrow \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix}$$

is an asymptotically valid size $\alpha$, asymptotically unbiased, and consistent test. It is thus seen that under assumptions that ensure the asymptotic normality of the estimator $\hat{\beta}$ and the consistency of $\hat{S}^2$, the Wald test of $H_0: \mathbf{R}\beta = \mathbf{r}$ versus $H_a: \mathbf{R}\beta \neq \mathbf{r}$ can be applied when $\varepsilon$ is *not* multivariate normally distributed.

A minor extension of the preceding argument allows the conclusion that the preceding test rule is appropriate for testing nonlinear hypotheses $H_0: \mathbf{R}(\beta) = \mathbf{r}$ versus $H_a: \mathbf{R}(\beta) \neq \mathbf{r}$. In particular, assuming that $\mathbf{R}(\beta)$ is continuously differentiable and contains no redundant constraints and that $\partial\mathbf{R}(\beta_0)/\partial\beta'$ has full row rank, it follows directly from Theorem 10.9 that an asymptotic size $\alpha$, asymptotically unbiased, and consistent test of $H_0: \mathbf{R}(\beta) = \mathbf{r}$ versus $H_a: \mathbf{R}(\beta) \neq \mathbf{r}$ is given by

$$w = (\mathbf{R}(\hat{\beta}) - \mathbf{r})'\left(\hat{s}^2\frac{\partial\mathbf{R}(\hat{\beta})'}{\partial\beta}(\mathbf{x}'\mathbf{x})^{-1}\frac{\partial\mathbf{R}(\hat{\beta})}{\partial\beta}\right)^{-1}(\mathbf{R}(\hat{\beta}) - \mathbf{r})\begin{pmatrix} \geq \\ < \end{pmatrix}\chi^2_{q;\alpha} \Rightarrow \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix}.$$

The asymptotically valid power function of the Wald test under sequences of local alternatives can be defined in terms of the noncentral $\chi^2$-distribution based on Theorem 10.10. We provide an example of this Wald test later in Ex. 10.19.

*Testing* $R(\beta) = r$, $R(\beta) \leq r$, *or* $R(\beta) \geq r$ *When R Is a Scalar Function: Asymptotic Normal Tests* In the case where $R(\beta)$ is a scalar function of the vector $\beta$, the Wald statistic of the preceding subsection specializes to

$$W = (R(\hat{\beta}) - r)^2 \Big/ \left[\hat{S}^2\frac{\partial R(\hat{\beta})'}{\partial\beta}(\mathbf{x}'\mathbf{x})^{-1}\frac{\partial R(\hat{\beta})}{\partial\beta}\right] \xrightarrow{d} \chi^2_1 \text{ under } H_0.$$

An alternative test statistic whose square is $W$ is given by

$$Z = (R(\hat{\beta}) - r) \Big/ \left[\hat{S}^2\frac{\partial R(\hat{\beta})'}{\partial\beta}(\mathbf{x}'\mathbf{x})^{-1}\frac{\partial R(\hat{\beta})}{\partial\beta}\right]^{1/2} \xrightarrow{d} N(0, 1) \text{ under } H_0.$$

[10] Asymptotic tests can be defined even when $n^{-1}\mathbf{x}'\mathbf{x} \not\to \mathbf{Q}$ or when the $\varepsilon_i$'s are not iid. See H. White (1984), *Asymptotic Theory for Econometricians*. Orlando: Academic Press.

The latter limiting distribution can be established via an application of Theorem 5.39 applied to the scalar function, $R(\hat{\beta})$, of the asymptotically normally distributed random vector $\hat{\beta}$ and through Slutsky's theorem. An asymptotic size $\alpha$, asymptotically unbiased, and consistent test of $H_0$: $R(\beta) = r$ versus $H_a$: $R(\beta) \neq r$ in terms of a critical region for the $Z$-statistic is given by

$$z\begin{bmatrix}\in\\ \notin\end{bmatrix}(-\infty,-z_{\alpha/2}]\cup[z_{\alpha/2},\infty)\Rightarrow\begin{bmatrix}\text{reject } H_0\\ \text{accept } H_0\end{bmatrix}.$$

Other asymptotic size $\alpha$, asymptotically unbiased, and consistent tests in terms of the $Z$-statistic are given as follows:

| Case | $H_0$ | $H_a$ | $C_r$ |
|---|---|---|---|
| (1) | $R(\beta) \leq r$ | $R(\beta) > r$ | $z \in [z_\alpha, \infty)$ |
| (2) | $R(\beta) \geq r$ | $R(\beta) < r$ | $z \in (-\infty, -z_\alpha]$ |

In order to test the *linear* hypothesis $H_0$: $R\beta = r$ versus $H_a$: $R\beta \neq \mathbf{r}$, the test is performed with $R$ replacing $\partial R(\hat{\beta})/\partial\beta'$ and $R\hat{\beta}$ replacing $R(\hat{\beta})$ in the definition of the $Z$-statistic. It should be noted that in practice, the $t$-distribution is sometimes used in place of the standard normal distribution to define the critical regions of the aforementioned tests. In the limit, there is no difference between the two procedures, since $t_v \rightarrow N(0,1)$ as $v \rightarrow \infty$. However, in small samples, the $t$-distribution has fatter tails, so that $t_\alpha > z_\alpha$, and thus the *actual size* of the aforementioned tests is smaller when based on $t_\alpha$ and $t_{\alpha/2}$ in place of $z_\alpha$ and $z_{\alpha/2}$, respectively. The use of the $t$-distribution can be viewed as a conservative policy toward rejection of $H_0$ in that the probability of type I error is less and stronger evidence will be required for rejection. On the other hand, the *actual power* of the test is also reduced when the $t$-distribution is utilized.

Asymptotically valid power calculations for local alternatives when one of the $z$-tests is used can be based on the standard normal distribution. In particular, for local alternatives of the form $R(\beta) - r = n^{-1/2}\phi$, it follows that

$$(R(\hat{\beta}) - r - n^{-1/2}\phi)/\left[\hat{S}^2\frac{\partial R(\hat{\beta})'}{\partial\beta}(\mathbf{x}'\mathbf{x})^{-1}\frac{\partial R(\hat{\beta})}{\partial\beta}\right]^{1/2} \xrightarrow{d} N(0,1).$$

Then assuming that $n^{-1}\mathbf{x}'\mathbf{x} \rightarrow \mathbf{Q}$, Slutsky's theorems can be used to show that

$$Z \xrightarrow{d} N\left(\phi/\left[\sigma^2\frac{\partial R(\beta)'}{\partial\beta}\mathbf{Q}^{-1}\frac{\partial R(\beta)}{\partial\beta}\right]^{1/2}, 1\right) = N\left(\frac{\phi}{\xi}, 1\right),$$

so that the asymptotic power function of the $z$-test is represented as

$$\pi_{C_r}(\phi) = \int_{x \in C_r} N\left(z; \frac{\phi}{\xi}, 1\right) dz.$$

Note that the Bonferroni approach to hypothesis testing can be applied as before using either Wald statistics or $t$-statistics except that now the bound on the level of type I error is only asymptotically valid.

**Example 10.19** A quadratic production function representing output in terms of two-variable inputs is specified as

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + [x_{t1} \;\; x_{t2}]\begin{bmatrix} \beta_3 & \beta_5 \\ \beta_5 & \beta_4 \end{bmatrix}\begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} + \varepsilon_t, \quad \text{for } t = 1, 2, \ldots, n,$$

where $y_t$ is output is period $t$, $x_{ti}$ is quantity of input $i$ used in period $t$, and the $\varepsilon_t$'s are iid with $\mathrm{E}\varepsilon_t = 0$ and $\mathrm{E}\varepsilon_t^2 = \sigma^2$, $\forall\, t$. Representing the production function in a more obvious GLM form, we have

$$y_t = \begin{bmatrix} x_{t1} & x_{t2} & x_{t1}^2 & x_{t2}^2 & 2x_{t1}x_{t2} \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + \varepsilon_t, \quad \text{for } t = 1, 2, \ldots, n; \text{ or}$$

$$\mathbf{Y} = \mathbf{x}_* \boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Assuming that $\mathbf{x}_*$ is a matrix with full column rank, the classical GLM assumptions hold. Assume further that $n^{-1}\mathbf{x}_*'\mathbf{x}_* \to \mathbf{Q}$, so that $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N([\mathbf{0}], \sigma^2\mathbf{Q}^{-1})$, and note that $n\hat{S}^2(\mathbf{x}_*'\mathbf{x}_*)^{-1}$ is a consistent estimator of $\sigma^2\mathbf{Q}^{-1}$.

Suppose 40 observations are used to generate the following least-squares–based estimates:

$$\mathbf{b} = [7.1 \quad 5.2 \quad -.11 \quad -.07 \quad -.02]', \qquad \hat{s}^2 = .6 \times 10^{-3},$$

$$(\mathbf{x}_*'\mathbf{x}_*)^{-1} = \begin{bmatrix} 9.571 & 6.463 & 7.187 & 4.400 & 8.247 \\ & 9.686 & 4.124 & 7.364 & 8.800 \\ & & 5.766 & 2.814 & 6.087 \\ & & & 5.945 & 6.914 \\ \text{(symmetric)} & & & & 11.254 \end{bmatrix}.$$

**a.** Test whether the production function is concave. Use an asymptotically valid level-.10 test.

**b.** Test whether the marginal products (in terms of expected output) of inputs $x_1$ and $x_2$ are identical at equal input levels. Use an asymptotically valid size .05 test.

**Answer:**

**a.** Note that a necessary and sufficient condition for the production function to be concave is that $\beta_3 \leq 0$ and $\beta_3\beta_4 - \beta_5^2 \geq 0$. We will test each hypothesis using a $Z$-test of asymptotic size-.05, so that based on the Bonferroni approach, the joint test will have asymptotic size $\leq$ .10 and thus asymptotic level .10.

Regarding $H_0\colon \beta_3 \leq 0$, the $Z$-statistic outcome is

$$z = \frac{\mathbf{Rb} - r}{[\hat{s}^2\mathbf{R}(\mathbf{x}_*'\mathbf{x}_*)^{-1}\mathbf{R}']^{1/2}} = \frac{-.11 - 0}{.0588} = -1.8707,$$

where $\mathbf{R} = [0 \;\; 0 \;\; 1 \;\; 0 \;\; 0]$ and $r = 0$. The critical region for the test is $C_r = [z_{.05}, \infty) = [1.645, \infty)$, and since $w \notin C_r$, $H_0$ is accepted.

Regarding $H_0\colon\ \beta_3\beta_4 - \beta_5^2 \geq 0$, first note that $\partial R(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \partial(\beta_3\beta_4 - \beta_5^2)/\partial\boldsymbol{\beta} = [0\ \ 0\ \ \beta_4\ \ \beta_3\ 2\beta_5]'$. Then $\hat{s}^2(\partial R(\hat{\boldsymbol{\beta}})/\partial\boldsymbol{\beta}')(\mathbf{x}_*'\mathbf{x}_*)^{-1}(\partial R(\hat{\boldsymbol{\beta}})/\partial\boldsymbol{\beta}) = .3996 \times 10^{-4}$, so that the $Z$-statistic outcome is

$$z = \frac{.0073 - 0}{(.3996 \times 10^{-4})^{1/2}} = 1.1548.$$

Since the critical region is $C_r = [-\infty, -z_{.05}] = (-\infty, -1.645]$ and $w \notin C_r$, $H_0$ is accepted. Overall, the joint null hypothesis of concavity of the production function is not rejected at the asymptotic .10 level of significance.

**b.** Note that

$$\frac{\partial \mathrm{E}Y_t}{\partial x_{t1}} = \beta_1 + 2\beta_3 x_{t1} + 2\beta_5 x_{t2},$$

$$\frac{\partial \mathrm{E}Y_t}{\partial x_{t2}} = \beta_2 + 2\beta_5 x_{t1} + 2\beta_4 x_{t2}.$$

The marginal products will be identical at equal input levels iff

$$\frac{\partial \mathrm{E}Y_t}{\partial x_{t1}} - \frac{\partial \mathrm{E}Y_t}{\partial x_{t2}} = \beta_1 - \beta_2 + 2(\beta_3 - \beta_4)\xi = 0 \quad \forall\, \xi = x_{t1} = x_{t2}.$$

Thus we must test $H_0\colon\ \beta_1 - \beta_2 = 0$ and $\beta_3 - \beta_4 = 0$ versus $H_a$: not $H_0$. The Wald statistic appropriate for this problem has the form

$$w = (\mathbf{Rb} - \mathbf{r})'\left[\hat{s}^2\mathbf{R}(\mathbf{x}_*'\mathbf{x}_*)^{-1}\mathbf{R}'\right]^{-1}(\mathbf{Rb} - \mathbf{r}).$$

Note that

$$\mathbf{R}\hat{\boldsymbol{\beta}} = \begin{bmatrix} b_1 - b_2 \\ b_3 - b_4 \end{bmatrix}, \qquad \mathbf{r} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \text{and } \mathbf{R} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{bmatrix},$$

and the Wald statistic outcome is 17,390.828. Since the critical region of the test is $C_r = [\chi^2_{2,.05}, \infty) = [5.991, \infty)$ and $17{,}390.828 \in C_r$, $H_0$ is (soundly) rejected at the asymptotic .05 level of significance. □

*Tests for $\sigma^2$: Asymptotic $\chi^2$ and Normal Tests* In order to test hypotheses concerning the magnitude of $\sigma^2$, consider using a Wald statistic. Assume the conditions of case (7) in Table 8.1 ensuring the asymptotic normality of $\hat{S}^2$ based on Theorem 8.8. Then $n^{1/2}(\hat{S}^2 - \sigma^2) \xrightarrow{d} N(0, (\mu_4' - \sigma^4))$, and it follows directly from Theorem 10.9 that a consistent, asymptotically unbiased, and asymptotic size $\alpha$ Wald test of $H_0\colon\ \sigma^2 = d$ versus $H_a\colon\ \sigma^2 \neq d$ is given by

$$w = \frac{n(\hat{s}^2 - d)^2}{\hat{\xi}} \begin{bmatrix} \in \\ \notin \end{bmatrix} [\chi^2_{1,\alpha}, \infty) \Rightarrow \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix},$$

where $\hat{\xi}$ is a consistent estimate of $\mathrm{var}(\varepsilon_i^2) = \mu_4' - \sigma^4$, such as $\hat{\xi} = n^{-1}\sum_{i=1}^n \hat{e}_i^4 - \hat{s}^4$, with $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{xb}$.

Asymptotic size $\alpha$, asymptotically unbiased, and consistent tests for one-sided hypotheses can be defined using an asymptotically valid $Z$-statistic as follows:

| Case | $H_0$ | $H_a$ | $C_r$ |
|---|---|---|---|
| (1) | $\sigma^2 \leq d$ | $\sigma^2 > d$ | $n^{1/2}(\hat{s}^2 - d)/\hat{\xi}^{1/2} \in [z_\alpha, \infty)$ |
| (2) | $\sigma^2 \geq d$ | $\sigma^2 < d$ | $n^{1/2}(\hat{s}^2 - d)/\hat{\xi}^{1/2} \in (-\infty, -z_\alpha]$ |

The reader will be asked to justify the asymptotic size and consistency of these tests in this chapter's problem section.

**Example 10.20** Recall Ex. 10.19, and test the null hypothesis $H_0: \sigma^2 \leq .25 \times 10^{-3}$ at the .05 level of significance.

**Answer**: Assuming the appropriate conditions for the asymptotic normality of $\hat{S}^2$ and the consistency of $n^{-1}\sum_{i=1}^n e_i^4 - \hat{S}^4$ for $\mu_4' - \sigma^4$, the $z$-statistic, with $n^{-1}\sum_{i=1}^n \hat{e}_i^4 = .87 \times 10^{-6}$, say, is

$$z = \frac{n^{1/2}(\hat{s}^2 - d)}{\hat{\xi}^{1/2}} = \frac{(40)^{1/2}(.6 \times 10^{-3} - .25 \times 10^{-3})}{(.51 \times 10^{-6})^{1/2}} = 3.0997.$$

The critical region of the test is $C_r = [z_{.05}, \infty) = [1.645, \infty)$, and since $3.0997 \in C_r$, $H_0$ is rejected. □

Note that in the preceding three subsections we have made no explicit assumptions regarding the functional forms of the PDFs associated with the error terms of the GLM. We proceeded entirely on the basis of moment assumptions. If a functional form for the PDF of the error term is assumed, then the GLR or LM hypothesis-testing procedures can be used to provide alternative statistical tests for all of the preceding hypotheses. The reader is encouraged to contemplate the possibilities (see R. Davidson and J. G. MacKinnon, *Estimation and Inference*, pp. 452–458, for some such tests).

## 10.7 Confidence Intervals and Regions

In this section we provide an introduction to the concepts of *confidence intervals* and *confidence regions.* We will approach the problem both from the perspective of the duality between hypothesis-testing and confidence-region (or interval) estimation and from the concept of *pivotal quantities.* The study of confidence-region or confidence-interval estimation encompasses its own theory and practice that could easily consume considerably more space if treated at a general level. In particular, there is an entire paradigm that encompasses properties of confidence regions and methods of estimating optimal confidence regions or intervals that parallels the theories developed for hypothesis testing. Moreover, there is a full duality between certain optimality properties of hypothesis tests and their counterpart for confidence-region or interval estimation. We will point out some of these parallels in our discussion, but we will not examine them in detail. Readers interested in more detail can begin their readings by examining the work of M. Kendall and A. Stuart (1979), *Advanced*

*Statistics*, Chapter 20, and E. Lehmann (1986), *Testing Statistical Hypotheses*, Chapter 5.

In the case of a single scalar parameter, $\Theta$, or a scalar function of parameters, $R(\Theta)$, a *confidence interval* is a *random* interval whose outcomes have a known probability of containing the true value of the parameter or function of parameters. In practice, a confidence interval will be defined by random variables that represent the upper and lower bounds of the interval, and thus outcomes of the random variables defining the bounds also define an outcome of the random interval. It is possible for one of the random variables defining the bounds of the confidence interval to be replaced by a constant, in which case the confidence interval is a *one-sided confidence interval* or a *confidence bound.* We provide a formal definition of what is meant by a confidence interval next.

**Definition 10.2**
***Confidence interval with confidence level*** $\gamma$

Let $f(\mathbf{x}; \Theta)$ for some $\Theta \in \Omega$ be the joint density of the random sample $X$, and let $R(\Theta)$ be some scalar function of the parameter vector $\Theta$. The following are **confidence intervals** for $R(\Theta)$ with confidence level $\gamma$:

**a.** two-sided: $(\ell(\mathbf{X}), u(\mathbf{X}))$ such that $P(\ell(\mathbf{x}) < R(\Theta) < u(\mathbf{x}); \Theta) \geq \gamma, \forall\, \Theta \in \Omega$;

**b.** one-sided lower: $(\ell(\mathbf{X}), \infty)$ such that $P(\ell(\mathbf{x}) < R(\Theta); \Theta) \geq \gamma, \forall\, \Theta \in \Omega$;

**c.** one-sided upper: $(-\infty, u(\mathbf{X}))$ such that $P(R(\Theta) < \mu(\mathbf{x}); \Theta) \geq \gamma, \forall\, \Theta \in \Omega$.

The random variables $\ell(\mathbf{X})$ and $u(\mathbf{X})$ are called the **lower** and **upper confidence limits**, respectively.

In any of the three cases in Def. 10.2, the smallest value (or infimum) of all of the probabilities for $\Theta \in \Omega$ is called the **confidence coefficient**.

From the definition, it is seen that the random variables representing the lower and/or upper confidence limits are to be chosen so that they have outcomes defining an interval containing the value of $R(\Theta)$ with probability $\geq \gamma$ for any value of $\Theta$, i.e., $\forall\, \Theta \in \Omega$. Note that the inequality is used here to accommodate discrete cases in which it may not be possible to achieve an exact level of probability $\gamma$ for all values of $\Theta$.[11] With continuous random variables, strict equality can be achieved. Note further that for any given outcome of a confidence interval, $R(\Theta)$ either *is* or *is not* in the interval, and so in practice we do not know whether or not $R(\Theta)$ is really contained in a *given* confidence-interval outcome. However, since the probability of the event that the confidence-interval outcome "covers" or contains $R(\Theta)$ is known to be $\geq \gamma$, we do know that in repeated sampling $\geq 100\gamma$ percent of the intervals generated *will* contain the value of $R(\Theta)$, on average. It is in this sense that we have "confidence" that a confidence interval contains the value of $R(\Theta)$.

[11] Analogous to our discussion of size in hypothesis testing, it is possible to overcome the difficulty if *randomization* of confidence intervals is allowed, which is similar to randomization of critical regions in hypothesis testing (recall Problem 9.8). This is rarely done in practice, and we will not pursue this topic further.

The concept of a *confidence region* generalizes the concept of confidence intervals to the case where either (1) the random set designed to contain a scalar $R(\Theta)$ with known probability is not in the form of an interval, or else (2) $\mathbf{R}(\Theta)$ is a *vector* function so that the random set used to contain $\mathbf{R}(\Theta)$ with known probability is inherently in noninterval form. The random set defining the confidence region will be designed so the probability that outcomes of the confidence region contain the value of $R(\Theta)$ is known or is at least lower bounded in the discrete case. We now provide the formal definition of the concept of a confidence region.

**Definition 10.3**
***Confidence region with confidence level $\gamma$***

Let $f(\mathbf{x}; \Theta)$ be the joint density of the random sample $\mathbf{X}$, and let $\mathbf{R}(\Theta)$ be some scalar or vector function of the parameter vector $\Theta$. Then a confidence region with confidence level $\gamma$ for $\mathbf{R}(\Theta)$ is defined by a random set $A(\mathbf{X})$ for which $P(\mathbf{R}(\Theta) \in A(\mathbf{x}); \Theta) \geq \gamma, \forall\, \Theta \in \Omega$.

As we remarked following Def. 10.2, the smallest value (or infimum) of $P(\mathbf{R}(\Theta) \in A(\mathbf{x}); \Theta)$ for $\Theta \in \Omega$ is called the confidence coefficient.

In practice the random set $A(\mathbf{X})$ will be defined using random variables that appear in the set-defining conditions of the set. As these random variables take outcomes equal to the various values in their range, different outcomes of the random set are defined. For example, one possibility is that $A(\mathbf{X})$ is a random open rectangle defined as $A(\mathbf{X}) = \{\mathbf{R}(\Theta): \ell_i(\mathbf{X}) < R_i(\Theta) < u_i(\mathbf{X}), i = 1, \ldots, m\}$ so that a given outcome of the confidence region for $\mathbf{R}(\Theta)$ would be $A(\mathbf{x}) = \times_{i=1}^{m}(\ell_i(\mathbf{x}), u_i(\mathbf{x}))$. Other examples of confidence sets will be presented subsequently.

Note that confidence sets can be thought of as an alternative or supplementary to point estimation where instead of only generating a best (in some sense) *point* estimate of the value of some unknown function of the parameters of a joint density function, a *set* of values is generated that, a priori, has a high probability of containing the unknown function of the parameters. When one is interested in knowing a likely range of values for the unknown $\mathbf{R}(\Theta)$, the confidence-set concept has obvious appeal. As further motivation for the use of confidence regions, recall that point estimators generally have a notable probability of generating point estimates that are literally wrong, and in fact the probability that a continuous point estimator will generate the true value of $\mathbf{R}(\Theta)$ is zero! Alternatively, a random confidence set has a high probability of generating an outcome that contains the true $\mathbf{R}(\Theta)$ by design.

### Defining Confidence Regions via Duality with Critical Regions

The logical next question to ask is how does one go about defining confidence intervals or confidence sets? As it turns out, there is a duality between confidence regions for $\mathbf{R}(\Theta)$ and critical regions for hypothesis tests about $\mathbf{R}(\Theta)$ in the sense that if a critical region has been defined for the hypothesis-testing problem, one has in effect already defined a confidence region for $\mathbf{R}(\Theta)$. The specific nature of this duality is presented in the following theorem.

**Theorem 10.11** (Duality Between Confidence and Critical Regions) *Let the random sample* $\mathbf{X}$ *have joint density function* $f(\mathbf{x};\Theta)$ *for some* $\Theta \in \Omega$, *and let* $C_r(\tau)$ *be a level (or size)* $\alpha$ *critical region for testing* $H_0\colon \mathbf{R}(\Theta) = \tau$ *versus one of the alternatives* $H_a\colon \mathbf{R}(\Theta) \neq \tau$ *or* $H_a\colon \mathbf{R}(\Theta) < \tau$ *or* $H_a\colon \mathbf{R}(\Theta) > \tau$, *for any* $\tau \in \Omega_{\mathbf{R}(\Theta)} = \{\tau\colon \tau = \mathbf{R}(\Theta), \Theta \in \Omega\}$. *Then the random set of* $\tau$ *values represented by*

$$A(\mathbf{X}) = \{\tau\colon \mathbf{X} \in \bar{C}_r(\tau),\ \tau \in \Omega_{\mathbf{R}(\Theta)}\}$$

*is a confidence region for* $\mathbf{R}(\Theta)$ *having confidence level (or coefficient)* $1 - \alpha$.

**Proof** Given the definitions of $C_r(\tau)$ and $A(\mathbf{x})$, it follows that $\mathbf{x} \in \bar{C}_r(\tau) \Leftrightarrow \tau \in A(\mathbf{x})$. To see this, first note that if $\mathbf{x} \in \bar{C}_r(\tau)$, then hypothesis $\mathbf{R}(\Theta) = \tau$ would be accepted by the test defined by $C_r(\tau)$. It follows that $\tau$ is in the set of null hypotheses that would be accepted on the basis of the sample outcome $\mathbf{x}$, this set being precisely $A(\mathbf{x})$. Thus, $\mathbf{x} \in \bar{C}_r(\tau) \Rightarrow \tau \in A(\mathbf{x})$. Alternatively, if $\tau \in A(\mathbf{x})$, then $\mathbf{x}$ is a sample outcome that results in $H_0\colon \mathbf{R}(\Theta) = \tau$ being acceptable on the basis of $C_r(\tau)$, so that $\mathbf{x} \in \bar{C}_r(\tau)$. Thus, $\tau \in A(\mathbf{x}) \Rightarrow \mathbf{x} \in \bar{C}_r(\tau)$.

From the equivalence of the two sets $\{\mathbf{x}\colon \mathbf{x} \in \bar{C}_r(\tau)\}$ and $\{\mathbf{x}\colon \tau \in A(\mathbf{x})\}$, it follows that

$$P(\tau \in A(\mathbf{x})) = P(\mathbf{x} \in \bar{C}_r(\tau)) \geq 1 - \alpha,$$

where the inequality follows from $C_r(\tau)$'s defining a level-$\alpha$ test so that a true $\Theta$ would be rejected with $\leq \alpha$ probability. Thus, $A(\mathbf{X})$ is a confidence region for $\tau = \mathbf{R}(\Theta)$ with confidence level $1-\alpha$. If the size of the test defined by $C_r(\tau)$ is $\alpha$, so that the *maximum* probability of rejecting a true $\Theta$ equals $\alpha$, then it follows that the smallest value of $P(\tau \in A(\mathbf{x}))$ is $1 - \alpha$, which equals the confidence coefficient. ■

Theorem 10.11 states that to generate an outcome of a confidence region for $\mathbf{R}(\Theta)$, one can begin with a hypothesis-testing procedure for testing the null hypothesis $H_0\colon \mathbf{R}(\Theta) = \tau$ against a one- or two-sided alternative hypothesis. Then a confidence-region outcome is the collection of all possible values of $\tau$ (i.e., values in $\Omega_{\mathbf{R}(\Theta)}$) that represent *acceptable* null hypotheses on the basis of the testing procedure applied to the sample outcome, $\mathbf{x}$. Some examples will clarify the mechanics of the procedure. We will see that one-sided and two-sided hypothesis tests can lead respectively to one- sided and two-sided confidence intervals.

**Example 10.21** Confidence Interval for $\mu$ in a Normal Distribution

Let $\mathbf{X}$ be a random sample of size 50 from an $N(\mu, \sigma^2)$ population distribution representing observations on fill levels of 16-oz. bottles of a certain brand of liquid detergent. A sample outcome resulted in $\bar{x} = 16.02$ and $s^2 = .0001$. Calculate a two-sided confidence-interval outcome having confidence coefficient .95 for the mean fill level of the bottles.

**Answer**: From Def. 9.15, we know that a UMPU level and size .05 test of $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$ is given by

$$\frac{\bar{x}-\mu_0}{(s^2/(n-1))^{1/2}} \begin{bmatrix}\in\\ \notin\end{bmatrix} (-\infty, -t_{.025}] \cup [t_{.025}, \infty) \Rightarrow \begin{bmatrix}\text{reject } H_0\\ \text{accept } H_0\end{bmatrix},$$

where $t_{.025}$ refers to a critical point of the $t$-distribution with $n-1$ degrees of freedom. Then on the basis of a given outcome $\mathbf{x}$ for the random sample, the collection of null hypotheses, $\mu_0$, that would be accepted on the basis of this test procedure is given by

$$A(\mathbf{x}) = \left\{\mu_0 : \bar{x} - t_{.025}\frac{s}{\sqrt{n-1}} < \mu_0 < \bar{x} + t_{.025}\frac{s}{\sqrt{n-1}}\right\}.$$

Since $t_{.025} = 2.262$ for a $t$-distribution with 9 degrees of freedom, it follows that the confidence-interval outcome is given by (16.017, 16.023). The mean fill level is a number within the interval with confidence level (and coefficient) .95. □

**Example 10.22** Confidence Interval for $\mathbf{R}\beta$ in the GLM

Let $Y_t = \beta_0 \prod_{i=1}^{3} x_{ti}^{\beta_i} e^{\varepsilon_t}$, $t = 1, \ldots, 25$, represent a random sample of observations on a production function based on three inputs $(x_{t1}, x_{t2}, x_{t3})$, and assume that the classical GLM assumptions apply to the natural logarithmic transformation of the production function,

$$\ln Y_t = \beta_0^* + \sum_{i=1}^{3} \beta_i \ln x_{ti} + \varepsilon_t.$$

Furthermore, let $\varepsilon_t \sim$ iid $N(0, \sigma^2)$. An outcome of the random sample resulted in the following estimator outcomes corresponding to the transformed production function:

$$\mathbf{b} = \begin{bmatrix} 2.33 \\ .17 \\ .60 \\ .18 \end{bmatrix} \quad \text{and} \quad s^2(\mathbf{x}_*'\mathbf{x}_*)^{-1} = \begin{bmatrix} .64 & .011 & .0001 & .0008 \\ & .0025 & .23 \times 10^{-6} & .3 \times 10^{-5} \\ & & .0003 & .47 \times 10^{-6} \\ \text{(symmetric)} & & & .0036 \end{bmatrix}$$

Calculate a two-sided confidence interval having confidence level .95 for the degree of homogeneity of the production function, $\mathbf{R}\beta = [0\ 1\ 1\ 1]\beta = \sum_{i=1}^{3} \beta_i$ (note that $\beta_0^*$ is the first entry in $\beta$). Also, calculate a one-sided *upper* confidence bound with confidence level .95 for the output elasticity with respect to input 1, $\mathbf{R}\beta = [0\ 1\ 0\ 0]\beta = \beta_1$.

**Answer**: Examine the problem of the two-sided confidence interval first. We know from Section 10.6 that a UMPU level and size $\alpha$ test of $H_0: \mathbf{R}\beta = \tau$ versus the *two-sided* alternative $H_a: \mathbf{R}\beta \neq \tau$ is given by

$$\frac{\mathbf{R}\mathbf{b} - \tau}{[\hat{s}^2\mathbf{R}(\mathbf{x}_*'\mathbf{x}_*)^{-1}\mathbf{R}']^{1/2}} \begin{bmatrix}\in\\ \notin\end{bmatrix} (-\infty, -t_{\alpha/2}] \cup [t_{\alpha/2}, \infty) \Rightarrow \begin{bmatrix}\text{reject } H_0\\ \text{accept } H_0\end{bmatrix}.$$

For a given outcome of $\hat{\beta}$ and $\hat{S}^2$, the set of null hypothesis values, $\tau$, that would be accepted on the basis of this test procedure is given by

$$A(\mathbf{y}) = \{\tau: \mathbf{Rb} - t_{\alpha/2}(\hat{s}^2\mathbf{R}(\mathbf{x}_*'\mathbf{x}_*)^{-1}\mathbf{R}')^{1/2} < \tau < \mathbf{Rb} + t_{\alpha/2}(\hat{s}^2\mathbf{R}(\mathbf{x}_*'\mathbf{x}_*)^{-1}\mathbf{R}')^{1/2}\}.$$

Since $t_{.025} = 2.08$ for a $t$-distribution with 21 degrees of freedom, $\mathbf{Rb} = .95$, and $\hat{s}^2\mathbf{R}(\mathbf{x}_*'\mathbf{x}_*)^{-1}\mathbf{R}' = .00641$, it follows that the two-sided confidence interval for the homogeneity of the production function is (.78, 1.12). We have confidence level (and confidence coefficient) .95 that the degree of homogeneity is some number in the interval.

Regarding the one-sided *upper* confidence bound for $\mathbf{R}\beta = \beta_1$, we know from Section 10.6 that a UMPU size $\alpha$ test of $H_0$: $\mathbf{R}\beta = \tau$ versus the (*lower*) one-sided alternative hypothesis $H_a$: $\mathbf{R}\beta < \tau$ is given by

$$\frac{\mathbf{Rb} - \tau}{[\hat{s}^2\mathbf{R}(\mathbf{x}_*'\mathbf{x}_*)^{-1}\mathbf{R}']^{1/2}} \begin{bmatrix} \in \\ \notin \end{bmatrix} (-\infty, -t_\alpha] \Rightarrow \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix}.$$

The set of acceptable null hypothesis values, $\tau$, on the basis of a given sample outcome and this test procedure is given by

$$A(\mathbf{y}) = \{\tau: \tau < \mathbf{Rb} + t_\alpha(\hat{s}^2\mathbf{R}(\mathbf{x}_*'\mathbf{x}_*)^{-1}\mathbf{R}')^{1/2}\}.$$

Since $t_{.05} = 1.721$ for a $t$-distribution with 21 degrees of freedom, $\mathbf{Rb} = .17$, and $\hat{s}^2\mathbf{R}(\mathbf{x}_*'\mathbf{x}_*)^{-1}\mathbf{R}' = .0025$, it follows that the one-sided upper confidence bound for the output elasticity of input 1 is $(-\infty, .26)$. We have confidence level (and confidence coefficient) .95 that the elasticity is some number in the interval. □

**Example 10.23** Confidence Interval for $\theta$ in an Exponential Distribution

Recall Ex. 10.6 in which operating lives of computer screens were being analyzed, and a UMPU level and size .05 test of $H_0: \theta = 1$ versus $H_a: \theta > 1$ on the basis of a random sample of size 10 was found. Retracing the development of the statistical test, it is seen that the UMPU level-.05 GLR test of $H_0: \theta = \theta_0$ versus $H_a: \theta > \theta_0$, for an *arbitrary* choice of $\theta_0 > 0$, can be represented in the form $\ln \lambda(\mathbf{x}) = 10\ln(\bar{x}/\theta_0) + 10(1 - \bar{x}/\theta_0) \leq \ln c$. Under $H_0: \theta = \theta_0$, $X_i$'s $\sim$ iid $\theta_0^{-1}\exp(-x_i/\theta_0)$, so that $\bar{x} \sim \text{Gamma}(10, \theta_0/10)$ and then $\bar{X}/\theta_0 \sim \text{Gamma}(10, .1)$. Note that the value of $\ln \lambda(\mathbf{x})$ attains its maximum value of zero when $\bar{x}/\theta_0 = 1$, and $\ln \lambda(\mathbf{x})$ strictly decreases for movements of $\bar{x}/\theta_0$ away from 1 in either direction. Since the one-sided nature of the GLR test is such that $c < 1 \Leftrightarrow \bar{x}/\theta_0 > 1$ (recall Ex. 10.6), the test can be performed in terms of the test statistic $\bar{x}/\theta_0$, and we seek a value of $d$ such that $P(\bar{x}/\theta_0 \geq d) = .05$. With the Gamma(10, .1) distribution of $\bar{x}/\theta_0$, the value of $d$ is found to be 1.57052 (compare to Ex. 10.6), and the UMPU level .05 test of $H_0: \theta = \theta_0$ versus $H_a: \theta > \theta_0$ is

$$\frac{\bar{x}}{\theta_0} \begin{bmatrix} \geq \\ < \end{bmatrix} 1.57052 \Rightarrow \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix}.$$

In order to calculate a one-sided *lower* confidence bound for $\theta$ having confidence level .95, note that the set of acceptable null hypothesis values, $\theta_0$, based on the preceding UMPU level-.05 testing procedure and a given sample outcome is

$$A(\mathbf{x}) = \left\{\theta : \theta > \frac{\bar{x}}{1.57052}\right\}.$$

Suppose that $\bar{x} = 1.37$. Then an outcome for the one-sided lower confidence bound is $(.87232, \infty)$. We have confidence level (and confidence coefficient) .95 that the true mean operating life of the computer screens is contained in the interval. □

Confidence intervals or regions can be based on hypothesis-testing procedures that are only asymptotically valid, in which case the confidence regions inherit asymptotic validity from the duality result of Theorem 10.11.[12] The use of asymptotic procedures can be especially convenient for simplifying cases where the joint density of the random sample is discrete. Of course, the simplification comes at the price of the confidence region's confidence level being only an approximation to the true confidence level.

**Example 10.24** Asymptotic Confidence Interval for $p$ of Bernoulli Distribution

A food processor has developed a new fat-free butter substitute and intends to use a random sample of consumers to determine the proportion of U.S. food consumers who prefer the taste of the new product to that of butter. Of 250 consumers who sampled the product, 97 preferred the taste of the new butter substitute. Calculate a confidence-interval outcome with confidence level .90 for the proportion of consumers who prefer the taste of the butter substitute. Given the small sample size relative to the size of the food-consuming public, you may treat the sample as having occurred with replacement.

**Answer**: Recall that $n^{1/2}(\bar{X} - p)/(p(1-p))^{1/2} \xrightarrow{d} N(0,1)$, by the LLCLT, when random sampling from the Bernoulli population distribution. Since the sample size $n = 250$ is large, we can consider using a Wald test based on a $\chi_1^2$ limiting distribution, or equivalently, a $Z$-statistic based on the $N(0,1)$ distribution to construct an asymptotic level-.10 consistent test of $H_0: p = p_0$ versus $H_a: p \neq p_0$. Adopting the latter approach, and realizing that $\bar{X}(1-\bar{X}) \xrightarrow{p} p(1-p)$, an asymptotically valid level-.10 consistent test is given by

$$\frac{(250)^{1/2}(\bar{x} - p_0)}{[\bar{x}(1-\bar{x})]^{1/2}} \begin{bmatrix} \in \\ \notin \end{bmatrix} (-\infty, -1.645] \cup [1.645, \infty) \Rightarrow \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix},$$

where $\int_{1.645}^{\infty} N(z; 0, 1)dz = .05$.

Now consider using duality to define a confidence interval for $p$ that has an asymptotic confidence level of .90. Based on a given outcome of the random sample, the collection of null hypotheses, $p_0$, that would be accepted on the

[12] We are suppressing a technical condition for this inheritance in that convergence of the test statistic's probability distribution to a limiting distribution should be uniform in $\Theta \in \Omega$. This will occur for the typical PDFs used in practice. For further details, see C. R. Rao, *Statistical Inference*, pp. 350–351.

basis of the preceding test procedure is given by

$$A(\mathbf{x}) = \left\{ p_0 : \bar{x} - 1.645\left(\frac{\bar{x}(1-\bar{x})}{250}\right)^{1/2} < p_0 < \bar{x} + 1.645\left(\frac{\bar{x}(1-\bar{x})}{250}\right)^{1/2} \right\}.$$

Since $\bar{x} = .388$, the confidence-interval outcome is (.337, .439). We have confidence at approximately (asymptotic) level (and approximate confidence coefficient) .90 that the true proportion of individuals preferring the butter substitute is contained in the interval. □

Although in principle a confidence region for two or more functions of parameters may have a myriad of shapes, the typical shape of a confidence region that is derived from duality with a hypothesis test based on the Wald statistic is an ellipse (two dimensions) or ellipsoid (three or more dimensions). A typical application is in the GLM.

**Example 10.25** Ellipsoid Confidence Region in the GLM

Revisit the production function problem in Ex. 10.22. Construct a confidence region for the three output elasticities $\beta_1, \beta_2$, and $\beta_3$, having confidence level .95.

**Answer**: From Section 10.6, we know that a level-.05 and consistent test of $H_0$: $\mathbf{R}\boldsymbol{\beta} = \boldsymbol{\tau}$ versus $H_a$: $\mathbf{R}\boldsymbol{\beta} \neq \boldsymbol{\tau}$ can be defined in terms of the Wald statistic $W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \boldsymbol{\tau})'(\hat{S}^2\mathbf{R}(\mathbf{x}_*'\mathbf{x}_*)^{-1}\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \boldsymbol{\tau})$ as

$$w\begin{bmatrix} \geq \\ < \end{bmatrix} qF_\alpha(q, n-k) \Rightarrow \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix}$$

where in the current application $q = 3$, $n - k = 21$, $\alpha = .05$, $F_{.05}(3, 21) = 3.07$, and

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Based on the sample outcome reported in Ex. 10.22, the set of acceptable null hypotheses, $\boldsymbol{\tau}$, based on the preceding test procedure is given by

$$A(\mathbf{y}) = \left\{ \boldsymbol{\tau} : \begin{bmatrix} .17 - \tau_1 \\ .60 - \tau_2 \\ .18 - \tau_3 \end{bmatrix}' \begin{bmatrix} .0025 & .23 \times 10^{-6} & .3 \times 10^{-5} \\ & .0003 & .47 \times 10^{-6} \\ \text{(symmetric)} & & .0036 \end{bmatrix}^{-1} \begin{bmatrix} .17 - \tau_1 \\ .60 - \tau_2 \\ .18 - \tau_3 \end{bmatrix} < 9.21 \right\}.$$

The confidence region is a three-dimensional ellipsoid with center at (.17, .60, .18), and its shape resembles a football. We have confidence at level .95 that the ellipsoid contains the true value of the output elasticities $\beta_1$, $\beta_2$, and $\beta_3$. □

### Properties of Confidence Regions

Given the complete duality between confidence regions and hypothesis tests, one might expect that there is also a duality between properties of hypothesis tests and properties of confidence regions. Such is indeed the case, and we will briefly discuss properties of confidence regions and their relationship with properties of hypothesis tests.

Recall that in our discussion of hypothesis tests we examined the properties of significance level, size, unbiasedness, uniformly most powerful, uniformly most powerful unbiased, and consistency. Each of these has a counterpart with respect to properties of confidence regions; the relationships are as follows:

| **Hypothesis-Test Property** | ⇔ | **Confidence-Region Property** |
|---|---|---|
| significance level, $\alpha$ | | confidence level, $1-\alpha$ |
| size, $\alpha$ | | confidence coefficient, $1-\alpha$ |
| unbiased | | unbiased |
| UMP (uniformly most powerful) | | UMA (uniformly most accurate) |
| UMPU (uniformly most powerful unbiased) | | UMAU (uniformly most accurate unbiased) |
| consistent | | consistent |

In deriving confidence regions via duality with hypothesis tests, each of the properties possessed by the hypothesis-testing procedure is transferred to the corresponding property of the confidence region. This begs the question regarding definitions of the latter four confidence-region properties. We examine each of the properties in turn.

A confidence region is said to be **unbiased** if the probability that the confidence region contains the *true* $\mathbf{R}(\Theta)$ is greater than or equal to the probability that it contains a *false* $\mathbf{R}(\Theta)$. The formal definition is as follows.

**Definition 10.4** ***Unbiased confidence region***

A confidence region for $\mathbf{R}(\Theta)$, $A(\mathbf{X})$, is unbiased iff

$$P(\mathbf{R}(\Theta) \in A(\mathbf{x}); \Theta) \geq P(\mathbf{R}(\Theta) \in A(\mathbf{x}); \Theta_*), \quad \forall\, \mathbf{R}(\Theta) \neq \mathbf{R}(\Theta_*).$$

Unbiasedness is a reasonable property for a confidence region to possess, since one would certainly desire a confidence region to contain the true value of $\mathbf{R}(\Theta)$ more often, or with higher probability, than false values.

A confidence region for $\mathbf{R}(\Theta)$ is **uniformly most accurate (UMA)** at confidence level $\gamma$ if it has the lowest probability of containing false values of $\mathbf{R}(\Theta)$ relative to any other confidence region for $\mathbf{R}(\Theta)$ with confidence level $\gamma$.

**Definition 10.5** ***Uniformly most accurate (UMA) level-$\gamma$ confidence region***

A confidence region for $\mathbf{R}(\Theta)$, $A(\mathbf{X})$, having confidence level $\gamma$ is uniformly most accurate iff

$$P(\mathbf{R}(\Theta) \in A(\mathbf{x}); \Theta_*) \leq P(\mathbf{R}(\Theta) \in A_*(\mathbf{x}); \Theta_*), \quad \forall\, \mathbf{R}(\Theta) \neq \mathbf{R}(\Theta_*),$$

where $A_*(\mathbf{X})$ is any confidence region for $\mathbf{R}(\Theta)$ having confidence level $\gamma$.

Essentially, a UMA confidence region "filters out" false values of $\mathbf{R}(\Theta)$ with probability as high or higher than any other confidence region of like confidence level, which is obviously a desirable property.

We noted in our discussion of hypothesis testing in Section 9.5 that UMP tests, and thus now UMA confidence regions, do not exist with any degree of generality. However, UMPU tests exist much more frequently, and thus so do UMAU confidence regions. A UMAU confidence region is simply a confidence region that exhibits the UMA property when compared *only* to other unbiased confidence regions.

**Definition 10.6** ***Uniformly most accurate unbiased (UMAU) confidence regions***

A confidence region for $R(\Theta)$, $A(\mathbf{X})$, having confidence level $\gamma$ is uniformly most accurate unbiased iff within the class of unbiased confidence regions with confidence level $\gamma$, $A(\mathbf{X})$ is UMA.

In comparing confidence *intervals* for $R(\Theta)$, it can be shown that a UMAU confidence interval with confidence level $\gamma$ has the smallest expected length of any other confidence interval for $R(\Theta)$ with confidence level $\gamma$. In terms of narrowing ignorance of the value of $R(\Theta)$, this is clearly a desirable property.

Finally, *consistency* of a confidence-region sequence means that as the sample size $n \to \infty$, the length or volume of the confidence regions shrinks to zero at any confidence level, with the true $\mathbf{R}(\Theta)$ remaining as the only point in the confidence region with probability → 1.

**Definition 10.7** ***Consistent level-γ confidence region sequence***

A confidence-region sequence for $\mathbf{R}(\Theta)$, $\{A_n(\mathbf{X})\}$, with all members of the sequence having confidence level $\gamma$, is a consistent confidence-region sequence iff

$$\lim_{n\to\infty} P(\mathbf{R}(\Theta) \in A_n(\mathbf{x}); \Theta_*) = 0, \quad \forall\, \mathbf{R}(\Theta) \neq \mathbf{R}(\Theta_*).$$

A consistent confidence-region sequence is such that as $n \to \infty$, all false values of $\mathbf{R}(\Theta)$ are ultimately "filtered out" with probability → 1.

Based on the duality between hypothesis tests and confidence regions, we can state that the confidence interval for $\mu$ in Ex. 10.21, for $\sum_{i=1}^{3} \beta_i$ and $\beta_1$ in Ex. 10.22, and the confidence interval for $\theta$ in Ex. 10.23 are all unbiased, UMAU, and consistent. The confidence interval for $p$ in Ex. 10.24 is approximately (asymptotically) unbiased and UMAU, and is also consistent. Finally, the ellipsoid confidence region for $(\beta_1, \beta_2, \beta_3)$ in Ex. 10.25 is unbiased and consistent.

### Confidence Regions from Pivotal Quantities

A method of defining a confidence region for $\mathbf{R}(\Theta)$ without invoking duality with a hypothesis test involves *pivotal quantities*, defined as follows.

**Definition 10.8** *Pivotal Quantities*

Let the random sample $\mathbf{X}$ have PDF $f(\mathbf{x}; \Theta)$. A function of $\mathbf{X}$ and $\mathbf{R}(\Theta)$, $Q = q(\mathbf{X}, \mathbf{R}(\Theta))$, whose probability distribution does not depend on the value of $\Theta \in \Omega$ is called a **pivotal quantity** for $\mathbf{R}(\Theta)$.

The fact that the pivotal quantity has a fixed probability distribution that does not depend on the parameter vector $\Theta$ allows a confidence region for $\mathbf{R}(\Theta)$ to be defined via the following method.

**Theorem 10.12** (Pivotal Quantity Method of Confidence-Region Construction) *Let* $Q = q(\mathbf{X},\mathbf{R}(\Theta))$ *be a pivotal quantity for* $\mathbf{R}(\Theta)$. *Define the values* $\ell$ *and* $\mu$ *so that* $P(\ell < q(\mathbf{x},\mathbf{R}(\Theta)) < \mu;\Theta) = \gamma$. *Then* $A(\mathbf{X}) = \{\mathbf{R}(\Theta)\colon \ell < q(\mathbf{X},\mathbf{R}(\Theta)) < \mu\}$ *defines a confidence region for* $\mathbf{R}(\Theta)$ *having confidence level (and coefficient)* $\gamma$.

**Proof** Since $Q$ is a pivotal quantity, $P(\ell < q(\mathbf{x}, \mathbf{R}(\Theta)) < \mu; \Theta) = \gamma$ holds for every $\Theta \in \Omega$ and, in particular, for the true $\Theta_0 \in \Omega$. Also note that for a given outcome of $\mathbf{x}$, $\ell < q(\mathbf{x}, \mathbf{R}(\Theta)) < \mu \Leftrightarrow \mathbf{R}(\Theta) \in A(\mathbf{x})$. It follows that $P(\mathbf{R}(\Theta) \in A(\mathbf{x}); \Theta) = \gamma$ $\forall\, \Theta \in \Omega$, and thus by Def. 10.3, $A(\mathbf{X})$ defines a confidence region for $\mathbf{R}(\Theta)$ with confidence level (and coefficient) $\gamma$. ■

Historically, the reason why the random variable $Q = q(\mathbf{X}; \mathbf{R}(\Theta))$ of Def. 10.8 and Theorem 10.12 was called a "pivotal" quantity is because when $R(\Theta)$ is a scalar many such random variables are such that the $\mathbf{x}$ argument in $\ell < q(\mathbf{x}, R(\Theta)) < \mu$ can be "pivoted" (or better, inverted) out of the center term to yield the alternative inequality representation $t_\ell(\mathbf{x}) < \mathbf{R}(\Theta) < t_\mu(\mathbf{x})$. The latter inequality defines the confidence interval $(t_\ell(\mathbf{X}), t_\mu(\mathbf{X}))$ for $\mathbf{R}(\Theta)$, which is the confidence region $A(\mathbf{X})$ of Def. 10.2. In practice, while it is often the case that such "pivoting" can be accomplished so that a confidence *interval* is defined, the characteristics of a pivotal quantity in Def. 10.8 do *not* guarantee that $q(\mathbf{x}, \mathbf{R}(\Theta))$ can be pivoted, in which case $A(\mathbf{X})$ may *not* be an *interval*. Of course, if $\mathbf{R}(\Theta)$ is a $(j \times 1)$ vector and $j \geq 2$, then obtaining a confidence *interval* will be neither possible nor relevant. The reader may find it interesting to know that *every* example of confidence regions examined heretofore can be motivated within the context of the pivotal quantity method, as will be seen in the next example.

**Example 10.26** Revisit Ex. 10.21 through Ex. 10.25. For each of the examples, Table 10.1 identifies the pivotal quantity that can be used to derive the confidence region for the respective $\mathbf{R}(\Theta)$ as defined in the example. In addition, the table identifies the fixed PDF for each pivotal quantity that applies regardless of the value of $\Theta \in \Omega$, the event for each pivotal quantity from which the confidence region is defined, and the resultant explicit form of the confidence region for $\mathbf{R}(\Theta)$. Note that in the case of Ex. 10.24, the table identifies an **asymptotic pivotal quantity**, meaning a random variable of the form $q(\mathbf{X}, \mathbf{R}(\Theta))$ whose limiting distribution does not depend on $\Theta \in \Omega$. In practice, such asymptotic pivotal quantities are motivated via central limit theory. □

**Table 10.1** Pivotal Quantities and Some Associated Confidence Regions

| Ex. | R(Θ) | Pivotal Quantity, $Q$ | PDF for $Q$, $\forall \ \Theta \in \Omega$ | Event for $Q$ | Level $1-\alpha$ Confidence Region, $A$ |
|---|---|---|---|---|---|
| 10.21 | $\mu$ | $\dfrac{\bar{X}-\mu}{S/\sqrt{n-1}}$ | $t$-distribution,<br>$n-1$ degrees of freedom | $-t_{\alpha/2} < Q < t_{\alpha/2}$ | $(\bar{X}-t_{\alpha/2}S/\sqrt{n-1}, \bar{X}+t_{\alpha/2}S/\sqrt{n-1})$ |
| 10.22 | $R\beta$ | $\dfrac{\mathbf{R}\hat{\beta}-\mathbf{R}\beta}{[\hat{S}^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{1/2}}$ | $t$-distribution,<br>$n-k$ degrees of freedom | $-t_{\alpha/2} < Q < t_{\alpha/2}$<br>$-t_\alpha < Q < \infty$ | $(\mathbf{R}\hat{\beta}-t_{\alpha/2}D, \mathbf{R}\hat{\beta}+t_{\alpha/2}D)$<br>where $D=[\hat{S}^2\mathbf{R}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{R}']^{1/2}$<br>$(-\infty, \mathbf{R}\hat{\beta}+t_\alpha D)$ |
| 10.23 | $\theta$ | $\bar{X}/\theta$ | Gamma$(n, n^{-1})$ | $Q < g_\alpha$ | $(\bar{X}/g_\alpha, \infty)^a$ |
| 10.24 | $p$ | $\dfrac{n^{1/2}(\bar{X}-p)}{[\bar{X}(1-\bar{X})]^{1/2}}$ | $N(0,1)$<br>(asymptotically) | $-z_{\alpha/2} < Q < z_{\alpha/2}$ | $(\bar{X}-z_{\alpha/2}D, \bar{X}+z_{\alpha/2}D)$<br>where $D=\left[\dfrac{\bar{X}(1-\bar{X})}{n}\right]^{1/2}$ |
| 10.25 | $\mathbf{R}\beta$ | $q_2^{-1}(\mathbf{R}\hat{\beta}-\mathbf{R}\beta)'$<br>$[\hat{S}^2\mathbf{R}(\mathbf{x}_*'\mathbf{x}_*)^{-1}\mathbf{R}']^{1/2}(\mathbf{R}\hat{\beta}-\mathbf{R}\beta)$ | $F$-distribution, $q$ and<br>$n-k$ degrees of freedom | $Q < F_\alpha(q, n-k)$ | ellipsoid with center $\mathbf{R}\hat{\beta}$ |

$g_\alpha$ is such that $\int_{g_\alpha}^{\infty}$ Gamma$(x; n, n^{-1})dx = \alpha$.

A notable practical difficulty in using pivotal quantities to define confidence regions is the problem of finding a pivotal quantity for a given $\mathbf{R}(\Theta)$. There is no general method that can be relied upon to define pivotal quantities, and indeed a pivotal quantity need not exist at all. However, for certain special but important classes of problems, pivotal quantities are readily available.

**Theorem 10.13** (Pivotal Quantities for Location–Scale Parameter Families of PDFs) *Let* $\mathbf{X}$ *be a random sample from a population distribution* $f(z;\Theta)$. *Let* $\hat{\Theta}$ *denote the MLE of* $\Theta$. *The following relationships exist between the functional form of* $f(z;\Theta)$ *and pivotal quantities for the elements of* $\Theta$, *where* $f_0(y)$ *denotes a PDF whose values do not depend on unknown parameters*:

1. *Location parameter family of PDFS*: $f(z;\Theta) = f_0(z - \Theta) \Rightarrow Q = \hat{\Theta} - \Theta$ *is a pivotal quantity for a scalar* $\Theta$.
2. *Scale parameter family of PDFs*: $f(z;\Theta) = \Theta^{-1}f_0(z/\Theta) \Rightarrow Q = \hat{\Theta}/\Theta$ *is a pivotal quantity for a scalar* $\Theta$.
3. *Location–scale family of PDFs*: $f(z;\Theta) = \Theta_2^{-1}f_0((z - \Theta_1)/\Theta_2) \Rightarrow Q_1 = (\hat{\Theta}_1 - \Theta_1)/\hat{\Theta}_2$ *and* $Q_2 = \hat{\Theta}_2/\Theta_2$ *are pivotal quantities for* $\Theta_1$ *and* $\Theta_2$ *in the* $(2 \times 1)$ *parameter vector* $\Theta$.

**Proof** C. E. Antle, and L. J. Bain (1969), "A property of maximum likelihood estimators of location and scale parameters." *SIAM Review*, 11, p. 251. ■

The theorem indicates that as long as random sampling is from a population distribution having one of the indicated general forms, there are specific functions of the MLEs that define pivotal quantities for the parameters of the population distribution.

A more general result that applies to random sampling from *any* continuous PDF having a scalar parameter $\Theta$ is as follows.

**Theorem 10.14** (Pivotal Quantities for Continuous Population PDFs) *Let* $\mathbf{X}$ *be a random sample from the continuous PDF* $f(z;\Theta)$, *where* $\Theta$ *is a scalar. Then*

$$-2\sum_{i=1}^{n} \ln F(X_i;\Theta) \sim \chi^2_{2n}$$

*is a pivotal quantity for* $\Theta$, *where* $F(z;\Theta)$ *is the CDF for* $Z$.

**Proof** If $Z \sim f(z;\Theta)$, then the probability integral transform of $Z$, $F(Z;\Theta)$, is distributed uniformly $(0,1)$ (recall Theorem 6.22). It follows that $W = -\ln F(Z;\Theta) \sim$ Exponential(1) (use the change-of-variables approach to derive the PDF of $W$). Then $-2\sum_{i=1}^{n} \ln F(X_i;\Theta)$ is 2 times the sum of $n$ independent Exponential(1) random variables, which has a $\chi^2$-distribution with $2n$ degrees of freedom (use the MGF approach to derive the PDF). Since $Q = q(X,\Theta) = -2\sum_{i=1}^{n} \ln F(X_i;\Theta) \sim \chi^2_{2n}$ $\forall\, \Theta \in \Omega$, it follows that $Q$ is a pivotal quantity for $\Theta$. ■

Theorem 10.14 implies that pivotal quantities *always* exist for a scalar $\Theta$ when random sampling is from a continuous PDF $f(x;\Theta)$. The reader will be asked to demonstrate that $-2\sum_{i=1}^{n}\ln[1-F(X_i;\Theta)] \sim \chi^2_{2n}$ is also a pivotal quantity for $\Theta$ (see Problem 10.21).

**Example 10.27** The proportion of the work day that a particular assembly line is stopped because of malfunctions on the line is the outcome of a random variable with PDF $f(z;\Theta) = \Theta z^{\Theta-1} I_{[0,1]}(z)$, for $\Theta > 0$. Based on a random sample of $n$ daily observations from $f(z;\Theta)$, define a confidence region for $\Theta$ having confidence coefficient .95.

**Answer**: To use Theorem 10.14, first note that the CDF of $Z$ is $F(b;\Theta) = b^{\Theta} I_{[0,1]}(b) + I_{(1,\infty)}(b)$. Then $-2\ln F(X_i;\Theta) = -2\Theta \ln X_i$, and thus

$$Q = -2\Theta \sum_{i=1}^{n} \ln X_i \sim \chi^2_{2n}$$

is a pivotal quantity for $\Theta$. The probability of the event $\{q\colon \chi^2_{2n;1-\alpha/2} < q < \chi^2_{2n;\alpha/2}\}$ is $1-\alpha$, and given a random sample outcome $\mathbf{x}$, the event can be pivoted to define a $(1-\alpha)$ level confidence interval for $\Theta$ as

$$\frac{\chi^2_{2n;1-\alpha/2}}{-2\sum_{i=1}^{n}\ln x_i} < \Theta < \frac{\chi^2_{2n;\alpha/2}}{-2\sum_{i=1}^{n}\ln x_i}.$$

(Note when "pivoting" that $-2\sum_{i=1}^{n}\ln x_i > 0$ in this case, and so the sense of the inequalities does not reverse in the pivot operation.) □

Regarding the choice of the pivotal quantity event from which the confidence region is defined, if the pivotal quantity is monotone in $R(\Theta)$ and a confidence interval or bound is desired, the following relationships are useful.

| Pivotal Quantity Event | Monotonicity of $q(\mathbf{x}, R(\Theta))$ | Confidence Interval |
|---|---|---|
| $\ell < q < \mu$ | increasing in $R(\Theta)$ | $t_1(x) < R(\Theta) < t_2(x)$ |
| | decreasing in $R(\Theta)$ | $t_1(x) < R(\Theta) < t_2(x)$ |
| $-\infty < q < \mu$ | increasing in $R(\Theta)$ | $-\infty < R(\Theta) < t(x)$ |
| | decreasing in $R(\Theta)$ | $t(x) < R(\Theta) < \infty$ |
| $\ell < q < \infty$ | increasing in $R(\Theta)$ | $t(x) < R(\Theta) < \infty$ |
| | decreasing in $R(\Theta)$ | $-\infty < R(\Theta) < t(x)$ |

Also note that often many choices of $\ell$ and $\mu$ are available such that $P(\ell < q < \mu) = \gamma$, e.g., an infinite number in the case when $q$ is a continuous random variable and $\gamma \in (0,1)$. If $q(\mathbf{x}, R(\Theta))$ is *monotone* in $R(\Theta)$, then it is desirable to choose $\ell$ and $\mu$ so that the length, or expected length, of the resultant confidence *interval* is minimized. However, in practice, a rule of thumb for two-sided confidence intervals is to choose $\ell$ and $\mu$ so as to define "equal-tail probabilities" in the definition of a $\gamma$-level confidence interval. That is, $\ell$ and $\mu$ are chosen such that $P(q < \ell) = P(q > \mu) = \alpha/2$, leading to a $\gamma = (1-\alpha)$-level confidence interval. Note that this principle was followed when defining the confidence interval in Ex. 10.27. The expected length of the confidence interval could have been reduced slightly by choosing the lower and upper $\chi^2$ values, say $\chi^2_{\alpha_\ell}$ and

$\chi^2_{\alpha_\mu}$, so that $\chi^2_{\alpha_\mu} - \chi^2_{\alpha_\ell}$ is minimized subject to the confidence-level condition $P(\chi^2_{\alpha_\ell} < q < \chi^2_{\alpha_\mu}) = \gamma$, which can be accomplished on a computer.

The pivotal quantity method does not guarantee that any optimal properties will apply to the confidence regions derived from it. However, the method is relatively straightforward, it applies in many cases of practical importance, and the confidence regions it defines are most often quite adequate for their intended purpose.

### Confidence Regions as Hypothesis Tests

We briefly note that since there is a full duality between confidence regions and hypothesis tests, one can pursue the duality in the reverse of what we have done heretofore and derive hypothesis-testing procedures from confidence regions. Effectively, a null hypothesis is accepted or rejected depending on whether or not (respectively) the hypothesized value of $\mathbf{R}(\Theta)$ is contained in the confidence region. It follows that the preceding pivotal quantity method can be used to define hypothesis tests. A hypothesis test is unbiased, UMP, UMPU, and/or consistent according to whether the confidence region is unbiased, UMA, UMAU, and/or consistent, respectively.

We will not pursue this further since for the approach to be of substantive practical use requires that we establish methods of deriving UMA, UMAU, and/or consistent critical regions independent of the hypothesis test to which it is dual. This requires further study of the theory of confidence-region estimation, which the reader can begin by referring to the readings suggested at the beginning of this section.

## 10.8 Nonparametric Tests of Distributional Assumptions

In this section we provide an introduction to nonparametric testing of the distributional assumptions underlying a point estimation, hypothesis-testing, or confidence-region estimation problem. By "nonparametric" we mean that the hypotheses under consideration are not defined in terms of the values of parameters, per se, as has been the case heretofore. Rather, hypotheses examined here will be more general and refer to functional forms of density functions and to whether random variables contained in a sample are iid. The tests we will introduce can be used in practice to address questions such as "could the random sample have come from a normal or exponential, or Beta, or ... distribution?" and "it is possible that the outcomes observed in the sample are outcomes of a random sample from some population distribution?"

Note that to a large degree the methods of statistical inference that we have examined so far required certain basic assumptions to hold (collectively representing the *maintained hypothesis*) before any analysis could proceed. It is clearly useful to have methods of assessing the validity of these assumptions when they are held only tentatively and/or without conviction. Two of the more pervasive assumptions made to this point have been the assumption of a specific functional form for the density of the random sample and the assumption that random variables in the sample are iid. We will focus on some nonparametric tests relating to these assumptions that have been useful in practice and also concentrate on scalar random variables. The field of nonparametric analysis is vast and growing. A useful place to begin additional reading and to seek additional references is J. D. Gibbons (1985), *Nonparametric Methods for Quantitative Analysis*, 2nd ed., Columbus, OH: American Science Press.

### Functional Forms of Probability Distributions

A number of testing procedures are available for testing hypotheses regarding the functional form of the joint density of the random sample (see M. Kendall and A. Stuart, vol. 2, *Advanced Statistics*, Chapter 30; and C. Huang and B. Bolch (1974), "On testing of regression disturbances for normality." *JASA*, pp. 330–335, for alternatives and references). We will examine the $\chi^2$ goodness-of-fit test because of its versatility, the Kolmogorov–Smirnov test because of its refinement in analyzing continuous distributions, and the Shapiro–Wilks test because of its superiority in testing the normality assumption.

*$\chi^2$ Goodness-of-Fit Test* The $\chi^2$ goodness-of-fit test is used to test the null hypothesis that a random sample is from a population distribution of the form $f(z; \Theta)$ where $\Theta$ is a $(k \times 1)$ vector. In particular, the hypotheses under consideration are $H_0$: $\mathbf{X} \sim \prod_{i=1}^n f(x_i; \Theta)$ versus $H_a$: $\mathbf{X} \not\sim \prod_{i=1}^n f(x_i; \Theta)$. The test procedure differs depending on whether the null hypothesis is simple, meaning that $f(z; \Theta)$ is fully specified (i.e., it does not contain any unknown parameters), or the null hypothesis is composite, that is, $\Theta$ in $f(z; \Theta)$ is left unspecified so that a *family* of density functions is specified as the null hypothesis.

We first examine the case where the null hypothesis is *simple*. Given the null hypothesis $H_0$: $Z \sim f(z; \Theta_0)$ for the population distribution where $\Theta_0$ is fixed and known, and assuming the range of $Z$ has been partitioned into $m$ subintervals $\Delta_i$, $i = 1, \ldots, m$, it follows that under $H_0$, $P(\Delta_i; \Theta_0) = \int_{z\in\Delta_i} f(z; \Theta_0)dz = p_i(\Theta_0)$, for $i = 1, \ldots, m$. For a random sample of size $n$, let $n_i$ represent the number of times an outcome occurs in interval $\Delta_i$, $i = 1, \ldots, m$. Then the probability distribution of $(N_1, \ldots, N_m)$ is multinomial with parameters $(p_1, \ldots, p_m)$, $\sum_{i=1}^m p_i = 1$ (recall Chapter 4).

Using the multivariate version of the LLCLT, it follows that $n^{1/2}(\bar{\mathbf{X}}_* - \mathbf{p}_*) \xrightarrow{d} N([\mathbf{0}], \boldsymbol{\Sigma}_*)$, where $\bar{\mathbf{X}}_* = n^{-1}(N_1, \ldots, N_{m-1})'$, $\mathbf{p}_* = (p_1, \ldots, p_{m-1})'$, and $\boldsymbol{\Sigma}_*$ is an $(m-1) \times (m-1)$ covariance matrix with $p_i(1-p_i)$, $i = 1, \ldots, m-1$, along the diagonal and $-p_ip_j$ in the off-diagonal entries. (Note that $n_m$ is determined by $n_m = n - \sum_{i=1}^{m-1} n_i$, and thus $n_m/n = 1 - \sum_{i=1}^{m-1} \bar{x}_{*i}$.) Then from Theorem 5.9,

$$W = n(\bar{\mathbf{X}}_* - p_*)'\boldsymbol{\Sigma}_*^{-1}(\bar{\mathbf{X}}_* - \mathbf{p}_*) = (\mathbf{N}_* - n\mathbf{p}_*)(n\boldsymbol{\Sigma}_*)^{-1}(\mathbf{N}_* - n\mathbf{p}_*) \xrightarrow{d} \chi^2_{m-1},$$

where $\mathbf{N}_* = (N_1, \ldots, N_{m-1})'$. It follows that a size $\alpha$ test of $H_0$: $Z \sim f(z; \Theta_0)$ versus $H_0$: $Z \not\sim f(z; \Theta_0)$ is given by

$$w\begin{bmatrix} \geq \\ < \end{bmatrix}\chi^2_{m-1,\alpha} \Rightarrow \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix},$$

where $(p_1, \ldots, p_{m-1})$ in $\mathbf{p}_*$ and in $\boldsymbol{\Sigma}_*$ are replaced by their values under $H_0$, i.e., $p_i = p_i(\Theta_0)$, $i = 1, \ldots, m-1$. The test is seen to reject $H_0$ when the probabilities of the interval events $i$, $i = 1, \ldots, m$, as hypothesized under $H_0$ are in conflict with their estimated values, given by the relative frequencies, $n_i/n$.

In practice, a statistic that is algebraically and numerically equivalent to $W$ is generally used to calculate the test outcome. In particular, some tedious but conceptually straightforward matrix algebra leads to

$$w = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}$$

(Kendall and Stuart, vol. 2, *Advanced Statistics*, p. 381).

Now suppose that the null hypothesis is *composite* as $H_0 \colon Z \sim f(z; \Theta)$, $\Theta \in \Omega$, versus $H_a \colon Z \not\sim f(z; \Theta)$, $\Theta \in \Omega$. How do we proceed in this case when $\Theta$ is left unspecified? The unknown parameters are estimated from the data, but how this is done has a substantial impact on the form of the test procedure. Consider two different forms of MLEs, the difference being the choice of likelihood function.

In one case, let the likelihood function be the multinomial distribution parameterized via $\Theta$ under $H_0$, i.e.,

$$L(\Theta; \mathbf{x}) = \binom{n}{n_1 \ldots n_m} \prod_{i=1}^{m} p_i(\Theta)^{n_i} \quad \text{for } n_i \in \{0, 1, \ldots, n\}, \quad \forall\, i, \sum_{i=1}^{n} n_i = n,$$

where $p_i(\Theta) = \int_{z \in \Delta_i} f(z; \Theta) dx \;\forall\, i$. After one obtains the maximum likelihood estimate of $\Theta$, $\hat{\theta}$, one proceeds by replacing the $p_i$'s by $\hat{p}_i = p_i(\hat{\theta})$ in the calculation of the $w$-statistic, obtaining

$$w = \sum_{i=1}^{m} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}.$$

Finally, the test rule is defined as

$$w \begin{bmatrix} \geq \\ < \end{bmatrix} \chi^2_{m-1-k;\alpha} \Rightarrow \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix}.$$

The reason that one degree of freedom is lost for each parameter estimated is somewhat involved. A detailed proof can be found in H. Cramér (1946), *Mathematical Methods of Statistics*, Princeton, NJ: Princeton Univ. Press.

An alternative ML procedure is to estimate $\Theta$ using the ML method applied to the likelihood function $L(\Theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \Theta)$, i.e., use the hypothesized population distribution *directly* to specify the likelihood function, and estimate $\Theta$ accordingly. One then computes $\hat{p}_i = \int_{z \in \Delta_i} f(z; \hat{\theta}) dz$ to calculate the $w$-statistic above. In this case, it turns out that $W$ does *not* have a limiting $\chi^2$-distribution at all. However, it can be shown that the distribution of $W$ in this case is bounded *between* a $\chi^2_{m-1}$ and a $\chi^2_{m-1-k}$ distribution (H. Chernoff and E. L. Lehmann (1954), "The use of maximum likelihood estimates in $\chi^2$ tests for goodness of fit." *Ann. Math. Statist.*, p. 579). Unfortunately, the limiting distribution is difficult to use, and in practice one can use $\chi^2_{m-1;\alpha}$ to define a size $\leq \alpha$ test, or else use $\chi^2_{m-1-k;\alpha}$ to define a test that is more likely to reject a true $H_0$ than the size $\alpha$ would indicate. The conservative (toward $H_0$) approach would be to utilize the latter critical value.

It can be shown that all of the preceding tests are *consistent* for any alternatives that imply multinomial probabilities different from those implied by $H_0$. On the other hand, the tests are generally *biased* to some degree.

One operational problem remains. How does one choose the sets $\Delta_i$, $i = 1, \ldots, m$? Rules of thumb developed from both empirical and theoretical considerations suggest that intervals be defined that have equal probability, i.e., $1/m$, under $f(x; \Theta_0)$ (simple $H_0$) or $f(x; \hat{\theta})$ (composite $H_0$). Furthermore, in order for the asymptotics to be a reasonable approximation, the intervals chosen should be such that $np_i(\Theta_0)$ or $np_i(\hat{\theta}) \geq 5 \ \forall \ i$.

**Example 10.28** Testing for an Exponential Population Distribution

Forty observations on the waiting times, in minutes, between customer arrivals at a service station/convenience store in a mid-size city were as follows:

| | | | |
|---|---|---|---|
| 1.37 | 1.96 | 0.74 | 0.42 |
| 0.12 | 0.61 | 1.98 | 1.76 |
| 1.73 | 3.32 | 1.44 | 2.46 |
| 0.34 | 2.31 | 2.14 | 2.11 |
| 2.84 | 2.47 | 1.25 | 0.66 |
| 2.23 | 1.11 | 1.73 | 0.26 |
| 1.77 | 1.35 | 2.91 | 0.93 |
| 1.50 | 2.72 | 1.73 | 0.59 |
| 0.36 | 0.24 | 2.68 | 0.30 |
| 0.10 | 2.75 | 1.68 | 0.88 |

Use a (approximate) size .05 $\chi^2$ goodness-of-fit test to test the null hypothesis that the waiting times are exponentially distributed versus some other family of distributions, i.e., $H_0: z \sim \theta^{-1}\exp(-z/\theta)I_{(0,\infty)}(x)$, $\theta > 0$, versus $H_a$: not $H_0$.

**Answer**: The ML estimate of $\theta$ based on the exponential population distribution assumption is $\hat{\theta} = \bar{x} = 1.494$. We let $m = 8$, and choose $\Delta_i$, $i = 1, \ldots, 8$, so that $P(z \in \Delta_i; \hat{\theta}) = (1/8) \ \forall \ i$. The boundaries of the intervals are found as follows:

$$.125 = \int_0^{c_1} (1.494)^{-1} \exp(-z/1.494) dz \Rightarrow c_1 = .1995,$$

$$.125 = \int_{.1995}^{c_2} (1.494)^{-1} \exp(-z/1.494) dz \Rightarrow c_2 = .4298,$$

$$\begin{aligned} &\cdot & c_3 &= .7022, \\ &\cdot & c_4 &= 1.0356, \\ &\cdot & c_5 &= 1.4654, \end{aligned}$$

$$c_6 = 2.0711,$$

$$.125 = \int_{2.0711}^{c_7} (1.494)^{-1} \exp(-z/1.494)dz \Rightarrow c_7 = 3.1067.$$

The intervals and the number of observations that actually occurred in each interval are as follows:

| $i$ | $\Delta_i$ | $n_i$ |
|---|---|---|
| 1 | (0, .1995] | 2 |
| 2 | (.1995, .4298] | 6 |
| 3 | (.4298, .7022] | 3 |
| 4 | (.7022, 1.0356] | 3 |
| 5 | (1.0356, 1.4654] | 5 |
| 6 | (1.4654, 2.0711] | 9 |
| 7 | (2.0711, 3.1067] | 11 |
| 8 | (3.1067, ∞) | 1 |

Since $n\hat{p}_i = 40(.125) = 5$, the value of the $\chi^2$ statistic is given by

$$w = \sum_{i=1}^{8} \frac{(n_i - 5)^2}{5} = 17.2.$$

Adopting a conservative (toward $H_0$) stance and using the critical value $\chi^2_{m-1-k;.05} = \chi^2_{6;05} = 12.6$, $w > 12.6$ and the exponential family of distributions is rejected. Note that if the more liberal critical value of $\chi^2_{m-1;.05} = \chi^2_{7;.05} = 14.1$ is used, $w > 14.1$ and $H_0$ is still rejected, and we know that the significance level of this test $\leq .05$. □

*Kolmogorov–Smirnov and Lilliefors Tests* The Kolmogorov–Smirnov (K–S) test is an alternative procedure for testing whether a random-sample outcome was drawn from a specified population distribution. The test procedure is based on the empirical distribution function and the Glivenko–Cantelli theorem (recall Chapter 6), which effectively indicated that the EDF would converge functionally across all points of comparison to the true population CDF associated with a random sample. The basic idea of the K–S test is to reject $H_0$ when there are significant discrepancies between an EDF and a hypothesized CDF as revealed through the probability integral transform of the data (recall Section 6.8).

Before examining the test procedure in more detail, we note some advantages and limitations of the K–S test relative to the aforementioned $\chi^2$ test. On the positive side, the K–S test is fully applicable in the case of small samples, whereas the $\chi^2$-test is strictly only an asymptotically valid test. The K–S test deals with sample observations directly, but the data must be summarized into categories for analysis via the $\chi^2$ approach, and information can potentially be lost in the categorization process. The K–S test assumes the population distribution is continuous and thus provides a more refined analysis specific to this case. Limitations include the fact that the K–S approach cannot be easily adjusted to allow for estimation of unknown parameters as in the $\chi^2$ case, so that in practice the K–S test is generally restricted to testing simple null hypotheses. Also, the applicability of K–S is limited to cases involving continuous distributions. We note that the K–S test has been modified by H. Lilliefors ("On the Kolmogorov-Smirnov test for normality with mean and variance unknown." *JASA*, 1967, pp. 399–402) to accommodate the estimation of $\mu$ and $\sigma^2$ when the null hypothesis is that of *normality*. We will examine this case later. Some work has also been done on extending the K–S approach to exponential and Weibull cases where parameters are unknown.[13]

In order to identify the test procedure, we begin with the null hypothesis for the population distribution $H_0 : Z \sim F_0(z)$ versus $H_a : Z \not\sim F_0(z)$, where $F_0(z)$ is a *completely* specified CDF. Let $X_1, \ldots, X_n$ be a random sample, and let $\hat{F}_n(z)$ be the EDF based on the random sample. The **Kolmogorov–Smirnov test statistic** is given by[14]

$$d_n = \sup_z |\hat{F}_n(z) - F_0(z)|.$$

The value of $d_n$ is seen to be the largest distance between the hypothesized CDF, $F_0(z)$, and the EDF estimate of the true CDF, $\hat{F}_n(z)$. The larger is $d_n$, the greater is the largest numerical discrepancy between the estimated and hypothesized CDFs. Note also that $y_0 = F_0(z)$ is effectively the probability integral transform of the random variable outcome $z$ as conjectured in $H_0$, while the EDF can be viewed as estimating outcomes from the true probability integral transform $y = F(z)$. Thus, an alternative interpretation of $d_n$ is the largest discrepancy between hypothesized and estimated probability integral transforms of $Z$.

It is useful to note for computational purposes that $d_n = \max(d_n^+, d_n^-)$, where

$$d_n^+ = \max_{i \in \{1,\ldots,n\}} (i/n - F_0(x_{(i)})),$$

$$d_n^- = \max_{i \in \{1,\ldots,n\}} (F_0(x_{(i)}) - (i-1)/n),$$

where $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are the random-sample outcomes ordered from lowest to highest (i.e., they are the *order statistics*). This computational approach makes the calculation of $d_n$ a simple matter on a computer through sorting and differencing operations.

In order to decide whether a discrepancy is significant, the sampling distribution of $D_n$ is required. The exact distribution of $D_n$ for $n \le 40$ has been tabled by J. D. Gibbons, *Nonparametric Methods*, p. 400. It was shown by A. N. Kolmogorov (*Giorn. Inst. Ital. Attuari*, 4 (1933), pp. 83–91) that the limiting CDF of $n^{1/2}D_n$ could be represented as

$$\lim_{n\to\infty} P(n^{1/2}d_n \le t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t^2},$$

[13] See M. A. Stephens (1974), "EDF statistics for goodness-of fit and some comparisons," *JASA*, p. 730; and M. Chandra, et al. (1981), "Kolmogorov statistics for tests of fit for the extreme-value and Weibull distributions," *JASA*, p. 729.

[14] As always, "sup" can be replaced by "max" when the maximum exists.

which results in the approximate critical points

| $\alpha$ | .10 | .05 | .01 |
|---|---|---|---|
| $d_{n;\alpha}$ | $1.224n^{-1/2}$ | $1.358n^{-1/2}$ | $1.628n^{-1/2}$ |

Stephens (see footnote 6) has analyzed adjustments to these approximate critical values that make them accurate for small $n$. The adjustments are to divide each of them by the factor $(1+.12n^{-1/2}+.11n^{-1})$.

A size $\alpha$ K–S test of $H_0: Z \sim F_0(z)$ versus $H_a: Z \not\sim F_0(z)$ is given as follows:

$$d_n \begin{bmatrix} \geq \\ < \end{bmatrix} d_{n;\alpha} \Rightarrow \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix}.$$

From the Glivenko–Cantelli theorem and the fact that $d_{n;\alpha} \to 0 \ \forall\, \alpha \in (0,1)$, it follows that the test is consistent.

**Example 10.29** Revisit Ex. 10.28, and examine the null hypothesis that the distribution of waiting times is uniform over the interval (0, 3.5), i.e., $H_0: Z \sim F_0(z)$ versus $H_a: Z \not\sim F_0(z)$ with $F_0(z) = z/3.5 I_{(0,3.5)}(z)$. The value of $d_n = \sup_{z\in(0,3.5)} |\hat{F}_n(z) - z/3.5| = .1443$, the maximum difference occurring for $z = 2.47$, which has an EDF value of .85 and a (hypothesized) CDF value of $2.47/3.5 = .7057$. Stephens' adjusted critical value for $n = 40$ and a size $\alpha = .05$ test is given by $d_{40;.05} = [1.358/(40)^{1/2}]/(1+.12/(40)^{1/2}+.11/40) = .2102$. Since $d_n = .1443 < d_{40;.05} = .2102$, we do *not* reject the hypothesis that waiting times are uniformly distributed over the interval (0, 3.5). □

Lilliefors provided an adjustment to the K–S test that justifies its use for testing the null hypothesis $H_0: Z \sim N(\mu, \sigma^2)$ versus $H_a: Z \not\sim N(\mu, \sigma^2)$ when $\mu$ and $\sigma^2$ are *unknown* and must be estimated from the data. In particular, he tabled the distribution of $D_n$ under the null hypothesis for $n = 1, \ldots, 30$, via Monte Carlo methods (see previous Lilliefors reference) and presented approximate critical values that are accurate for $n > 30$ as follows:

| $\alpha$ | .10 | .05 | .01 |
|---|---|---|---|
| $d^*_{n;\alpha}$ | $.805n^{-1/2}$ | $.886n^{-1/2}$ | $1.031n^{-1/2}$ |

The test proceeds by first unbiasedly estimating $\mu$ and $\sigma^2$ using the estimates $\bar{x}$ and $\hat{\sigma}^2 = ns^2/(n-1)$ (using the sample variance itself would lead to the same test size asymptotically). Then $N(\bar{x}, \hat{\sigma}^2)$ is used for the distribution under the null hypothesis, and $d_n = \sup_z |\hat{F}_n(z) - \int_{-\infty}^{z} N(z; \bar{x}, \hat{\sigma}^2)dz|$ is calculated. Finally, the K–S test is performed as before but using Lilliefors' critical values, $d^*_{n;\alpha}$, for $d_{n;\alpha}$.

**Example 10.30** Revisit Ex. 10.28, and use the K–S test, as adjusted by Lilliefors, to test the hypothesis $H_0: Z \sim N(\mu, \sigma^2)$ versus $H_a: Z \not\sim N(\mu, \sigma^2)$ for some $\mu$ and $\sigma^2$. Use $\alpha = .05$.

**Answer**: The estimates of $\mu$ and $\sigma^2$ are respectively $\hat{\mu} = \bar{x} = 1.494$ and $\hat{\sigma}^2 = .8338$. Then using $N(1.494, .8338)$ for the distribution of $Z$ under the null hypothesis, the value of the K–S statistic can be calculated as

$$d_n = \sup_z |\hat{F}_n(z) - \int_{-\infty}^{(z-1.494)/.9131} N(z; 0, 1)dz | = .0955,$$

which occurs for $z = .74$, in which case the EDF value is .30 and

$$\int_{-\infty}^{.74} N(z; 1.494, .8338)dz = .2045.$$

Using Lilliefors' critical value of $.886/(40)^{1/2} = .1401$, the hypothesis of normality *cannot* be rejected with this sample outcome. □

While the Lilliefors K–S test is usually very effective in detecting a false $H_0$, in this case it has failed. It turns out that the waiting-times data were indeed generated by a uniform distribution (based on uniform random numbers generated by a computer), and so the Lilliefors test failed to reject a false $H_0$. Note that for large enough $n$, we would have rejected the hypothesis, since this K–S test is consistent. We note that Lilliefors' own Monte Carlo calculations indicated that among all of the alternative distributions he examined, his test had the *lowest* power against the uniform distribution, which may be why the test had difficulty in the preceding example.

*Shapiro–Wilk Test* The **Shapiro–Wilk (SW) test** was designed *specifically* tc test the hypothesis that a random sample is from an $N(\mu, \sigma^2)$ population distribution, where the null and alternative hypotheses are $H_0: Z \sim N(\mu, \sigma^2)$ and $H_a: Z \not\sim N(\mu, \sigma^2)$, respectively, for some (unspecified) $\mu$ and $\sigma^2$. The test procedure has been shown to perform especially well in comparison to other tests of normality when applied to the calculated residuals $\hat{\mathbf{e}}_i = \mathbf{y} - \mathbf{xb}$ of a least-squares estimate of the GLM $\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The null hypothesis in this case is that $\varepsilon_i \sim N(0, \sigma^2)\ \forall\ i$ (see C. Huang and B. Bolch (1974), *Testing Disturbances*). The procedure fairs well in more general settings also. The SW test was originally devised by S. Shapiro and M. Wilk in "An analysis of variance test for normality." *Biometrika*, (1965), pp. 591–611.

The test statistic devised by SW for testing the normality hypothesis is calculated as

$$w = \left[\sum_{i=1}^{m} a_{n-i+1}(x_{(n-i+1)} - x_{(i)})\right]^2 / ns^2, \qquad m = \begin{cases} n/2 & \text{if } n \text{ is even,} \\ (n-1)/2 & \text{if } n \text{ is odd,} \end{cases}$$

where $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ are the sample observations ordered from smallest to largest (i.e., the *order statistics*), $s^2$ is the sample variance, and the $a_i$'s are coefficients that are tabulated in the article by SW, pp. 603–604. The distribution of the test statistic is quite involved, but it has been tabulated by SW for sample sizes $\leq 50$ (p. 605 of SW article) and for various levels of tail probabilities. The test rule is given by

$$w\begin{bmatrix}\leq\\>\end{bmatrix}w_{n;1-\alpha}\Rightarrow\begin{bmatrix}\text{reject } H_0\\ \text{accept } H_0\end{bmatrix}.$$

The test procedure can be applied for small samples and is also a consistent test of $H_0$ versus $H_a$.

The heuristic motivation for the test lies in the fact that under the null hypothesis of normality, both the numerator and denominator of the SW statistic can be shown to be estimating the same quantity, $\sigma^2$, apart from constants. Under the alternative hypothesis, the denominator will still (apart from $n^{-1}$) be estimating $\sigma^2$, but such will not be the case for the numerator (in general). The difference in behavior of the numerator and denominator under $H_a$ is exploited by SW in the design of the test procedure, and further motivation can be found in their article.

**Example 10.31** The relationship between varying levels of advertising expenditures and the quantity demanded of a product being advertised was studied over a 20-week period, and 20 weekly observations were used to estimate the GLM $Y = \mathbf{x}\beta+\varepsilon$, where an outcome of $\mathbf{Y}$ represents the vector of 20 observations on quantities sold, and $\mathbf{x}$ represents a matrix of variables explaining the demand for the product, including the level of advertising expenditures. The calculated residuals, $\hat{e}$, from the least-squares fit of the model yielded the following values, $\hat{e}_*$, ordered from lowest to highest rowwise:

| | | | |
|---|---|---|---|
| −7.0890 | −5.6489 | −5.1510 | −4.9672 |
| −4.8390 | −3.9812 | −3.9161 | −3.6296 |
| −3.5914 | −3.2087 | −2.9827 | −2.7810 |
| −1.2857 | −1.1390 | 0.8321 | 5.3253 |
| 6.2581 | 11.3091 | 14.4254 | 16.0611 |

Use the SW test to assess the hypothesis that $H_0: \varepsilon = N([0], \sigma^2\mathbf{I})$ versus $H_a$: not $H_0$. Use $\alpha = .05$.

**Answer**: In this case, $ns^2 = \sum_{i=1}^n(\hat{e}_i - \bar{\hat{e}})^2 = \sum_{i=1}^n \hat{e}_i^2 = 906.0822$ (since $\bar{\hat{e}} = 0$). Since $n$ is even, $m = n/2 = 10$. To calculate the numerator of the SW statistic, we need the appropriate values of $a_{n-i+1} = a_{21-i}$, $i = 1, \ldots, 10$, which are given in SW's table as

| | |
|---|---|
| $a_{20} = .4734,$ | $a_{15} = .1334,$ |
| $a_{19} = .3211,$ | $a_{14} = .1013,$ |
| $a_{18} = .2565,$ | $a_{13} = .0711,$ |
| $a_{17} = .2085,$ | $a_{12} = .0422,$ |
| $a_{16} = .1686,$ | $a_{11} = .0140,$ |

Then

$$\sum_{i=1}^{10} a_{21-i}(x_{(21-i)} - x_{(i)}) = .4734(16.0611 + 7.0890) + .3211(14.4254 + 5.6489)$$

$$
\begin{aligned}
&+ .2565(11.3091 + 5.1510) + .2085(6.2581 + 4.9672) \\
&\quad\vdots \qquad\qquad\qquad\qquad\quad \vdots \\
&+ .0422(-2.7810 + 3.5914) + .0140(-2.9827 + 3.2087) \\
&= 26.8090.
\end{aligned}
$$

The critical value of $w$ is found from SW's table to be $w_{20;.05} = .905$. Then since $w = (26.8090)^2/906.0822 = .7932 \leq .905$, $H_0: \varepsilon \sim N([0], \sigma^2\mathbf{I})$ is *rejected* at the .05 level. □

### IID Assumption

We will examine a test of the hypothesis that $(X_1, \ldots, X_n)$ is a random sample from some population distribution, the alternative hypothesis being that $(X_1, \ldots, X_n)$ is a sample from some composite experiment. That is, $H_0$: $X_i$'s are iid versus $H_a$: $X_i$'s are *not* iid. The test we will present, the **Wald–Wolfowitz (WW) runs test**, depends on the concept of a *run* in the outcome of a random sample. To be able to define a run, we need to categorize the data according to some dichotomous criteria, which results in the sample outcomes' being transformed into a collection of iid Bernoulli outcomes if $H_0$ is true. A **run** is defined to be a succession of one or more identical values (1's or 0's) preceded and followed by a different value or else no value at all if the run occurs at the beginning or end of the sample sequence. For example in the sequence of 0's and 1's given by

0 0 1 1 1 0 1 0 1 1 0 0 1 1,

there are eight runs, which we differentiate by vertical lines as

00 | 111 | 0 | 1 | 0 | 11 | 00 | 11.

The basic idea underlying the test of the iid assumption based on runs is that if the random variables are truly iid, then neither too few nor too many runs should be observed in any given outcome of the sample. Too few runs could be indicative of grouping, clustering, or trending. Too many runs could indicate a systematic alternating pattern. The Wald–Wolfowitz runs test exploits this idea rigorously.

*Wald–Wolfowitz (WW) Runs Test* The WW test is concerned with testing whether or not $(X_1, \ldots, X_n)$ might be considered a random sample from some population distribution. According to some dichotomous characteristic, we can transform the outcome of $X_i$ into a 1 or 0, and we can apply the same dichotomous characterization to all of the $x_i$'s. Note that the characteristic can be something inherent to the experiment, such as the male/female characteristic in a case where $X_i$'s refer to some measurement of consumer's response in an (alleged) random sample of consumers. Alternatively, a dichotomy could be imposed on random variables whose outcomes are in the form of some numerical response by assigning a 1 to values exceeding a specified value (e.g., the median value) and a 0 to values below the specified value.

However the dichotomy is defined, note that if there are $n_1$ responses equaling 1 and $n_0$ responses equaling 0, then there are $\binom{n_1+n_0}{n_1}$ different ways to rearrange the $n_1$ 1's and $n_0$ 0's. If the $X_i$'s are truly iid (i.e., $H_0$ is true), then each of these rearrangements is equally likely, and the probability of each rearrangement is $\binom{n_1+n_0}{n_1}^{-1}$, by classical probability.

Now let $w_1$ and $w_0$ represent the number of runs involving 1's and involving 0's, respectively. By the definition of a run, it must be the case that $|w_1 - w_0| \leq 1$. Examine the case where the total number of runs, $w = w_1 + w_0$, is *even*, so that $w_1 = w_0 = w/2$, and consider the definition of sequences consisting of $n_1$ 1's and $n_0$ 0's that represent $w_1$ runs of 1's and $w_0$ runs of 0's. How many different ways can we define the sequence of $w_1$ sets of 1's in the collection of $w$ runs? This number is equivalent to the number of different ways $w_1 - 1 = w/2 - 1$ vertical lines can be inserted in the $n_1 - 1$ spaces between the $n_1$ 1's (recall our previous use of vertical lines to delineate runs) and equals $\binom{n_1-1}{w/2-1}$. Similarly, the number of different ways we can define the sequence of $w_0$ sets of 0's in the collection of $w$ runs is $\binom{n_0-1}{w/2-1}$. Finally, since the sequence of $w$ runs can begin with either a group of 0's or a group of 1's, there are $2\binom{n_1-1}{w/2-1}\binom{n_0-1}{w/2-1}$ different ways of obtaining $w$ runs when $w$ is even, and thus

$$P(w; n_1, n_0) = \frac{2\binom{n_1-1}{w/2-1}\binom{n_0-1}{w/2-1}}{\binom{n_0+n_1}{n_1}}, \quad \text{for even } w.$$

Now examine the case were $w = w_1 + w_0$ is *odd*, so that either $w_1 = (w+1)/2$ and $w_0 = (w-1)/2$ if the sequence begins and ends with a group of 1's, or else $w_1 = (w-1)/2$ and $w_0 = (w+1)/2$ if the sequence begins and ends with a group of 0's. Using logic analogous to the case where $w$ was even, we have that the number of different ways of obtaining $w$ runs when $w$ is odd is given by the numerator of

$$P(w; n_1, n_0) = \frac{\binom{n_1-1}{(w-1)/2}\binom{n_0-1}{(w-3)/2} + \binom{n_1-1}{(w-3)/2}\binom{n_0-1}{(w-1)/2}}{\binom{n_0+n_1}{n_1}}, \quad \text{for odd } w.$$

We define $P(w; n_1, n_0) = 0$ in all other cases.

The discrete density function $P(w; n_1, n_0)$ can be used to define upper and lower critical values, and thus upper and lower rejection regions for the test statistic outcome $w$. As usual, the size of the test will be determined by the choice of critical values and the associated upper and lower tail probabilities as indicated by $P(w; n_1, n_0)$.

It can be shown that the mean and variance of $P(w; n_1, n_0)$ are respectively given by

$$\mathrm{E}(W) = \frac{2n_1n_0}{n_1+n_0} + 1 \quad \text{and} \quad \mathrm{var}(W) = \frac{2n_1n_0(2n_1n_0 - n_1 - n_0)}{(n_1+n_0)^2(n_1+n_0-1)}.$$

Furthermore, as $n_1 \to \infty$ and $n_0 \to \infty$, it follows that $[W - \mathrm{E}(W)]/[\mathrm{var}(W)]^{1/2} \xrightarrow{d} N(0, 1)$, allowing asymptotically valid size $\alpha$ tests of the iid assumption to be defined in terms of standard normal critical values. The rate of convergence to the limiting distribution is very rapid, so that for $n_1 > 10$ and $n_0 > 10$, the normal distribution approximation is very good. A table of critical values based on $P(w; n_1, n_0)$ for $n \leq 20$ has been published by Swed and Eisenhart (1943), *Annals of Mathematical Statistics.* (14), pp. 66–87.

A size $\alpha = \alpha_\ell + \alpha_h$ runs test can be defined as follows:

$$w \begin{bmatrix} \in \\ \notin \end{bmatrix} [0, w_{n_1,n_0;1-\alpha_\ell}] \cup [w_{n_1,n_0;\alpha_h}, n] \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix}.$$

It should be noted that the test will have no power against alternatives to $H_0$ that do not alter the equally likely character of the different rearrangements of $n$, 1's and $n_0$ 0's. For example, if the $X_i$'s were independent $N(0, \sigma_i^2)$ random variables and outcomes were dichotomized on the basis of runs above and below the median of 0, then the aforementioned null distribution of $W$ would nevertheless hold, and the $WW$ test would have no power for recognizing the heteroskedasticity of the $X_i$'s. For this reason, not rejecting $H_0$ with the $WW$ test should be considered a necessary, but not sufficient condition for concluding that the $X_i$'s are iid.

**Example 10.32** Test whether the observations on the waiting times between customers given in Ex. 10.28 might be viewed as iid observations from some population distribution. Use a test of size .05.

**Answer**: The observations occurred sequentially rowwise in the data matrix of Ex. 10.28, and a median of the observation is $(1.50 + 1.68)/2 = 1.59$. Classifying the sample outcomes as 1's or 0's according to whether the outcome is > or < 1.59 yields the following, with runs delineated by vertical lines

0 | 1 | 0000 | 1111 | 0 | 1 | 0 | 11111 | 00 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 00 | 11 | 000 | 1 | 00 | 11 | 0.

In this case, there are $w = 23$ runs. Since $n_1$ and $n_0$ are each > 10, we use the normal approximation to the distribution of $W$ to conduct the test. Note that $n_1 = n_0 = 20$ in this case, so that

$$\mathrm{E}(W) = \frac{2(20)(20)}{40} + 1 = 21$$

$$\mathrm{var}(W) = \frac{2(20)(20)(2(20)(20) - 20 - 20)}{(20 + 20)^2(20 + 20 - 1)} = 9.7436.$$

The size $\alpha$ runs test in terms of the asymptotic normal distribution for $W$ is

$$z = \frac{w - E(W)}{(\mathrm{var}(W))^{1/2}} \begin{bmatrix} \in \\ \notin \end{bmatrix} (-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty) \Rightarrow \begin{bmatrix} \text{reject } H_0 \\ \text{accept } H_0 \end{bmatrix}.$$

Since in the case at hand $z = (23 - 21)/(9.7436)^{1/2} = .6407$ and $z_{.025} = 1.96$, then $z \notin (-\infty, -1.96] \cup [1.96, \infty)$, and the iid hypothesis is *not* rejected. □

One might contemplate using the runs test in the GLM context to check the validity of the iid assumption for the disturbances. Unfortunately, even if the $\varepsilon$ vector of $\mathbf{Y} = \mathbf{x}\beta + \varepsilon$ consists of iid random variables, $\mathbf{e} = \mathbf{Y} - \mathbf{x}\hat{\beta}$ does not, since $\mathbf{Cov}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')$. However, as $n$ increases, $\mathbf{e} = (\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\varepsilon \approx \varepsilon$ if $(\mathbf{x}'\mathbf{x})^{-1} \rightarrow [\mathbf{0}]$, and so the runs test may serve as an approximate test for large samples. The properties of the runs test in this context are not well established.

## 10.9 Noncentral $\chi^2$- and $F$-Distributions

### Family Name: Noncentral $\chi^2$-Distribution

*Parameterization:* $v$ (degrees of freedom) and $\lambda$ (noncentrality parameter)

*Density Definition:* $f(x; v, \lambda) = \sum_{j=0}^{\infty}(\lambda^j/j!)e^{-\lambda}h(x; v + 2j)I_{(0,\infty)}(x)$, where $h(x; v + 2j)$ is the (central) $\chi^2$-density with $v + 2j$ degrees of freedom.

*Moments:* $\mu = v + 2\lambda$, $\sigma^2 = 2(v + 4\lambda)$, $\mu_3 = 8(v + 6\lambda)$

*MGF:* $M_X(t) = (1 - 2t)^{-v/2}\exp[\frac{2\lambda t}{1-2t}]$ for $t < 1/2$

The noncentral $\chi^2$-distribution is the distribution of $Y = \sum_{i=1}^{v} Z_i^2$, where the $Z_i$'s are independent normally distributed random variables with unit variances and means $\phi_i$, $i = 1, \ldots, v$, i.e., $Z_i \sim N(\phi_i, 1)$, $i = 1, \ldots, v$, with $Z_1, \ldots, Z_v$ being independent random variables. The density definition can be found via the change-of-variables approach of Section 6.6. The noncentrality parameter $\lambda$ is related to the means of the $Z_i$'s as $\lambda = \phi'\phi/2$.

Note that the density definition can be recognized as a Poisson-weighted sum of (central) $\chi^2_{v+2j}$ density functions, $j = 0, 1, 2, \ldots$, and in effect can be thought of as the expected value of a $\chi^2$-density with $v + 2J$ degrees of freedom, the "random variable" $J$ having a Poisson density. This type of density definition, where the parameters of one density function family are effectively being treated as random variables and an expectation is taken with respect to another family of density functions, produces what is known as a **mixture distribution**. That is, members of one family are effectively being "mixed" using weights, applied to parameters, that are provided by another density family.

When $\lambda = 0$, the *central* $\chi^2$-distribution is defined, which we discussed in Chapter 4. When $\lambda > 0$, the mean of the distribution moves to the right, and the variance and skewness increase so that the density has less height and has a fatter, more pronounced right tail compared to the central $\chi^2$-density. In particular, for any $c < \infty$,

$$\lim_{\lambda \to \infty} P(x \geq c; \lambda) = \lim_{\lambda \to \infty} \int_c^{\infty} f(x; v, \lambda)dx = 1,$$

and $P(x \geq c; \lambda)$ is monotonically increasing as a function of $\lambda$.

Our principal use of the noncentral $\chi^2$-distribution is in analyzing the power function of statistical tests based on $\chi^2$-statistics whose distribution under $H_a$

is noncentral $\chi^2$. Power in these cases is generally represented as a function of the noncentrality parameter as

$$\pi_{C_r}(\lambda) = \int_{x \in C_r} f(x; v, \lambda) dx.$$

Examples are given by the GLR, LM, and Wald statistics, all of which have asymptotically valid noncentral $\chi^2$-distributions under local alternative hypotheses in $H_a$.

Integration of the noncentral $\chi^2$-distribution is straightforwardly accomplished on personal computers using software such as GAUSS, SAS, or Mathematica. The reader must be warned that the parameterization of noncentrality is *not* standard in the literature. Other parameterizations include noncentrality equal to $2\lambda$ or $\sqrt{2\lambda}$. The parameterization we use is more prevalent in the econometrics literature.

For further information, see N. L. Johnson and S. Kotz (1970), *Continuous Univariate Distributions*, II. New York: Wiley, pp. 130–148.

### Family Name: Noncentral $F$-Distribution

*Parameterization:* $v_1$ (numerator degrees of freedom), $v_2$ (denominator degrees of freedom), $\lambda$ (noncentrality parameter)

*Density Definition:*

$$f(x; v_1, v_2, \lambda) = \sum_{j=0}^{\infty} \left(\frac{\lambda^j}{j!}\right) e^{-\lambda} \left[\frac{(v_1/v_2)(v_1 x/v_2)^{(v_1/2)+j-1}}{B((v_1/2)+j, v_2/2)(1+v_1 x/v_2)^{.5(v_1+v_2)+j}}\right] I_{(0,\infty)}(x),$$

where $B(a,b)$ is the Beta function (recall the Beta distribution discussion in Chapter 4).

*Moments:*

$$\mu = \frac{v_2(v_1+2\lambda)}{v_1(v_2-2)}, \quad \text{for } v_2 > 2,$$

$$\sigma^2 = 2\left(\frac{v_2}{v_1}\right)^2 \frac{(v_1+2\lambda)^2 + (v_1+4\lambda)(v_2-2)}{(v_2-2)^2(v_2-4)}, \quad \text{for } v_2 > 4$$

$\mu_3 = g(v_1, v_2, \lambda)$ increasing in $\lambda$ and $v_1$, decreasing in $v_2$, for $v_2 > 6$.

*MGF:* Does not exist.

The noncentral $F$-distribution is the distribution of $Y = (Z_{v_1}/v_1)/(Z_{v_2}/v_2)$, where $Z_{v_1} \sim \chi^2_{v_1}(\lambda)$ and $Z_{v_2} \sim \chi^2_{v_2}$ are independent noncentral and central $\chi^2$ random variables with $v_1$ and $v_2$ degrees of freedom, respectively. The density of $Y$ can be established via transformation from the joint density of $(Z_{v_1}, Z_{v_2})$, which is the product of a noncentral and a central $\chi^2$-density.

When $\lambda = 0$, $Y$ is the ratio of two independent central $\chi^2$-random variables, each divided by their degrees of freedom, so that $Y$ has the (central) $F$-distribution as we derived in Section 6.7. When $\lambda > 0$, the mean of the distribution moves to the right, and the variance and skewness also increase so that the density has less height and has a fatter, more pronounced right tail in comparison to the central $F$-distribution. In particular, for $c < \infty$, $\lim_{\lambda \to \infty} P(x \geq c; \lambda) = \lim_{\lambda \to \infty} \int_c^\infty f(x; v_1, v_2, \lambda) dx = 1$, and $P(x \geq c; \lambda)$ is monotonically increasing as a function of $\lambda$.

As in the case of the noncentral $\chi^2$-distribution, our principal use of the noncentral $F$-distribution is in analyzing the power function of statistical tests based on test statistics that have a noncentral $F$-distribution when $H_a$ is true. Power in these cases is generally represented as a function of the noncentrality parameter as

$$\pi_{C_r}(\lambda) = \int_{x \in C_r} f(x; v_1, v_2, \lambda) dx.$$

An example of the use of the noncentral $F$-distribution was given in the discussion of testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ in the GLM context under normality, in which case the power function was representable in terms of a noncentral $F$-distribution.

Integration of the noncentral $F$-distribution is straightforwardly accomplished on personal computers using software such as GAUSS, SAS, and Mathematica. The warning regarding the parameterization of noncentrality given in our discussion of the noncentral $\chi^2$-distribution applies equally well here. Further information can be found in N. L. Johnson and S. Kotz, *Continuous Distributions*, pp. 189–200.

## Key Words, Phrases, and Symbols

heuristic principle of test construction
generalized likelihood ratio (GLR) test
asymptotically unbiased tests
restricted likelihood function
local alternatives
Pittman drift
asymptotic power
asymptotically equivalent tests
Lagrange multiplier (LM) test
scoring test
Wald test
noncentral $\chi^2$-distribution
noncentral $F$-distribution
noncentral $t$-distribution
equal-tails test
confidence interval, confidence region
confidence coefficient
confidence level
duality between confidence and critical regions
upper and lower confidence intervals
asymptotic confidence intervals and regions
unbiased confidence region
uniformly most accurate (UMA)
uniformly most accurate unbiased (UMAU)
consistent confidence region
pivotal quantity
pivotal quantity method
asymptotic pivotal quantity
$\chi^2$ goodness-of-fit test
Kolmogorov–Smirnov test statistic
Lilliefors test
Shapiro–Wilks test
Wald–Wolfowitz (WW) runs test
run
mixture distribution

## Problems

**1.** A manufacturer of breakfast cereals has been accused of systematically underfilling its cereal packages. The manufacturer claims that the complaint must stem from settling of the cereal that occurs in shipment and that the air space at the top of the package is normal settling of the contents. It claims that the filling process is normally distributed and that the system is *under control* at a mean fill rate of $\mu = EZ = 16.03$ ounces and a standard deviation of $\sigma = .01$, so that it is highly improbable that a package is filled below its stated contents of 16 ounces. It wants you to test the hypothesis that the filling system is under control at a level of significance equal to .10. It provides you with the following random-sample outcome of 40 observations on fill weights:

| | | | |
|---|---|---|---|
| 15.97 | 15.75 | 15.90 | 15.87 |
| 15.96 | 15.90 | 16.05 | 16.04 |
| 16.13 | 15.92 | 15.70 | 15.89 |
| 15.79 | 15.74 | 15.88 | 15.86 |
| 16.01 | 15.89 | 15.83 | 15.97 |
| 15.92 | 15.88 | 15.84 | 15.95 |
| 15.82 | 16.07 | 16.01 | 16.04 |
| 15.92 | 15.81 | 15.71 | 15.95 |
| 15.88 | 15.81 | 15.85 | 15.84 |
| 15.79 | 16.03 | 15.80 | 15.80 |

a. Define a size .10 GLR test of the hypothesis $H_0: \mu = \mu_0, \sigma = \sigma_0$, versus $H_a$: not $H_0$. Test the hypothesis that the filling process is under control. You may use the asymptotic distribution of the GLR if you wish.

b. Define a size .10 LM test of the same hypothesis as in (a), and test the hypothesis at significance level .10. Are the two tests different? Are the two test decisions in agreement?

c. Test the two hypotheses $H_0: \mu = 16.03$ and $H_0: \sigma = .01$ *individually*, using size .05 tests. Use whatever test procedures you feel are appropriate. Interpret the outcomes of the tests individually. Interpret the tests jointly using a Bonferroni approach.

*d. Can you define a Wald test for the hypothesis in (a)? If so, perform a Wald test of the joint null hypothesis at significance level .10.

**2. Testing for Differences in Two Populations: Variances**

After the firm in Problem 10.1 was inspected by the Dept. of Weights and Measures, a new random sample of 40 observations on fill weights was taken the next day, resulting in the following summary statistics:

$m'_1 = 16.02753$ and $s^2 = .9968 \times 10^{-2}$.

You may continue to assume normality for the population distribution.

a. Define a GLR size .05 test of $H_0: \sigma_1^2 \leq \sigma_2^2$ versus $H_a$: not $H_0$, where $\sigma_1^2$ and $\sigma_2^2$ refer to the variances of the populations from which the first and second samples were taken. Test the hypothesis.

b. Repeat (a) for the hypothesis $H_0: \sigma_1^2 \geq \sigma_2^2$ versus $H_a$: not $H_0$.

c. Define a GLR size .05 test of $H_0: \sigma_1^2 = \sigma_2^2$ versus $H_a$: not $H_0$. Either go to the computer and perform this test, or else define an approximation to this test based on "equal tails" and perform the test.

*d. Define the power function for each of the hypothesis-testing procedures you defined above. With the aid of a computer, plot the power functions and interpret their meanings.

**3. Testing for Differences in Two Populations: Means (Equal Variances)**

Referring to the data obtained in both preceding problems, consider testing whether the two population means are the same. *Assume that the population variances are identical in the two cases—did you find any evidence to contradict this in your answer to Problem 10.2 above?*

a. Define a GLR size .05 test of $H_0: \mu_1 \leq \mu_2$ versus $H_a$: not $H_0$, where $\mu_1$ and $\mu_2$ refer to the means of the populations from which the first and second samples were taken. Test the hypothesis.

b. Repeat (a) for the hypothesis $H_0: \mu_1 \geq \mu_2$ versus $H_a$: not $H_0$.

c. Define a GLR size .05 test of $H_0: \mu_1 = \mu_2$ versus $H_a$: not $H_0$. Test the hypothesis at the .10 level of significance.

*d. Define the power function for each of the hypothesis-testing procedures you defined above. With the aid of a computer, plot the power functions and interpret their meanings.

***4. Testing for Differences in Two Populations: Means (Unequal Variances—The Behrens–Fisher Problem)**

Consider the hypotheses in Problem 10.3 in the case where it is *not* assumed that the variances of the two populations are equal.

a. Attempt to find a GLR test of the hypothesis $H_0 : \mu_1 = \mu_2$ versus $H_a$: not $H_0$ based on a $t$-statistic similar to what you found in Problem 10.3. Is this approach valid? The problem of testing hypotheses concerning the means of two normal population distributions when it is not assumed that the variances are equal is known as the Behrens–Fisher problem, which as of yet has no universally accepted solution.

One *reasonable* solution to this problem is to use the statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{(s_1^2/(n_1 - 1) + s_2^2/(n_2 - 1))^{1/2}}$$

where $n_1$ and $n_2$ refer to the respective sample sizes of samples from the two populations. It can be shown that a test of size $\leq \alpha$ of any of the hypotheses concerning the means of the two population distributions given in Problem 10.3(a)–(c) can be performed by defining critical regions as follows:

| $H_0$ | $H_a$ | $C_r$ |
|---|---|---|
| $\mu_1 = \mu_2$ | $\mu_1 \neq \mu_2$ | $(-\infty, -t_{\alpha/2}(m)] \cup [t_{\alpha/2}(m), \infty)$ |
| $\mu_1 \geq \mu_2$ | $\mu_1 < \mu_2$ | $(-\infty, -t_\alpha(m)]$ |
| $\mu_1 \leq \mu_2$ | $\mu_1 > \mu_2$ | $[t_\alpha(m), \infty)$ |

The value of $t_\alpha(m)$ is a typical critical value of the student $t$-distribution found in $t$-tables, except the degrees-of-freedom parameter $m = \min(n_1, n_2) - 1$. The procedure is sometimes referred to as the *Hsu procedure.* Further discussion of the Behrens–Fisher problem can be found in H. Scheffé (1970), "Practical solutions of the Behrens–Fisher problems." *JASA*, (65), pp. 1501–1508.

b. Using the Hsu procedure discussed above, test the hypothesis of thc equality of the means in the population distributions referred to in Problems 10.1 and 10.2.

**5.** Regarding the asymptotically valid Wald and $Z$-tests for the value of $\sigma^2$ in the case of nonnormally distributed populations and the GLM as discussed at the end of Section 10.6, justify their asymptotic size and consistency.

**6.** Recall the analysis of the demand for textiles in the Netherlands performed by renown econometrician Henri Theil, as discussed in Problem 8.9. With reference to the least-squares–based estimate of the constant elasticity demand function generated from the data in the problem, respond to the following, using size .10 tests when testing is needed.

a. Test the hypothesis that the matrix of explanatory variables is significant in explaining changes in the logarithm of textile demand.

b. Test whether income is significant in explaining textile consumption. *Plot the power function of the test, and interpret its meaning.

c. Test whether price is significant in explaining textile consumption.

d. Test whether the price elasticity is inelastic. Test whether the income elasticity is inelastic.

e. Is the demand equation homogeneous degree zero in price and income? *Plot the power function of the test, and interpret its meaning.

f. Calculate a confidence interval that has a .95 confidence coefficient for the income elasticity. Interpret the meaning of this confidence interval.

g. Calculate a confidence interval that has a .95 confidence coefficient for the price elasticity. Interpret the meaning of this confidence interval.

*h. Calculate a confidence region having confidence coefficient .90 for the income and price elasticities. With the aid of a computer, graph the confidence region. Superimpose the intervals you calculated in (f) and (g) on this graph. The regions are different—interpret the difference.

*i. Test the least-squares residuals for normality, preferably using the Shapiro–Wilks test (you'll need tables for this), or else use the $\chi^2$ square goodness-of-fit test.

**7. Testing the Equality of Two Exponential Population Distributions**

The Reliable Computer Co. is considering purchasing CPU chips from one of two different suppliers to use in the production of personal computers. It has two bids from the suppliers, with supplier number one offering the lower bid. Before making a purchase decision, you want to test the durability of the chips, and you obtain a random sample of 50 chips from each supplier. It is known that both CPUs have operating lives that are exponentially distributed, and you want to test the equality of the expected operating lives of the chips.

a. Define a size .10 GLR test of the equality of the means of the two population distributions, i.e., a test of $H_0: \theta_1 = \theta_2$ versus $H_a$: not $H_0$.

b. The respective sample means of operating lives for the two samples were $\bar{x}_1 = 24.23$ and $\bar{x}_2 = 18.23$. Conduct the test of the null hypothesis. Does the test outcome help you decide from which supplier to purchase the chips?

c. How would your test rule change if you wanted to test a one-sided hypothesis concerning the means of the population distributions?

d. Consider using the LM test procedure for this problem. What is the test rule? Can you perform the test with the information provided?

e. Consider using the WALD test procedure for this problem. What is the test rule? Can you perform the test with the information provided?

*__8.__ The Huntington Chemical Co. has been accused of dumping improperly treated waste in a local trout lake, and litigation may be imminent. The central allegation is that the trout population has been severely reduced by the company's dumping practices. The company counters with the claim that at least 100,000 trout still remain in the lake, and since this is a substantial number for a lake of its size, the firm is innocent of all charges. In order to investigate the size of the trout population, 500 trout are captured, tagged, and then returned to the lake. After the fish have redistributed themselves in the lake, 500 fish are captured, and the number of tagged fish are observed. Letting $\bar{x}$ represent the proportion of tagged fish, can a size $\leq .10$ test of the company's claim be defined? Would an outcome of $\bar{x} = .14$ support or contradict the company's claim?

**9.** Use a Wald–Wolfowitz runs test to assess whether the data provided by the cereal manufacturer in Problem 10.1 might be the outcome of a random sample from some population distribution. The observations occurred sequentially *rowwise*. Use a size .05 test.

**10.** Test whether the data of the cereal manufacturer provided in Problem 10.1 can be interpreted as an outcome of a random sample from the specific normal population distribution claimed, i.e., test $H_0: Z \sim N(16.03, .0001)$ versus $H_a$: not $H_0$. Perform another test that assesses the more general hypothesis that the data are the outcome of a random sample from *some* normal population distribution, i.e., test $H_0: Z \sim N(\mu, \sigma^2)$ versus $H_a$: not $H_0$. Use size .05 tests.

**11. Paired Comparisons of Means**
A family counseling service offers a one-day program of training in meal planning, which they claim is effective in reducing cholesterol levels of individuals completing the course. In order to assess the effectiveness of the program, a consumer advocacy group randomly samples the cholesterol levels of 50 program participants both one day before and one month after the course is taken. It then summarizes the pairs of observations on the individuals by reporting the sample mean and sample standard deviation of the *differences* between the before and after observations of the 50 program participants. Their finding were $\bar{d} = -11.73$ and $s = 3.89$.

a. Assuming that the pairs of observations are iid outcomes from some bivariate normal population distribution with *before* and *after* means $\mu_b$ and $\mu_a$, define the appropriate likelihood function for the mean $\mu = \mu_a - \mu_b$ and variance $\sigma^2$ of the population distribution of *differences* in pairs of cholesterol measurements. Use this likelihood function to define a GLR size .05 test of the hypothesis that the meal planning program has no effect, i.e., a test for $H_0: \mu = 0$ versus $H_a$: not $H_0$. Test the hypothesis.

b. Can you define an LM test of the null hypothesis in (a)? Can you test the hypothesis with the information available? If so, perform the test—if not, what other information would you need?

c. Describe how you might test the hypothesis that the observations on paired differences are from a normal population distribution. What information would you need to test the hypothesis?

d. Suppose that normality of the observations is *not* assumed. Can you define another test of the effectiveness of the meal-planning program? If so, use the test to assess the effectiveness of the meal-planning program. Discuss any differences that are required in the interpretation of the outcome of this test compared to the test in (a) that you could perform if the observations were normally distributed.

**12.** The production function for a commodity can be approximated over a restricted range of relevant input levels by a linear relationship of the form $\mathbf{Y} = \mathbf{x}\beta + \varepsilon$. The columns of $\mathbf{x}$ include, in order, a column of 1's and three column vectors representing observations on labor, energy, and capital input levels. Thirty observations on weekly production levels yielded the following summary statistics:

$$\mathbf{b} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} = \begin{bmatrix} 13.792 \\ 3.005 \\ 1.327 \\ 6.385 \end{bmatrix}, \qquad \hat{s}^2 = 3.775,$$

$$(\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} 4.002 & -.136 & -.191 & -.452 \\ & .020 & .006 & -.007 \\ & & .041 & .001 \\ \text{(symmetric)} & & & .099 \end{bmatrix}.$$

In answering the following questions, be sure to clearly define any assumptions that you make.

a. Test the joint significance of the input variables for explaining the expected level of production using a size .05 test.

b. Test the significance of the input variables individually. Which input variables contribute significantly to the explanation of the expected level of output? Be sure to explain the basis for your conclusion.

c. Define confidence-interval outcomes for each of the marginal products of the inputs. Use .95 confidence coefficients. What do these confidence intervals mean?

d. Test the hypothesis that expected output $\leq 35$ when labor, energy, and capital are applied at levels 8, 4, and 3, respectively. Use a size .05 test.

e. Calculate a confidence-interval outcome for the expected level of production at the input levels indicated in (d). Use a .95 confidence coefficient. What is the meaning of this confidence interval?

g. Test the hypothesis that the variance of the production process $\geq 6$ at a .10 level of significance. Assume normality in conducting this test. What information would you need to conduct an asymptotically valid test of the hypothesis in the absence of normality?

**13.** The production department of a paper products manufacturer is analyzing the frequency of breakdowns in a certain type of envelope machine as it contemplates future machinery repair and replacement policy. It has been suggested that the number of breakdowns per day is a Poisson-distributed random variables, and the department intends to test this conjecture. Forty days' worth of observations on the number of daily machine breakdowns yielded the following observations, which occurred in sequence rowwise:

| | | | |
|---|---|---|---|
| 4 | 8 | 5 | 4 |
| 3 | 5 | 8 | 2 |
| 7 | 4 | 6 | 8 |
| 7 | 9 | 8 | 8 |
| 6 | 2 | 5 | 0 |
| 8 | 1 | 7 | 3 |
| 6 | 2 | 0 | 2 |
| 0 | 4 | 3 | 5 |
| 1 | 3 | 9 | 5 |
| 3 | 4 | 3 | 0 |

a. Test whether the observations might be a random-sample outcome from some population distribution. Use a .10-level runs test.

b. Test the hypothesis that the observations are from a Poisson population distribution. Use a size .10 test.

c. Test the hypothesis that the observations are from a uniform population distribution. Use a size .10 test.

d. Test the hypothesis that the expected number of daily breakdowns for this equipment $\geq 8$, using a significance level of .10. Use whatever test procedure you feel is appropriate. Based on this test procedure, calculate a .90-level confidence interval for the expected number of daily breakdowns.

**14.** The relationship between sales and level of advertising expenditure is hypothesized to be quadratic over the relevant range of expenditure levels being examined, i.e., $Y_t = \beta_1 + \beta_2 a_t + \beta_3 a_t^2 + \varepsilon_t$, where $y_t$ is the level of sales in period $t$, $a_t$ is the level of advertising expenditure in period $t$, and the $\varepsilon_t$'s are disturbance terms assumed to be iid normally distributed. The least-squares–based estimate of the relationship using 30 periods' worth of observations resulted in the following:

$$\mathbf{b} = (103.27 \quad 2.71 \quad -.13), \qquad \hat{s}^2 = 1.27,$$

$$(\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} 111.17 & 1.71 & .013 \\ & .64 & .0017 \\ \text{(symmetric)} & & .004 \end{bmatrix}.$$

a. Test the hypothesis that advertising expenditure has a significant impact on the expected sales level. Use any significance level you feel is appropriate.

b. Test whether the relationship between advertising expenditures and sales is actually a linear as opposed to a quadratic relationship.

c. Define a confidence interval with confidence coefficient .95 for the expected level of sales expressed as a function of advertising level $a_t$. Plot the confidence interval as a function of advertising expenditures (use expected sales on the vertical axis and advertising expenditures on the horizontal axis). Interpret the plot.

d. Test the hypothesis that the level of advertising expenditure that maximizes expected sales $\geq 15$.

*e. Calculate a confidence interval with confidence coefficient .95 for the level of advertising expenditure that maximizes expected sales.

**15.** An analyst is investigating the effect of certain policy events on common stock prices in a given industry. In an attempt to isolate abnormal from normal returns of firms in the industry, the following returns-generating equation was estimated via least squares: $R_t = \beta_1 + \beta_2 R_{mt} + \varepsilon_t$, where $R_t$ is the actual return by a firm on day $t$ and $R_{mt}$ is the return on a large portfolio of stocks designed to be representative of market return on day $t$. Abnormal returns are defined to be $\text{AR}_t = \varepsilon_t$ and are estimated from an estimate of the returns-generating equation as

$$\widehat{\text{AR}}_t = \text{R}_t - \hat{\text{R}}_t, \quad \text{where } \hat{\text{R}}_t = b_1 + b_2\text{R}_{mt}.$$

Thus, the estimated abnormal returns are effectively the estimated residuals from the least-squares estimate of the returns-generating equation. A summary of the estimation results from an analysis of 43 observations on a given firm is given below:

$$\mathbf{b} = [.03\ 1.07]', \quad (\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} .0001 & .2 \times 10^{-4} \\ .2 \times 10^{-4} & .81 \end{bmatrix},$$

$$\hat{s}^2 = 1.0752,$$

$$\hat{\mathbf{e}} = \begin{array}{cccc} -0.44 & 1.49 & -0.46 & -0.90 \\ 0.73 & -0.45 & -1.04 & -0.24 \\ -0.68 & -0.40 & -0.070 & -2.44 \\ -1.21 & 1.53 & 0.21 & -0.30 \\ -0.23 & -0.85 & 0.55 & -0.58 \\ 0.030 & 0.16 & -0.74 & -0.30 \\ 1.52 & 2.39 & 0.28 & 0.66 \\ -0.10 & -0.75 & 1.59 & 0.97 \\ -2.5 & -0.62 & -1.24 & -0.24 \\ 0.14 & 0.76 & 0.84 & 0.16 \\ 0.18 & -0.99 & -1.74 & \end{array}$$

a. Use a runs test to provide an indication of whether the residuals of the returns-generating equation, and hence the abnormal returns, can be viewed as iid from some population distribution. Use a size .05 test. Comment on the appropriateness of the runs test in this application.

b. Test to see whether the abnormal returns can be interpreted as a random sample from some normal population distribution of returns. Use a size .05 level of test.

c. Assuming normality, test whether the expected daily return of the firm is proportional to market return. Use a size .05 test.

d. Define and test an hypothesis that will assess whether the expected return of the firm is greater than market return for any $R_{mt} \geq 0$. Use a size .05 test.

**16. Testing for Independence in a Bivariate Normal Population Distribution**

Let $(X_i, Y_i)$, $i = 1, \ldots, n$, be a random sample from some bivariate normal population distribution $N(\mu, \Sigma)$ where $\sigma_1^2 = \sigma_2^2$. In this case, we know that the bivariate random variables are independent iff the correlation between them is zero. Consider testing the hypothesis of independence $H_0: \rho = 0$ versus the alternative hypothesis of dependence $H_a: \rho \neq 0$.

a. Define a size .05 GLR test of the independence of the two random variables. You may use the limiting distribution of the GLR statistic to define the test.

b. In a sample of 50 observations, it was found that $s_x^2 = 5.37$, $s_y^2 = 3.62$, and $s_{xy} = .98$. Is this sufficient to reject the hypothesis of independence based on the asymptotically valid test above?

c. Show that you can transform the GLR test into a test involving a critical region for the test statistic $w = s_{xy}/[(s_x^2 + s_y^2)/2]$.

*d. Derive the sampling distribution of the test statistic $W$ defined in (c) under $H_0$. Can you define a size .05 critical region for the test statistic? If so, test the hypothesis using the exact (as opposed to asymptotic) size .05 GLR test.

**17. Test of the Equality of Proportions**

Let $X_1$ and $X_2$ be independent random variables with binomial distributions Binomial$(n_i, p_i)$, $i = 1, 2$, where the $n_i$'s are assumed known. Consider testing the null hypothesis that the proportions (or probabilities) $p_1$ and $p_2$ are equal, i.e., the hypothesis is $H_0: p_1 = p_2$ versus $H_a: p_1 \neq p_2$.

a. Define an asymptotically valid size $\alpha$ Wald-type test of $H_0$ versus $H_a$.

b. If $n_1 = 45$, $n_2 = 34$, $x_1 = 19$, and $x_2 = 24$, is the hypothesis of equality of proportions rejected if $\alpha = .10$?

c. Define an asymptotically valid, size GLR-type test of $H_0$ versus $H_a$. Repeat part (b).

**18. Testing Whether Random Sample Is from a Discrete Uniform Population Distribution**
Let $X_i$, $i = 1, \ldots, n$ be a random sample from some discrete integer-valued population distribution. Let the range of the random variables be $1, 2, \ldots, m$.

a. Describe how you would use the $\chi^2$ goodness-of-fit size $\alpha$ test to assess the hypothesis that the population distribution was a discrete uniform distribution with support $1, 2, \ldots, m$.

b. The Nevada Gaming Commission has been called in to investigate whether the infamous Dewey, Cheatum & Howe Casino is using fair dice in its crap games. One of the alleged crooked dice is tossed 240 times, and the following outcomes were recorded:

| $X$ | Frequency |
|---|---|
| 1 | 33 |
| 2 | 44 |
| 3 | 47 |
| 4 | 39 |
| 5 | 31 |
| 6 | 46 |

Does this die appear to be fair? What size test will you use? Why?

**19.** Describe how you would test one-sided and two-sided hypotheses, using both finite sample and asymptotically valid testing procedures, in each of the following cases. Use whatever procedures you feel are appropriate.

a. Testing hypotheses about the mean of a binomial distribution based on a random sample of size 1 from the binomial distribution.

b. Testing hypotheses about the mean of a Poisson distribution based on a random sample of size $n$ from a Poisson population distribution.

**20.** The weekly proportion of storage tank capacity that is utilized at a regional hazardous liquid waste receiving site is an outcome of a Beta$(\Theta, 1)$ random variable. At the end of each week the storage tank is emptied. A random sample of a year's worth of observations of capacity utilization at the site produced the following in sequence rowwise:

| | | | |
|---|---|---|---|
| 0.148 | 0.501 | 0.394 | 0.257 |
| 0.759 | 0.763 | 0.092 | 0.155 |
| 0.409 | 0.257 | 0.586 | 0.919 |
| 0.278 | 0.076 | 0.019 | 0.513 |
| 0.123 | 0.390 | 0.030 | 0.082 |
| 0.935 | 0.607 | 0.075 | 0.729 |
| 0.664 | 0.802 | 0.338 | 0.539 |
| 0.055 | 0.984 | 0.269 | 0.069 |
| 0.679 | 0.382 | 0.549 | 0.028 |
| 0.770 | 0.296 | 0.105 | 0.465 |
| 0.194 | 0.675 | 0.696 | 0.068 |
| 0.091 | 0.132 | 0.156 | 0.050 |
| 0.477 | 0.754 | 0.164 | 0.527 |

a. Define a GLR size $\alpha$ test of $H_0: \Theta = \Theta_0$ versus $H_0: \Theta \neq \Theta_0$.

b. Test the hypothesis that the expected weekly storage tank capacity utilized is equal to .5 at a significance level of .05. If the exact distribution of the GLR test appears to be intractable, then use an asymptotic test.

c. Test that the observations might be interpreted as a random sample from some population distribution. Use a size .05 runs test.

d. Test that the observations can be interpreted as a random sample from a Beta$(\Theta, 1)$ population distribution. Use a size .05 test.

e. Would the use of an LM test be tractable here? If so, perform an LM test of the hypothesis.

**21.** Follow the proof of Theorem 10.14 to demonstrate that when random sampling from the continuous PDF $f(z; \Theta)$ with scalar parameter $\Theta$,

$$-2\sum_{i=1}^{n} \ln[1 - F(X_i; \Theta)] \sim \chi^2_{2n}$$

is a pivotal quantity for $\Theta$.

**22.** Regional daily demand for gasoline in the summer driving months is assumed to be the outcome of an $N(\mu, \sigma^2)$ random variable. Assume you have 40 iid daily observations on daily demand for gasoline with quantity demanded measured in millions of gallons, and $\bar{x} = 43$ and $s^2 = 2$.

a. Show that $N(\mu, \sigma^2)$ is a location–scale parameter family of PDFs (recall Theorem 10.13).

b. Define a pivotal quantity for $\mu$, and use it to define a .95-level confidence interval for $\mu$. Also, define a .95 lower confidence bound for $\mu$.

c. Define a pivotal quantity for $\sigma^2$, and use it to define a .95-level confidence interval for $\sigma^2$.

*d. A colleague claims that mean daily gasoline demand is only 37 million gallons. Is your answer to (b) consistent with this claim? Explain.

**23.** The population distribution of income in a populous developing country is assumed to be given (approximately) by the continuous PDF $f(x;\Theta) = \Theta(1 + x)^{-(\Theta+1)} I_{(0,\infty)}(x)$, where $\Theta > 0$. A summary measure of a random-sample outcome of size 100 from the population distribution is given by $\sum_{i=1}^{100} \ln(1 + x_i) = 40.54$, where $x_i$ is measured in thousands of units of the developing country's currency.

a. Define a pivotal quantity for the parameter $\Theta$ (Hint: Theorem 10.13 and Problem 10.21 might be useful here.)

b. Define a general expression for a level-$\gamma$ confidence interval for $\Theta$ based on $n$ iid observations from $f(x;\Theta)$.

c. Calculate an outcome of a .90-level confidence interval for $\Theta$ based on the expression derived in (b). Note: For large degrees of freedom > 30,

$$\chi^2_{v;\alpha} \approx v\left[1 - \frac{2}{9v} + z_\alpha\left(\frac{2}{9v}\right)^{1/2}\right]^3,$$

where $\int_{z_\alpha}^{\infty} N(x; 0, 1)dx = \alpha$, is a very good approximation to critical values of the $\chi^2$-distribution (M. Abramowitz and I. A. Stegun (1972), *Handbook of Mathematical Functions*. New York: Dover, p. 941).

# A Math Review: Sets, Functions, Permutations, Combinations, and Notation

## A.1 Introduction

In this appendix we review basic results concerning set theory, relations and functions, combinations and permutations, and summation and integration notation. We also review the meaning of the terms *definition, axiom, theorem, corollary, and lemma*, which are labels that are affixed to a myriad of statements and results that constitute the theory of probability and mathematical statistics. The topics reviewed in this appendix constitute basic foundational material on which our study of mathematical statistics will be based. Additional mathematical results, often of a more advanced nature, are introduced throughout the text as the need arises.

## A.2 Definitions, Axioms, Theorems, Corollaries, and Lemmas

The development of the theory of probability and mathematical statistics involves a considerable number of statements consisting of definitions, axioms, theorems, corollaries, and lemmas. These terms will be used for organizing the various statements and results we will examine into these categories:

1. descriptions of meaning;
2. statements that are acceptable as true without proof;

3. formulas or statements that require proof of validity;
4. formulas or statements whose validity follows immediately from other true formulas or statements; and
5. results, generally from other branches of mathematics, whose primary purpose is to facilitate the proof of validity of formulas or statements in mathematical statistics.

More formally, we present the following meaning of the terms:

**Definition** A statement of the meaning of a word, word group, sign, or symbol.

**Axiom (or postulate)** A statement that has found general acceptance, or is thought to be worthy thereof, on the basis of an appeal to intrinsic merit or self-evidence and thus requires no proof of validity.

**Theorem (or proposition)** A formula or statement that is deduced from other proved or accepted formulas or statements and whose validity is thereby proved.

**Corollary** A formula or statement that is immediately deducible from a proven theorem and that requires little or no additional proof of validity.

**Lemma** An auxiliary proposition that has been proved either by the user or elsewhere and that is stated for the expressed purpose of immediate use in the proof of another proposition.

Thus, in the development of the theory of probability and mathematical statistics, axioms are the fundamental truths that are to be accepted at face value and not proven. Theorems and their corollaries are statements deducible from the fundamental truths and other proven statements and thus are *derived truths*. Lemmas represent results that facilitate proofs of other statements that are of more fundamental interest.

We elaborate on the concept of a lemma, since our discussions will implicitly rely on lemmas more than any other type of statement, but we will generally choose not to exhaustively catalogue lemmas in the discussions. What constitutes a lemma and what does not depends on the problem context or on one's point of view. A fundamental integration result from calculus could technically be referred to as a lemma when used in a proof of a statement in mathematical statistics, while in the study of calculus, it might be referred to as a theorem to be proved in and of itself. Since our study will require numerous auxiliary results from algebra, calculus, and matrix theory, exhaustively cataloguing these results as lemmas would be cumbersome and, more importantly, not necessary given the prerequisites assumed for this course of study, namely, a familiarity with the basic concepts of algebra, univariate and multivariate calculus, and an introduction to matrix theory. We will have occasion to state a number of lemmas, but we will generally reserve this label for more exotic mathematical results that fall outside the realm of mathematics encompassed by the prerequisites.

## A.3 Elements of Set Theory

In the study of probability and mathematical statistics, sets are the fundamental objects to which probability will be assigned, and it is important that the concept of a set, and operations on sets, be well understood. In this section we review some basic properties of and operations on sets. This begs the following question: What is meant by the term *set*? In modern axiomatic developments of set theory, the concept of a set is taken to be primitive and incapable of being defined in terms of more basic ideas. For our purposes, a more intuitive notion of a set will suffice, and we avoid the complexity of an axiomatic development of the theory (see Marsden,[1] Appendix A, for a brief introduction to the axiomatic development). We base our definition of a set on the intuitive definition originally proposed by the founder of set theory, Georg Cantor (1845–1918).[2]

**Definition A.1**
***Set***

A **set** is a collection of objects with the following characteristics:

1. All objects in the collection are *clearly defined*, so that it is evident which objects are members of the collection and which are not;
2. All objects are *distinguishable*, so that objects in the collection do not appear more than once;
3. *Order is irrelevant* regarding the listing of objects in the collection, so two collections that contain the same objects but are listed in different order are nonetheless the *same set*; and
4. *Objects in the collection can be sets themselves*, so that a *set of sets* can be defined.

The objects in the collection of objects comprising a set are its **elements**. The term **members** is also used to refer to the objects in the collection. In order to signify that an object belongs to a given set, the symbol $\in$, will be used in an expression such as $\mathbf{x} \in \mathbf{A}$, which is to be read "$x$ is an element (or member) of the set $A$." If an object is *not* a member of a given set, then a slash will be used as $x \notin A$ to denote that "$x$ is not an element (or member) of the set A." Note that the slash symbol, /, is used to indicate negation of a relationship. The characteristics of sets presented in Def. A.1 will be clarified and elaborated upon in examples and discussions provided in subsequent subsections.

[1] J.E. Marsden, (1974), *Elementary Classical Analysis*, San Francisco: Freeman and Co.

[2] Cantor's original text reads: "Unter einer 'Menge' verstehen wir jede Zusammenfassung $M$ von bestimmten wohlunterschiedenen Objekten $m$ unserer Anschauung oder unseres Denkens (welche die 'Elemente' von $M$ genannt werden) zu einem ganzen" (*Collected Papers*, p. 282). Our translation of Cantor's definition is, "By set we mean any collection, $M$, of clearly defined, distinguishable objects, $m$ (which will be called elements of $M$), which from our perspective or through our reasoning we understand to be a whole."

### Set-Defining Methods

Three basic methods are used in defining the objects in the collection constituting a given set: (1) **exhaustive listing**; (2) **verbal rule**; and (3) **mathematical rule.** An *exhaustive listing* requires that each and every object in a collection be individually identified either numerically, if the set is a collection of numbers, or by an explicit verbal description, if the collection is not of numbers. The object descriptions are conventionally separated by commas, and the entire group of descriptions is enclosed in brackets. The following are examples of sets defined by an exhaustive listing of the objects that are elements of the set:

**Example A.1** $S_1 = \{\text{HEAD}, \text{TAIL}\}$

Here $S_1$ is the set of possible occurrences when tossing a coin into the air and observing its resting position. Note the set can be equivalently represented as $S_1 = \{\text{TAIL}, \text{HEAD}\}$. □

**Example A.2** $S_2 = \{1, 2, 3, 4, 5, 6\}$

Here $S_2$ is the set of positive integers from 1 to 6. Note that the set $S_2$ can be equivalently represented by listing the positive integers 1 to 6 in any order. □

A *verbal rule* is a verbal statement of characteristics that only the objects that are elements of a given set possess and that can be used as a test to determine set membership. The general form of the verbal rule is {$x$: *verbal statement*}, which is to be read "the collection of all $x$ for which *verbal statement* is true." The following are examples of sets described by verbal rules:

**Example A.3** $S_3 = \{x : x \text{ is a college student}\}$

Here $S_3$ is the set of college students. An individual is an element of the set $S_3$ **iff** (if and only if) he or she is a college student. □

**Example A.4** $S_4 = \{x : x \text{ is a positive integer}\}$

Here $S_4$ is the set of positive integers 1,2,3, ... A number is an element of the set $S_4$ iff it is a positive integer. □

A *mathematical rule* is of the same general form as a verbal rule, except the verbal statement is replaced by a mathematical expression of some type. The general form of the mathematical rule is {$x$: *mathematical expression*}, which is to be read "the collection of all $x$ for which *mathematical expression* is true." The following are examples of sets described by mathematical rules:

**Example A.5** $S_5 = \{x : x = 2k + 1, k = 0, 1, 2, 3, \ldots\}$

Here $S_5$ is the set of odd positive integers. A number is an element of the set $S_5$ iff the number is equal to $2k + 1$ for some choice of $k = 0, 1, 2, 3, \ldots$. □

**Example A.6** $S_6 = \{x : 0 \leq x \leq 1\}$

Here $S_6$ is the set of numbers greater than or equal to 0 but less than or equal to 1. A number is an element of the set $S_6$ iff it is neither less than 0 nor greater than 1. □

The choice of method for describing the objects that constitute elements of a set depends on what is convenient and/or feasible for the case at hand. For example, exhaustive listing of the elements in set $S_6$ is impossible. On the other hand, there is some discretion that can be exercised, since, for example, a verbal rule could have adequately described the set $S_5$, say as $S_5 = \{x :$ $x$ is an odd positive integer$\}$. A mixing of the basic methods might also be used, such as $S_5 = \{x : x = 2k + 1,$ $k$ is zero or a positive integer$\}$. One can choose whatever method appears most useful in a given problem context.

Note that although our preceding examples of verbal and mathematical rules treat $x$ as inherently one-dimensional, a vector interpretation of $x$ is clearly permissible. For example, we can represent the set of points on the boundary or interior of a circle centered at (0,0) and having radius 1 as

$$S_7 = \left\{(x_1, x_2) : x_1^2 + x_2^2 \leq 1\right\},$$

or we can represent the set of input–output combinations associated with a two-input Cobb–Douglas production function as

$$S_8 = \left\{(y, x_1, x_2) : y = b_0 x_1^{b_1} x_2^{b_2}, x_1 \geq 0, x_2 \geq 0\right\}$$

for given numerical values of $b_0$, $b_1$, and $b_2$. Of course, the entries in the $x$-vector need not be numbers, as in the set

$$S_9 = \{(x_1, x_2) : x_1 \text{ is an economist}, x_2 \text{ is an accountant}\}.$$

### Set Classifications

Sets are classified according to the number of elements they contain and whether the elements are countable. We differentiate between sets that have a finite number of elements and sets whose elements are infinite in number, referring to a set of the former type as a **finite set** and a set of the latter type as an **infinite set**. In terms of countability, sets are classified as being either countable or uncountable. Note that when we count objects, we intuitively place the objects in a one-to-one correspondence with the positive integers, i.e., we identify objects one by one and count "1, 2, 3, 4, . . . ." Thus, a **countable set** is one whose elements can be placed in a one-to-one correspondence with some or all of the positive integers–any other set is referred to as an **uncountable set**.

A finite set is, of course, always countable, and thus it would be redundant to use the term "countable finite set." Sets are thus either **finite, countably infinite**, or **uncountably infinite**. Of the sets $S_1$ through $S_9$ described earlier, $S_1$, $S_2$, $S_3$, and $S_9$ are finite, $S_4$ and $S_5$ are countably infinite, and $S_6$, $S_7$, and $S_8$ are uncountably infinite (why?).

### Special Sets, Set Operations, and Set Relationships

We now proceed to a number of definitions and illustrations establishing relationships between sets, mathematical operations on sets, and the notions of the universal and empty sets.

**Definition A.2**
***Subset***

$A$ is a **subset** of $B$, denoted as $A \subset B$ and read *$A$ is contained in $B$*, iff every element of $A$ is also an element of $B$.

**Definition A.3**
***Equality of sets***

Set $A$ is equal to set $B$, denoted as $A = B$, iff every element of $A$ is also an element of $B$, and every element of $B$ is also an element of $A$, i.e., iff $A \subset B$ and $B \subset A$.

**Definition A.4**
***Universal set***

The set containing all objects under consideration in a given problem setting, and from which all subsets are extracted, is the **universal set.**

**Definition A.5**
***Empty or null set***

The set containing no elements, denoted by $\emptyset$, is called the **empty**, or **null set.**

**Definition A.6**
***Set difference***

Given any two subsets $A$ and $B$ of a universal set, the set of all elements in $A$ that are *not* in $B$ is called the **set difference** between $A$ and $B$ and is denoted by $A - B$. If $A \subset B$, then $A - B = \emptyset$

**Definition A.7**
***Complement***

Let $A$ be a subset of a universal set, $\Omega$. The **complement** of the set $A$ is the set of all elements in $\Omega$ that are not in $A$, and is denoted by $\bar{A}$. Equivalently, $\bar{A} = \Omega - A$.

**Definition A.8**
***Union***

Let $A$ and $B$ be any two subsets of a universal set, $\Omega$. Then the **union** of the sets $A$ and $B$ is the set of all elements in $\Omega$ that are in *at least one* of the sets $A$ or $B$; it is denoted by $A \cup B$.

**Definition A.9**
***Intersection***

Let $A$ and $B$ be any two subsets of a specified universal set, $\Omega$. Then the **intersection** of the sets $A$ and $B$ is the set of all elements in $\Omega$ that are *in both* sets $A$ and $B$, it is denoted by $A \cap B$

**Definition A.10**
***Mutually exclusive (or disjoint) sets***

Subsets $A$ and $B$ of a universal set, $\Omega$, are said to be **mutually exclusive** or **disjoint** sets iff they have no elements in common, i.e., iff $A \cap B = \emptyset$.

We continue to use the slash, /, to indicate negation of a relationship (recall that / was previously used to indicate the negation of $\in$). Thus, $A \not\subset B$ denotes that $A$ is not a subset of $B$, and $A \neq B$ denotes that $A$ is not equal to $B$. We note here (and we shall state later as a theorem) that it is a logical requirement that $\emptyset$ is a subset of any set $A$, since if $\emptyset$ does not contain any elements, it cannot be the case that $\emptyset \not\subset A$, since the negation of $\subset$ would require the existence of an element in $\emptyset$ that was not in $A$.

**Example A.7** Let the universal set be defined as $\Omega = \{x\colon\ 0 \le x \le 1\}$, and define three additional sets as

$$A = \{x:\ 0 \le x \le .5\}, \qquad B = \{x:\ .25 \le x \le .75\} \qquad \text{and,} \qquad C = \{x:\ .75 < x \le 1\}.$$

Then we can establish the following set relationships:

$$\begin{aligned}
\bar{B} &= \{x:\ 0 \le x < .25 \text{ or } .75 < x \le 1\},\\
A \cup C &= \{x:\ 0 \le x \le .5 \text{ or } .75 < x \le 1\},\\
\bar{B} &\subset A \cup C,\\
C \cap A &= C \cap B = \emptyset,\\
\bar{C} &= A \cup B = \{x:\ 0 \le x \le .75\},\\
A - B &= \{x:\ 0 \le x < .25\},\\
A \cap B &= \{x:\ .25 \le x \le .5\}.
\end{aligned}$$

□

Note that although our definitions of subset, equality of sets, set difference, complement, union, and intersection explicitly involve only two sets, $A$ and $B$, it is implicit that the concepts can be applied to more complicated expressions involving an arbitrary number of sets. For example, since $A \cap B$ is itself a set, we can form its intersection with a set $C$ as $(A \cap B) \cap C$, or form the set difference, $(A \cap B) - C$, or establish that $(A \cap B) =$ or $\neq C$, and so on. The point is that the concepts apply to sets, which themselves may have been constructed from other sets via various set operations.

**Example A.8** Let $\Omega$, $A$, $B$, and $C$ be defined as in Ex. A.7. Then the following set relationships can be established:

$$\begin{aligned}
A \cup B \cup C &= \Omega = \{x:\ 0 \le x \le 1\},\\
(A \cup C) \cap B &= \{x:\ .25 \le x \le .5\},\\
(A \cap B) \cap C &= \overline{(A \cup B)} \cap \bar{C} = \emptyset,\\
(B \cup C) - A &= \{x:\ .5 < x \le 1\},\\
((A \cup B) - (A \cap B)) &\subset \bar{C} \cap (\bar{A} \cup \bar{B}).
\end{aligned}$$

Can $\subset$ be replaced with $=$ in the last relationship? □

It is sometimes useful to conceptualize set relationships through illustrations called *Venn diagrams* (named after the nineteenth-century English logician, John Venn). In a Venn diagram, the universal set is generally denoted by a rectangle, with subsets of the universal set represented by various geometric shapes located within the bounds of the rectangle. Figure A.1 uses Venn diagrams to illustrate the set relationships defined previously.

### Rules Governing Set Operations

Operations on sets must satisfy a number of basic rules. We state these basic rules as theorems, although we will not take the time to prove them here. The reader may wish to verify the plausibility of some of the theorems through the use of Venn diagrams. One of DeMorgan's laws will be proved to illustrate the formal proof method for the interested reader.

**Theorem A.1** (Idempotency Laws)

$A \cup A = A \quad \textit{and} \quad A \cap A = A$

**Theorem A.2** (Commutative Laws)

$A \cup B = B \cup A \quad \textit{and} \quad A \cap B = B \cap A$

**Theorem A.3** (Associative Laws)

$(A \cup B) \cup C = A \cup (B \cup C) \quad \textit{and} \quad (A \cap B) \cap C = A \cap (B \cap C)$

**Theorem A.4** (Distributive Laws)

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad \textit{and} \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

**Theorem A.5** (Identity Elements $\cap$ and $\cup$)

$A \cap \Omega = A$ (*$\Omega$ is the identity element for* $\cap$)

$A \cup \emptyset = A$ (*$\emptyset$ is the identity element for* $\cup$)

**Theorem A.6** (Intersection and Union of Complements)

$A \cup \bar{A} = \Omega \quad \textit{and} \quad A \cap \bar{A} = \emptyset$

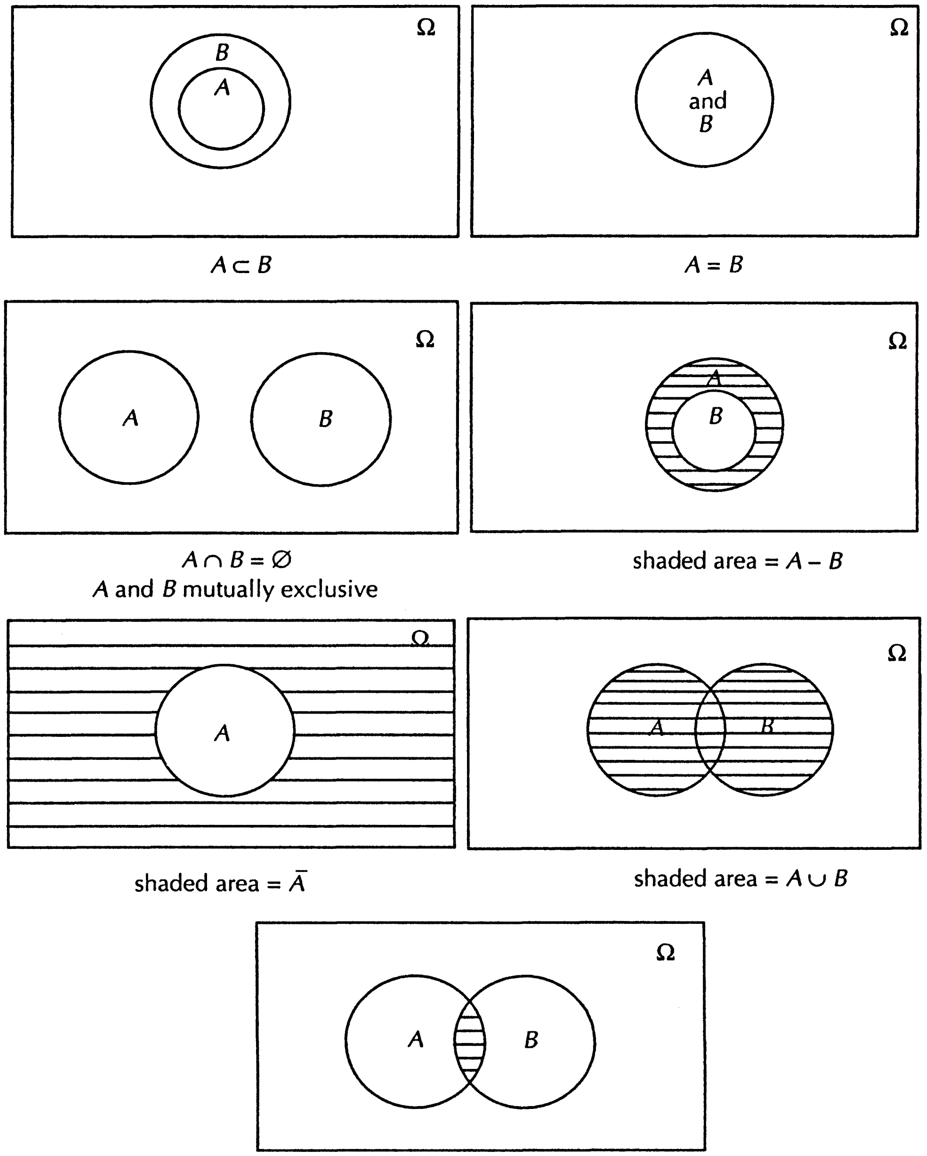
**Theorem A.7** (Complements of Complements)

$(\bar{\bar{A}}) = A$

**Theorem A.8** (Intersection with the Null Set)

$A \cap \emptyset = \emptyset$

**Theorem A.9** (Null Set as a Subset) *If $A$ is any set, then $\emptyset \subset A$*



$A \subset B$

$A = B$

$A \cap B = \varnothing$
$A$ and $B$ mutually exclusive

shaded area = $A - B$

shaded area = $\bar{A}$

shaded area = $A \cup B$

shaded area = $A \cap B$

**Figure A-1**
Venn diagrams illustrating set relationships.

**Theorem A.10** (DeMorgan's Laws)

$(\overline{A \cup B}) = \bar{A} \cap \bar{B}$ *and* $(\overline{A \cap B}) = \bar{A} \cup \bar{B}$

**Example A.9** Formal Proof of $(\overline{A \cap B}) = \bar{A} \cup \bar{B}$

By definition of the equality of sets, two sets are equal iff each is contained in the other. We first demonstrate that $\overline{A \cap B} \subset \bar{A} \cup \bar{B}$. By definition, $x \in \overline{A \cap B}$

implies that $x \notin A \cap B$. Suppose $x \notin \bar{A} \cup \bar{B}$. This implies $x \notin \bar{A}$ and $x \notin \bar{B}$, which implies $x \in A$ and $x \in B$, i.e., $x \in A \cap B$, a contradiction. Therefore, if $x \in \overline{A \cap B}$, then $x \in \bar{A} \cup \bar{B}$, which implies $\overline{A \cap B} \subset \bar{A} \cup \bar{B}$. We next demonstrate that $\bar{A} \cup \bar{B} \subset \overline{A \cap B}$. Let $x \in \bar{A} \cup \bar{B}$. Then $x \notin A \cap B$, for if it were, then $x \in A$ and $x \in B$, contradicting that $x$ belongs to at least one of $\bar{A}$ and $\bar{B}$. However, $x \notin A \cap B$ implies $x \in \overline{A \cap B}$, and thus $\bar{A} \cup \bar{B} \subset \overline{A \cap B}$. □

We remind the reader that since the sets used in Theorems A.1–A.10. could themselves be the result of set operations applied to other sets, the theorems are extendable in a myriad of ways to involve an arbitrary number of sets. For example, in the first of DeMorgan's laws listed in Theorem A.10, if $A = C \cup D$ and $B = E \cup F$, then by substitution,

$$(\overline{C \cup D \cup E \cup F}) = (\overline{C \cup D}) \cap (\overline{E \cup F}).$$

Then by applying Theorem A.10 to both $(C \cup D)$ and $(E \cup F)$, we obtain a generalization of DeMorgan's law as

$$(\overline{C \cup D \cup E \cup F}) = \bar{C} \cap \bar{D} \cap \bar{E} \cap \bar{F}.$$

Given the wide range of extensions that are possible, Theorems A.1–A.10 provide a surprisingly broad conceptual foundation for applying the rules governing set operations.

*Some Useful Set Notation* Situations sometimes arise in which one is required to denote the union or intersection of a large number of sets. A convenient notation that represents such unions or intersections quite efficiently is available. Two types of notations are generally used, and they are differentiated on the basis of whether the union or intersection is of sets identified by a natural sequence of integer subscripts or whether the sets are identified by subscripts, say $i$'s, that are elements of some set of subscripts, $I$, called an **index set**.

**Definition A.11**
***Multiple union notation***

**a.** $\cup_{i=1}^{n} A_i = A_1 \cup A_2 \cup A_3 \ \dots \ \cup A_n$.

**b.** $\cup_{i \in I} A_i =$ union of all sets $A_i$ for which $i \in I$.

**Definition A.12**
***Multiple intersection notation***

**a.** $\cap_{i=1}^{n} A_i = A_1 \cap A_2 \cap A_3 \ \dots \ \cap A_n$.

**b.** $\cap_{i \in I} A_i =$ intersection of all sets $A_i$ for which $i \in I$.

**Example A.10** Let the universal set be defined as $\Omega = \{x\colon\ 0 \leq x \leq 1\}$, and examine the following subsets of $\mathbf{\Omega}$:

$A_1 = \{x\colon\ 0 \leq x \leq .25\}$, $\quad A_2 = \{x\colon\ 0 \leq x \leq .5\}$,

$A_3 = \{x\colon\ 0 \leq x \leq .75\}$, $\quad A_4 = \{x\colon\ .75 \leq x \leq 1\}$.

Define the index sets $I_1$ and $I_2$ as

$I_1 = \{1,3\}$ and $I_2 = \{1,3,4\}$.

Then,

$$\cup_{i=1}^{4} A_i = \cup_{i=2}^{4} A_i = \cup_{i=3}^{4} A_i = \{x\colon 0 \leq x \leq 1\} = \Omega,$$

$$\cup_{i \in I_1} A_i = A_1 \cup A_3 = \{x\colon 0 \leq x \leq .75\},$$

$$\cup_{i \in I_2} A_i = A_1 \cup A_3 \cup A_4 = \{x\colon 0 \leq x \leq 1\} = \Omega,$$

$$\cap_{i=1}^{4} A_i = \cap_{i=2}^{4} A_i = \emptyset,$$

$$\cap_{i=3}^{4} A_i = \{.75\},$$

$$\cap_{i \in I_1} A_i = A_1 \cap A_3 = \{x\colon 0 \leq x \leq .25\},$$

$$\cap_{i \in I_2} A_i = A_1 \cap A_3 \cap A_4 = \emptyset. \qquad \square$$

Whenever a set $A$ is an interval subset of the *real line* (where the **real line** refers to all of the numbers between $-\infty$ and $\infty$), the set can be indicated in abbreviated form by the standard notation for intervals, stated in the following definition.

**Definition A.13**
***Interval set notation***

Let $a$ and $b$ be two numbers on the real line for which $a < b$. Then the following four sets, called intervals with endpoints $a$ and $b$, can be defined as

a. **Closed interval:**

$$[a, b] = \{x\colon a \leq x \leq b\},$$

b. **Half-open (or half-closed) intervals:**

$$(a, b] = \{x\colon a < x \leq b\}, \text{ and}$$

$$[a, b) = \{x\colon a \leq x < b\},$$

c. **Open interval:**

$$(a, b) = \{x\colon a < x < b\}.$$

Note that *weak inequalities*, $x \leq$ or $\leq x$, are signified by brackets ] or [, respectively. *Strong inequalities*, $x <$ *or* $< x$, are signified by parentheses,) or (, respectively. Note further that whether the interval set contains its endpoints determines whether the set is closed.

As we have already done, $(x, y)$ will also be used to denote coordinates in the two-dimensional plane. The context of the discussion will make clear whether we are referring to an open interval $(a, b)$ or a pair of coordinates $(x, y)$.

## A.4 Relations, Point Functions, and Set Functions

The concepts of point function and set function are central to our discussion of probability and statistics. We will see that probabilities can be represented by set functions and that in a large number of cases of practical interest, set functions can in turn be represented by a summation or integration operation applied to point functions. While readers may be somewhat familiar with point functions from introductory courses in algebra, the concept of a set function may not be familiar. We will review both function concepts within the broader context of the theory of relations. The relations context facilitates the presentation of a very general definition of "function" in which inputs into and outputs from the function may be objects of any kind, including, but not limited to, numbers. The relations context also facilitates a demonstration of the significant similarities between the concept of a set function and the more familiar point function concept.

### Cartesian Product

The concept of a relation can be made clear once we define what is meant by the Cartesian product of two sets $A$ and $B$, named after the French mathematician René Descartes (1596–1650).

**Definition A.14** ***Cartesian product of A and B***

Let $A$ and $B$ be two sets. Then the **Cartesian product** of $A$ and $B$, denoted as $A \times B$, is the set of ordered pairs

$A \times B = \{(x, y) \colon x \in A, y \in B\}$.

In words, $A \times B$ is the set of all possible pairs $(x, y)$ such that $x$ is an element of the set $A$ and $y$ is an element of the set $B$. Note carefully that the pairs are *ordered* in the sense that the first object in the pair must come from set $A$ and the second object from set $B$.

**Example A.11** Let $A = \{x \colon 1 \leq x \leq 2\}$ and $B = \{y \colon 2 \leq y \leq 4\}$. Then $A \times B = \{(x, y) \colon 1 \leq x \leq 2$ and $2 \leq y \leq 4\}$ (see Figure A.2). □

**Example A.12** Let $A = \{x \colon x$ is a man$\}$ and $B = \{y \colon y$ is a woman$\}$. Then $A \times B = \{(x, y) \colon x$ is a man and $y$ is a woman$\}$, which is the set of all possible man–woman pairings. □

We will have use for a more general notion of Cartesian product involving more than just two sets. The extension is given in the next definition.

**Definition A.15** ***Cartesian product (general)***

Let $A_1, \ldots, A_n$ be $n$ sets. Then the Cartesian product of $A_1, \ldots, A_n$ is the set of ordered $n$-tuples

$\times_{i=1}^{n} A_i = A_1 \times A_2 \times \ldots \times A_n = \{(x_1, \ldots, x_n) \colon x_i \in A_i, i = 1, \ldots, n\}$.

In words, $\times_{i=1}^{n} A_i$ is the set of all possible $n$-tuples $(x_1, \ldots, x_n)$ such that $x_1$ is an element of set $A_1$, $x_2$ is an element of set $A_2$, and so on. Note that should the

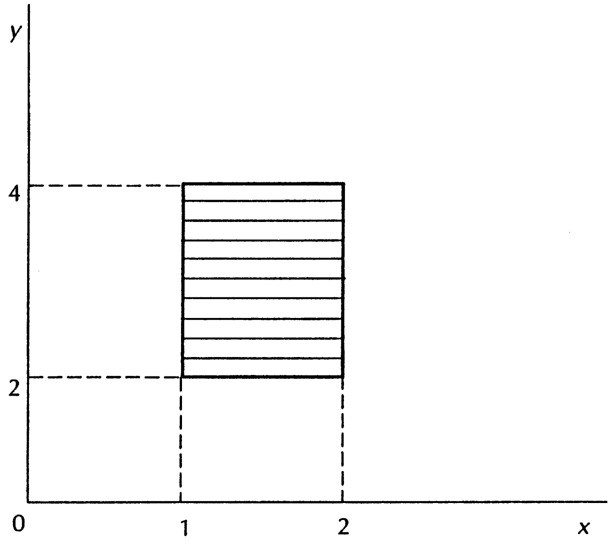


**Figure A-2**
*A* × *B* = shaded area.

need arise, a general Cartesian product of sets could also be represented by the notation $\times_{i \in I} A_i$, where here the product is taken over all sets having subscript $i$ in the index set $I$ (recall Def. A.11 and Def. A.12, and the use of index set notation).

In certain cases, we may be interested in forming a Cartesian product of a set $A$ with itself. While we might represent such a Cartesian product by the notation

$$\times_{i=1}^{n} A = \{(x_1, \ldots, x_n) : x_i \in A, i = 1, \ldots, n\},$$

such a Cartesian product is generally denoted by $A^n$, and so, for example, $A^2 = A \times A$.

### Relation (Binary)

We now define what we mean by the term *binary relation.*

**Definition A.16**
***Binary Relation***

Any subset of the Cartesian product $A \times B$ is a **binary relation from $A$ to $B$.**

Note that the adjective *binary* signifies that only two sets are involved in the relation. Relations that involve more than two sets can be defined, but for our purposes the concept of a binary relation will suffice.[3] Henceforth, we will use the word *relation* to mean binary relation. We should also mention that in the case where $B = A$, we will simply remain consistent with Def. A.16 and

[3] A higher-order relation could be defined by taking a subset of the Cartesian product $\times_{i=1}^{n} A_i$, for example.

refer to a subset of $A \times A$ as a *relation from A to A*, although in this special case some authors prefer to call the subset of $A \times A$ a relation *on* $A$.

Now let $S \subset A \times B$. Thus, by definition, $S$ is a relation from $A$ to $B$. (We emphasize at this point that the choice of the letter $S$ is quite arbitrary, and we could just as well have chosen any other letter to represent a subset of $A \times B$ defining a relation from $A$ to $B$.) If $(x, y) \in S$, we say that **$x$ is in the relation $S$ to $y$** or that **$x$ is $S$-related to $y$**. An alternative notation for $(x, y) \in S$ is $xSy$. Also, we use $S\colon A \twoheadrightarrow B$ as an abbreviation for "the relation $S$ from $A$ to $B$."

As it now stands, the concept of a relation no doubt appears quite abstract. However, in practice, it is the context provided by the definition of the subset $S$ and the definitions of the sets $A$ and $B$ that provide intuitive meaning to $xSy$. That is, $x$ will be $S$-related to $y$ because of some property satisfied by the $(x, y)$ pair, the property being indicated in the set definition of $S$. The real-world objects being related will be clearly identified in the set definitions of $A$ and $B$. Some examples will clarify the intuitive side of the relation concept.

**Example A.13** Let $A = [0, \infty)$, and form the Cartesian product $A^2 = \{(x, y)\colon x \in A \text{ and } y \in A\}$. The set $A^2$ can thus be interpreted as the nonnegative (or first) quadrant of the Euclidean plane. Then $S = \{(x, y)\colon x \geq y, (x, y) \in A^2\}$ is a relation from $A$ to $A$ representing the set of points in the nonnegative quadrant for which the first coordinate has a value greater than or equal to the value of the second coordinate. The defining property of the relation $S$ is "$\geq$." This is displayed in Figure A.3. □

**Example A.14** Let $A = \{x\colon x \text{ is an employed U.S. citizen}\}$ and $B = \{y\colon y \text{ is a U.S. corporation}\}$. Then $A \times B = \{(x, y)\colon x \text{ is an employed U.S. citizen and } y \text{ is a U.S. corporation}\}$ is the set of all possible pairings of employed U.S. citizens with U.S. corporations.

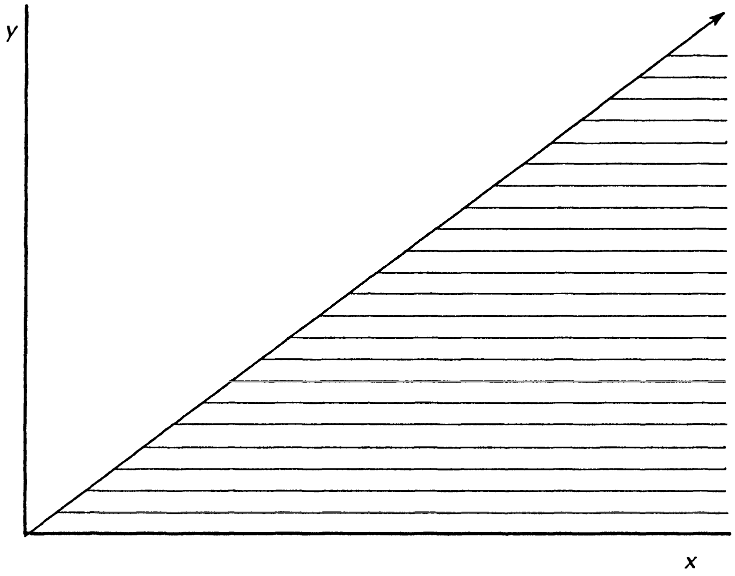


**Figure A-3**
$xSy$ in shaded area.

The relation $S = \{(x, y): x \text{ is employed by } y, (x, y) \in A \times B\}$ from $A$ to $B$ is the collection of U.S. citizens who are employed by U.S. corporations paired with their respective corporate affiliation. The defining property of the relation is the phrase "is employed by," and $xSy$ iff $x$ is a U.S. citizen employed by a U.S. corporation, and $y$ is his or her corporate affiliation. □

### Function

We are now in a position to define what is meant by the concept of a function. As indicated in the following definition, a function is simply a special type of relation. In the definition we introduce the symbol $\forall$ , which means **for every** or **for all**, and the symbol $\exists$, which means **there exists**.

**Definition A.17** ***Function***

A function from $A$ to $B$ is a relation $S: A \to B$ such that $\forall\ a \in A\ \exists$ one unique $b \in B$ such that $(a, b) \in S$.

A relation satisfying the above condition will often be given a special symbol to distinguish the relation as a function. A popular symbol used for this purpose is "$f$," where $f: A \to B$ is a common notation for designating the **function $f$ from $A$ to $B$**. As we had remarked when choosing a symbol to represent a relation, the choice of the letter $f$ is arbitrary, and when it is convenient or useful, any other letter or symbol could be used to depict a subset of $A \times B$ that represents a function from $A$ to $B$. In the text we will often have occasion to use a variety of letters to designate various functions of interest.

The unique element $b \in B$ that the function $f: A \to B$ associates with a given element $x \in A$ is called **the image of x under $f$** and is represented symbolically by the notation $\boldsymbol{f(x)}$. If $f(x)$ is a real number, the image of $x$ under $f$ is alternatively referred to as **the value of the function $f$ at $x$**. In the following example, we use $\boldsymbol{R}$ to denote the set of real numbers $(-\infty, \infty)$, i.e., $R$ stands for the **real line**. Furthermore, the nonnegative subset of the real line is represented by $R_{\geq 0}, = [0, \infty)$.

**Example A.15** Let $f: R \to R_{\geq 0}$ be defined by $f = \{(x, y): y = x^2, x \in R\}$. The *image of* $-2$ *under* $f$ is $f(-2) = 4$. The *value of the function $f$ at 3* is $f(3) = 9$. □

Associated with a given function, $f$, are two important sets called the **domain** and **range** of the function.

**Definition A.18** ***Domain and range of a function***

The **domain** of a function $f: A \to B$ is defined as $D(f) = A$. The **range** of $f$ is defined as $R(f) = \{y: y = f(x), x \in A\}$.

Thus, the domain of a function $f: A \to B$ is simply the set $A$ of the Cartesian product $A \times B$ associated with the function. The range of $f$ is the collection of all elements in $B$ that are images of the elements in $A$ under the function $f$. It

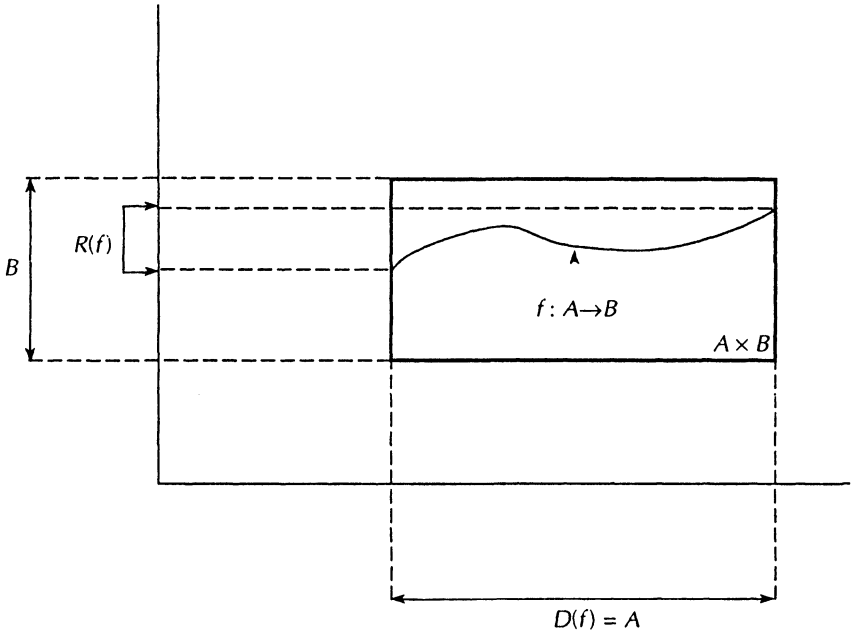


**Figure A-4**
Function with domain $D(f)$ and range $R(f)$

follows that $R(f) \subset B$. Figure A.4 provides a pictorial example of a function, including its *domain* and *range.*

Note that the concept of a function is completely general regarding the nature of the elements of the sets $D(f) = A$, $R(f)$, and $B$. The elements can be numbers, or other objects, or the elements can be sets themselves. For our work, it will suffice to deal only with *real-valued functions,* meaning that $R(f)$ is a set of real numbers.

**Definition A.19**
***Real-valued function***

A function $f: A \rightarrow B$ such that $R(f) \subset R$ is called a **real-valued function.**

The function defined in Ex. A.15 is a real-valued function. The following is another example.

**Example A.16** Examine the Cobb–Douglas production function $f: R^2_{\geq 0} \rightarrow R_{\geq 0}$ defined as $f = \{((x_1, x_2), y): y = 10x_1^2 x_2, (x_1, x_2) \in R^2_{\geq 0}\}$. Interpreting $(x_1, x_2)$ as inputs into a production process, and $y$ as the output of the process, we see that the domain of the production function is $D(f) = R^2_{\geq 0}$, i.e., any nonnegative level of the input pair $(x_1, x_2)$ is an admissible input level. The associated range of the production function is $R(f) = R_{\geq 0}$, i.e., any nonnegative level of output, $y$, is possible. Since $R(f) = [0, \infty) \subset R$, the production function is a real-valued function. □

In some cases the relation from $A$ to $B$ that defines the function $f: A \rightarrow B$ also defines a function from $B$ to $A$. This relates to the concept of an **inverse**

**function**. In particular, if for each element $y \in B$ there exists precisely one element $x \in A$ whose image under $f$ is $y$, then such an *inverse function* exists.

**Definition A.20**
***Inverse function of f***

Let $f\colon A \to B$ be a function from $A$ to $B$. If $R(f) = B$ and $\forall\, y \in B\ \exists$ a unique $x \in A$ such that $y = f(x)$, then the relation $\{(y,x)\colon y = f(x), y \in B\}$ is a function from $B$ to $A$ called the **inverse function of** $f$ and denoted by $f^{-1}\colon B \to A$.

Note that neither of the functions in Ex. A.15 or Ex. A.16 is such that an inverse function exists. In Ex. A.15, the uniqueness condition of Def. A.20 is violated since $\forall\, y \neq 0$ there exist two values of $x$ for which $y = x^2$, namely $x = \pm\sqrt{y}$. For example, when $y = 4$, $x = 2$ and $-2$ are each such that $y = x^2$. The reader can verify that Ex. A.16 also violates the uniqueness condition where an infinite number of $(x_1, x_2)$ values satisfy $y = 10x_1^2x_2$ for a fixed value of $y$ (defining level sets or isoquants of the production function). Also, note that an inverse function does not exist for the function illustrated in Figure A.4.

As an example of a function for which an inverse function does exist, consider the following.

**Example A.17**

Let $f\colon R \to R_+$ be defined as

$f = \{(x,y)\colon y = e^x, x \in R\}$, where $R_+ = (0,\infty)$.

Note that the *inverse function* can be represented as

$f^{-1} = \{(y,x)\colon x = \ln y, y \in R_+\}$,

so that $f^{-1}\colon R_+ \to R$. It is clear that $\forall\, y \in R_+$, $\exists$ one and only one $x \in R$ such that $x = \ln y$. □

The final concept concerning functions that we will review here is the **inverse image** of $y \in R(f)$ or of $H \subset R(f)$. The inverse image of $y$ is the set of domain elements $x \in D(f)$ such that $y = f(x)$, i.e., the collection of all $x$ values in the domain of $f$ whose image under the function $f$ is $y$. The inverse image of $y$ can be represented as the set $\{x\colon f(x) = y\}$, and when the inverse function exists, the inverse image of $y$ can be represented as the value $f^{-1}(y)$. In Ex. A.17, the inverse image of 5 is $f^{-1}(5) = \ln(5) = 1.6094$; in Ex. A.15, the inverse image of 4 is $\{-2, 2\}$; and in Ex. A.16, the inverse image of 3 is the isoquant $\{(x_1,x_2)\colon 3 = 10x_1^2x_2, (x_1,x_2) \in R^2_{\geq 0}\}$. Similarly, the inverse image of $H \subset R(f)$ is the set of $x$ values in the domain of $f$ whose images under $f$ equal some $y \in H$, i.e., the inverse image of $H$ is $\{x\colon f(x) = y, y \in H\}$, and if the inverse function exists, $\{x\colon x = f^{-1}(y), y \in H\}$.

### Real-Valued Point Versus Set Functions

Two types of functions—point functions and set functions—are utilized extensively in modern discussions of probability and mathematical statistics. The reader should already have considerable experience with the application of real-valued point functions, since this type of function is the one that appears in elementary algebra and calculus courses and is central to discussions of utility, demand, production, and supply that the reader has encountered in his or her study of economic theory. Specifically, a **real-valued point function** is a real-valued function whose domain consists of a collection of points, where points are represented by coordinate vectors in $R^n$. We have encountered examples of this type of function previously in Ex. A.15 through Ex. A.17. A typical ordered pair associated with a real-valued point function is of the form $(\mathbf{x}, y)$, where $\mathbf{x}$ is a vector in $R^n$ and $y$ is a real number in $R$.

A set function is more general than a point function in that its domain consists of a collection of *sets* rather than a collection of points.[4] A typical ordered pair belonging to a **real-valued set function** would have the form $(A, y)$, where $A$ is a set of some type of objects and $y$ is a real number in $R$. If the sets in the domain of the set function are contained in $R^n$, i.e., they are sets of real numbers, then a real-valued set function assigns a real number to each *set* of points in its domain, whereas a real-valued point function assigns a real number to each *point* in its domain. A pictorial illustration of a set function contrasted with a point function is given in Figure A.5.

Examples of set functions are presented below.

**Example A.18** Let $\Omega = \{1, 2, 3\}$, and let $A$ be the collection of all of the subsets of $\Omega$, i.e., $A = \{A_1, A_2, \ldots, A_8\}$, where $A_1 = \{1\}, A_2 = \{2\}, A_3 = \{3\}, A_4 = \{1, 2\}, A_5 = \{1, 3\}$, $A_6 = \{2, 3\}$, $A_7 = \{1, 2, 3\}$, and $A_8 = \emptyset$.

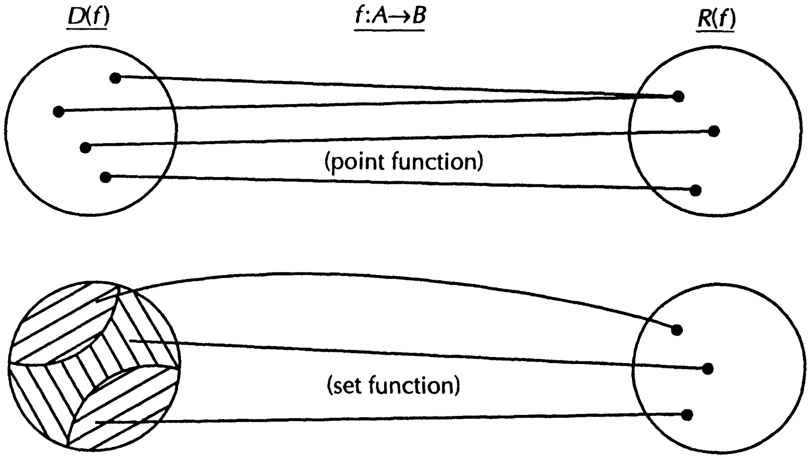


**Figure A-5**
Point versus set function.

[4] Note that, in a sense, a point function can be viewed as a special case of a set function, since points can be interpreted as singleton (single-element) sets. The set function concept is introduced to accommodate the case where one or more sets in its domain are *not* singleton.

The following is a real-valued set function $f\colon A \rightarrow R$:

$$f = \{(A_i, y)\colon y = \sum_{x \in A_i} x, A_i \subset A\},$$

where $\sum_{x \in A_i} x$ signifies the sum of the numerical values of all of the elements in the set $A_i$, and $\sum_{x \in \emptyset} x$ is defined to be zero. The range of the set function is $R(f) = \{0, 1, 2, 3, 4, 5, 6\}$, and the domain is the set of sets $D(f) = A$. The function can be represented in tabular form as follows:

| $A_i$ | $f(A_i)$ |
|---|---|
| $A_1$ | 1 |
| $A_2$ | 2 |
| $A_3$ | 3 |
| $A_4$ | 3 |
| $A_5$ | 4 |
| $A_6$ | 5 |
| $A_7$ | 6 |
| $A_8$ | 0 |

□

**Example A.19** Let $A = \{A_r\colon A_r = \{(x, y)\colon x^2 + y^2 \leq r^2\}, r \in [0, 1]\}$, so that $A$ is a set of sets, where the typical element $A_r$ represents the set of points in $R^2$ that are on the boundary and in the interior of a circle centered at (0,0) with radius $r$. The following is a real-valued set function $f\colon A \rightarrow R$:

$$f = \{(A_r, y)\colon y = \pi r^2, A_r \subset A\}.$$

Note that the set function assigns a real number representing the area to each set, $A_r$. The assignment is made for circles having a radius anywhere from 0 to 1. The range of the set function is given by $R(f) = [0, \pi]$, and the domain is the *set of sets* $D(f) = A$. □

A special type of set function called the size-of-set function will prove to be quite useful.

**Definition A.21**
***Size of set function***

Let $A$ be any set of objects. The **size-of-set function,** $N$, is the set function that assigns to the set $A$ the number of elements that are in set $A$, i.e., $N(A) = \sum_{x \in A} 1$.[5]

Applying the size-of-set function in Ex. A.18, note that $N(A) = 8$. In Ex. A.19, note that $N(A) = \infty$.

Another special (point) function that will be useful in our study is the **indicator function**, defined as follows.

[5] Note that $\sum_{x \in A} 1$ signified that a collection of 1's are being summed together, the number in the collection being equal to the number of elements in the set $A$. If $A = \emptyset$, effectively no 1's are being added together, and thus $N(\emptyset) = 0$.

**Definition A.22**
***Indicator function***

Let $A$ be any subset of some universal set $\Omega$. The indicator function, denoted by $I_A$, is a real-valued function with domain $\Omega$ and range $\{0, 1\}$ such that

$$I_A(x) = \left\{ \begin{matrix} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{matrix} \right\}.$$

Note that the indicator function *indicates* the set $A$ by assigning the number 1 to any $x$ that is an element of $A$, while assigning zero to any $x$ that is not an element of $A$. The main use of the indicator function is notational efficiency in defining functions, as the following example illustrates.

**Example A.20** Let the function $f\colon R \to R$ be defined by

$$f(x) = \begin{bmatrix} 0 & \text{for } x \in (-\infty, 0] \\ x & \text{for } x \in (0, 2] \\ 3 - x & \text{for } x \in (2, 3] \\ 0 & \text{for } x \in (3, \infty) \end{bmatrix}.$$

Utilizing the indicator function, we can alternatively represent $f(x)$ as

$$f(x) = xI_{(0,2]}(x) + (3 - x)I_{(2,3]}(x).$$ □

As a final note on the use of functions, we (as do the vast majority of other authors) will generally use a shorthand method for defining functions by simply specifying the relationship between elements in the domain of a function and their respective images in the range of the function. For example, we would define the function in Ex. A.19 by $f(A_r) = \pi r^2$ for $A_r \subset A$, or define the function in Ex. A.15 by $f(x) = x^2$ for $x \in R$. In all cases, the reader should remember that a function is a set of ordered pairs $(x, f(x))$, or $(A, f(A))$. The reader will sometimes find in the literature phrases like **the function $f(x)$** or **the set function $f(A)$**. Literally speaking, such phrases are inconsistent, because $f(x)$ and $f(A)$ are not functions but rather *images* of elements in the domain of the respective functions. In fact, the reader should *not* take such phrases literally, but rather interpret these phrases as shorthand for phrases such as *the function whose values are given by $f(x)$* or *the set function whose values are given by $f(A)$*.

## A.5 Combinations and Permutations

In a number of situations involving probability assignments, it will be useful to have an efficient method for counting the number of *different* ways a group of $r$ objects can be selected from a group of $n$ distinct objects, $n \geq r$. Obviously, if two groups of $r$ objects do not contain the same $r$ objects, they must be considered *different*. But what if two groups of $r$ objects do contain the same objects, except the objects in the group are arranged in different orders? Are the two groups to be considered *different*? If difference in order constitutes difference in groups, then we are dealing with the notion of **permutations**. On the other hand, if the order of listing the objects in a group is not used as a basis for distinguishing between groups, then we are dealing with the notion of **combinations**.

In order to establish a formula for determining the number of permutations of $n$ distinct objects taken $r$ at a time, the following example is suggestive.

**Example A.21** Examine the number of different ways a group of three letters can be selected from the letters $a, b, c, d$, where difference in order of listing is taken to mean difference in groups. Note that the first letter can be chosen in four different ways. After we have chosen one of the letters for the first selection, the second selection can be any of the remaining three letters. Finally, after we have chosen two letters in the first two selections, there are then two letters left to be potentially chosen for the third selection. Thus, there are $4 \cdot 3 \cdot 2 = 24$ different ways of selecting a group of 3 letters from the letters $a, b, c, d$ if difference in the order of listing constitutes difference in groups. (The reader should attempt to list the 24 different groups.) □

The logic of the preceding example can be applied to establish a general formula for determining the number of permutations of $n$ distinct objects taken $r$ at a time:

$$(n)_r = \frac{n!}{(n-r)!} = n(n-1)(n-2)\cdots(n-r+1),$$

where ! denotes the **factorial operation**, i.e.,

$$n! = n(n-1)(n-2)(n-3)\ldots 1.^6$$

Thus, for example, $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$. In Ex. A.21, $n = 4$ and $r = 3$, so that $(4)_3 = 4!/1! = 24$.

In order to establish a formula for determining the number of combinations of $n$ distinct objects taken $r$ at a time, we return to Ex. A.21.

**Example A.21** (continued)

Examine the number of different ways a group of three letters can be selected from the letters $a, b, c, d$, where difference in order of listing does *not* imply the groups are different. Recall that we discovered that there were 24 permutations of the 4 letters $a, b, c, d$ selected 3 at a time. Now note that any 3 letters, say $a, b, c$, can be arranged in $(3)_3 = 6$ different orders, which represents "overcounting" from the combinations point of view. Reducing the number of permutations by the degree of "overcounting" results in the number of combinations, i.e., there are $24/6 = 4$ combinations of the 4 letters taken 3 at a time, namely $(a, b, c)$, $(a, b, d)$, $(a, c, d)$, and $(b, c, d)$. □

In the preceding example, the number of permutations of $n$ $(= 4)$ objects taken $r$ $(= 3)$ at a time was reduced by a factor of $r!$ $(= 3!)$, where the latter value

[6] By definition, we take $0!=1$.

represents the number of possible permutations of $r$ objects. This suggests the general formula for the number of combinations of $n$ objects taken $r$ at a time:

$$\binom{n}{r} = \frac{(n)_r}{r!} = \frac{n!}{(n-r)!r!}.$$

In Ex. A.21, we have

$$\binom{4}{3} = \frac{4!}{1!3!} = 4$$

as the appropriate number of combinations.

The concept of combinations is useful in determining the number of subsets that can be constructed from a finite set $A$. Note that in counting the number of subsets, changes in the order of listing set elements do not produce a different set, e.g., the sets $\{a, b, c\}$ and $\{c, a, b\}$ are the same set of letters (recall the definition of a set). Then the total number of subsets of a set $A$ containing $n$ elements is given by the number of different subsets defined by taking no elements (i.e., the null set) plus the number of different subsets defined by taking one element, plus the number of different subsets defined by taking two elements, ..., and finally, the number of different subsets defined by taking all $n$ elements (i.e., the set $A$ itself). Thus, the total number of different subsets of $A$ can be written as

$$\sum_{r=0}^{n} \binom{n}{r} = \sum_{r=0}^{n} \frac{n!}{(n-r)!r!}.$$

This sum can be greatly simplified by recalling that

$$(x+y)^n = \sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r}, \quad n = 1, 2, \ldots,$$

which is the **binomial theorem**. Then letting $x = y = 1$, we have that

$$2^n = \sum_{r=0}^{n} \binom{n}{r},$$

so that $2^n$ is the number of different subsets contained in a set $A$ that has $n$ elements.

**Example A.22** In Ex. A.18, recall that we identified a total of eight subsets of the set $\Omega = \{1, 2, 3\}$. This is the number of subsets we would expect from our discussion above, i.e., since $n = 3$, there are $2^3 = 8$ subsets of $\Omega$. □

It should be noted that $\binom{n}{r}$ is *defined* to be 0 whenever $n < r$, or whenever $n$ and/or $r < 0$ or are not integer valued. The rationale for $\binom{n}{r} = 0$ in each of these cases is that there is no way to define subsets of size $r$ from a collection of $n$ objects for the designated values of $n$ and $r$.

When $n$ is large, the calculation of $n!$ needed in the previous formulas pertaining to numbers of permutations or combinations can be quite formidable.

A result known as Stirling's formula can provide a useful approximation to $n!$ for large $n$:

**Definition A.23** ***Stirling's formula***[7]

$n! \approx (2\pi)^{1/2} n^{n+.5} e^{-n}$ for large $n$.[8]

A logical question to ask regarding the use of Stirling's formula is how large is "large $n$"? Stirling's formula invariably underestimates $n!$, but the percentage error $\leq 1\%$ for $n \geq 10$, and it monotonically decreases as $n \to \infty$.

## A.6 Summation, Integration and Matrix Differentiation Notation

We will use a number of variations on summation and integration notation in this text. The meaning of the various types of notation are presented in Table A.1.

We illustrate the use of some of the notation in the following example.

**Example A.23** Summation Notation

Let $A_1 = \{1,2,3\}$, $A_2 = \{2,4,6\}$, $A = A_1 \times A_2$, $B = \{(x_1,x_2)\colon x_1 \in A_1, x_2 \in \{x_1, x_1+1, \ldots, 3x_1\}\}$, $y = (y_1, y_2, \ldots, y_n)$, and $f(x_1,x_2) = x_1 + 2x_2^2$. Then

$$\sum_{x \in A_1} x = 1+2+3 = 6, \qquad \sum_{x \in A_2} x^2 = 2^2 + 4^2 + 6^2 = 56,$$

$$\sum_{i \in A_2} y_i = y_2 + y_4 + y_6, \qquad \sum_{i \in A_1} y_i = \sum_{i=1}^{3} y_i = y_1 + y_2 + y_3,$$

$$\sum_{x_1 \in A_1} \sum_{x_2 \in A_2} f(x_1,x_2) = \sum_{x_1 \in A_1} \sum_{x_2 \in A_2} (x_1 + 2x_2^2) = 354,$$

$$\sum\sum_{(x_1,x_2) \in A} f(x_1,x_2) = \sum_{x_1 \in A_1} \sum_{x_2 \in A_2} f(x_1,x_2) = 354,$$

$$\sum\sum_{(x_1,x_2) \in B} f(x_1,x_2) = \sum_{x_1=1}^{3} \sum_{x_2=x_1}^{3x_1} (x_1 + 2x_2^2) = 802.$$

□

**Example A.24** Integration Notation

Let $A_1 = [0,3]$, $A_2 = [2,4]$, $A = A_1 \times A_2$, $B = \{(x_1,x_2)\colon x_1 \in A_1, 0 < x_2 < x_1^2\}$, and $f(x_1,x_2) = x_1 x_2^2$. Then

$$\int_{x \in A_1} 2x\,dx = \int_0^3 2x\,dx = \frac{2x^2}{2} \Big|_0^3 = 9,$$

[7] See W. Feller (1968), *An Introduction to the Theory of Probability and Its Applications*, 3rd ed. pp. 52–54.

[8] Note that $\approx$ means "approximately equal to."

**Table A.1** Summation and Integration Notation

| Notation | Definition |
|---|---|
| $\sum_{i=\ell}^{n} x_i$ | Sum the values of $x_\ell, x_{\ell+1}, \ldots, x_n$, i.e., $x_\ell + x_{\ell+1} + \cdots + x_n$. |
| $\sum_{i \in I} x_i$ | Sum the values of the $x_i$'s, for $i \in I$. |
| $\sum_{x \in A} x$ | Sum the values of $x \in A$. |
| $\sum_{x=a}^{b} x$ | Sum the values of $x$ in the sequence of integers from $a$ to $b$, i.e., $a + (a+1) + (a+2) + \cdots + b$. |
| $\sum_{i=\ell}^{n} \sum_{j=k}^{m} x_{ij}$ | Sum the values of the $x_{ij}$'s for $i = \ell, \ell + 1, \ldots, n$ and $j = k, k + 1, \ldots, m$. |
| $\sum_{i \in I} \sum_{j \in J} x_{ij}$ or $\sum \sum_{(i,j) \in A} x_{ij}$ | Sum the values of the $x_{ij}$'s for $i \in I$ and $j \in J$, or for $(i, j) \in A$. |
| $\sum_{x_1 \in A_1} \cdots \sum_{x_n \in A_n} f(x_1, \ldots, x_n)$ | Sum the values of $f(x_1, \ldots, x_n)$ for $x_i \in A_i$, $i = 1, \ldots, n$. |
| $\sum \cdots \sum_{(x_1, \ldots, x_n) \in A} f(x_1, \ldots, x_n)$ | Sum the values of $f(x_1, \ldots, x_n)$ for $(x_1, \ldots, x_n) \in A$. |
| $\sum_{x_1=a_1}^{b_1} \cdots \sum_{x_n=a_n}^{b_n} f(x_1, \ldots, x_n)$ | Sum the values of $f(x_1, \ldots, x_n)$ for $x_i$ in the sequence of integers $a_i$ to $b_i$, $i = 1, \ldots, n$. |
| $\int_a^b f(x)dx$ | Integral of the function $f(x)$ from $a$ to $b$ ($a$ can be $-\infty$ and/or $b$ can be $\infty$). |
| $\int_{x \in A} f(x)dx$ | Integral of the function $f(x)$ over the set of points $A$. |
| $\int_{x_1 \in A_1} \cdots \int_{x_n \in A_n} f(x_1, \ldots, x_n)dx_n \ldots dx_1$ | Iterated integral of the function $f(x_1, \ldots, x_n)$ over the points $x_i \in A_i$, $i = 1, \ldots, n$. |
| $\int \cdots \int_{(x_1, \ldots, x_n) \in A} f(x_1, \ldots, x_n)dx_1 \ldots dx_n$ | Multiple integral of the function $f(x_1, \ldots, x_n)$ over the points $(x_1, \ldots, x_n) \in A$. |
| $\int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} f(x_1, \ldots, x_n)dx_n \ldots dx_1$ | Iterated integral of the function $f(x_1, \ldots, x_n)$ for $x_i$ in the (open, half open-half closed, or closed) interval $a_i$ to $b_i$, for $i = 1, \ldots, n$. |

$$\int_{x \in A_1 \cap A_2} x^2\, dx = \int_2^3 x^2\, dx = \frac{x^3}{3} \Big|_2^3 = \frac{19}{3},$$

$$\begin{aligned}\int_{x_1 \in A_1} \int_{x_2 \in A_2} f(x_1, x_2) dx_2\, dx_1 &= \int_0^3 \int_2^4 f(x_1, x_2) dx_2\, dx_1 \\ &= \int_0^3 \frac{x_1 x_2^3}{3} \Big|_2^4 dx_1 = \int_0^3 \frac{56}{3} x_1\, dx_1 \\ &= \frac{56 x_1^2}{6} \Big|_0^3 = 84,\end{aligned}$$

$$\iint\limits_{(x_1, x_2) \in A} f(x_1, x_2) dx_1\, dx_2 = \int_{x_1 \in A_1} \int_{x_2 \in A_2} f(x_1, x_2) dx_2\, dx_1 = 84,$$

$$\iint\limits_{(x_1,x_2)\in B} f(x_1,x_2)dx_1\,dx_2 = \int_0^3 \int_0^{x_1^2} f(x_1,x_2)dx_2\,dx_1$$

$$= \int_0^3 \frac{x_1 x_2^3}{3} \Big|_0^{x_1^2} dx_1 = \int_0^3 \frac{x_1^7}{3} dx_1$$

$$= \frac{x_1^8}{24} \Big|_0^3 = 273.375. \qquad \square$$

Regarding matrix differentiation notation, we utilize the following conventions. Let $g(\mathbf{x})$ and $\mathbf{y}(\mathbf{x})$ be a scalar and $n \times 1$ vector function of the $k \times 1$ vector $\mathbf{x}$, respectively. Then

| Derivative | Matrix Dimension | $(i, j)$th Entry |
|---|---|---|
| $\dfrac{\partial g(\mathbf{x})}{\partial \mathbf{x}}$ | $k \times 1$ | $\dfrac{\partial g}{\partial x_i}$ |
| $\dfrac{\partial g(\mathbf{x})}{\partial \mathbf{x}'} = \dfrac{\partial g(\mathbf{x})'}{\partial \mathbf{x}}$ | $1 \times k$ | $\dfrac{\partial g}{\partial x_j}$ |
| $\dfrac{\partial g(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'}$ | $k \times k$ | $\dfrac{\partial^2 g}{\partial x_i \partial x_j}$ |
| $\dfrac{\partial \mathbf{y}(\mathbf{x})}{\partial \mathbf{x}}$ | $k \times n$ | $\dfrac{\partial y_j}{\partial x_i}$ |
| $\dfrac{\partial \mathbf{y}(\mathbf{x})}{\partial \mathbf{x}'} = \dfrac{\partial \mathbf{y}(\mathbf{x})'}{\partial \mathbf{x}}$ | $n \times k$ | $\dfrac{\partial y_i}{\partial x_j}$ |

## Key Words, Phrases, and Symbols

definition
axiom (or postulate)
theorem (or proposition)
corollary
lemma
set
element
∈
negation, /
exhaustive listing
verbal rule
mathematical rule
finite set
iff (if and only if)
infinite set
countable set
uncountable set
subset
equality of sets, =
universal set
empty or null set, ∅
set difference, −
complement, $\bar{A}$
union, ∪
intersection, ∩
mutually exclusive (disjoint)
Venn diagram
idempotency laws
commutative laws
associative laws
distributive laws
identity elements
intersection and union of complements
complements of complements
intersection with null set
null set as a subset
DeMorgan's laws
index set
multiple union notation
multiple intersection notation
real line
interval set notation
closed, open, half-open-half-closed intervals
Cartesian product
binary relation from $A$ to $B$, $S: A \to B$
$xSy$
∀ (for every)
∃ (there exists)

function from $A$ to $B$, $f\colon A \to B$
real-valued function
$A^2 = A \times A$
$R_{\geq 0}$, $R_+$
image of $x$ under $f$
inverse function, $f^{-1}\colon B \to A$
inverse image

point function
set function
size-of-set function
indicator function, $I_A(x)$
"the function $f(x)$"
"the set function $f(A)$"

permutations, $(n)_r$
combinations, $\binom{n}{r}$
binomial theorem
Stirling's formula
summation notation
integration notation

## Problems

**1.** Using either an exhaustive listing, verbal rule, or mathematical rule, define the following sets:

a. the set of all senior citizens receiving social security payments in the United States;

b. the set of all positive numbers that are positive integer powers of the number 10 (i.e., $10^1$, $10^2$, etc.);

c. the set of all possible outcomes resulting from rolling a red and a green die and calculating the values of $y - x$, where $y$ = number of dots on the red die and $x$ = number of dots on the green die;

d. the set of all two-tuples $(x_1, x_2)$ where $x_1$ is any real number and $x_2$ is related to $x_1$ by raising the number $e$ to the power $x_1$.

**2.** Label the sets you have identified in Problem 1. as being either finite, countably infinite, or uncountably infinite, and explain your choice.

**3.** For each set below, state whether the set is finite, countably infinite, or uncountably infinite.

a. $S = \{x\colon x$ is a U.S. citizen who has purchased a Japanese car during the past year$\}$.

b. $S = \{(x, y)\colon y \leq x^2, x$ is a positive integer, $y \in R_{\geq 0}\}$.

c. $S = \{p\colon p$ is the price of a quart of milk sold at a retail store in the United States on Friday, September 13, 1991$\}$.

d. $S = \{x\colon x = 2y, y$ is a positive integer$\}$.

**4.** Let the universal set be $\Omega = [0, 10]$, and define the following subsets of $\Omega$:

$$A = [0, 2), \quad B = [2, 7], \quad C = [5, 6], \quad D = \{2\},$$

$$E = \{x\colon x = y^{-1}, y \text{ is an even positive integer } \geq 4\}.$$

a. Define the following sets:

$$A \cup B, \quad A \cap B, \quad \overline{A \cup C}, \quad (A \cup D) \cap B,$$

$$B - C, \quad A \cap E, \quad \bar{D} \cap B.$$

b. For each of the sets in (a), indicate whether the set is finite, countably infinite, or uncountably infinite.

**5.** Let the universal set be defined by $\Omega = [-5, 5]$, and define the following subsets of $\Omega$:

$$A_1 = [-2, 1),$$

$$A_2 = (1, 2),$$

$$A_3 = [2, 5],$$

$$A_4 = [-5, -2].$$

Also, define an index set $I = \{1, 3, 4\}$.

a. Define $\cup_{i \in I} A_i$.

b. Define $\cup_{i=1}^{4} A_i$.

c. Define $A_1 \cap A_2$.

d. Define $A_4 - A_1$.

e. Define $\bar{A}_4$.

**6.** Define the universal set, $\Omega$, as

$$\Omega = \{x\colon 0 \leq x \leq 5 \text{ or } 10 \leq x \leq 20\},$$

and define the following subsets of $\Omega$ as

$$A_1 = \{x\colon 0 \leq x < 2.5\},$$

$$A_2 = \{x\colon 15 < x \leq 20\},$$

$$A_3 = \{x\colon 2.5 \leq x \leq 5 \text{ or } 10 \leq x \leq 20\},$$

$$A_4 = \{x\colon 0 \leq x \leq 5 \text{ or } 10 \leq x \leq 15\}.$$

In addition, define the following two index sets as

$$I_1 = \{1, 3\}, \quad I_2 = \{1, 4\}.$$

Define the following sets:

a. $\cup_{i \in I_1} A_i$

b. $\cap_{i=1}^{4} A_i$

c. $\cap_{i \in I_2} A_i$

d. $\cap_{i=1}^{2} A_i$

e. $A_1 - A_2$

f. $A_4 - A_3$

g. $\bar{A}_3$

h. $A_2 - \cup_{i \in I_1} A_i$

**7.** In each situation below indicate whether the relation is a function. If so, determine the domain and range of the function.

a. $A = [0, 10]$, $B = [0, \ln(11)]$, $S = \{(x, y): y = \ln(1 + x), (x, y) \in A \times B\}$.

c. $A = R^2_{\geq 0}$, $B = [0, \infty)$, $S = \{((x_1, x_2), y): y = 5x_1x_2^2, ((x_1, x_2), y) \in A \times B\}$.

**8.** For each relation below, state whether the relation is a function, and state whether an inverse function exists. Explicitly define the inverse function if it exists.

a. Let $P = \{\$.01, \$.02, \ldots, \$1.00, \$1.01, \ldots\}$ represent a set of possible prices for a given commodity, and let $Q = [0, \infty)$ represent possible levels of quantity demanded. Define $S: P \rightarrow Q$ as

$$S = \{(p, q): q = 20p^{-1.5}, (p, q) \in P \times Q\}.$$

b. Let $A = \{D: D = \{(x_1, x_2): x_1 \in [a_1, b_1], x_2 \in [a_2, b_2]\}, a_1 < b_1, a_2 < b_2\}$ be a set of rectangular sets. Define $S: A \rightarrow R$ as

$$S = \{(D, y): y = \text{ area of } D, (D, y) \in A \times R_+\}.$$

**9.** Let $A = (0, \infty)$, and examine the following relation on $A$:

$$S = \{(x, y): y = 2 + 3x, (x, y) \in A^2)\}.$$

a. Is $2S8$?

b. Is $S$ a function?

c. Does an inverse function exist?

d. If you can, define $f(2)$ and $f^{-1}(5)$.

e. What is $D(f)$? What is $R(f)$?

**10.** Define a universal set as

$$\Omega = \{x: 0 \leq x \leq 5\},$$

and consider the set function

$$P(A) = .5 \int_{x \in A} x dx + 12.5,$$

where the domain of the set function is all subsets $A \subset \Omega$ of the form $A = [a, b]$, for $0 \leq a \leq b \leq 5$.

a. What is the image of the set $A = [0, 2]$ under $P$?

b. What is the image of $\Omega$ under $P$?

c. What is the image of the set $A = [3, 3]$ under $P$?

d. What is the range of the set function?

e. What is the inverse image of 2?

**11.** Define a set function that will assign the appropriate area to all rectangles of the form $[x_1, x_2] \times [y_1, y_2]$, $x_2 \geq x_1$ and $y_2 \geq y_1$, contained in $R^2$. Be sure to identify the domain and range of the set function.

**12.** A statistics class has 20 students in attendance in a room where 25 desks are available for the students. How many different ways can the students leave 5 desks unoccupied?

**13.** There are 15 students in an econometrics class that you are attending.

a. How many different ways can a three-person committee be formed to give a class report?

b. Of the number of possible three-person committees indicated in (a), how many involve you?

**14.** Competing for the title of Miss America are 50 contestants from each of the 50 states plus 1 contestant from the District of Columbia. How many different ways can the contestants be assigned the titles of Miss America, 1st runner up, ..., 4th runner up?

**15.** Let $A_1 = \{x: x \text{ is a positive integer}\}$, $A_2 = \{1, 2, 3, 4, 5\}$, $B = \{(x_1, x_2): (x_1, x_2) \in A_1 \times A_2, x_1 \leq x_2\}$, and $y_i = i^2$. Calculate the values of the following sums.

a. $\sum_{x \in A_2} x$

b. $\sum_{i \in A_2} y_i$

c. $\sum_{x_1 \in A_1} \sum_{x_2 \in A_2} (1/2)^{x_1} x_2^2$

d. $\sum_{x \in A_1 - A_2} (1/3)^x$

e. $\sum \sum_{(x_1, x_2) \in B} (x_1 + x_2)$

**16.** Let $A_1 = [0, \infty)$, $A_2 = [1, 10]$, and $B = \{(x_1, x_2): (x_1, x_2) \in A_1 \times A_2, x_2 > x_1\}$. Calculate the values of the following integrals.

a. $\int_{x \in A_1} (1/2) e^{-x/2} dx$

b. $\int_{x_1 \in A_1} \int_{x_2 \in A_2} x_2 e^{-x_1} dx_2 \, dx_1$

c. $\iint_{(x_1, x_2) \in B} (x_1 + x_2) dx_1 \, dx_2$

d. $\int_0^2 \int_{x_2 \in A_1 \cap A_2} x_1 x_2^2 \, dx_2 \, dx_1$

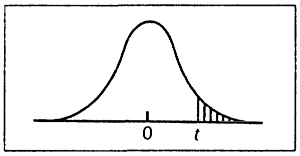
# B Useful Tables

**Table B.1** Cumulative Normal Distribution $F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$

| $x$ | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| .0 | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 | .5239 | .5279 | .5319 | .5359 |
| .1 | .5398 | .5438 | .5478 | .5517 | .5557 | .5596 | .5636 | .5675 | .5714 | .5753 |
| .2 | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 | .6026 | .6064 | .6103 | .6141 |
| .3 | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 | .6406 | .6443 | .6480 | .6517 |
| .4 | .6554 | .6591 | .6628 | .6664 | .6700 | .6736 | .6772 | .6808 | .6844 | .6879 |
| .5 | .6915 | .6950 | .6985 | .7019 | .7054 | .7088 | .7123 | .7157 | .7190 | .7224 |
| .6 | .7257 | .7291 | .7324 | .7357 | .7389 | .7422 | .7454 | .7486 | .7517 | .7549 |
| .7 | .7580 | .7611 | .7642 | .7673 | .7704 | .7734 | .7764 | .7794 | .7823 | .7852 |
| .8 | .7881 | .7910 | .7939 | .7967 | .7995 | .8023 | .8051 | .8078 | .8106 | .8133 |
| .9 | .8159 | .8186 | .8212 | .8238 | .8264 | .8289 | .8315 | .8340 | .8365 | .8389 |
| 1.0 | .8413 | .8438 | .8461 | .8485 | .8508 | .8531 | .8554 | .8577 | .8599 | .8621 |
| 1.1 | .8643 | .8665 | .8686 | .8708 | .8729 | .8749 | .8770 | .8790 | .8810 | .8830 |
| 1.2 | .8849 | .8869 | .8888 | .8907 | .8925 | .8944 | .8962 | .8980 | .8997 | .9015 |
| 1.3 | .9032 | .9049 | .9066 | .9082 | .9099 | .9115 | .9131 | .9147 | .9162 | .9177 |
| 1.4 | .9192 | .9207 | .9222 | .9236 | .9251 | .9265 | .9279 | .9292 | .9306 | .9319 |
| 1.5 | .9332 | .9345 | .9357 | .9370 | .9382 | .9394 | .9406 | .9418 | .9429 | .9441 |
| 1.6 | .9452 | .9463 | .9474 | .9484 | .9495 | .9505 | .9515 | .9525 | .9535 | .9545 |
| 1.7 | .9554 | .9564 | .9573 | .9582 | .9591 | .9599 | .9608 | .9616 | .9625 | .9633 |
| 1.8 | .9641 | .9649 | .9656 | .9664 | .9671 | .9678 | .9686 | .9693 | .9699 | .9706 |
| 1.9 | .9713 | .9719 | .9726 | .9732 | .9738 | .9744 | .9750 | .9756 | .9761 | .9767 |
| 2.0 | .9772 | .9778 | .9783 | .9788 | .9793 | .9798 | .9803 | .9808 | .9812 | .9817 |
| 2.1 | .9821 | .9826 | .9830 | .9830 | .9834 | .9838 | .9846 | .9850 | .9854 | .9857 |
| 2.2 | .9861 | .9864 | .9868 | .9871 | .9875 | .9878 | .9881 | .9884 | .9887 | .9890 |
| 2.3 | .9893 | .9896 | .9898 | .9901 | .9904 | .9906 | .9909 | .9911 | .9913 | .9916 |
| 2.4 | .9918 | .9920 | .9922 | .9925 | .9927 | .9929 | .9931 | .9932 | .9934 | .9936 |
| 2.5 | .9938 | .9940 | .9941 | .9943 | .9945 | .9946 | .9948 | .9949 | .9951 | .9952 |
| 2.6 | .9953 | .9955 | .9956 | .9957 | .9959 | .9960 | .9961 | .9962 | .9963 | .9964 |
| 2.7 | .9965 | .9966 | .9967 | .9968 | .9969 | .9970 | .9971 | .9972 | .9973 | .9974 |
| 2.8 | .9974 | .9975 | .9976 | .9977 | .9978 | .9979 | .9979 | .9979 | .9980 | .9981 |
| 2.9 | .9981 | .9982 | .9982 | .9983 | .9984 | .9984 | .9985 | .9985 | .9986 | .9986 |
| 3.0 | .9987 | .9987 | .9987 | .9988 | .9988 | .9989 | .9989 | .9989 | .9990 | .9990 |
| 3.1 | .9990 | .9991 | .9991. | .9991 | .9992 | .9992 | .9992 | .9992 | .9993 | .9993 |
| 3.2 | .9993 | .9993 | .9994 | .9994 | .9994 | .9994 | .9994 | .9995 | .9995 | .9995 |
| 3.3 | .9995 | .9995 | .9995 | .9996 | .9996 | .9996 | .9996 | .9996 | .9996 | .9997 |
| 3.4 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9998 |

(Source: Reprinted, by permission of the publisher, from A. M. Mood, F. A. Graybill, and D. C. Boes, *Introduction to the Theory of Statistics*, 3d ed., New York: McGraw-Hill, 1974, p. 552.)

**Table B.2** Student's *t* Distribution
The first column lists the number of degrees of freedom ($\nu$). The headings of the other columns give probabilities ($P$) for $t$ to exceed the entry value. Use symmetry for negative $t$ values.



| $P$<br>$\nu$ | .10 | .05 | .025 | .01 | .005 |
|---|---|---|---|---|---|
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 |
| 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 |
| 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 |
| 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 |
| 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 |
| 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 |
| 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 |
| 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 |
| 30 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 |
| 40 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 |
| 60 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 |
| 120 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 |
| $\infty$ | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 |

(Source: Reprinted, by permission of the publisher, from P. G. Hoel, *Introduction to Mathematical Statistics,* 4th ed., New York: John Wiley and Sons, Inc., 1971, p. 393.)

**Table B.3** Chi-square Distribution

The first column lists the number of degrees of freedom ($v$). The headings of the other columns give probabilities ($P$) for the $\chi^2_v$ random variable to exceed the entry value.

| $v$ \ $P$ | 0.995 | 0.990 | 0.975 | 0.950 | 0.900 | 0.750 |
|---|---|---|---|---|---|---|
| 1 | $392704 \times 10^{-10}$ | $157088 \times 10^{-9}$ | $982069 \times 10^{-9}$ | $393214 \times 10^{-8}$ | 0.0157908 | 0.1015308 |
| 2 | 0.0100251 | 0.0201007 | 0.0506356 | 0.102587 | 0.210720 | 0.575364 |
| 3 | 0.0717212 | 0.114832 | 0.215795 | 0.351846 | 0.584375 | 1.212534 |
| 4 | 0.206990 | 0.297110 | 0.484419 | 0.710721 | 1.063623 | 1.92255 |
| 5 | 0.411740 | 0.554300 | 0.831211 | 1.145476 | 1.61031 | 2.67460 |
| 6 | 0.675727 | 0.872085 | 1.237347 | 1.63539 | 2.20413 | 3.45460 |
| 7 | 0.989265 | 1.239043 | 1.68987 | 2.16735 | 2.83311 | 4.25485 |
| 8 | 1.344419 | 1.646482 | 2.17973 | 2.73264 | 3.48954 | 5.07064 |
| 9 | 1.734926 | 2.087912 | 2.70039 | 3.32511 | 1.16816 | 5.89883 |
| 10 | 2.15585 | 2.55821 | 3.24697 | 3.94030 | 4.86518 | 6.73720 |
| 11 | 2.60321 | 3.15347 | 3.81575 | 4.57481 | 5.57779 | 7.58412 |
| 12 | 3.07382 | 3.57056 | 4.40379 | 5.22603 | 6.30380 | 8.43842 |
| 13 | 3.56503 | 4.10691 | 5.00874 | 5.89186 | 7.04150 | 9.29906 |
| 14 | 4.07468 | 4.66043 | 5.62872 | 6.57063 | 7.78953 | 10.1653 |
| 15 | 4.60094 | 5.22935 | 6.26214 | 7.26094 | 8.54675 | 11.0365 |
| 16 | 5.14224 | 5.81221 | 6.90766 | 7.96164 | 9.31223 | 11.9122 |
| 17 | 5.69724 | 6.40776 | 7.56418 | 8.67176 | 10.0852 | 12.7919 |
| 18 | 6.26481 | 7.01491 | 8.23075 | 9.39046 | 10.8649 | 13.6753 |
| 19 | 6.84398 | 7.63273 | 8.90655 | 10.1170 | 11.6509 | 14.5620 |
| 20 | 7.43386 | 8.26040 | 9.59083 | 10.8508 | 12.4426 | 15.4518 |
| 21 | 8.03366 | 8.89720 | 10.28293 | 11.5613 | 13.2396 | 16.3444 |
| 22 | 8.64272 | 9.54279 | 10.3923 | 12.3380 | 14.0415 | 17.2396 |
| 23 | 9.26042 | 10.19567 | 11.6885 | 13.0905 | 14.8479 | 18.1373 |
| 24 | 9.88623 | 10.8564 | 12.4011 | 13.8484 | 15.6587 | 19.0372 |
| 25 | 10.5197 | 11.5240 | 13.1197 | 14.6114 | 16.4734 | 19.9393 |
| 26 | 11.1603 | 12.1981 | 13.8439 | 15.3791 | 17.2919 | 20.8434 |
| 27 | 11.8076 | 12.8786 | 14.5733 | 16.1513 | 18.1138 | 21.7494 |
| 28 | 12.4613 | 13.5648 | 15.3079 | 16.9279 | 18.9392 | 22.6572 |
| 29 | 13.1211 | 14.2565 | 16.0471 | 17.7083 | 19.7677 | 23.5666 |
| 30 | 13.7867 | 14.9535 | 16.7908 | 18.4926 | 20.5992 | 24.4776 |
| 40 | 20.7065 | 22.1643 | 24.4331 | 26.5093 | 29.0505 | 33.6603 |
| 50 | 27.9907 | 29.7067 | 32.3574 | 34.7642 | 37.6886 | 42.9421 |
| 60 | 35.5346 | 37.4848 | 40.4817 | 43.1879 | 46.4589 | 52.2938 |
| 70 | 43.2752 | 45.4418 | 48.7576 | 51.7393 | 55.3290 | 61.6983 |
| 80 | 51.1720 | 53.5400 | 57.1532 | 60.3915 | 64.2778 | 71.1445 |
| 90 | 59.1963 | 61.7541 | 65.6466 | 69.1260 | 73.2912 | 80.6247 |
| 100 | 67.3276 | 70.0648 | 74.2219 | 77.9295 | 82.3581 | 90.1332 |

Table B.3 (*continued*)

| $P$<br>$\nu$ | 0.500 | 0.250 | 0.100 | 0.050 | 0.025 | 0.010 | 0.005 |
|---|---|---|---|---|---|---|---|
| 1 | 0.454937 | 1.32330 | 2.70554 | 3.84146 | 5.02389 | 6.63490 | 7.87944 |
| 2 | 1.38629 | 2.77259 | 4.60517 | 5.99147 | 7.37776 | 9.21034 | 10.5966 |
| 3 | 2.36597 | 1.10835 | 6.25139 | 7.81473 | 9.34840 | 11.3449 | 12.8381 |
| 4 | 3.35670 | 5.38527 | 7.77944 | 9.48773 | 11.1433 | 13.2767 | 14.8602 |
| 5 | 4.35146 | 6.62568 | 9.23635 | 11.0705 | 12.8325 | 15.0863 | 16.7496 |
| 6 | 5.34812 | 7.84080 | 10.6446 | 12.5916 | 14.4494 | 16.8119 | 18.5476 |
| 7 | 6.34581 | 9.03715 | 12.0170 | 14.0671 | 16.0128 | 18.4753 | 20.2777 |
| 8 | 7.34412 | 10.2188 | 13.3616 | 15.5073 | 17.5346 | 20.0902 | 21.9550 |
| 9 | 8.34283 | 11.3887 | 14.6837 | 16.9190 | 19.0228 | 21.6660 | 23.5893 |
| 10 | 9.34182 | 12.5489 | 15.9871 | 18.3070 | 20.4831 | 23.2093 | 25.1882 |
| 11 | 10.3410 | 13.7007 | 17.2750 | 19.6751 | 21.9200 | 24.7250 | 26.7569 |
| 12 | 11.3403 | 14.8454 | 18.5494 | 21.0261 | 23.3367 | 26.2170 | 28.2995 |
| 13 | 12.3398 | 15.9839 | 19.8119 | 22.3621 | 24.7356 | 27.6883 | 29.8194 |
| 14 | 13.3393 | 17.1170 | 21.0642 | 23.6848 | 26.1190 | 29.1413 | 31.3193 |
| 15 | 14.3389 | 18.2451 | 22.3072 | 24.9958 | 27.4884 | 30.5779 | 32.8013 |
| 16 | 15.3385 | 19.3688 | 23.5418 | 26.2962 | 28.8454 | 31.9999 | 34.2672 |
| 17 | 16.3381 | 20.4887 | 24.4690 | 27.5871 | 30.1910 | 33.4087 | 35.7185 |
| 18 | 17.3379 | 21.6049 | 25.9894 | 28.8693 | 31.5264 | 34.8053 | 37.1564 |
| 19 | 18.3376 | 22.7178 | 27.2036 | 30.1435 | 32.8523 | 36.1908 | 38.5822 |
| 20 | 19.3374 | 23.8277 | 28.4120 | 31.4104 | 34.1696 | 37.5662 | 39.9968 |
| 21 | 20.3372 | 24.9348 | 29.6151 | 32.6705 | 35.4789 | 38.9321 | 41.4010 |
| 22 | 21.3370 | 26.0393 | 30.8133 | 33.9244 | 36.7807 | 40.2894 | 42.7956 |
| 23 | 22.3369 | 27.1413 | 32.0069 | 35.1725 | 38.0757 | 41.6384 | 44.1813 |
| 24 | 23.3367 | 28.2412 | 33.1963 | 36.4151 | 39.3641 | 42.9798 | 45.5585 |
| 25 | 24.3366 | 29.3389 | 34.3816 | 37.6525 | 40.6465 | 44.3141 | 46.9278 |
| 26 | 25.3364 | 30.4345 | 35.5631 | 38.8852 | 41.9232 | 45.6417 | 48.2899 |
| 27 | 26.3363 | 31.5284 | 36.7412 | 40.1133 | 43.1944 | 46.9630 | 49.6449 |
| 28 | 27.3363 | 32.6205 | 37.9159 | 41.3372 | 44.4607 | 48.2782 | 50.9933 |
| 29 | 28.3362 | 33.7109 | 39.0875 | 42.5569 | 45.7222 | 49.5879 | 52.3356 |
| 30 | 29.3360 | 34.7998 | 40.2560 | 43.7729 | 46.9792 | 50.8922 | 53.6720 |
| 40 | 39.3354 | 45.6160 | 51.8050 | 55.7585 | 59.3417 | 63.6907 | 66.7659 |
| 50 | 49.3349 | 56.3336 | 63.1671 | 67.5048 | 71.4202 | 76.1539 | 79.4900 |
| 60 | 59.3347 | 66.9814 | 74.3970 | 79.0819 | 83.2976 | 88.3794 | 91.9517 |
| 70 | 69.3344 | 77.5766 | 85.5271 | 90.5312 | 95.0231 | 100.425 | 104.215 |
| 80 | 79.3343 | 88.1303 | 96.5782 | 101.879 | 106.629 | 112.329 | 116.321 |
| 90 | 89.3342 | 98.6499 | 107.565 | 113.145 | 118.136 | 124.116 | 128.299 |
| 100 | 99.3341 | 109.141 | 118.498 | 124.342 | 129.561 | 135.807 | 140.169 |

(Source: Reprinted, by permission of the Biometrika Trustees from C. M. Thompson, "Tables of Percentage Points of the $\chi^2$ Distribution," *Biometrika* 32 (1941): 188–189.)

**Table B.4** *F*-Distribution: 5% Points

The first column lists the number of denominator degrees of freedom ($v_2$). The headings of the other columns list the numerator degrees of freedom ($v_1$). The table entry is the value of $c$ for which $P(F_{v_1,v_2} \geq c) = .05$.

| $v_2$ \ $v_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 161.45 | 199.50 | 215.71 | 224.58 | 230.16 | 233.99 | 236.77 | 238.88 | 240.54 |
| 2 | 18.513 | 19.000 | 19.164 | 19.247 | 19.296 | 19.330 | 19.353 | 19.371 | 19.385 |
| 3 | 10.128 | 9.5521 | 9.2766 | 9.1172 | 9.0135 | 8.9406 | 8.8868 | 8.8452 | 8.8123 |
| 4 | 7.7086 | 6.9443 | 6.5914 | 6.3883 | 6.3560 | 6.1631 | 6.0942 | 6.0410 | 5.9988 |
| 5 | 6.6079 | 5.7861 | 5.4095 | 5.1922 | 5.0503 | 4.9503 | 4.8759 | 4.8183 | 4.7725 |
| 6 | 5.9874 | 5.1433 | 4.7571 | 4.5337 | 1.3874 | 4.2839 | 4.2066 | 4.1468 | 4.0990 |
| 7 | 5.5914 | 4.7374 | 4.3468 | 4.1203 | 3.9715 | 3.8660 | 3.7870 | 3.7257 | 3.6767 |
| 8 | 5.3177 | 4.4590 | 4.0662 | 3.8378 | 3.6875 | 3.5806 | 3.5005 | 3.4381 | 3.3881 |
| 9 | 5.1174 | 4.2565 | 3.8626 | 3.6331 | 3.4817 | 3.3738 | 3.2927 | 3.2296 | 3.1789 |
| 10 | 4.9646 | 4.1028 | 3.7083 | 3.4780 | 3.3258 | 3.2172 | 3.1355 | 3.0717 | 3.0204 |
| 11 | 4.8443 | 3.9823 | 3.5874 | 3.3567 | 3.2039 | 3.0946 | 3.0123 | 2.9480 | 2.8962 |
| 12 | 4.7472 | 3.8856 | 3.4903 | 3.2592 | 3.1059 | 2.9961 | 3.9134 | 2.8486 | 2.7964 |
| 13 | 4.6672 | 3.8056 | 3.4105 | 3.1791 | 3.0254 | 2.9153 | 2.8321 | 2.7669 | 2.7144 |
| 14 | 4.6001 | 3.7389 | 3.3439 | 3.1122 | 2.9582 | 2.8477 | 2.7642 | 2.6987 | 2.6458 |
| 15 | 4.5431 | 3.6823 | 3.2874 | 3.0556 | 2.9013 | 2.7905 | 2.7066 | 2.6408 | 2.5876 |
| 16 | 4.4940 | 3.6337 | 3.2389 | 3.0069 | 2.8524 | 2.7413 | 2.6572 | 2.5911 | 2.5377 |
| 17 | 4.4513 | 3.5915 | 3.1968 | 2.9647 | 2.8100 | 2.6987 | 2.6143 | 2.5480 | 2.4943 |
| 18 | 4.4139 | 3.5546 | 3.1599 | 2.9277 | 2.7729 | 2.6613 | 2.5767 | 2.5102 | 2.4563 |
| 19 | 4.3808 | 3.5219 | 3.1274 | 2.8951 | 2.7401 | 2.6283 | 2.5435 | 2.4768 | 2.4227 |
| 20 | 4.3513 | 3.4928 | 3.0984 | 2.8661 | 2.7109 | 2.5990 | 2.5140 | 2.4471 | 2.3928 |
| 21 | 4.3248 | 3.4668 | 3.0725 | 2.8401 | 2.6848 | 2.5727 | 2.4876 | 2.4205 | 2.3661 |
| 22 | 4.3009 | 3.4434 | 3.0491 | 2.8167 | 2.6613 | 2.5491 | 2.4638 | 2.3965 | 2.3419 |
| 23 | 4.2793 | 3.4221 | 3.0280 | 2.7955 | 2.6400 | 2.5277 | 2.4422 | 2.3748 | 2.3201 |
| 24 | 4.2597 | 3.4028 | 3.0088 | 3.7763 | 2.6207 | 2.5082 | 2.4226 | 2.3551 | 2.3002 |
| 25 | 4.2417 | 3.3852 | 2.9912 | 2.7587 | 2.6030 | 2.4904 | 2.4047 | 2.3371 | 2.2821 |
| 26 | 4.2252 | 3.3690 | 2.9751 | 2.7426 | 2.5868 | 2.4741 | 2.3883 | 2.3205 | 2.2655 |
| 27 | 4.2100 | 3.3541 | 2.9604 | 2.7278 | 2.5719 | 2.4591 | 2.3732 | 2.3053 | 2.2501 |
| 28 | 4.1960 | 3.3404 | 2.9467 | 2.7141 | 2.5581 | 2.4453 | 2.3593 | 2.2913 | 2.2360 |
| 29 | 4.1830 | 3.3277 | 2.9340 | 2.7014 | 2.5454 | 2.4324 | 2.3463 | 2.2782 | 2.2229 |
| 30 | 4.1709 | 3.3158 | 2.9223 | 2.6896 | 2.5336 | 2.4205 | 2.3343 | 2.2662 | 2.2107 |
| 40 | 4.0848 | 3.2317 | 2.8387 | 2.6060 | 2.4495 | 2.3359 | 2.2490 | 2.1802 | 2.1240 |
| 60 | 4.0012 | 3.1504 | 2.7581 | 2.5252 | 2.3683 | 2.2540 | 2.1665 | 2.0970 | 2.0401 |
| 120 | 3.9201 | 3.0718 | 2.6802 | 2.4472 | 2.2900 | 2.1750 | 2.0867 | 2.0164 | 1.9588 |
| ∞ | 3.8415 | 2.9957 | 2.6049 | 2.3719 | 2.2141 | 2.0986 | 2.0096 | 1.9354 | 1.8799 |

**Table B.4** *F*-Distribution: 5% Points (*continued*)

| $v_1$<br>$v_2$ | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 241.88 | 243.91 | 245.95 | 248.01 | 249.05 | 250.09 | 251.14 | 252.20 | 253.25 | 254.32 |
| 2 | 19.396 | 19.413 | 19.429 | 19.446 | 19.454 | 19.462 | 19.471 | 19.479 | 19.487 | 19.496 |
| 3 | 8.7855 | 8.7446 | 8.7029 | 8.6602 | 8.6385 | 8.6166 | 8.5944 | 8.5720 | 8.5494 | 8.5265 |
| 4 | 5.9644 | 5.9117 | 5.8578 | 5.8025 | 5.7744 | 5.7459 | 5.7170 | 5.6878 | 5.6581 | 5.6281 |
| 5 | 4.7351 | 4.6777 | 4.6188 | 4.5581 | 4.5272 | 4.4957 | 4.4638 | 4.4314 | 4.3984 | 4.3650 |
| 6 | 4.0600 | 3.9999 | 3.9381 | 3.8742 | 3.8415 | 3.8082 | 3.7743 | 3.7398 | 3.7047 | 3.6688 |
| 7 | 3.6365 | 3.5747 | 3.5108 | 3.4445 | 3.4105 | 3.3758 | 3.3404 | 3.3043 | 3.2674 | 3.2298 |
| 8 | 3.3472 | 3.2840 | 3.2184 | 3.1503 | 3.1152 | 3.0794 | 3.0428 | 3.0053 | 2.9669 | 2.9276 |
| 9 | 3.1373 | 3.0729 | 3.0061 | 2.9365 | 2.9005 | 2.8637 | 2.8259 | 2.7872 | 2.7475 | 2.7067 |
| 10 | 2.9782 | 2.9130 | 2.8450 | 2.7740 | 2.7372 | 2.6996 | 2.6609 | 2.6211 | 2.5801 | 2.5379 |
| 11 | 2.8536 | 2.7876 | 2.7186 | 2.6464 | 2.6090 | 2.5705 | 2.5309 | 2.4901 | 2.4480 | 2.4045 |
| 12 | 2.7534 | 2.6866 | 2.6169 | 2.5436 | 2.5055 | 2.4663 | 2.4259 | 2.3842 | 2.3410 | 2.2962 |
| 13 | 2.6710 | 2.6037 | 2.5331 | 2.4589 | 2.4202 | 2.3803 | 2.3392 | 2.2966 | 2.2524 | 2.2064 |
| 14 | 2.6021 | 2.5342 | 2.4630 | 2.3879 | 2.3487 | 2.3082 | 2.2664 | 2.2230 | 2.1778 | 2.1307 |
| 15 | 2.5437 | 2.4753 | 2.4035 | 2.3275 | 2.2878 | 2.2468 | 2.2043 | 2.1601 | 2.1141 | 2.0658 |
| 16 | 2.4935 | 2.4247 | 2.3522 | 2.2756 | 2.2354 | 2.1938 | 2.1507 | 2.1058 | 2.0589 | 2.0096 |
| 17 | 2.4499 | 2.3807 | 2.3077 | 2.2304 | 2.1898 | 2.1477 | 2.1040 | 2.0584 | 2.0107 | 1.9604 |
| 18 | 2.4117 | 2.3421 | 2.2686 | 2.1906 | 2.1497 | 2.1071 | 2.0629 | 2.0166 | 1.9681 | 1.9168 |
| 19 | 2.3779 | 2.3080 | 2.2341 | 2.1555 | 2.1141 | 2.0712 | 2.0264 | 1.9796 | 1.9302 | 1.8780 |
| 20 | 2.3479 | 2.2776 | 2.2033 | 2.1242 | 2.0825 | 2.0391 | 1.9938 | 1.9464 | 1.8963 | 1.8432 |
| 21 | 2.3210 | 2.2504 | 2.1757 | 2.0960 | 2.0540 | 2.0102 | 1.9645 | 1.9165 | 1.8657 | 1.8117 |
| 22 | 2.2967 | 2.2258 | 2.1508 | 2.0707 | 2.0283 | 1.9842 | 1.9380 | 1.8895 | 1.8380 | 1.7831 |
| 23 | 2.2747 | 2.2036 | 2.1282 | 1.0476 | 2.0050 | 1.9605 | 1.9139 | 1.8649 | 1.8128 | 1.7570 |
| 24 | 2.2547 | 2.1834 | 2.1077 | 2.0267 | 1.9838 | 1.9390 | 1.8920 | 1.8424 | 1.7897 | 1.7331 |
| 25 | 2.2365 | 2.1649 | 2.0889 | 2.0075 | 1.9643 | 1.9192 | 1.8718 | 1.8217 | 1.7684 | 1.7110 |
| 26 | 2.2197 | 2.1479 | 2.0716 | 1.9898 | 1.9464 | 1.9010 | 1.8533 | 1.8027 | 1.7488 | 1.6906 |
| 27 | 2.2043 | 2.1323 | 2.0558 | 1.9736 | 1.9299 | 1.8842 | 1.8361 | 1.7851 | 1.7307 | 1.6717 |
| 28 | 2.1900 | 2.1179 | 2.0411 | 1.9586 | 1.9147 | 1.8687 | 1.8203 | 1.7689 | 1.7138 | 1.6541 |
| 29 | 2.1768 | 2.1045 | 2.0245 | 1.9446 | 1.9005 | 1.8543 | 1.8055 | 1.7537 | 1.6981 | 1.6377 |
| 30 | 2.1646 | 2.0921 | 2.0148 | 1.9317 | 1.8874 | 1.8409 | 1.7918 | 1.7396 | 1.6835 | 1.6223 |
| 40 | 2.0772 | 2.0035 | 1.9245 | 1.8389 | 1.7929 | 1.7444 | 1.6928 | 1.6373 | 1.5766 | 1.5089 |
| 60 | 1.9926 | 1.9174 | 1.8364 | 1.7480 | 1.7001 | 1.6491 | 1.5943 | 1.5343 | 1.4673 | 1.3893 |
| 120 | 1.9105 | 1.8337 | 1.7505 | 1.6587 | 1.6084 | 1.5543 | 1.4952 | 1.4290 | 1.3519 | 1.2539 |
| ∞ | 1.8307 | 1.7522 | 1.6664 | 1.5705 | 1.5173 | 1.4591 | 1.3940 | 1.3180 | 1.2214 | 1.0000 |

(Source: Reprinted, by permission of the Biometrika Trustees from M. Merrington and C. M. Thompson, "Tables of Percentage Points of the Inverted Beta(*F*) Distribution," *Biometrika* 33 (1943): 80–81.)

**Table B.5** *F*-Distribution: 1% Points

The first column lists the number of denominator degrees of freedom ($v_2$). The headings of the other columns list the numerator degrees of freedom ($v_1$). The table entry is the value of $c$ for which $P(F_{v_1,v_2} \geq c) = .01$.

| $v_1$ $v_2$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 4052.2 | 4999.5 | 5403.3 | 5624.6 | 5763.7 | 5859.0 | 5928.3 | 5981.6 | 6022.5 |
| 2 | 98.503 | 99.000 | 99.166 | 99.249 | 99.299 | 99.332 | 99.356 | 99.374 | 99.388 |
| 3 | 34.116 | 30.817 | 29.457 | 28.710 | 28.237 | 27.911 | 27.672 | 27.489 | 27.345 |
| 4 | 21.198 | 18.000 | 16.694 | 15.977 | 15.522 | 15.207 | 14.976 | 14.799 | 14.659 |
| 5 | 16.258 | 13.274 | 12.060 | 11.392 | 10.967 | 10.672 | 10.456 | 10.289 | 10.158 |
| 6 | 13.745 | 10.925 | 9.7795 | 9.1483 | 8.7459 | 8.4661 | 8.2600 | 8.1016 | 7.9761 |
| 7 | 12.246 | 9.5466 | 8.4513 | 7.8467 | 7.4604 | 7.1914 | 6.9928 | 6.8401 | 6.7188 |
| 8 | 11.259 | 8.6491 | 7.5910 | 7.0060 | 6.6318 | 6.3707 | 6.1776 | 6.0289 | 5.9106 |
| 9 | 10.561 | 8.0215 | 6.9919 | 6.4221 | 6.0569 | 5.8018 | 5.6129 | 5.4671 | 5.3511 |
| 10 | 10.044 | 7.5594 | 6.5523 | 5.9943 | 5.6363 | 5.3858 | 5.2001 | 5.0567 | 4.9424 |
| 11 | 9.6460 | 7.2057 | 6.2167 | 5.6683 | 5.3160 | 5.0692 | 4.8861 | 4.7445 | 4.6315 |
| 12 | 9.3302 | 6.9266 | 5.9526 | 5.4119 | 5.0643 | 4.8206 | 4.6395 | 4.4994 | 4.3875 |
| 13 | 9.0738 | 6.7010 | 5.7394 | 5.2053 | 4.8616 | 4.6204 | 4.4410 | 4.3021 | 4.1911 |
| 14 | 8.8616 | 6.5149 | 5.5639 | 5.0354 | 4.6950 | 4.4558 | 4.2779 | 4.1399 | 4.0297 |
| 15 | 8.6831 | 6.3589 | 5.4170 | 4.8932 | 4.5556 | 4.3183 | 4.1415 | 4.0045 | 3.8948 |
| 16 | 8.5310 | 6.2262 | 5.2922 | 4.7726 | 4.4374 | 4.2016 | 4.0259 | 3.8896 | 3.7804 |
| 17 | 8.3997 | 6.1121 | 5.1850 | 4.6690 | 4.3359 | 4.1015 | 3.9267 | 3.7910 | 3.6822 |
| 18 | 8.2854 | 6.0129 | 5.0919 | 4.5790 | 4.2479 | 4.0146 | 3.8406 | 3.7054 | 3.5971 |
| 19 | 8.1850 | 5.9259 | 5.0103 | 4.5003 | 4.1708 | 3.9386 | 3.7653 | 3.6305 | 3.5225 |
| 20 | 8.0906 | 5.8489 | 4.9382 | 4.4307 | 4.1027 | 3.8714 | 3.6987 | 3.5644 | 3.4567 |
| 21 | 8.0166 | 5.7804 | 4.8740 | 4.3688 | 4.0421 | 3.8117 | 3.6396 | 3.5056 | 3.3981 |
| 22 | 7.9454 | 5.7190 | 4.8166 | 4.3134 | 3.9880 | 3.7583 | 3.5867 | 3.4530 | 3.3458 |
| 23 | 7.8811 | 5.6637 | 4.7649 | 4.2635 | 3.9392 | 3.7102 | 3.5390 | 3.4057 | 3.2986 |
| 24 | 7.8229 | 5.6131 | 4.7181 | 4.2184 | 3.8951 | 3.6667 | 3.4959 | 3.3629 | 3.2560 |
| 25 | 7.7689 | 5.5680 | 4.6755 | 4.1774 | 3.8550 | 3.6272 | 3.4568 | 3.3239 | 3.2172 |
| 26 | 7.7213 | 5.5263 | 4.6366 | 4.1400 | 3.8183 | 3.5911 | 3.4210 | 3.2884 | 3.1818 |
| 27 | 7.6767 | 5.4881 | 4.6009 | 4.1056 | 3.7848 | 3.5580 | 3.3882 | 3.2558 | 3.1494 |
| 28 | 7.6356 | 5.4529 | 4.5681 | 4.0740 | 3.7539 | 3.5276 | 3.3581 | 3.2259 | 3.1195 |
| 29 | 7.5976 | 5.4205 | 4.5378 | 4.0449 | 3.7254 | 3.4995 | 3.3302 | 3.1982 | 3.0920 |
| 30 | 7.5625 | 5.3904 | 4.5097 | 4.0179 | 3.6990 | 3.4735 | 3.3045 | 3.1726 | 3.0665 |
| 40 | 7.3141 | 5.1785 | 4.3126 | 3.8283 | 3.5138 | 3.2910 | 3.1238 | 2.9930 | 2.8876 |
| 60 | 7.0771 | 4.9774 | 4.1259 | 3.6491 | 3.3389 | 3.1187 | 2.9530 | 2.8233 | 2.7185 |
| 120 | 6.8510 | 4.7865 | 3.9493 | 3.4796 | 3.1735 | 2.9559 | 2.7918 | 2.6629 | 2.5586 |
| $\infty$ | 6.6349 | 4.6052 | 3.7816 | 3.3192 | 3.0173 | 2.8020 | 2.6393 | 2.5113 | 2.4073 |

**Table B.5** *F*-Distribution: 1% Points (*continued*)

| $v_2$ \ $v_1$ | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 6055.8 | 6106.3 | 6157.3 | 6208.7 | 6234.6 | 6260.7 | 6286.8 | 6313.0 | 6339.4 | 6366.0 |
| 2 | 99.399 | 99.416 | 99.449 | 99.458 | 99.458 | 99.466 | 99.474 | 99.483 | 99.491 | 99.501 |
| 3 | 27.229 | 27.052 | 26.872 | 26.690 | 26.598 | 26.505 | 26.411 | 26.316 | 26.221 | 26.125 |
| 4 | 14.546 | 14.374 | 14.198 | 14.020 | 13.929 | 13.838 | 13.745 | 13.652 | 13.558 | 13.463 |
| 5 | 10.051 | 9.8883 | 9.7222 | 9.5527 | 9.4665 | 9.3793 | 9.2912 | 9.2020 | 9.1118 | 9.0204 |
| 6 | 7.8741 | 7.7183 | 7.5590 | 7.3958 | 7.3127 | 7.2285 | 7.1432 | 7.0568 | 6.9690 | 6.8801 |
| 7 | 6.6201 | 6.4691 | 6.3143 | 6.1554 | 6.0743 | 5.9921 | 5.9084 | 5.8236 | 5.7372 | 5.6495 |
| 8 | 5.8143 | 5.6668 | 5.5151 | 5.3591 | 5.2793 | 5.1981 | 5.1156 | 5.0316 | 4.9460 | 4.8588 |
| 9 | 5.2565 | 5.1114 | 4.9621 | 4.8080 | 4.7290 | 4.6486 | 4.5667 | 4.4831 | 4.3978 | 4.3105 |
| 10 | 4.8492 | 4.7059 | 4.5582 | 4.4054 | 4.3269 | 4.2469 | 4.1653 | 4.0819 | 3.9965 | 3.9090 |
| 11 | 4.5393 | 4.3974 | 4.2509 | 4.0990 | 4.0209 | 3.9411 | 3.8596 | 3.7761 | 3.6904 | 3.6025 |
| 12 | 4.2961 | 4.1553 | 4.0096 | 3.8584 | 3.7805 | 3.7008 | 3.6192 | 3.5355 | 3.4494 | 3.3608 |
| 13 | 4.1003 | 3.9603 | 3.8154 | 3.6646 | 3.5868 | 3.5070 | 3.4253 | 3.3413 | 3.2548 | 3.1654 |
| 14 | 3.9394 | 3.8001 | 3.6557 | 3.5052 | 3.4274 | 3.3476 | 3.2656 | 3.1813 | 3.0942 | 3.0040 |
| 15 | 3.8049 | 3.6662 | 3.5255 | 3.3719 | 3.2940 | 3.2141 | 3.1319 | 3.0471 | 2.9595 | 2.8684 |
| 16 | 3.6909 | 3.5527 | 3.4089 | 3.2588 | 3.1808 | 3.1007 | 3.0182 | 2.9330 | 2.8447 | 2.7528 |
| 17 | 3.5931 | 3.4552 | 3.3117 | 3.1615 | 3.0835 | 3.0032 | 2.9205 | 2.8348 | 2.7459 | 2.6530 |
| 18 | 3.5082 | 3.3706 | 3.2273 | 3.0771 | 2.9990 | 2.9185 | 3.8354 | 2.7493 | 2.6597 | 2.5660 |
| 19 | 3.4338 | 3.2965 | 3.1533 | 3.0031 | 2.9249 | 2.8442 | 2.7608 | 2.6742 | 2.5839 | 2.4893 |
| 20 | 3.3682 | 3.2311 | 3.0880 | 2.9377 | 2.8594 | 2.7785 | 2.6947 | 2.6077 | 2.5168 | 2.4212 |
| 21 | 3.3098 | 3.1729 | 3.0299 | 2.8796 | 2.8011 | 2.7200 | 2.6359 | 2.5484 | 2.4568 | 2.3603 |
| 22 | 3.2576 | 3.1209 | 2.9780 | 2.8274 | 2.7488 | 2.6675 | 2.5831 | 2.4951 | 2.4029 | 2.3055 |
| 23 | 3.2106 | 3.0740 | 2.9311 | 2.7805 | 2.7017 | 2.6202 | 2.5355 | 2.4471 | 2.3542 | 2.2559 |
| 24 | 3.1681 | 3.0316 | 2.8887 | 2.7380 | 2.6591 | 2.5773 | 2.4923 | 2.4035 | 2.3099 | 2.2107 |
| 25 | 3.1294 | 2.9331 | 2.8502 | 2.6993 | 2.6203 | 2.5383 | 2.4530 | 2.3667 | 2.2695 | 2.1694 |
| 26 | 3.0941 | 2.9576 | 2.8150 | 2.6640 | 2.5848 | 2.5026 | 2.4170 | 2.3273 | 2.2325 | 2.1315 |
| 27 | 3.0618 | 2.2956 | 2.7827 | 2.6316 | 2.5522 | 2.4699 | 2.3840 | 2.2938 | 2.1984 | 2.0965 |
| 28 | 3.0320 | 2.8959 | 2.7530 | 2.6017 | 2.5223 | 2.4397 | 2.3535 | 2.2629 | 2.1670 | 2.0642 |
| 29 | 3.0045 | 2.8685 | 2.7256 | 2.5742 | 2.4946 | 2.4118 | 2.3253 | 2.2344 | 2.1378 | 2.0342 |
| 30 | 2.9791 | 2.8431 | 2.7002 | 2.5487 | 2.4689 | 2.3680 | 2.2992 | 2.2079 | 2.1107 | 2.0062 |
| 40 | 2.8005 | 2.6648 | 2.5216 | 2.3689 | 2.2880 | 2.2034 | 2.1142 | 2.0194 | 1.9172 | 1.8047 |
| 60 | 2.6318 | 2.4961 | 2.3523 | 2.1978 | 2.1154 | 2.0285 | 1.9360 | 1.8363 | 1.7263 | 1.6006 |
| 120 | 2.4721 | 2.3363 | 2.1915 | 2.0346 | 1.9500 | 1.8600 | 1.7628 | 1.6557 | 1.5330 | 1.3805 |
| ∞ | 2.3209 | 2.1848 | 2.0385 | 1.8783 | 1.7908 | 1.6964 | 1.5923 | 1.4730 | 1.3246 | 1.0000 |

(Source: Reprinted, by permission of the Biometrika Trustees from M. Merrington and C. M. Thompson, "Tables of Percentage Points of the Inverted Beta(*F*) Distribution," *Biometrika* 33 (1943): 84–85.)

# Index