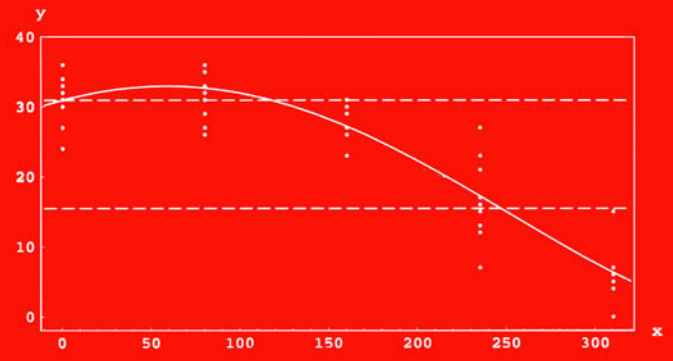
Mathematica Laboratories for Mathematical Statistics

Emphasizing Simulation and Computer Intensive Methods

Jenny A. Baglivo



ASA
SIAM

Mathematica Laboratories for Mathematical Statistics

ASA-SIAM Series on Statistics and Applied Probability

The ASA-SIAM Series on Statistics and Applied Probability is published jointly by the American Statistical Association and the Society for Industrial and Applied Mathematics. The series consists of a broad spectrum of books on topics in statistics and applied probability. The purpose of the series is to provide inexpensive, quality publications of interest to the intersecting membership of the two societies.

Editorial Board

Baglivo, J. A., *Mathematica Laboratories for Mathematical Statistics: Emphasizing Simulation and Computer Intensive Methods*

Lee, H. K. H., *Bayesian Nonparametrics via Neural Networks*

O'Gorman, T. W., *Applied Adaptive Statistical Methods: Tests of Significance and Confidence Intervals*

Ross, T. J., Booker, J. M., and Parkinson, W. J., eds., *Fuzzy Logic and Probability Applications: Bridging the Gap*

Nelson, W. B., *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*

Mason, R. L. and Young, J. C., *Multivariate Statistical Process Control with Industrial Applications*

Smith, P. L., *A Primer for Sampling Solids, Liquids, and Gases: Based on the Seven Sampling Errors of Pierre Gy*

Meyer, M. A. and Booker, J. M., *Eliciting and Analyzing Expert Judgment: A Practical Guide*

Latouche, G. and Ramaswami, V., *Introduction to Matrix Analytic Methods in Stochastic Modeling*

Peck, R., Haugh, L., and Goodman, A., *Statistical Case Studies: A Collaboration Between Academe and Industry, Student Edition*

Peck, R., Haugh, L., and Goodman, A., *Statistical Case Studies: A Collaboration Between Academe and Industry*

Barlow, R., *Engineering Reliability*

Czitrom, V. and Spagon, P. D., *Statistical Case Studies for Industrial Process Improvement*

Mathematica Laboratories for Mathematical Statistics

Emphasizing Simulation and Computer Intensive Methods

Jenny A. Baglivo
Boston College
Chestnut Hill, Massachusetts

siam
Society for Industrial and Applied Mathematics
Philadelphia, Pennsylvania

ASA
American Statistical Association
Alexandria, Virginia

The correct bibliographic citation for this book is as follows: Baglivo, Jenny A., *Mathematica Laboratories for Mathematical Statistics: Emphasizing Simulation and Computer Intensive Methods,* ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2005.

10 9 8 7 6 5 4 3 2 1

This work was supported by Boston College through its faculty research programs and by the National Science Foundation through its Division of Undergraduate Education (NSF DUE 9555178).

Library of Congress Cataloging-in-Publication Data

Baglivo, Jenny A. (Jenny Antoinette)
Mathematica laboratories for mathematical statistics : emphasizing simulation and computer intensive methods / Jenny A. Baglivo.
p. cm. — (ASA-SIAM series on statistics and applied probability)
Includes bibliographical references and index.
ISBN 0-89871-566-0 (pbk.)
1. Mathematical statistics—Computer simulation. 2. Mathematica (Computer file)
I. Title. II. Series.

QA276.4.B34 2005
519.5'01'13—dc22 2004056579

Dedicated to the memory of

Anthony and Philomena Baglivo

Marion Ferri

This page intentionally left blank

Contents

This page intentionally left blank

Preface

There is no doubt that the computer has revolutionized the practice of statistics in recent years. Computers allow us to analyze data more quickly using classical techniques, to analyze much larger data sets, to replace classical data analytic methods—whose assumptions may not be met—with more flexible computer intensive approaches, and to solve problems with no satisfactory classical solution.

Nor is there doubt that undergraduate mathematics and statistics courses could benefit from the integration of computer technology. Computer laboratories can be used to illustrate and reinforce important concepts; allow students to simulate experiments and visualize their results; and allow them to compare the results of classical methods of data analysis with those using alternative techniques. The problem is how best to introduce these techniques in the curriculum.

This book introduces an approach to incorporating technology in the mathematical statistics sequence, with an emphasis on simulation and computer intensive methods. The printed book is a concise introduction to the concepts of probability theory and mathematical statistics. The accompanying electronic materials are a series of in-class and take-home computer laboratory problems designed to reinforce the concepts and to apply the techniques in real and realistic settings.

The laboratory materials are written as *Mathematica* Version 5 notebooks [112] and are designed so that students with little or no experience in *Mathematica* will be able to complete the work. *Mathematica* notebooks contain text, data, computations, and graphics; they are particularly well suited for presenting concepts and problems and for writing solutions.

Laboratory problems, custom tools designed to enhance the capabilities of *Mathematica*, an introduction to using *Mathematica* for probability and statistics, and additional materials are included in an accompanying CD. An instructor's CD is available to those who adopt the book. The instructor's CD contains complete solutions to all laboratory problems, instructor guides, and hints on developing additional tools and laboratory problems.

The materials are written to be used in the mathematical statistics sequence given at most colleges and universities (two courses of four semester hours each or three courses of three semester hours each). Multivariable calculus and familiarity with the basics of set theory, vectors and matrices, and problem solving using a computer are assumed. The order of topics generally follows that of a standard sequence. Chapters 1 through 5 cover concepts in probability. Chapters 6 through 10 cover introductory mathematical statistics. Chapters 11 and 12 are on permutation

and bootstrap methods. In each case, problems are designed to expand on ideas from previous chapters so that instructors could choose to use some of the problems earlier in the course. Permutation and bootstrap methods also appear in the later chapters. Chapters 13, 14, and 15 are on multiple sample analysis, linear least squares, and analysis of contingency tables, respectively. References for specialized topics in Chapters 10 through 15 are given at the beginning of each chapter.

The materials can also be used profitably by statistical practitioners or consultants interested in a computer-based introduction to mathematical statistics, especially to computer intensive methods.

Laboratory problems

Each chapter has a main laboratory notebook, containing between five and seven problems, and a series of additional problem notebooks. The problems in the main laboratory notebook are for basic understanding and can be used for in-class work or assigned for homework. The additional problem notebooks reinforce and/or expand the ideas from the main laboratory notebook and are generally longer and more involved.

There are a total of 238 laboratory problems. Each main laboratory notebook and many of the problem notebooks contain examples for students to work before starting the assigned problems. One hundred twenty-three examples and problems use simulation, permutation, and bootstrap methods. One hundred twenty-five problems use real data.

Many problems are based on recent research reports or ongoing research—for example, analyses of the spread of an infectious disease in the cultured oyster population in the northeastern United States [18], [42], [100]; analyses of the ecological effects of the introduction of the Asian shore crab to the eastern United States [19], [20]; comparison of modeling strategies for occurrences of earthquakes in southern California [35]; comparison of spatial distributions of earthquakes [60] and of animal species [105]; comparison of treatments for multiple sclerosis [63], [8]; and analyses of associations between cellular telephone use and car accidents [88], between genetics and longevity [114], and between incidence of childhood leukemia and distance to a hazardous waste site [111]. Whimsical examples include comparisons of world-class sprinters [108] and of winning baseball players and teams [98].

Note to the student

Concepts from probability and statistics are used routinely in fields as diverse as actuarial science, ecology, economics, engineering, genetics, health sciences, marketing, and quality management. The ideas discussed in each chapter of the text will give you a basic understanding of the important concepts. The last section in each chapter outlines the laboratory problems.

Although formal proofs are not emphasized, the logical progression of the ideas in a proof is given whenever possible. Comments, including reminders about topics from calculus and pointers to where concepts will be applied, are enclosed in boxes throughout the text.

The accompanying CD contains two folders:

1. The `PDFFiles` folder contains documents in Acrobat PDF format. You will need a current copy of Adobe Acrobat Reader to open and print these files. Adobe Acrobat Reader is available for free from `adobe.com`.

2. The `MMAFiles` folder contains *Mathematica* files. You will need a copy of *Mathematica* Version 5 to work with these files.

The `PDFFiles` folder includes two appendices to the printed text and 15 laboratory workbooks. Appendix A is an introduction to the *Mathematica* commands used in the laboratory problems. Print Appendix A and keep it for reference. Appendix B contains tables of probabilities and quantiles suitable for solving problems when you are not using the computer. Print Appendix B and keep it for reference. There is one laboratory workbook for each chapter of the text. Print the ones you need for your course.

The `MMAFiles` folder includes 15 folders of laboratory problems and a folder of customized tools (`StatTools`). The `StatTools` folder should be placed in the user base directory or other appropriate directory on your system. Consult the online help within the *Mathematica* system for details, or speak to your instructor.

Note to the instructor

The material in the text is sufficient to support a problem-oriented mathematical statistics sequence, where the computer is used throughout the sequence. In fact, the first lab can be scheduled after three or four class meetings. Students are introduced to parametric, nonparametric, permutation, and bootstrap methods and will learn about data analysis, including diagnostic methods. (See the chapter outlines below.)

The text does not include exercises intended to be done by hand. You will need to supplement the text with by-hand exercises from other books or with ones that you design yourself. Suggestions for by-hand exercises that complement certain laboratory problems are given in the instructor's CD.

In addition, the printed text does not include *Mathematica* commands. Step-by-step instructions for using *Mathematica* commands are given in examples in the electronic materials. Online help is available, and Appendix A on the CD can be used as a reference.

Chapter outlines

Chapter 1 covers counting methods, axioms of probability, conditional probability, and independence. The first laboratory session is intended to be scheduled early in the term, as soon as the counting methods, axioms, and first examples are discussed. Students become familiar with using *Mathematica* commands to compute and graph binomial coefficients and hypergeometric probabilities (called "urn probabilities" in the lab) and get an informal introduction to maximum likelihood and likelihood ratio methods using custom tools. The additional problem notebooks reinforce these ideas

and include problems on frequency generating functions, conditional probability, and independence.

Chapters 2 and 3 are on discrete and continuous families of probability distributions, respectively. In the laboratory sessions, students become familiar with using *Mathematica* commands for computing probabilities and pseudorandom samples from univariate distributions, and with using custom tools for graphing models and samples. The additional problem notebooks reinforce these ideas, give students an informal introduction to goodness-of-fit, and include problems on probability generating functions, bivariate distributions, and transformations.

Chapter 4 is on mathematical expectation. In the laboratory and additional problem notebooks, students work with *Mathematica* commands for model and sample summaries, use sample summaries to estimate unknown parameters, apply the Chebyshev and Markov inequalities, and work with conditional expectations.

Chapter 5 is on limit theorems. In the laboratory session, students use custom tools to study sequences of running sums and averages, and answer a variety of questions on exact and approximate distributions of sums. The additional problem notebooks reinforce and expand on these ideas, and include several problems on probability and moment generating functions.

Chapter 6 serves as a transition from probability to statistics. The chi-square, Student t, and f ratio distributions are defined, and several applications are introduced, including the relationship of the chi-square distribution to the sampling distribution of the sample variance of a random sample from a normal distribution and the application of the chi-square distribution to the multinomial goodness-of-fit problem. In the laboratory session, students become familiar with chi-square and multinomial distributions, and use a custom tool for carrying out a goodness-of-fit analysis using Pearson's test (including analysis of standardized residuals). The additional problem notebooks contain simulation studies and applications of Pearson's goodness-of-fit test, and introduce students to minimum chi-square and method of moments estimates. The chapter is intended to precede formal statistical inference.

Chapters 7 and 8 are on estimation theory and hypothesis testing theory, respectively. In the first laboratory session, students become familiar with *Mathematica* commands for constructing confidence intervals for normal means and variances, and use custom tools to study the concepts of confidence interval and maximum likelihood estimation. In the second laboratory session, students become familiar with *Mathematica* commands for carrying out tests for normal means and variances, construct power curves, use a custom tool to construct tests and compute power at fixed alternatives, and compute sample sizes. The additional problem notebooks reinforce and expand on these ideas, contain simulation studies, introduce the idea of inverting tests to produce confidence intervals, and include applications of the likelihood ratio goodness-of-fit test.

Chapter 9 is on order statistics and quantiles. In the laboratory session, students apply custom tools for visualizing order-statistic distributions, for quantile estimation, and for constructing box plots in a variety of problems. The additional problem notebooks reinforce and expand on these ideas, introduce probability plots, study order statistics for uniform models, and contain simulation studies.

Chapter 10 is on parametric and nonparametric two sample analysis. In the laboratory session, students apply *Mathematica* commands for analyzing independent random samples from normal distributions and custom tools for the Wilcoxon rank sum test in a variety of problems. Normal probability plots of standardized observations are used to determine whether parametric methods should be used. The additional problem notebooks reinforce and expand on these ideas, contain simulation studies, introduce custom tools for quantile-quantile plots and inverting the Wilcoxon rank sum test under the shift model, and consider the randomization model for two sample analysis.

Chapter 11 is an introduction to permutation analysis, using nonparametric analyses of two samples and paired samples as first examples. In the laboratory session, students apply the rank sum, Smirnov, correlation, and signed rank tests in a variety of problems. The additional problem notebooks introduce a variety of different applications of permutation methods (using a variety of different test statistics) and use frequency generating functions to construct certain permutation distributions. Custom tools are used throughout, including tools for signed rank analyses, for constructing random reorderings of data, and for visualizing random reorderings of data.

Chapter 12 is an introduction to parametric and nonparametric bootstrap analysis. In the laboratory and additional problem notebooks, students consider the performance of the bootstrap and apply bootstrap estimation and testing methods in a variety of situations. Custom tools are used to construct random resamples, to visualize random resamples, to summarize the results of bootstrap analyses, and to construct approximate bootstrap confidence intervals using Efron's BC_a method in the nonparametric setting.

Chapter 13 is on parametric, nonparametric, and permutation methods for analysis of multiple samples. In the laboratory session, students use simulation to study analysis of variance for one-way layouts and blocked designs and to study Kruskal–Wallis and Friedman tests and apply these techniques in a variety of situations. Normal probability plots of standardized residuals are used to check analysis of variance assumptions. The additional problem notebooks reinforce these ideas and contain simulation studies and problems on analysis of variance in the balanced two-way layout setting. Custom tools are used throughout, including tools for analysis of variance, Bonferroni analysis, and Kruskal–Wallis and Friedman tests.

Chapter 14 is on linear least squares, including simple and multiple linear regression, permutation and bootstrap methods, and regression diagnostics. In the laboratory session, students use simulation to study the components of a linear regression analysis and apply the techniques in a variety of situations. The additional problem notebooks reinforce these ideas and contain problems on goodness-of-fit for simple linear models, analysis of covariance, model building, and locally weighted regression. Custom tools are provided for permutation analysis of slope in the simple linear setting, locally weighted regression, and diagnostic plots.

Chapter 15 is on large sample and small sample analyses of contingency tables, including diagnostic methods. In the laboratory session, students apply custom tools for large sample analyses of I-by-J tables and for constructing large sample confidence intervals for odds ratios to data from four studies. The additional problem

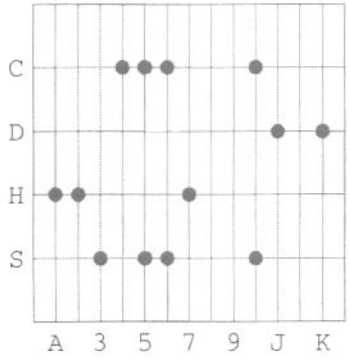
notebooks reinforce these ideas, consider the relationship between odds ratios and risk ratios, introduce McNemar's test for paired samples, and contain problems on permutation methods for fourfold and I-by-J tables.

Acknowledgments

This work was supported by Boston College through its faculty research programs and by the National Science Foundation through its Division of Undergraduate Education (NSF DUE 9555178). Boston College provided generous released time over several years while the materials were in development. NSF provided summer support for me, stipends for six additional faculty members, and generous support for an assistant.

Boston College Professors Dan Chambers and Charlie Landraitis, College of St. Catherine Professor Adele Rothan, C.S.J., and Stetson University Professor Erich Freedman used earlier versions of the laboratory materials in their classes and provided helpful comments and suggestions. Mt. Holyoke College Professor George Cobb and Harvard University Professor Marcello Pagano provided guidance on project design. I consulted with Boston College Professor Peg Kenney on assessment issues. University of Ottawa Professor John Nash and Boston College Professor Rob Gross provided interesting problem ideas and expert LaTeX advice. Ms. Sarah Quebec worked with me as an undergraduate and masters student at Boston College and then as a research assistant on this project; her thoughtful comments helped shape the final product. The comments provided by students in my classes were uniformly helpful in improving the laboratory materials. I extend a warm thank you to SIAM's editorial team, especially Linda Thiel, and to the reviewers of the text and laboratory materials.

Data sets were kindly provided by Fairfield University Biology Professor Diane Brousseau, Boston College Geophysics Professors John Ebel and Alan Kafka, Dr. Susan Ford (Haskin Shellfish Laboratory, Rutgers University), and Dr. Roxana Smolowitz (Marine Biological Laboratories, University of Pennsylvania).



Chapter 1
Introductory Probability Concepts

Probability is the study of random phenomena. Probability theory can be applied, for example, to study games of chance (e.g., roulette games, card games), occurrences of catastrophic events (e.g., tornados, earthquakes), survival of animal species, and changes in stock and commodity markets.

This chapter introduces probability theory. The first three sections are concerned with the definitions, axioms, and properties of probability and with counting methods used to compute probabilities. The concepts of conditional probability and independence are introduced in Sections 4 and 5, respectively. Section 6 outlines the laboratory problems for this chapter.

1.1 Definitions

The term *experiment* (or *random experiment*) is used in probability theory to describe a procedure whose outcome is not known in advance with certainty. Further, experiments are assumed to be repeatable (at least in theory) and to have a well-defined set of possible outcomes.

The *sample space* $\mathcal{S}$ is the set of all possible outcomes of an experiment. An *event* is a subset of the sample space. A *simple event* is an event with a single outcome. Events are usually denoted by capital letters (A, B, C, ...) and outcomes by lowercase letters (x, y, z, ...). If $x \in A$ is observed, then A is said to have *occurred*. The favorable outcomes of an experiment form the event of interest.

Each repetition of an experiment is called a *trial*. *Repeated trials* are repetitions of the experiment using the specified procedure, with the outcomes of the trials having no influence on one another.

Example: Coin-tossing experiment

For example, suppose you toss a fair coin 5 times and record h (for heads) or t (for tails) each time. The sample space for this experiment is the collection of $32 = 2^5$

sequences of 5 h's or t's:

$$\mathcal{S} = \{hhhhh, hhhht, hhhth, hhthh, hthhh, thhhh, hhhtt, hhtht, hthht, thhht, hhtth, hthth, thhth, htthh, ththh, tthhh, ttthh, tthth, thtth, httth, tthht, ththt, httht, thhtt, hthtt, hhttt, htttt, thttt, tthtt, tttht, tttth, ttttt\}.$$

If you are interested in getting exactly 5 heads, then the event of interest is the simple event $A = \{hhhhh\}$. If you are interested in getting exactly 3 heads, then the event of interest is

$$A = \{hhhtt, hhtht, hthht, thhht, hhtth, hthth, thhth, htthh, ththh, tthhh\}.$$

1.2 Kolmogorov axioms

The basic rules (or axioms) of probability were introduced by A. Kolmogorov in the 1930's. Let $A \subseteq \mathcal{S}$ be an event, and let $P(A)$ be the probability that A will occur.

A *probability distribution*, or simply a *probability*, on a sample space $\mathcal{S}$ is a specification of numbers $P(A)$ satisfying the following axioms:

1. $P(\mathcal{S}) = 1$.
2. If A is an event, then $0 \leq P(A) \leq 1$.
3. If A_1 and A_2 are *disjoint events* (that is, if $A_1 \cap A_2 = \emptyset$), then
$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$
3′. More generally, if $A_1, A_2, \ldots$ are *pairwise disjoint events* (that is, if $A_i \cap A_j = \emptyset$ when $i \neq j$), then
$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots.$$
If the sequence of events is infinite, then the right-hand side is understood to be the sum of a convergent infinite series.

Since $\mathcal{S}$ is the set of all possible outcomes, an outcome in $\mathcal{S}$ is certain to occur; the probability of an event that is certain to occur must be 1 (axiom 1). Probabilities must be between 0 and 1 (axiom 2), and probabilities must be additive when events are pairwise disjoint (axiom 3).

Relative frequencies

The probability of an event can be written as the limit of relative frequencies. That is, if $A \subseteq \mathcal{S}$ is an event, then

$$P(A) = \lim_{n \to \infty} \frac{\#(A)}{n},$$

where $\#(A)$ is the number of occurrences of event A in n repeated trials of the experiment. If $P(A)$ is the probability of event A, then $nP(A)$ is the *expected number* of occurrences of event A in n repeated trials of the experiment.

Example: Equally likely outcomes

If $\mathcal{S}$ is a finite set with N elements, A is a subset of $\mathcal{S}$ with n elements, and each outcome is equally likely, then

$$P(A) = \frac{\text{Number of elements in } A}{\text{Number of elements in } \mathcal{S}} = \frac{|A|}{|\mathcal{S}|} = \frac{n}{N}.$$

For example, if you toss a fair coin 5 times and record heads or tails each time, then the probability of getting exactly 3 heads is $10/32 = 0.3125$. Further, in 2000 repetitions of the experiment you expect to observe exactly 3 heads:

$$2000P(\text{exactly 3 heads}) = 2000(0.3125) = 625 \text{ times}.$$

Example: Geometric sequences and series

Geometric sequences and series are used often in probability. A typical setup is as follows: the sample space $\mathcal{S}$ is a countably infinite set of outcomes,

$$\mathcal{S} = \{x_o,\ x_1,\ x_2, \ldots, x_n, \ldots\},$$

and the probabilities of the simple events form a geometric sequence,

$$P(\{x_n\}) = (1-p)^n p, \qquad n = 0, 1, \ldots,$$

where p is a proportion ($0 < p < 1$). The sum of the sequence is 1.

For example, if you toss a fair coin until you get tails and record the sequence of h's and t's, then the sample space is

$$\mathcal{S} = \{t,\ ht,\ hht,\ hhht,\ hhhht, \ldots\}.$$

The probabilities of the simple events form a geometric sequence with $p = 1/2$. Further, the probability that tails is observed in three or fewer tosses is

$$P(\{t, ht, hht\}) = P(\{t\}) + P(\{ht\}) + P(\{hht\}) = \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} = \tfrac{7}{8} = 0.875.$$

Recall that the n^{th} partial sum of the geometric sequence $a, ar, ar^2, \ldots$ is

$$S_n = a + ar + ar^2 + \cdots + ar^{n-1} = \frac{a(1-r^n)}{(1-r)} \quad \text{when } r \neq 1$$

and that its sum is

$$a + ar + ar^2 + \cdots = \lim_{n\to\infty} S_n = \frac{a}{(1-r)} \quad \text{when } |r| < 1.$$

In the application above, $a = p$, $r = 1 - p$, and the sum is 1.

Properties following from the axioms

Properties following from the Kolmogorov axioms include the following:

1. *Complement rule.* Let $A^c = \mathcal{S} - A$ be the *complement* of A in $\mathcal{S}$. Then

$$P(A^c) = 1 - P(A).$$

   In particular, $P(\emptyset) = 0$.

2. *Subset rule.* If A is a subset of B, then $P(A) \leq P(B)$.

3. *Inclusion-exclusion rule.* If A and B are events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

To demonstrate the complement rule, note that the sample space $\mathcal{S}$ can be written as the disjoint union of A and A^c:

$$\mathcal{S} = A \cup A^c, \text{ where } A \cap A^c = \emptyset.$$

Thus, $P(\mathcal{S}) = P(A) + P(A^c)$ by axiom 3. Since $P(\mathcal{S}) = 1$ by axiom 1, the additive rule then implies that $P(A^c) = 1 - P(A)$.

To demonstrate the subset rule, note that event B can be written as the disjoint union of A and $B \cap A^c$:

$$B = A \cup (B \cap A^c), \text{ where } A \cap (B \cap A^c) = \emptyset.$$

Thus, $P(B) = P(A) + P(B \cap A^c)$ by axiom 3. Since $P(B \cap A^c) \geq 0$ by axiom 2, the additive rule then implies that $P(A) \leq P(B)$.

To demonstrate the inclusion-exclusion rule, note that event $A \cup B$ can be written as the disjoint union of A and $B \cap A^c$,

$$A \cup B = A \cup (B \cap A^c), \text{ where } A \cap (B \cap A^c) = \emptyset,$$

and that event B can be written as the disjoint union of $B \cap A$ and $B \cap A^c$,

$$B = (B \cap A) \cup (B \cap A^c), \text{ where } (B \cap A) \cap (B \cap A^c) = \emptyset.$$

Axiom 3 applied twice implies the inclusion-exclusion rule:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B \cap A^c) \\ &= P(A) + \Big(P(B) - P(B \cap A)\Big) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$

1.3 Counting methods

Methods for counting the number of outcomes in a sample space or event are important in probability. The multiplication rule is the basic counting formula.

Theorem 1.1 (Multiplication Rule). *If an operation consists of r steps of which the first can be done in n_1 ways, for each of these the second can be done in n_2 ways, for each of the first and second steps the third can be done in n_3 ways, etc., then the entire operation can be done in $n_1 \times n_2 \times \cdots \times n_r$ ways.*

Two special cases of the multiplication rule are as follows:

1. *Sampling with replacement.* For a set of size n and a sample of size r, there are a total of $n^r = n \times n \times \cdots \times n$ ordered samples, if duplication is allowed.
2. *Sampling without replacement.* For a set of size n and a sample of size r, there are a total of

$$\frac{n!}{(n-r)!} = n \times (n-1) \times \cdots \times (n-r+1)$$

ordered samples, if duplication is not allowed.

If n is a positive integer, the notation $n!$ ("n factorial") is used for the product

$$n! = n \times (n-1) \times \cdots \times 1.$$

For convenience, 0! is defined to equal 1 ($0! = 1$).

Example: Birthday problem

For example, suppose there are r unrelated people in a room, none of whom was born on February 29 of a leap year. You would like to determine the probability that at least two people have the same birthday.

(i) You ask for, and record, each person's birthday. There are

$$365^r = 365 \times 365 \times \cdots \times 365$$

possible outcomes, where an outcome is a sequences of r responses.

(ii) Consider the event "everyone has a different birthday." The number of outcomes in this event is

$$\frac{365!}{(365-r)!} = 365 \times 364 \times \cdots \times (365-r+1).$$

(iii) Suppose that each sequence of birthdays is equally likely. The probability that at least two people have a common birthday is 1 minus the probability that everyone has a different birthday, or

$$1 - \frac{365 \times 364 \times \cdots \times (365-r+1)}{365^r}.$$

In particular, if $r = 25$ and A is the event "at least two people have a common birthday," then $P(A) \approx 0.57$.

1.3.1 Permutations and combinations

A *permutation* is an ordered subset of r distinct objects out of a set of n objects. A *combination* is an unordered subset of r distinct objects out of the n objects. By the multiplication rule (Theorem 1.1), there are a total of

$$ {}_nP_r = \frac{n!}{(n-r)!} = n \times (n-1) \times \cdots \times (n-r+1) $$

permutations of r objects out of n objects. Since each unordered subset corresponds to $r!$ ordered subsets (the r chosen elements are permuted in all possible ways), there are a total of

$$ {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{(n-r)!\,r!} = \frac{n \times (n-1) \times \cdots \times (n-r+1)}{r \times (r-1) \times \cdots \times 1} $$

combinations of r objects out of n objects.

For example, there are a total of 5040 ordered subsets of size 4 from a set of size 10 and a total of $5040/24 = 210$ unordered subsets.

The notation $\binom{n}{r}$ (read "n choose r") is used to denote the total number of combinations. Special cases are as follows:

$$ \binom{n}{0} = \binom{n}{n} = 1 \quad \text{and} \quad \binom{n}{1} = \binom{n}{n-1} = n. $$

Further, since choosing r elements to form a subset is equivalent to choosing the remaining $n - r$ elements to form the complementary subset,

$$ \binom{n}{r} = \binom{n}{n-r} \quad \text{for } r = 0, 1, \ldots, n. $$

Example: Simple urn model

Suppose there are M special objects in an urn containing a total of N objects. In a subset of size n chosen from the urn, exactly m are special.

(i) *Unordered subsets.* There are a total of

$$ \binom{M}{m} \times \binom{N-M}{n-m} $$

unordered subsets with exactly m special objects (and exactly $n - m$ other objects). If each choice of subset is equally likely, then for each m

$$ P(m \text{ special objects}) = \frac{\binom{M}{m} \times \binom{N-M}{n-m}}{\binom{N}{n}}. $$

(ii) *Ordered subsets.* There are a total of

$$\binom{n}{m} \times {}_M P_m \times {}_{N-M} P_{n-m}$$

ordered subsets with exactly m special objects. (The positions of the special objects are selected first, followed by the special objects to fill these positions, followed by the nonspecial objects to fill the remaining positions.) If each choice of subset is equally likely, then for each m

$$P(m \text{ special objects}) = \frac{\binom{n}{m} \times {}_M P_m \times {}_{N-M} P_{n-m}}{{}_N P_n}.$$

Interestingly, $P(m$ special objects) is the same in both cases. For example, let $N = 25$, $M = 10$, $n = 8$, and $m = 3$. Then, using the first formula,

$$P(3 \text{ special objects}) = \frac{\binom{10}{3} \times \binom{15}{5}}{\binom{25}{8}} = \frac{120 \times 3003}{1081575} = \frac{728}{2185} \approx 0.333.$$

Using the second formula, the probability is

$$P(3 \text{ special objects}) = \frac{\binom{8}{3} \times {}_{10}P_3 \times {}_{15}P_5}{{}_{25}P_8} = \frac{56 \times 720 \times 360360}{43609104000} = \frac{728}{2185} \approx 0.333.$$

Binomial coefficients

The quantities $\binom{n}{r}$, $r = 0, 1, \ldots, n$, are often referred to as the *binomial coefficients* because of the following theorem.

Theorem 1.2 (Binomial Theorem). *For all numbers x and y and each positive integer n,*

$$(x + y)^n = \sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r}.$$

The idea of the proof is as follows. The product on the left can be written as a sequence of n factors:

$$(x + y)^n = (x + y) \times (x + y) \times \cdots \times (x + y).$$

The product expands to 2^n summands, where each summand is a sequence of n letters (one from each factor). For each r, exactly $\binom{n}{r}$ sequences have r copies of x and $n - r$ copies of y.

1.3.2 Partitioning sets

The multiplication rule (Theorem 1.1) can be used to find the number of *partitions* of a set of n elements into k distinguishable subsets of sizes $r_1, r_2, \ldots, r_k$.

Specifically, r_1 of the n elements are chosen for the first subset, r_2 of the remaining $n - r_1$ elements are chosen for the second subset, etc. The result is the product of the numbers of ways to perform each step:

$$\binom{n}{r_1} \times \binom{n-r_1}{r_2} \times \cdots \times \binom{n-r_1-\cdots-r_{k-1}}{r_k}.$$

The product simplifies to

$$\frac{n!}{r_1!\, r_2! \cdots r_k!} \quad \text{and is denoted by} \quad \binom{n}{r_1, r_2, \ldots, r_k}$$

(read "n choose $r_1, r_2, \ldots, r_k$"). For example, there are a total of

$$\binom{15}{5, 5, 5} = \frac{15!}{5!\, 5!\, 5!} = 756,756$$

ways to partition the members of a class of 15 students into recitation sections of size 5 each led by Joe, Sally, and Mary, respectively. (The recitation sections are distinguished by their group leaders.)

Permutations of indistinguishable objects

The formula above also represents the number of ways to permute n objects, where the first r_1 are indistinguishable, the next r_2 are indistinguishable, ..., the last r_k are indistinguishable. The computation is done as follows: r_1 of the n positions are chosen for the first type of object, r_2 of the remaining $n - r_1$ positions are chosen for the second type of object, etc.

Multinomial coefficients

The quantities $\binom{n}{r_1, r_2, \ldots, r_k}$ are often referred to as the *multinomial coefficients* because of the following theorem.

Theorem 1.3 (Multinomial Theorem). *For all numbers $x_1, x_2, \ldots, x_k$ and each positive integer n,*

$$(x_1 + x_2 + \cdots + x_k)^n = \sum_{(r_1, r_2, \ldots, r_k)} \binom{n}{r_1, r_2, \ldots, r_k} x_1^{r_1} x_2^{r_2} \cdots x_k^{r_k},$$

where the sum is over all k-tuples of nonnegative integers with $\sum_i r_i = n$.

The idea of the proof is as follows. The product on the left can be written as a sequence of n factors,

$$(x_1 + x_2 + \cdots + x_k)^n = (x_1 + x_2 + \cdots + x_k)$$

$$\times (x_1 + x_2 + \cdots + x_k) \times \cdots \times (x_1 + x_2 + \cdots + x_k).$$

The product expands to k^n summands, where each summand is a sequence of n letters (one from each factor). For each $r_1, r_2, \ldots, r_k$, exactly $\binom{n}{r_1, r_2, \ldots, r_k}$ sequences have r_1 copies of x_1, r_2 copies of x_2, etc.

1.3.3 Generating functions

The *generating function* of the sequence $a_0, a_1, a_2, \ldots$ is the formal power series whose coefficients are the given sequence:

$$\mathrm{GF}(t) = a_0 + a_1 t + a_2 t^2 + \cdots = \sum_{i=0}^{\infty} a_i t^i.$$

If $a_n \neq 0$ and $a_i = 0$ for $i > n$ for some n, then the generating function reduces to a polynomial of degree n. For example, the generating function of the sequence of binomial coefficients is the polynomial

$$\mathrm{GF}(t) = \sum_{i=0}^{n} \binom{n}{i} t^i = (t+1)^n.$$

The following important property of generating functions can be proven using series (or polynomial) multiplication.

Theorem 1.4 (Convolution Theorem). *If* $\mathrm{GF}_1(t)$ *is the generating function of the sequence* $a_0, a_1, a_2, \ldots$, *and* $\mathrm{GF}_2(t)$ *is the generating function of the sequence* $b_0, b_1, b_2, \ldots$, *then* $\mathrm{GF}_1(t)\mathrm{GF}_2(t)$ *is the generating function of the sequence whose* k^{th} *term is*

$$c_k = a_0 b_k + a_1 b_{k-1} + a_2 b_{k-2} + \cdots + a_k b_0.$$

The convolution theorem can be applied to counting problems. For example, suppose an urn contains 10 slips of paper—four slips with the number 1 written on each, five slips with the number 2, and one slip with the number 3. The urn is sampled with replacement twice; the ordered pair of numbers and their sum are recorded. Among the 100 ordered pairs, the frequency with which the sum of k appears is the coefficient of t^k in the following polynomial expansion:

$$\left(4t + 5t^2 + t^3\right)^2 = 16t^2 + 40t^3 + 33t^4 + 10t^5 + t^6.$$

For example, a sum of 4 can be obtained in 33 ways: 25 ways from ordered pairs of the form (2,2), 4 ways from ordered pairs of the form (1,3), and 4 ways from ordered pairs of the form (3,1).

The polynomial above is called the *frequency generating function* (FGF) of the sequence of sums. More generally, $\left(4t + 5t^2 + t^3\right)^r$ is the FGF of the sequence of sums when the urn is sampled with replacement r times. That is, the coefficient of t^k is the number of times a sum of k appears among the 10^r ordered sequences.

1.4 Conditional probability

Assume that A and B are events and that $P(B) > 0$. Then the *conditional probability* of A given B is defined as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Event B is often referred to as the *conditional sample space*. $P(A|B)$ is the relative "size" of A within B.

For example, suppose that 40% of the adults in a certain population smoke cigarettes and that 28% smoke cigarettes and have respiratory problems. Then

$$\text{P(Respiratory problems | Smoker)} = \frac{0.28}{0.40} = 0.70$$

is the probability that an adult has respiratory problems given that the adult is a smoker (70% of smokers have respiratory problems).

Note that if the sample space $\mathcal{S}$ is finite and each outcome is equally likely, then the conditional probability of A given B simplifies to the following:

$$P(A|B) = \frac{|A \cap B|/|\mathcal{S}|}{|B|/|\mathcal{S}|} = \frac{|A \cap B|}{|B|} = \frac{\text{Number of elements in } A \cap B}{\text{Number of elements in } B}.$$

Multiplication rule for probability

Assume that A and B are events with positive probability. Then the definition of conditional probability implies that the probability of the intersection, $A \cap B$, can be written as a product of probabilities in two different ways:

$$P(A \cap B) = P(B) \times P(A|B) = P(A) \times P(B|A).$$

More generally, the following theorem holds.

Theorem 1.5 (Multiplication Rule for Probability). *If $A_1, A_2, \ldots, A_k$ are events and $P(A_1 \cap A_2 \cap \cdots A_{k-1}) > 0$, then*

$$\begin{aligned} &P(A_1 \cap A_2 \cap \cdots \cap A_k) \\ &\quad = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times \cdots \times P(A_k|A_1 \cap A_2 \cap \cdots \cap A_{k-1}). \end{aligned}$$

For example, suppose that 4 slips of paper are sampled without replacement from a well-mixed urn containing 25 slips of paper: 15 slips with the letter X written on each and 10 slips of paper with the letter Y written on each. Then the probability of observing the sequence $XYXX$ is $P(XYXX) =$

$$P(X) \times P(Y|X) \times P(X|XY) \times P(X|XYX) = \frac{15}{25} \times \frac{10}{24} \times \frac{14}{23} \times \frac{13}{22} \approx 0.09.$$

(The probability of choosing an X slip is 15/25; with an X removed from the urn, the probability of drawing a Y slip is 10/24; with an X and Y removed from the urn, the probability of drawing an X slip is 14/23; with two X slips and one Y slip removed from the urn, the probability of drawing an X slip is 13/22.)

1.4.1 Law of total probability

The law of total probability can be used to write an unconditional probability as the weighted average of conditional probabilities. Specifically, the following theorem holds.

Theorem 1.6 (Law of Total Probability). *Let $A_1, A_2, \ldots, A_k$ and B be events with nonzero probability. If $A_1, A_2, \ldots, A_k$ are pairwise disjoint with union $\mathcal{S}$, then*

$$P(B) = P(A_1) \times P(B|A_1) + P(A_2) \times P(B|A_2) + \cdots + P(A_k) \times P(B|A_k).$$

To demonstrate the law of total probability, note that if $A_1, A_2, \ldots, A_k$ are pairwise disjoint with union $\mathcal{S}$, then the sets $B \cap A_1, B \cap A_2, \ldots, B \cap A_k$ are pairwise disjoint with union B. Thus, axiom 3 and the definition of conditional probability imply that

$$P(B) = \sum_{j=1}^{k} P(B \cap A_j) = \sum_{j=1}^{k} P(A_j)P(B|A_j).$$

For example, suppose that 70% of smokers and 15% of nonsmokers in a certain population of adults have respiratory problems. If 40% of the population smoke cigarettes, then

$$\text{P(Respiratory problems)} = 0.40(0.70) + 0.60(0.15) = 0.37$$

is the probability of having respiratory problems.

Law of average conditional probabilities

The law of total probability is often called the *law of average conditional probabilities*. Specifically, $P(B)$ is the weighted average of the collection of conditional probabilities $\{P(B|A_j)\}$, using the collection of unconditional probabilities $\{P(A_j)\}$ as weights.

In the respiratory problems example above, 0.37 is the weighted average of 0.70 (the probability that a smoker has respiratory problems) and 0.15 (the probability that a nonsmoker has respiratory problems).

1.4.2 Bayes rule

Bayes rule, proven by the Reverend T. Bayes in the 1760's, can be used to update probabilities given that an event has occurred. Specifically, the following theorem holds.

Theorem 1.7 (Bayes Rule). *Let $A_1, A_2, \ldots, A_k$ and B be events with nonzero probability. If $A_1, A_2, \ldots, A_k$ are pairwise disjoint with union $\mathcal{S}$, then*

$$P(A_j|B) = \frac{P(A_j) \times P(B|A_j)}{\sum_{i=1}^{k} P(A_i) \times P(B|A_i)}, \quad j = 1, 2, \ldots, k.$$

Bayes rule is a restatement of the definition of conditional probability: the numerator in the formula is $P(A_j \cap B)$, the denominator is $P(B)$ (by the law of total probability), and the ratio is $P(A_j|B)$.

For example, suppose that 2% of the products assembled during the day shift and 6% of the products assembled during the night shift at a small company are defective and need reworking. If the day shift accounts for 55% of the products assembled by the company, then

$$\text{P(Day shift | Defective)} = \frac{0.55(0.02)}{0.55(0.02) + 0.45(0.06)} \approx 0.289$$

is the probability that a product was assembled during the day shift given that the product is defective. (Approximately 28.9% of defective products are assembled during the day shift.)

In applications, the collection of probabilities $\{P(A_j)\}$ are often referred to as the *prior* probabilities (the probabilities *before* observing an outcome in B), and the collection of probabilities $\{P(A_j|B)\}$ are often referred to as the *posterior* probabilities (the probabilities *after* event B has occurred).

1.5 Independent events

Events A and B are said to be *independent* if

$$P(A \cap B) = P(A) \times P(B).$$

Otherwise, A and B are said to be *dependent*.

If A and B are independent and have positive probabilities, then the multiplication rule for probability implies that

$$P(A|B) = P(A) \text{ and } P(B|A) = P(B).$$

(The relative size of A within B is the same as its relative size within $\mathcal{S}$; the relative size of B within A is the same as its relative size within $\mathcal{S}$.)

If A and B are independent, $0 < P(A) < 1$, and $0 < P(B) < 1$, then A and B^c are independent, A^c and B are independent, and A^c and B^c are independent.

More generally, events $A_1, A_2, \ldots, A_k$ are said to be *mutually independent* if

- for each pair of distinct indices (i_1, i_2), $P(A_{i_1} \cap A_{i_2}) = P(A_{i_1}) \times P(A_{i_2})$;
- for each triple of distinct indices (i_1, i_2, i_3),

$$P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) = P(A_{i_1}) \times P(A_{i_2}) \times P(A_{i_3});$$

- and so forth.

For example, suppose that 4 slips of paper are sampled with replacement from a well-mixed urn containing 25 slips of paper: 15 slips with the letter X written on

each and 10 slips of paper with the letter Y written on each. Then the probability of observing the sequence $XYXX$ is

$$P(XYXX) = P(X) \times P(Y) \times P(X) \times P(X) = \left(\tfrac{15}{25}\right)^3 \left(\tfrac{10}{25}\right) = \tfrac{54}{625} = 0.0864.$$

Further, since $\binom{4}{3} = 4$ sequences have exactly 3 X's, the probability of observing a sequence with exactly 3 X's is $4(0.0864) = 0.3456$.

1.5.1 Repeated trials and mutual independence

As stated at the beginning of the chapter, the term *experiment* is used in probability theory to describe a procedure whose outcome is not known in advance with certainty. Experiments are assumed to be repeatable and to have a well-defined set of outcomes. Repeated trials are repetitions of an experiment using the specified procedure, with the outcomes of the trials having no influence on one another. The results of repeated trials of an experiment are mutually independent.

1.6 Laboratory problems

The first set of laboratory problems introduce basic *Mathematica* commands and reinforce introductory probability concepts.

1.6.1 Laboratory: Introductory concepts

In the main laboratory notebook (Problems 1 to 6) you are asked to compute and graph binomial coefficients; choose random subsets and compute probabilities related to the subsets; choose random card hands and compute probabilities related to the hands; and estimate unknown parameters in the simple urn model using an event of maximum probability or a range of events with probability ratio greater than or equal to a fixed constant.

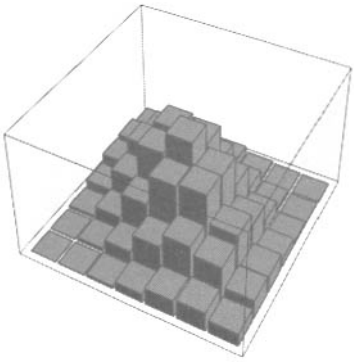
Note that computer algorithms called *pseudorandom* number generators are used to simulate the results of experiments, such as choosing random subsets or choosing random card hands. Computer simulation will be used in many laboratory problems in this book.

1.6.2 Additional problem notebooks

Problems 7 and 8 use frequency generating functions to compute probabilities related to roulette and dice games, respectively. Problem 9 applies the simple urn model to state lottery games.

Problems 10, 11, and 12 are additional applications of the simple urn model and model estimation. Problem 10 uses data from a study on mercury contamination in Maine lakes [54]. Problem 11 uses data from a study on estimating the size of a fish population [92]. Problem 12 uses data from a breast cancer study [9], [47]. Note that Problems 11 and 12 are applications of a method known in ecology as the *capture-recapture* method.

Problems 13, 14, and 15 involve computing and graphing probabilities. Problem 13 uses conditional probabilities to study polygraph tests [45]. Problem 14 uses mutually independent events to study best-of-seven series. Problem 15 uses mutually independent events to study an alternating shots game.



Chapter 2
Discrete Probability Distributions

Researchers use random variables to describe the numerical results of experiments. For example, if a fair coin is tossed five times and the total number of heads is recorded, then a random variable whose values are 0, 1, 2, 3, 4, 5 is used to give a numerical description of the results.

This chapter focuses on discrete random variables and their probability distributions. The first two sections give the important definitions and example families of distributions. Section 3 generalizes the ideas to joint distributions. Section 4 outlines the laboratory problems.

2.1 Definitions

A *random variable* is a function from the sample space of an experiment to the real numbers. The *range* of a random variable is the set of values the random variable assumes. Random variables are usually denoted by capital letters (X, Y, Z, ...) and their values by lowercase letters (x, y, z, ...).

If the range of a random variable is a finite or countably infinite set, then the random variable is said to be *discrete*; if the range is an interval or a union of intervals, the random variable is said to be *continuous*; otherwise, the random variable is said to be *mixed*.

If X is a discrete random variable, then $P(X = x)$ is the probability that an outcome has value x. Similarly, $P(X \leq x)$ is the probability that an outcome has value x or less, $P(a < X < b)$ is the probability that an outcome has value strictly between a and b, and so forth.

Example: Coin-tossing experiment

For example, suppose that you toss a fair coin eight times and record the sequence of heads and tails. Let X equal the difference between the number of heads and the number of tails in the sequence. Then X is a discrete random variable whose range

is $-8, -6, -4, \ldots, 8$. Further, $P(X \geq 3) = P(X = 4, 6, 8)$ equals the probability of 6 or more heads:

$$\binom{8}{6}\left(\frac{1}{2}\right)^8 + \binom{8}{7}\left(\frac{1}{2}\right)^8 + \binom{8}{8}\left(\frac{1}{2}\right)^8 = \frac{37}{256} \approx 0.1445.$$

(There are a total of 256 sequences, 37 of which have either 6, 7, or 8 heads.)

2.1.1 PDF and CDF for discrete distributions

If X is a discrete random variable, then the *frequency function* (FF) or *probability density function* (PDF) of X is defined as follows:

$$f(x) = P(X = x) \text{ for all real numbers } x.$$

PDFs satisfy the following properties:

1. $f(x) \geq 0$ for all real numbers x.
2. $\sum_{x \in R} f(x)$ equals 1, where R is the range of X.

Since $f(x)$ is the probability of an event, and events have nonnegative probabilities, $f(x)$ must be nonnegative for each x (property 1). Since the events $X = x$ for $x \in R$ are mutually disjoint with union S (the sample space), the sum of the probabilities of these events must be 1 (property 2).

The *cumulative distribution function* (CDF) of the discrete random variable X is defined as follows:

$$F(x) = P(X \leq x) \text{ for all real numbers } x.$$

CDFs satisfy the following properties:

1. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.
2. If $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$.
3. $F(x)$ is right continuous. That is, for each a, $\lim_{x \to a^+} F(x) = F(a)$.

$F(x)$ represents cumulative probability, with limits 0 and 1 (property 1). Cumulative probability increases with increasing x (property 2) and has discrete jumps at values of x in the range of the random variable (property 3).

Plotting PDF and CDF functions

The PDF of a discrete random variable is represented graphically by using a plot of pairs $(x, f(x))$ for $x \in R$, or by using a *probability histogram*, where area is used to represent probability. The CDF of a discrete random variable is represented graphically as a *step* function, with steps of height $f(x)$ at each $x \in R$.

For example, the left plot in Figure 2.1 is the probability histogram for the difference between the number of heads and the number of tails in eight tosses of

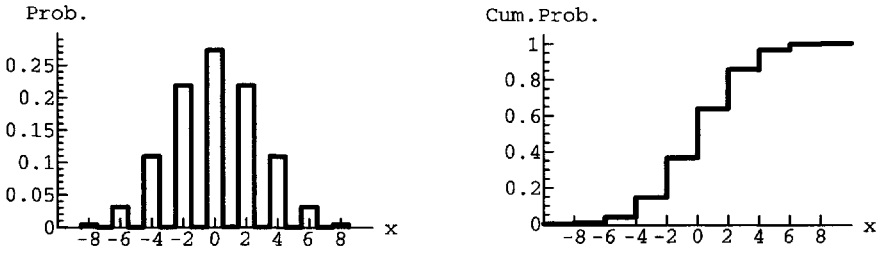


Figure 2.1. *Probability histogram (left plot) and CDF (right plot) for the difference between the number of heads and the number of tails in eight tosses of a fair coin.*

a fair coin. For each x in the range of the random variable, a rectangle with base equal to the interval $[x - 0.50, x + 0.50]$ and with height equal to $f(x)$ is drawn. The total area is 1.0. The right plot is a representation of the CDF. Note that $F(x)$ is nondecreasing, $F(x) = 0$ when $x < -8$, and $F(x) = 1$ when $x > 8$. Steps occur at $x = -8, -6, \ldots, 6, 8$.

2.2 Univariate distributions

This section defines several important families of distributions and states properties of these distributions.

2.2.1 Example: Discrete uniform distribution

Let n be a positive integer. The random variable X is said to be a *discrete uniform random variable*, or to have a *discrete uniform distribution*, with parameter n when its PDF is as follows:

$$f(x) = \frac{1}{n} \text{ when } x = 1, 2, \ldots, n \text{ and 0 otherwise.}$$

For example, if you roll a fair six-sided die and let X equal the number of dots on the top face, then X has a discrete uniform distribution with $n = 6$.

2.2.2 Example: Hypergeometric distribution

Let N, M, and n be integers with $0 < M < N$ and $0 < n < N$. The random variable X is said to be a *hypergeometric random variable*, or to have a *hypergeometric distribution*, with parameters n, M, and N, when its PDF is

$$f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

for integers, x, between $\max(0, n + M - N)$ and $\min(n, M)$ (and zero otherwise).

Hypergeometric distributions are used to model *urn* experiments. Suppose there are M special objects in an urn containing a total of N objects. Let X be the number of special objects in a subset of size n chosen from the urn. If the choice of each subset is equally likely, then X has a hypergeometric distribution. See the simple urn model example on page 6.

2.2.3 Distributions related to Bernoulli experiments

A *Bernoulli experiment* is an experiment with two possible outcomes. The outcome of chief interest is often called "success" and the other outcome "failure." Let p equal the probability of success.

Imagine repeating a Bernoulli experiment n times. The *expected number* of successes in n independent trials of a Bernoulli experiment with success probability p is np.

For example, suppose that you roll a fair six-sided die and observe the number on the top face. Let success be a 1 or 4 on the top face and failure be a 2, 3, 5, or 6 on the top face. Then $p = 1/3$ is the probability of success. In 600 trials of the experiment, you expect 200 successes.

Example: Bernoulli distribution

Suppose that a Bernoulli experiment is run once. Let X equal 1 if a success occurs and 0 if a failure occurs. Then X is said to be a *Bernoulli random variable*, or to have a *Bernoulli distribution*, with parameter p. The PDF of X is as follows:

$$f(1) = p,\ f(0) = 1 - p, \text{ and } f(x) = 0 \text{ otherwise.}$$

Example: Binomial distribution

Let X be the number of successes in n independent trials of a Bernoulli experiment with success probability p. Then X is said to be a *binomial random variable*, or to have a *binomial distribution*, with parameters n and p. The PDF of X is as follows:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ when } x = 0, 1, 2, \ldots, n \text{ and 0 otherwise.}$$

For each x, $f(x)$ is the probability of the event "exactly x successes in n independent trials." (There are a total of $\binom{n}{x}$ sequences with exactly x successes and $n-x$ failures; each sequence has probability $p^x(1-p)^{n-x}$.)

For example, if $x = 2$ and $n = 5$, then

$$f(2) = P(\{SSFFF, SFSFF, SFFSF, SFFFS, FSSFF, FSFSF, FSFFS, FFSSF, FFSFS, FFFSS\}) = 10p^2(1-p)^3,$$

where S represents success and F represents failure.

The binomial theorem (Theorem 1.2) can be used to demonstrate that the sum of probabilities $f(0) + f(1) + \cdots + f(n)$ equals one.

A Bernoulli random variable is a binomial random variable with $n = 1$.

Example: Geometric distribution on 0, 1, 2, ...

Let X be the number of failures before the first success in a sequence of independent Bernoulli experiments with success probability p. Then X is said to be a *geometric random variable*, or to have a *geometric distribution*, with parameter p. The PDF of X is as follows:

$$f(x) = (1-p)^x p \text{ when } x = 0, 1, 2, \ldots \text{ and 0 otherwise.}$$

For each x, $f(x)$ is the probability of the sequence of x failures (F) followed by a success (S). For example, $f(5) = \mathrm{P}(\{FFFFFS\}) = (1-p)^5 p$.

The probabilities $f(0), f(1), \ldots$ form a geometric sequence whose sum is 1.

An alternative definition of the geometric random variable is as follows: X is the trial number of the first success in a sequence of independent Bernoulli experiments with success probability p. In this case,

$$f(x) = (1-p)^{x-1} p \text{ when } x = 1, 2, 3, \ldots \text{ and 0 otherwise.}$$

In particular, the range is now the positive integers.

Example: Negative binomial distribution on 0, 1, 2, ...

Let r be a positive integer and X be the number of failures before the r^{th} success in a sequence of independent Bernoulli experiments with success probability p. Then X is said to be a *negative binomial random variable*, or to have a *negative binomial distribution*, with parameters r and p. The PDF of X is as follows:

$$f(x) = \binom{x+r-1}{r-1}(1-p)^x p^r \text{ when } x = 0, 1, 2, \ldots \text{ and 0 otherwise.}$$

For each x, $f(x)$ is the probability of the event "exactly x failures and r successes in $x+r$ trials, with the last trial a success." (There are a total of $\binom{x+r-1}{r-1}$ sequences with exactly x failures and r successes, with the last trial a success; each sequence has probability $(1-p)^x p^r$.)

For example, if $r = 3$ and $x = 2$, then

$$f(2) = P(\{FFSSS, FSFSS, FSSFS, SFFSS, SFSFS, SSFFS\}) = 6(1-p)^2 p^3,$$

where S represents success and F represents failure.

An alternative definition of the negative binomial random variable is as follows: X is the trial number of the r^{th} success in a sequence of independent Bernoulli trials with success probability p. In this case,

$$f(x) = \binom{x-1}{r-1}(1-p)^{x-r} p^r \text{ when } x = r, r+1, r+2, \ldots \text{ and 0 otherwise.}$$

For each x, $f(x)$ is the probability of the event "exactly $x-r$ failures and r successes in x trials, with the last trial a success."

2.2.4 Simple random samples

Suppose that an urn contains N objects. A *simple random sample* of size n is a sequence of n objects chosen *without* replacement from the urn, where the choice of each sequence is equally likely.

Let M be the number of special objects in the urn and X be the number of special objects in a simple random sample of size n. Then X has a hypergeometric distribution with parameters n, M, N. Further, if N is very large, then binomial probabilities can be used to approximate hypergeometric probabilities.

Theorem 2.1 (Binomial Approximation). *If N is large, then the binomial distribution with parameters n and $p = M/N$ can be used to approximate the hypergeometric distribution with parameters n, M, N. Specifically,*

$$P(x \text{ special objects}) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \approx \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for each } x.$$

Note that if X is the number of special objects in a sequence of n objects chosen *with* replacement from the urn and if the choice of each sequence is equally likely, then X has a binomial distribution with parameters n and $p = M/N$. The theorem says that if N is large, then the model where sampling is done *with* replacement can be used to approximate the model where sampling is done *without* replacement.

Survey analysis

Simple random samples are used in surveys. If the survey population is small, then hypergeometric distributions are used to analyze the results. If the survey population is large, then binomial distributions are used to analyze the results, even though each person's opinion is solicited at most once.

For example, suppose that a surveyor is interested in determining the level of support for a proposal to change the local tax structure and decides to choose a simple random sample of size 10 from the registered voter list. If there are a total of 120 registered voters, one-third of whom support the proposal, then the probability that exactly 3 of the 10 chosen voters support the proposal is

$$P(X = 3) = \frac{\binom{40}{3}\binom{80}{7}}{\binom{120}{10}} \approx 0.27.$$

If there are thousands of registered voters, the probability is

$$P(X = 3) \approx \binom{10}{3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^7 \approx 0.26.$$

Note that in the approximation you do not need to know the exact number of registered voters.

2.2.5 Example: Poisson distribution

Let λ be a positive real number. The random variable X is said to be a *Poisson random variable*, or to have a *Poisson distribution*, with parameter λ if its PDF is as follows:

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!} \text{ when } x = 0, 1, 2, \ldots \text{ and 0 otherwise.}$$

Recall that the Maclaurin series for $y = e^x$ is as follows:

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \text{ for all real numbers } x.$$

Thus, the sequence $f(0), f(1), \ldots$ has sum 1:

$$\sum_{x=0}^{\infty} f(x) = \sum_{x=0}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda}\left(\sum_{x=0}^{\infty}\frac{\lambda^x}{x!}\right) = e^{-\lambda}\left(e^{\lambda}\right) = 1.$$

The idea for the Poisson distribution comes from a limit theorem proven by the mathematician S. Poisson in the 1830's.

Theorem 2.2 (Poisson Limit Theorem). *Let λ be a positive real number, n a positive integer, and $p = \lambda/n$. Then*

$$\lim_{n\to\infty}\binom{n}{x}p^x(1-p)^{n-x} = e^{-\lambda}\frac{\lambda^x}{x!} \quad \textit{when } x = 0, 1, 2, \ldots.$$

Theorem 2.2 can be used to estimate binomial probabilities when the number of trials is large and the probability of success is small. For example, if $n = 10000$, $p = 2/5000$, and $x = 3$, then the probability of 3 successes in 10000 trials is

$$\binom{10000}{3}\left(\frac{2}{5000}\right)^3\left(\frac{4998}{5000}\right)^{9997} \approx 0.195386$$

and Poisson's approximation is

$$e^{-4}\frac{4^3}{3!} \approx 0.195367 \quad \text{using } \lambda = np = 4.$$

The values are very close.

Poisson process

Events occurring in time are said to be generated by an (approximate) *Poisson process* with rate λ when the following conditions are satisfied:

1. The number of events occurring in disjoint subintervals of time are independent of one another.

2. The probability of one event occurring in a sufficiently small subinterval of time is proportional to the size of the subinterval. If h is the size, then the probability is λh.

3. The probability of two or more events occurring in a sufficiently small subinterval of time is virtually zero.

In this definition, λ represents the average number of events per unit time.

If events follow an (approximate) Poisson process and X is the number of events observed in one unit of time, then X has a Poisson distribution with parameter λ.

Typical applications of Poisson distributions include the numbers of cars passing an intersection in a fixed period of time during a workday or the number of phone calls received in a fixed period of time during a workday.

The definition of Poisson process allows you to think of the PDF of X as the limit of a sequence of binomial PDFs. The observation interval is subdivided into n nonoverlapping subintervals; the i^{th} Bernoulli trial results in success if an event occurs in the i^{th} subinterval, and failure otherwise. If n is large enough, then the probability that two or more events occur in one subinterval can be assumed to be zero.

The idea of a Poisson process can be generalized to include events occurring over regions of space instead of intervals of time. ("Subregions" take the place of "subintervals" in the conditions above. In this case, λ represents the average number of events per unit area or per unit volume.)

2.3 Joint distributions

A probability distribution describing the joint variability of two or more random variables is called a *joint distribution*.

For example, if X is the height (in feet), Y is the weight (in pounds), and Z is the serum cholesterol level (in mg/dL) of a person chosen from a given population, then we may be interested in describing the joint distribution of the triple (X, Y, Z).

A *bivariate distribution* is the joint distribution of a pair of random variables.

2.3.1 Bivariate distributions; marginal distributions

Assume that X and Y are discrete random variables. The *joint frequency function* (joint FF) or *joint probability density function* (joint PDF) of (X, Y) is defined as follows:

$$f(x, y) = \mathrm{P}(X = x, Y = y) \text{ for all real pairs } (x, y),$$

where the comma is understood to mean the intersection of the events. The notation $f_{XY}(x, y)$ is sometimes used to emphasize the two random variables.

If X and Y are discrete random variables, then the *joint cumulative distribution function* (joint CDF) of (X, Y) is defined as follows:

$$F(x, y) = \mathrm{P}(X \le x, Y \le y) \text{ for all real pairs } (x, y),$$

where the comma is understood to mean the intersection of the events. The notation $F_{XY}(x, y)$ is sometimes used to emphasize the two random variables.

Table 2.1. *A discrete bivariate distribution.*

	$y=0$	$y=1$	$y=2$	$y=3$	
$x=0$	0.05	0.04	0.01	0.00	0.10
$x=1$	0.04	0.16	0.10	0.10	0.40
$x=2$	0.01	0.09	0.20	0.10	0.40
$x=3$	0.00	0.01	0.03	0.06	0.10
	0.10	0.30	0.34	0.26	1.00

The *marginal frequency function* (marginal FF) or *marginal probability density function* (marginal PDF) of X is

$$f_X(x) = \sum_y P(X = x, Y = y) \text{ for all real numbers } x,$$

where the sum is taken over all y in the range of Y. The marginal FF or marginal PDF of Y is defined similarly:

$$f_Y(y) = \sum_x P(X = x, Y = y) \text{ for all real numbers } y,$$

where the sum is taken over all x in the range of X.

Example: Finite joint distribution

For example, Table 2.1 displays the joint distribution of a pair of random variables with values 0, 1, 2, 3. The marginal distribution of X is given in the right column:

$$f_X(0) = f_X(3) = 0.1,\ f_X(1) = f_X(2) = 0.4, \text{ and } f_X(x) = 0 \text{ otherwise.}$$

Similarly, the marginal distribution of Y is given in the bottom row. Further,

$$P(X > Y) = \sum_{x=1}^{3} \sum_{y=0}^{x-1} f(x, y) = 0.04 + 0.10 + 0.04 = 0.18.$$

2.3.2 Conditional distributions; independence

Let X and Y be discrete random variables. If $P(Y = y) \neq 0$, then the *conditional frequency function* (conditional FF) or *conditional probability density function* (conditional PDF) of X given $Y = y$ is defined as follows:

$$f_{X|Y=y}(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} \text{ for all real numbers } x.$$

Similarly, if $P(X = x) \neq 0$, then the conditional PDF of Y given $X = x$ is

$$f_{Y|X=x}(y|x) = P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} \text{ for all real numbers } y.$$

Note that in the first case, the conditional sample space is the collection of outcomes with $Y = y$; in the second case, it is the collection of outcomes with $X = x$.

Conditional PDFs are often used as weights in weighted averages. See Chapter 4.

Given the joint distribution in Table 2.1, for example, the conditional PDF of Y given $X = 1$ is as follows:

$$f_{Y|X=1}(0|1) = 0.1, f_{Y|X=1}(1|1) = 0.4, f_{Y|X=1}(2|1) = f_{Y|X=1}(3|1) = 0.25$$

and is equal to zero otherwise.

Independent random variables

The discrete random variables X and Y are said to be *independent* if

$$f(x, y) = f_X(x)f_Y(y) \text{ for all real pairs } (x, y).$$

Otherwise, X and Y are said to be *dependent*.

X and Y are independent if the probability of the intersection is equal to the product of the probabilities

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

for all events of interest (for all x, y).

The random variables X and Y whose distributions are shown in Table 2.1 are dependent. For example, $f(1, 2) \neq f_X(1)f_Y(2)$.

2.3.3 Example: Bivariate hypergeometric distribution

Let n, M_1, M_2, and M_3 be positive integers with $n \leq M_1 + M_2 + M_3$. The random pair (X, Y) is said to have a *bivariate hypergeometric distribution* with parameters n and (M_1, M_2, M_3) if its joint PDF has the form

$$f(x, y) = \frac{\binom{M_1}{x}\binom{M_2}{y}\binom{M_3}{n-x-y}}{\binom{M_1+M_2+M_3}{n}}$$

when x and y are nonnegative integers with $x \leq \min(n, M_1)$, $y \leq \min(n, M_2)$, and $\max(0, n - M_3) \leq x + y \leq \min(n, M_3)$ and is equal to zero otherwise.

Bivariate hypergeometric distributions are used to model *urn* experiments. Specifically, suppose that an urn contains N objects, M_1 of type 1, M_2 of type 2, and M_3 of type 3 ($N = M_1 + M_2 + M_3$). Let X equal the number of objects of type 1 and Y equal the number of objects of type 2 in a subset of size n chosen from the urn. If each choice of subset is equally likely, then (X, Y) has a bivariate hypergeometric distribution with parameters n and (M_1, M_2, M_3).

Marginal and conditional distributions

If (X, Y) has a bivariate hypergeometric distribution, then X has a hypergeometric distribution with parameters n, M_1, and N; and Y has a hypergeometric distribution with parameters n, M_2, and N. In addition, each conditional distribution is hypergeometric.

2.3.4 Example: Trinomial distribution

Let n be a positive integer, and let p_1, p_2, and p_3 be positive proportions with sum 1. The random pair (X, Y) is said to have a *trinomial distribution* with parameters n and (p_1, p_2, p_3) when its joint PDF has the form

$$f(x, y) = \binom{n}{x, y, n-x-y} p_1^x p_2^y p_3^{n-x-y}$$

when $x = 0, 1, \ldots, n$; $y = 0, 1, \ldots, n$; $x + y \leq n$ and is equal to zero otherwise.

Trinomial distributions are used to model experiments with exactly three outcomes. Specifically, suppose that an experiment has three outcomes which occur with probabilities p_1, p_2, and p_3, respectively. Let X be the number of occurrences of outcome 1 and Y be the number of occurrences of outcome 2 in n independent trials of the experiment. Then (X, Y) has a trinomial distribution with parameters n and (p_1, p_2, p_3).

Marginal and conditional distributions

If (X, Y) has a trinomial distribution, then X has a binomial distribution with parameters n and p_1, and Y has a binomial distribution with parameters n and p_2. In addition, each conditional distribution is binomial.

2.3.5 Survey analysis

The results of Section 2.2.4 can be generalized. In particular, trinomial probabilities can be used to approximate bivariate hypergeometric probabilities when N is large enough, and each family of distributions can be used in survey analysis.

For example, suppose that a surveyor is interested in determining the level of support for a proposal to change the local tax structure and decides to choose a simple random sample of size 10 from the registered voter list. If there are a total of 120 registered voters, where one-third support the proposal, one-half oppose the proposal, and one-sixth have no opinion, then the probability that exactly 3 support, 5 oppose, and 2 have no opinion is

$$P(X = 3, Y = 5) = \frac{\binom{40}{3}\binom{60}{5}\binom{20}{2}}{\binom{120}{10}} \approx 0.088.$$

If there are thousands of registered voters, then the probability is

$$P(X = 3, Y = 5) \approx \binom{10}{3, 5, 2} \left(\frac{1}{3}\right)^3 \left(\frac{1}{2}\right)^5 \left(\frac{1}{6}\right)^2 \approx 0.081.$$

As before, you do not need to know the exact number of registered voters when you use the trinomial approximation.

2.3.6 Discrete multivariate distributions

A *multivariate distribution* is the joint distribution of k random variables.

Ideas studied in the bivariate case ($k = 2$) can be generalized to the case where $k > 2$. In particular, if $X_1, X_2, \ldots, X_k$ are discrete random variables, then the following hold:

1. The *joint frequency function* (joint FF) or *joint probability density function* (joint PDF) of $(X_1, X_2, \ldots, X_k)$ is defined as follows:

$$f_{\underline{X}}(x_1, x_2, \ldots, x_k) = P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$$

for all real k-tuples $(x_1, x_2, \ldots, x_k)$, where $\underline{X} = (X_1, X_2, \ldots, X_k)$ and commas are understood to mean the intersection of events.

2. The random variables $X_1, X_2, \ldots, X_k$ are said to be *mutually independent* (or *independent*) if

$$f_{\underline{X}}(x_1, x_2, \ldots, x_k) = f_1(x_1) f_2(x_2) \cdots f_k(x_k)$$

for all real k-tuples $(x_1, x_2, \ldots, x_k)$, where $f_i(x_i) = P(X_i = x_i)$ for $i = 1, 2, \ldots, k$. (The probability of the intersection is equal to the product of the probabilities for all events of interest.)

If the discrete random variables $X_1, X_2, \ldots, X_k$ are mutually independent and have a common distribution (each marginal PDF is the same), then $X_1, X_2, \ldots, X_k$ are said to be a *random sample* from that distribution.

2.3.7 Probability generating functions

Let X be a discrete random variable with values in the nonnegative integers and $p_i = P(X = i)$ for each i. The *probability generating function* of X is the following formal power series:

$$\text{PGF}(t) = p_0 + p_1 t + p_2 t^2 + \cdots = \sum_{i=0}^{\infty} p_i t^i.$$

If the range of X is finite, then the probability generating function reduces to a polynomial. For example, the probability generating function of a binomial random variable is the polynomial

$$\text{PGF}(t) = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} t^i = (pt + (1-p))^n.$$

An important property of probability generating functions is the following convolution theorem.

Theorem 2.3 (Convolution Theorem). *Let* $\mathrm{PGF}_1(t)$ *be the probability generating function of* X_1 *and* $\mathrm{PGF}_2(t)$ *be the probability generating function of* X_2. *If* X_1 *and* X_2 *are independent, then the probability generating function of the sum* $W = X_1 + X_2$ *is*

$$\mathrm{PGF}(t) = \mathrm{PGF}_1(t)\mathrm{PGF}_2(t).$$

Note that since

$$P(W = k) = \sum_{i=0}^{k} P(X_1 = i)P(X_2 = k - i) \text{ for each } k,$$

the convolution theorem follows from Theorem 1.4.

Corollary 2.4. *More generally, if* $X_1, X_2, \ldots, X_n$ *are mutually independent random variables whose values are in the nonnegative integers and* W *is their sum, then the probability generating function of* W *is*

$$\mathrm{PGF}(t) = \mathrm{PGF}_1(t)\mathrm{PGF}_2(t)\cdots\mathrm{PGF}_n(t),$$

where $\mathrm{PGF}_i(t)$ *is the probability generating function of* X_i.

2.4 Laboratory problems

The laboratory problems for this chapter introduce *Mathematica* commands for working with discrete probability distributions and reinforce ideas about discrete distributions.

2.4.1 Laboratory: Discrete models

In the main laboratory notebook (Problems 1 to 7), you are asked to compute probabilities using the PDF and CDF functions, use graphs to describe distributions, compute and summarize simulated random samples from distributions, and use graphs to compare simulated random samples to distributions. Binomial, Poisson, geometric, negative binomial, and hypergeometric models are used.

Note that the graphical method used to display samples is called an *empirical histogram* (or *histogram*). To construct a histogram, (1) the range of observed values is subdivided into a certain number of subintervals and the number of observations in each subinterval is counted; and (2) for each subinterval, a rectangle with base equal to the subinterval and with area equal to the proportion of observations in that subinterval is drawn. The sum of the areas is 1.

2.4.2 Additional problem notebooks

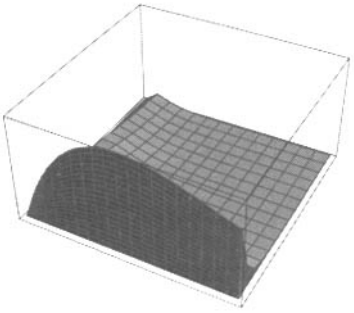
Problems 8 and 9 consider expected numbers in negative binomial and Poisson distributions. Both problems use a political campaign setting.

Problem 10 is an inverse problem. For a binomial random variable with success probability p and a fixed x_o, find the smallest n so that $P(X \geq x_o) \geq 0.95$. The problem uses a college admissions setting.

Problems 11 and 12 are informal *goodness-of-fit* problems. Relative errors (defined as the ratio of the difference between observed and expected numbers to an expected number) are computed using the given sample data and are compared to relative errors computed using simulated data. Problem 11 compares data from a study on the numbers of boys and girls in German families with exactly eight children to a binomial distribution [46], [80]. Problem 12 compares data from a study on outbreaks of war over a 400-year period to a Poisson distribution [91], [65].

Problem 13 uses independence and probability generating functions to construct the distribution of the number of school-age children in a town and to answer questions about the distribution.

Problems 14 and 15 consider the trinomial and bivariate hypergeometric distributions, respectively. In the introduction and initial parts of each problem, you are asked to graph distributions, compute and summarize simulated random samples, and work with conditional distributions (Y given $X = x$). The last part of Problem 14 applies trinomial models in a veterinary science setting; the last part of Problem 15 applies bivariate hypergeometric models in a company benefits setting.



Chapter 3
Continuous Probability Distributions

Researchers use random variables to describe the numerical results of experiments. In the continuous setting, the possible numerical values form an interval or a union of intervals. For example, a random variable whose values are the positive real numbers might be used to describe the lifetimes of individuals in a population.

This chapter focuses on continuous random variables and their probability distributions. The first two sections give the important definitions and example families of distributions. Section 3 generalizes the ideas to joint distributions. Section 4 outlines the laboratory problems.

3.1 Definitions

Recall that a random variable is a function from the sample space of an experiment to the real numbers and that the random variable X is said to be *continuous* if its range is an interval or a union of intervals.

3.1.1 PDF and CDF for continuous random variables

If X is a continuous random variable, then the *cumulative distribution function* (CDF) of X is defined as follows:

$$F(x) = P(X \leq x) \text{ for all real numbers } x.$$

CDFs satisfy the following properties:

1. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.
2. If $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$.
3. $F(x)$ is continuous.

$F(x)$ represents cumulative probability, with limits 0 and 1 (property 1). Cumulative probability increases with increasing x (property 2). For continuous random variables, the CDF is continuous (property 3).

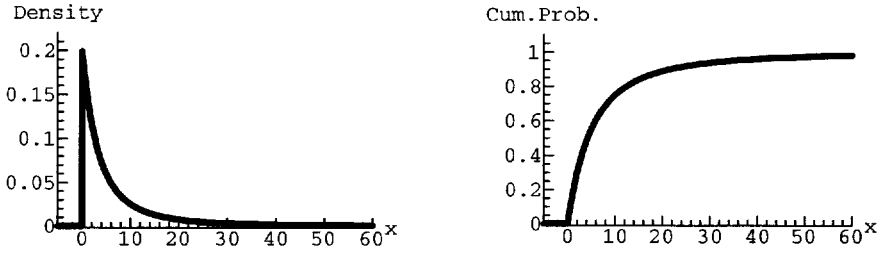


Figure 3.1. *PDF (left plot) and CDF (right plot) for a continuous random variable with range $x \geq 0$.*

The *probability density function* (PDF) (or *density function*) of the continuous random variable X is defined as follows:

$$f(x) = \frac{d}{dx}F(x) \text{ whenever the derivative exists.}$$

PDFs satisfy the following properties:

1. $f(x) \geq 0$ whenever it exists.
2. $\int_R f(x)dx$ equals 1, where R is the range of X.

$f(x)$ represents the rate of change of probability; the rate must be nonnegative (property 1). Since $f(x)$ is the derivative of the CDF, the area under $y = f(x)$ must be 1 (property 2).

If R is the range of X and I is an interval, then the probability of the event "the value of X is in the interval I" is computed by finding the area under the density curve for $x \in I \cap R$:

$$P(X \in I) = \int_{I \cap R} f(x)dx.$$

Note, in particular, that if $a \in R$, then $P(X = a) = 0$ since the area under the curve over an interval of length zero is zero.

Example: Distribution on the nonnegative real numbers

For example, let X be a continuous random variable whose range is the nonnegative real numbers and whose PDF is

$$f(x) = \frac{200}{(10+x)^3} \text{ when } x \geq 0 \text{ and 0 otherwise.}$$

The left part of Figure 3.1 is a plot of the density function of X, and the right part is a plot of its CDF:

$$F(x) = 1 - \frac{100}{(10+x)^2} \text{ when } x \geq 0 \text{ and 0 otherwise.}$$

Further, for this random variable, the probability that X is greater than 8 is

$$P(X > 8) = \int_8^\infty f(x)dx = 1 - F(8) = \frac{100}{18^2} \approx 0.3086.$$

3.1.2 Quantiles; percentiles

Assume that $0 < p < 1$. The p^{th} *quantile* (or $100p^{\text{th}}$ *percentile*) of the X distribution (when it exists) is the point, x_p, satisfying the equation

$$P(X \leq x_p) = p.$$

To find x_p, solve the equation $F(x) = p$ for x.

Important special cases of quantiles are as follows:

1. The *median* of the X distribution is the 50^{th} percentile.
2. The *quartiles* of the X distribution are the 25^{th}, 50^{th}, and 75^{th} percentiles.
3. The *deciles* of the X distribution are the 10^{th}, 20^{th}, ..., 90^{th} percentiles.

The median is a measure of the *center* (or *location*) of a distribution. Another measure of the center of a distribution is the mean, introduced in Chapter 4.

The *interquartile range* (IQR) is the difference between the 75^{th} and 25^{th} percentiles: $\text{IQR} = x_{0.75} - x_{0.25}$.

The IQR is a measure of the *scale* (or *spread*) of a distribution. Another measure of the scale of a distribution is the standard deviation, introduced in Chapter 4.

For the distribution displayed in Figure 3.1, a general formula for the p^{th} quantile is $x_p = -10 + 10/\sqrt{1-p}$, the median is (approximately) 4.142, and the IQR is (approximately) 8.453.

3.2 Univariate distributions

This section defines several important families of distributions and states properties of these distributions.

3.2.1 Example: Uniform distribution

Let a and b be real numbers with $a < b$. The continuous random variable X is said to be a *uniform random variable*, or to have a *uniform distribution*, on the interval $[a, b]$ when its PDF is as follows:

$$f(x) = \frac{1}{b-a} \text{ when } a \leq x \leq b \text{ and 0 otherwise.}$$

Note that the open interval (a, b), or one of the half-closed intervals $[a, b)$ or $(a, b]$, can be used instead of $[a, b]$ as the range of a uniform random variable.

Uniform distributions have constant density over an interval. That density is the reciprocal of the length of the interval. If X is a uniform random variable on the interval $[a, b]$, and $[c, d] \subseteq [a, b]$ is a subinterval, then

$$P(c \le X \le d) = \int_c^d \frac{1}{b-a}\,dx = \frac{d-c}{b-a} = \frac{\text{Length of } [c, d]}{\text{Length of } [a, b]}.$$

Also, $P(c < X < d) = P(c < X \le d) = P(c \le X < d) = (d - c)/(b - a)$.

Computer commands that return "random" numbers in the interval [0, 1] are simulating random numbers from the uniform distribution on the interval [0, 1].

3.2.2 Example: Exponential distribution

Let λ be a positive real number. The continuous random variable X is said to be an *exponential random variable*, or to have an *exponential distribution*, with parameter λ when its PDF is as follows:

$$f(x) = \lambda e^{-\lambda x} \text{ when } x \ge 0 \text{ and 0 otherwise.}$$

Note that the interval $x > 0$ can be used instead of the interval $x \ge 0$ as the range of an exponential random variable.

Exponential distributions are often used to represent the time that elapses before the occurrence of an event—for example, the time that a machine component will operate before breaking down.

Relationship with the Poisson process

If events occurring over time follow an approximate Poisson process with rate λ, where λ is the average number of events per unit time, then the time between successive events has an exponential distribution with parameter λ. To see this, note the following:

1. If you observe the process for t units of time and let Y equal the number of observed events, then Y has a Poisson distribution with parameter λt. The PDF of Y is as follows:

$$P(Y = y) = e^{-\lambda t}\frac{(\lambda t)^y}{y!} \text{ when } y = 0, 1, 2, \ldots \text{ and 0 otherwise.}$$

2. An event occurs, the clock is reset to time 0, and X is the time until the next event occurs. Then X is a continuous random variable whose range is $x > 0$. Further,

$$P(X > t) = P(0 \text{ events in the interval [0,t]}) = P(Y = 0) = e^{-\lambda t}$$

and $P(X \le t) = 1 - e^{-\lambda t}$.

3. Since $f(t) = \frac{d}{dt}P(X \le t) = \frac{d}{dt}\left(1 - e^{-\lambda t}\right) = \lambda e^{-\lambda t}$ when $t > 0$ (and 0 otherwise) is the same as the PDF of an exponential random variable with parameter λ, X has an exponential distribution with parameter λ.

3.2.3 Euler gamma function

Let r be a positive real number. The *Euler gamma function* is defined as follows:

$$\Gamma(r) = \int_{x=0}^{\infty} x^{r-1} e^{-x} dx.$$

If r is a positive integer, then $\Gamma(r) = (r-1)!$. Thus, the gamma function is said to *interpolate* the factorials.

The property $\Gamma(r) = (r-1)!$ for positive integers can be proven using induction. To start the induction, you need to demonstrate that $\Gamma(1) = 1$. To prove the induction step, you need to use integration by parts to demonstrate that $\Gamma(r+1) = r \times \Gamma(r)$.

3.2.4 Example: Gamma distribution

Let α and β be positive real numbers. The continuous random variable X is said to be a *gamma random variable*, or to have a *gamma distribution*, with parameters α and β when its PDF is as follows:

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \text{ when } x > 0 \text{ and 0 otherwise.}$$

Note that if $\alpha \geq 1$, then the interval $x \geq 0$ can be used instead of the interval $x > 0$ as the range of a gamma random variable.

This function is a valid PDF since $f(x) \geq 0$ for all x and

$$\begin{aligned} \int_{x=0}^{\infty} f(x)dx &= \tfrac{1}{\Gamma(\alpha)} \int_{x=0}^{\infty} \left(\tfrac{x}{\beta}\right)^{\alpha-1} e^{-(x/\beta)} \tfrac{dx}{\beta} \\ &= \tfrac{1}{\Gamma(\alpha)} \int_{u=0}^{\infty} u^{\alpha-1} e^{-u}\, du \quad (\text{where } u = x/\beta) \\ &= \tfrac{1}{\Gamma(\alpha)} \times \Gamma(\alpha) = 1. \end{aligned}$$

(By dividing by the gamma function $\Gamma(\alpha)$, the area under the curve becomes 1.)

The parameter α is called a *shape* parameter; gamma distributions with the same value of α have the same shape. The parameter β is called a *scale* parameter; for fixed α, as β changes, the scale on the vertical and horizontal axes change, but the shape remains the same.

If the shape parameter $\alpha = 1$, then the gamma distribution is the same as the exponential distribution with $\lambda = 1/\beta$.

Relationship with the Poisson process

If events occurring over time follow an approximate Poisson process with rate λ, where λ is the average number of events per unit time and if r is a positive integer, then the time until the r^{th} event occurs has a gamma distribution with $\alpha = r$ and $\beta = 1/\lambda$. To see this, note the following:

1. If you observe the process for t units of time and let Y equal the number of observed events, then Y has a Poisson distribution with parameter λt. The PDF of Y is as follows:

$$P(Y = y) = e^{-\lambda t}\frac{(\lambda t)^y}{y!} \text{ when } y = 0, 1, 2, \ldots \text{ and 0 otherwise.}$$

2. Let X be the time you observe the r^{th} event, starting from time 0. Then X is a continuous random variable whose range is $x > 0$. Further, $P(X > t)$ is the same as the probability that there are fewer than r events in the interval $[0, t]$. Thus,

$$P(X > t) = P(Y < r) = \sum_{y=0}^{r-1} e^{-\lambda t}\frac{(\lambda t)^y}{y!}$$

and $P(X \leq t) = 1 - P(X > t)$.

3. The PDF $f(t) = \frac{d}{dt}P(X \leq t)$ is computed using the product rule:

$$\begin{aligned}
\frac{d}{dt}\left[1 - e^{-\lambda t}\left(\sum_{y=0}^{r-1}\frac{\lambda^y t^y}{y!}\right)\right] &= \lambda e^{-\lambda t}\left(\sum_{y=0}^{r-1}\frac{\lambda^y t^y}{y!}\right) - e^{-\lambda t}\left(\sum_{y=1}^{r-1}\frac{\lambda^y y t^{y-1}}{y!}\right) \\
&= \lambda e^{-\lambda t}\left[\left(\sum_{y=0}^{r-1}\frac{\lambda^y t^y}{y!}\right) - \left(\sum_{y=1}^{r-1}\frac{\lambda^{y-1} t^{y-1}}{(y-1)!}\right)\right] \\
&= \lambda e^{-\lambda t}\left[\frac{\lambda^{r-1} t^{r-1}}{(r-1)!}\right] \\
&= \frac{\lambda^r}{(r-1)!} t^{r-1} e^{-\lambda t}.
\end{aligned}$$

Since $f(t)$ is the same as the PDF of a gamma random variable with parameters $\alpha = r$ and $\beta = 1/\lambda$, X has a gamma distribution with parameters $\alpha = r$ and $\beta = 1/\lambda$.

3.2.5 Distributions related to Poisson processes

In summary, three distributions are related to Poisson processes:

1. If X is the number of events occurring in a fixed period of time, then X is a Poisson random variable with parameter λ, where λ equals the average number of events for that fixed period of time. The probability that exactly x events occur in that interval is

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \text{ when } x = 0, 1, 2, \ldots \text{ and 0 otherwise.}$$

2. If X is the time between successive events, then X is an exponential random variable with parameter λ, where λ is the average number of events per unit time. The CDF of X is

$$F(x) = 1 - e^{-\lambda x} \text{ when } x > 0 \text{ and 0 otherwise.}$$

3. If X is the time to the r^{th} event, then X is a gamma random variable with parameters $\alpha = r$ and $\beta = 1/\lambda$, where λ is the average number of events per unit time. The CDF of X is

$$F(x) = 1 - \sum_{y=0}^{r-1} e^{-\lambda x} \frac{(\lambda x)^y}{y!} \text{ when } x > 0 \text{ and 0 otherwise.}$$

If $\alpha = r$ and r is not too large, then gamma probabilities can be computed by hand using the formula for the CDF above. Otherwise, the computer can be used to find probabilities for gamma distributions.

3.2.6 Example: Cauchy distribution

Let a be a real number and b be a positive real number. The continuous random variable X is said to be a *Cauchy random variable*, or to have a *Cauchy distribution*, with parameters a and b when its PDF is as follows:

$$f(x) = \frac{b}{\pi(b^2 + (x-a)^2)} \text{ for all real numbers } x.$$

Recall that the family of antiderivatives of $f(x)$ above is as follows:

$$\int f(x)\,dx = \frac{1}{\pi} \arctan\left(\frac{x-a}{b}\right) + C.$$

Using this fact, it is easy to demonstrate that $f(x)$ is a valid PDF.

The parameter a is called the *center* of the Cauchy distribution since the graph of the Cauchy PDF is symmetric around $x = a$. The median of the Cauchy distribution is a. The parameter b is called the *scale* (or spread) of the Cauchy distribution. The IQR of the Cauchy distribution is $2b$.

3.2.7 Example: Normal or Gaussian distribution

Let μ be a real number and σ be a positive real number. The continuous random variable X is said to be a *normal random variable*, or to have a *normal distribution*, with parameters μ and σ when its PDF is as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \text{ for all real numbers } x,$$

where exp() is the exponential function. The normal distribution is also called the *Gaussian distribution*, in honor of the mathematician Carl Friedrich Gauss.

The graph of the PDF of a normal random variable is the famous *bell-shaped curve*. The parameter μ is called the *mean* (or center) of the normal distribution, since the graph of the PDF is symmetric around $x = \mu$. The median of the normal distribution is μ. The parameter σ is called the *standard deviation* (or spread) of the normal distribution. The IQR of the normal distribution is (approximately) 1.35σ.

Normal distributions have many applications. For example, normal random variables can be used to model measurements of manufactured items made under strict controls; normal random variables can be used to model physical measurements (e.g., height, weight, blood values) in homogeneous populations.

The PDF of the normal distribution cannot be integrated in closed form. Most computer programs provide functions to compute probabilities and quantiles of the normal distribution.

Standard normal distribution

The continuous random variable Z is said to be a *standard normal random variable*, or to have a *standard normal distribution*, when Z is a normal random variable with $\mu = 0$ and $\sigma = 1$. The PDF and CDF of Z have special symbols:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \text{ and } \Phi(z) = \int_{-\infty}^{z} \phi(t)dt, \quad \text{where } z \text{ is a real number.}$$

The notation z_p is used for the p^{th} quantile of the Z distribution.

A table of cumulative probabilities of the standard normal random variable, suitable for doing problems by hand without using the computer, is given in Appendix B on the CD. The table can be used to estimate quantiles and to find probabilities and quantiles of other normal distributions.

3.2.8 Example: Laplace distribution

Let μ be a real number and β be a positive real number. The continuous random variable X is said to be a *Laplace random variable*, or to have a *Laplace distribution*, with parameters μ and β when its PDF is as follows:

$$f(x) = \frac{1}{2\beta} \exp\left(\frac{-|x-\mu|}{\beta}\right) \quad \text{for all real numbers } x,$$

where exp() is the exponential function.

The parameter μ is called the *mean* (or center) of the Laplace distribution, since the graph of the PDF is symmetric around $x = \mu$. The median of the Laplace distribution is μ. The parameter β is called the *scale* (or spread) of the Laplace distribution. The IQR of the Laplace distribution is $2\ln(2)\beta \approx 1.39\beta$.

3.2.9 Transforming continuous random variables

If X and $Y = g(X)$ are continuous random variables, then the CDF of X can be used to determine the CDF and PDF of Y.

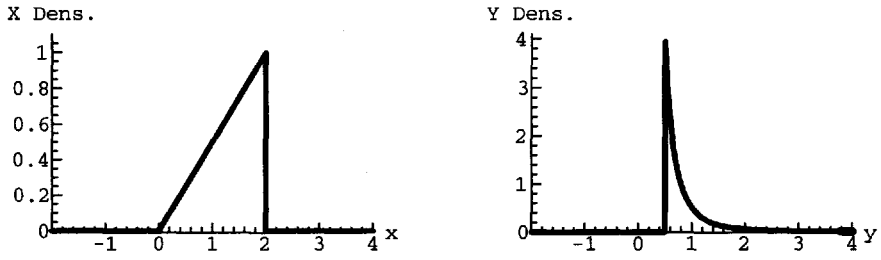


Figure 3.2. *PDFs of a continuous random variable with values on* $0 < x < 2$ *(left plot) and of its reciprocal with values on* $y > 0.5$ *(right plot).*

Example: Reciprocal transformation

For example, let X be a continuous random variable whose range is the open interval $(0, 2)$ and whose PDF is

$$f_X(x) = \frac{x}{2} \text{ when } 0 < x < 2 \text{ and 0 otherwise,}$$

and let Y equal the reciprocal of X, $Y = 1/X$. Then for $y > 1/2$,

$$P(Y \le y) = P\left(\frac{1}{X} \le y\right) = P\left(X \ge \frac{1}{y}\right) = \int_{x=1/y}^{2} \frac{x}{2} dx = 1 - \frac{1}{4y^2}$$

and $\frac{d}{dy}P(Y \le y) = 1/(2y^3)$. Thus, the PDF of Y is

$$f_Y(y) = \frac{1}{2y^3} \text{ when } y > 1/2 \text{ and 0 otherwise,}$$

and the CDF of Y is

$$F_Y(y) = 1 - \frac{1}{4y^2} \text{ when } y > 1/2 \text{ and 0 otherwise.}$$

The left part of Figure 3.2 is a plot of the density function of X, and the right part is a plot of the density function of Y.

Example: Square transformation

Let Z be the standard normal random variable and W be the square of Z, $W = Z^2$. Then for $w > 0$,

$$P(W \le w) = P(Z^2 \le w) = P(-\sqrt{w} \le Z \le \sqrt{w}) = \Phi(\sqrt{w}) - \Phi(-\sqrt{w}).$$

Since the graph of the PDF of Z is symmetric around $z = 0$, $\Phi(-z) = 1 - \Phi(z)$ for every z. Thus, the CDF of W is

$$F_W(w) = 2\Phi(\sqrt{w}) - 1 \text{ when } w > 0 \text{ and 0 otherwise,}$$

and the PDF of W is

$$f_W(w) = \frac{d}{dw} F_W(w) = \frac{1}{\sqrt{2\pi w}} e^{-w/2} \text{ when } w > 0 \text{ and 0 otherwise.}$$

$W = Z^2$ is said to have a *chi-square distribution* with one degree of freedom. The chi-square family of distributions is introduced in Chapter 6.

Location-scale distributions

The continuous uniform, Cauchy, normal, and Laplace distributions are examples of *location-scale* families of distributions. In each case, if X is a member of the family and Y is a linear transformation of X,

$$Y = mX + b \text{ for some real numbers } m \neq 0 \text{ and } b,$$

then Y is a member of the same family of distributions.

Note, in particular, that if X is a normal random variable with parameters μ and σ, and Z is the standard normal random variable, then $X = \sigma Z + \mu$.

3.3 Joint distributions

Recall that a probability distribution describing the joint variability of two or more random variables is called a *joint distribution* and that a *bivariate distribution* is the joint distribution of a pair of random variables.

3.3.1 Bivariate distributions; marginal distributions

Assume that X and Y are continuous random variables. The *joint cumulative distribution function* (joint CDF) of the random pair (X, Y) is defined as follows:

$$F(x, y) = \mathrm{P}(X \leq x, Y \leq y) \text{ for all real pairs } (x, y),$$

where the comma is understood to mean the intersection of the events. The notation $F_{XY}(x, y)$ is sometimes used to emphasize the two random variables.

The *joint probability density function* (joint PDF) (or *joint density function*) of the random pair (X, Y) is defined as follows:

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y) = \frac{\partial^2}{\partial y \partial x} F(x, y)$$

whenever F has continuous second partial derivatives. The notation $f_{XY}(x, y)$ is sometimes used to emphasize the two random variables.

The *marginal probability density function* (marginal PDF) (or *marginal density function*) of X is

$$f_X(x) = \int_y f(x, y) dy \text{ for } x \text{ in the range of } X$$

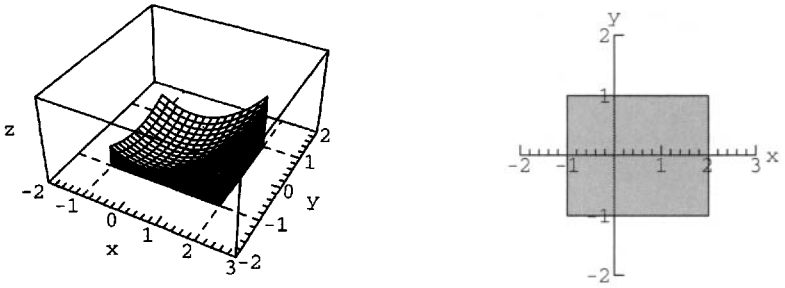


Figure 3.3. *Joint PDF for a random pair (left plot) and region of nonzero density in the xy-plane (right plot).*

and 0 otherwise, where the integral is taken over all y in the range of Y. The marginal PDF (or marginal density function) of Y is defined similarly:

$$f_Y(y) = \int_x f(x, y)dx \text{ for } y \text{ in the range of } Y$$

and 0 otherwise, where the integral is taken over all x in the range of X.

Example: Joint distribution on a rectangle

For example, assume that the continuous pair (X, Y) takes values in the rectangle $[-1, 2] \times [-1, 1]$ with joint PDF

$$f(x, y) = \frac{1}{8}(x^2 + y^2) \text{ when } -1 \le x \le 2, -1 \le y \le 1 \text{ and 0 otherwise.}$$

The left part of Figure 3.3 is a graph of the surface $z = f(x, y)$, and the right part shows the region of nonzero density in the xy-plane. The total volume under the surface and above the rectangular region in the xy-plane is 1.

For this random pair, the marginal PDF of X is

$$f_X(x) = \int_{y=-1}^{1} f(x, y)dy = \frac{1}{4}x^2 + \frac{1}{12} \text{ when } -1 \le x \le 2 \text{ and 0 otherwise,}$$

and the marginal PDF of Y is

$$f_Y(y) = \int_{x=-1}^{2} f(x, y)dx = \frac{3}{8}(y^2 + 1) \text{ when } -1 \le y \le 1 \text{ and 0 otherwise.}$$

Further, the probability that X is greater than Y is

$$P(X > Y) = \int_{y=-1}^{1} \int_{x=y}^{2} f(x, y)dxdy = \frac{5}{6} \approx 0.833.$$

3.3.2 Conditional distributions; independence

Let X and Y be continuous random variables with joint PDF $f(x, y)$ and marginal PDFs $f_X(x)$ and $f_Y(y)$, respectively. If $f_Y(y) \neq 0$, then the *conditional probability density function* (conditional PDF) (or *conditional density function*) of X given $Y = y$ is defined as follows:

$$f_{X|Y=y}(x|y) = \frac{f(x, y)}{f_Y(y)} \text{ for all real numbers } x.$$

Similarly, if $f_X(x) \neq 0$, then the conditional PDF (or conditional density function) of Y given $X = x$ is

$$f_{Y|X=x}(y|x) = \frac{f(x, y)}{f_X(x)} \text{ for all real numbers } y.$$

The function $f_{X|Y=y}(x|y)$ is a valid PDF since $f_{X|Y=y}(x|y) \geq 0$ and

$$\int_x f_{X|Y=y}(x|y)\, dx = \frac{1}{f_Y(y)} \int_x f(x, y)\, dx = \frac{1}{f_Y(y)} f_Y(y) = 1.$$

Similarly, the function $f_{Y|X=x}(y|x)$ is a valid PDF.

Conditional PDFs are often used as weights in weighted averages. See Chapter 4.

For the joint distribution pictured in Figure 3.3, for example, the area under the curve $z = f(x, 0)$ is $f_Y(0) = 3/8$, and the conditional distribution of X given $Y = 0$ has PDF

$$f_{X|Y=0}(x|0) = \frac{f(x, 0)}{f_Y(0)} = \frac{1}{3}x^2 \text{ when } -1 \leq x \leq 2 \text{ and 0 otherwise.}$$

Independent random variables

The continuous random variables X and Y are said to be *independent* if the joint CDF equals the product of the marginal CDFs for all real pairs

$$F(x, y) = F_X(x)F_Y(y) \text{ for all } (x, y)$$

or, equivalently, if the joint PDF equals the product of the marginal PDFs for all real pairs

$$f(x, y) = f_X(x)f_Y(y) \text{ for all } (x, y).$$

Otherwise, X and Y are said to be *dependent*.

The random variables X and Y whose joint distribution is shown in Figure 3.3 are dependent. For example, $f(0, 0) \neq f_X(0)f_Y(0)$.

3.3.3 Example: Bivariate uniform distribution

Let R be a region of the plane with finite positive area. The continuous random pair (X, Y) is said to have a *bivariate uniform distribution* on the region R if its joint PDF is as follows:

$$f(x, y) = \frac{1}{\text{Area of R}} \text{ when } (x, y) \in R \text{ and 0 otherwise.}$$

Bivariate uniform distributions have constant density over the region of nonzero density. That constant density is the reciprocal of the area. If A is a subregion of R ($A \subseteq R$), then

$$P((X, Y) \in A) = \frac{\text{Area of A}}{\text{Area of R}}.$$

That is, the probability of the event "the point (x, y) is in A" is the ratio of the area of the subregion to the area of the full region.

If (X, Y) has a bivariate uniform distribution, then X and Y may *not* be uniform random variables. Consider, for example, the bivariate uniform distribution on the triangle with vertices $(0, 0)$, $(1, 0)$, $(1, 1)$. Since the area of the triangle is 1/2, the PDF of X is

$$f_X(x) = \int_{y=0}^{x} 2dy = 2x \text{ when } 0 \le x \le 1 \text{ and 0 otherwise.}$$

Similarly, $f_Y(y) = 2y$ when $0 \le y \le 1$ and 0 otherwise. Since the PDFs are not constant on each range, the random variables X and Y are not uniformly distributed.

3.3.4 Example: Bivariate normal distribution

Let μ_x and μ_y be real numbers, σ_x and σ_y be positive real numbers, and ρ be a number in the interval $-1 < \rho < 1$. The random pair (X, Y) is said to have a *bivariate normal distribution* with parameters μ_x, μ_y, σ_x, σ_y, and ρ when its joint PDF is as follows:

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(\frac{-\text{term}}{2(1-\rho^2)}\right) \quad \text{for all real pairs } (x, y),$$

where

$$\text{term} = \left(\frac{x-\mu_x}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2$$

and exp() is the exponential function.

Standard bivariate normal distribution

The random pair (X, Y) is said to have a *standard bivariate normal distribution* with parameter ρ when (X, Y) has a bivariate normal distribution with $\mu_x = \mu_y = 0$ and $\sigma_x = \sigma_y = 1$. The joint PDF of (X, Y) is as follows:

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-(x^2 - 2\rho xy + y^2)}{2(1-\rho^2)}\right) \quad \text{for all real pairs } (x, y).$$

Marginal and conditional distributions

If (X, Y) has a bivariate normal distribution, then each marginal and conditional distribution is normal. In particular, if (X, Y) has a standard bivariate normal distribution, then

(i) X and Y are standard normal random variables,
(ii) the conditional distribution of Y given $X = x$ is normal with parameters $\mu = \rho x$ and $\sigma = \sqrt{1-\rho^2}$, and
(iii) the conditional distribution of X given $Y = y$ is normal with parameters $\mu = \rho y$ and $\sigma = \sqrt{1-\rho^2}$.

If (X, Y) has a bivariate normal distribution, then

$$(U, V) = \left(\frac{X-\mu_x}{\sigma_x}, \frac{Y-\mu_y}{\sigma_y}\right)$$

has a standard bivariate normal distribution.

3.3.5 Transforming continuous random variables

If X, Y, and $W = g(X, Y)$ are continuous random variables, then the joint PDF of (X, Y) can be used to determine the PDF and CDF of W.

Example: Product transformation

For example, let X be the length, Y be the width, and $W = XY$ be the area of a random rectangle. Specifically, assume that X is a uniform random variable on the interval $(0, 2)$, Y is a uniform random variable on the interval $(0, 1)$, and X and Y are independent.

Since the joint PDF of (X, Y) is

$$f(x, y) = f_X(x)f_Y(y) = 1/2 \text{ when } 0 < x < 2; 0 < y < 1 \text{ and } 0 \text{ otherwise,}$$

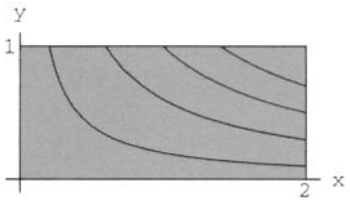
(X, Y) has a bivariate uniform distribution. For $0 < w < 2$,

$$P(W \le w) = P(XY \le w) = \frac{w}{2} + \frac{1}{2}\int_{x=w}^{2} \frac{w}{x}dx = \frac{1}{2}(w + w\ln(2) - w\ln(w))$$

and $\frac{d}{dw}P(W \le w) = \frac{1}{2}(\ln(2) - \ln(w)) = \frac{1}{2}\ln(2/w)$. Thus,

$$f_W(w) = \frac{1}{2}\ln(2/w) \text{ when } 0 < w < 2 \text{ and } 0 \text{ otherwise.}$$

The left part of Figure 3.4 shows the region of nonzero joint density with contours corresponding to $w = 0.2, 0.6, 1.0, 1.4$ superimposed. The right part is a plot of the density function of $W = XY$.



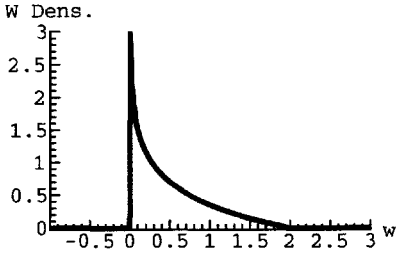


Figure 3.4. *The region of nonzero density for a bivariate uniform distribution on the rectangle* $(0, 2) \times (0, 1)$ *(left plot) and the PDF of the product of the coordinates with values in* $(0, 2)$ *(right plot). Contours for products equal to* 0.2, 0.6, 1.0, *and* 1.4 *are shown in the left plot.*

3.3.6 Continuous multivariate distributions

Recall that a *multivariate distribution* is the joint distribution of k random variables.

Ideas studied in the bivariate case ($k = 2$) can be generalized to the case where $k > 2$. In particular, if $X_1, X_2, \ldots, X_k$ are continuous random variables, then the following hold:

1. The *joint cumulative distribution function* (joint CDF) of $(X_1, X_2, \ldots, X_k)$ is defined as follows:

$$F_{\underline{X}}(x_1, x_2, \ldots, x_k) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_k \leq x_k)$$

for all real k-tuples $(x_1, x_2, \ldots, x_k)$, where $\underline{X} = (X_1, X_2, \ldots, X_k)$ and commas are understood to mean the intersection of events.

2. The *joint probability density function* (joint PDF) (or *joint density function*)

$$f_{\underline{X}}(x_1, x_2, \ldots, x_k)$$

is obtained from the joint CDF by taking multiple partial derivatives.

3. The random variables $X_1, X_2, \ldots, X_k$ are said to be *mutually independent* (or *independent*) if

$$F_{\underline{X}}(x_1, x_2, \ldots, x_k) = F_1(x_1)F_2(x_2)\cdots F_k(x_k)$$

or, equivalently, if

$$f_{\underline{X}}(x_1, x_2, \ldots, x_k) = f_1(x_1)f_2(x_2)\cdots f_k(x_k)$$

for all real k-tuples $(x_1, x_2, \ldots, x_k)$, where $F_i(x_i)$ is the CDF and $f_i(x_i)$ is the PDF of X_i for $i = 1, 2, \ldots, k$.

If the continuous random variables $X_1, X_2, \ldots, X_k$ are mutually independent and have a common distribution (each marginal distribution is the same), then $X_1, X_2, \ldots, X_k$ are said to be a *random sample* from that distribution.

3.4 Laboratory problems

The laboratory problems for this chapter introduce *Mathematica* commands for working with continuous probability distributions and reinforce ideas about continuous distributions.

3.4.1 Laboratory: Continuous models

In the main laboratory notebook (Problems 1 to 7), you are asked to compute probabilities using the PDF and CDF functions, use graphs to describe distributions, compute and summarize simulated random samples from distributions, and use graphs to compare simulated random samples to distributions. Normal, exponential, gamma, and uniform models are used. A relationship between the geometric and exponential distributions is demonstrated graphically.

3.4.2 Additional problem notebooks

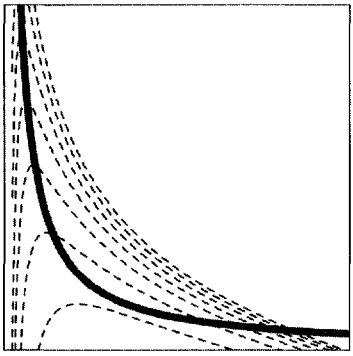
Problem 8 considers the inverse relationship between the average rate of a Poisson process and the expected waiting time until the r^{th} event occurs. The problem uses a backup system setting.

Problem 9 considers Cauchy random variables as transformations of uniform random variables. Simulation, graphing, and calculus techniques are used.

Problems 10 and 11 are informal *goodness-of-fit* problems. Relative errors (defined as the ratio of the difference between observed and expected numbers to an expected number) are computed using the given sample data and are compared to relative errors computed using simulated data. Problem 10 compares data from a genetics study to a uniform distribution [70], [79]. Problem 11 compares data from an IQ study to a normal distribution [106], [80].

Problems 12 through 15 consider transformations of the form $W = g(X, Y)$, where X and Y are independent continuous random variables. Simulation, graphing, and calculus techniques are used. Problem 12 (on ratios of exponentials) and Problem 13 (on differences of exponentials) apply the ideas to a waiting time setting. Problem 14 is on sums of gamma random variables. Problem 15 is on sums of uniform random variables.

Problem 16 considers the standard bivariate normal distribution. You are asked to describe how the joint PDF changes as the parameter ρ varies; compute simulated random samples, compare the samples to the joint distribution, and summarize the samples; and compute joint probabilities.



Chapter 4
Mathematical Expectation

Mathematical expectation generalizes the idea of a weighted average, where probability distributions are used as the weights.

The first two sections of this chapter define mathematical expectation for univariate distributions and introduce several important special cases. Sections 3 and 4 extend the ideas to bivariate and multivariate distributions. Section 5 outlines the laboratory problems for the chapter.

4.1 Definitions and properties

This section considers mathematical expectation for random variables and real-valued functions of random variables, and properties of expectation.

4.1.1 Discrete distributions

Let X be a discrete random variable with range R and PDF $f(x)$. The *mean* or *expected value* or *expectation* of X is defined as

$$E(X) = \sum_{x \in R} x\, f(x),$$

provided that $\sum_{x \in R} |x|\, f(x) < \infty$ (that is, provided that the series *converges absolutely*). Similarly, if $g(X)$ is a real-valued function of X, then the *mean* or *expected value* or *expectation* of $g(X)$ is

$$E(g(X)) = \sum_{x \in R} g(x)\, f(x),$$

provided that $\sum_{x \in R} |g(x)|\, f(x) < \infty$.

If the range of X is infinite, the absolute convergence of a series is not guaranteed. In cases where a series with absolute values diverges, we say that the expectation is *indeterminate*.

Example: Hypergeometric distribution

For example, assume you have 3 dimes and 5 nickels in your pocket. You choose a subset of 4 coins; let X equal the number of dimes in the subset and

$$g(X) = 10X + 5(4 - X) = 20 + 5X$$

be the total value (in cents) of the chosen coins. If each choice of subset is equally likely, then X has a hypergeometric distribution with parameters $n = 4$, $M = 3$, and $N = 8$; the expected number of dimes in the subset is

$$E(X) = \sum_{x=0}^{3} x\, f(x) = 0\left(\frac{5}{70}\right) + 1\left(\frac{30}{70}\right) + 2\left(\frac{30}{70}\right) + 3\left(\frac{5}{70}\right) = 1.5,$$

and the expected total value of the chosen coins is

$$E(g(X)) = 20\left(\frac{5}{70}\right) + 25\left(\frac{30}{70}\right) + 30\left(\frac{30}{70}\right) + 35\left(\frac{5}{70}\right) = 27.5 \text{ cents}.$$

Note that $E(g(X)) = g(E(X))$.

Example: Greatest integer transformation

Let U be a continuous uniform random variable on the open interval $(0, 1)$, and let X be the greatest integer less than or equal to the reciprocal of U (the "floor" of the reciprocal of U), $X = \lfloor 1/U \rfloor$. For x in the positive integers,

$$P(X = x) = P\left(\frac{1}{x+1} < U \leq \frac{1}{x}\right) = \frac{1}{x(x+1)};$$

$P(X = x) = 0$ otherwise. The expectation of X is

$$E(X) = \sum_{x=1}^{\infty} x\, \frac{1}{x(x+1)} = \sum_{x=1}^{\infty} \frac{1}{(x+1)}.$$

Since the series diverges, the expectation is indeterminate.

The left part of Figure 4.1 is a plot of the PDF of U with the vertical lines $u = 1/2, 1/3, 1/4, \ldots$ superimposed. The right part is a probability histogram of the distribution of $X = \lfloor 1/U \rfloor$. Note that the area between $u = 1/2$ and $u = 1$ on the left is $P(X = 1)$, the area between $u = 1/3$ and $u = 1/2$ is $P(X = 2)$, etc.

4.1.2 Continuous distributions

Let X be a continuous random variable with range R and PDF $f(x)$. The *mean* or *expected value* or *expectation* of X is defined as

$$E(X) = \int_{x \in R} x\, f(x)\, dx,$$

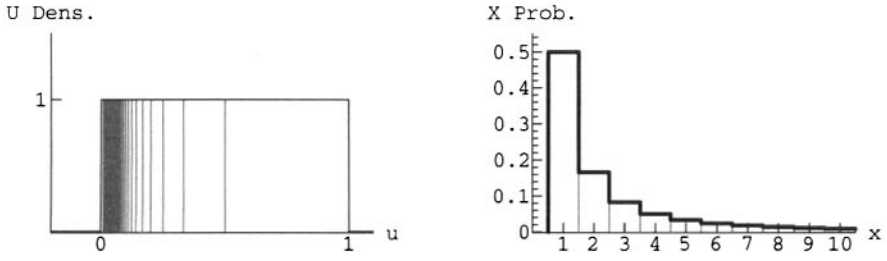


Figure 4.1. *PDF of the uniform random variable on the open interval* $(0, 1)$ *(left plot) and of the floor of its reciprocal (right plot). The range of the floor of the reciprocal is the positive integers. Vertical lines in the left plot are drawn at* $u = 1/2, 1/3, 1/4, \ldots$.

provided that $\int_{x \in R} |x| \, f(x) \, dx < \infty$ (that is, provided that the integral *converges absolutely*). Similarly, if $g(X)$ is a real-valued function of X, then the *mean* or *expected value* or *expectation* of $g(X)$ is

$$E(g(X)) = \int_{x \in R} g(x) \, f(x) \, dx,$$

provided that $\int_{x \in R} |g(x)| \, f(x) \, dx < \infty$.

Note that the absolute convergence of an integral is not guaranteed. In cases where an integral with absolute values diverges, we say that the expectation is *indeterminate*.

Example: Triangular distribution

For example, let X be a continuous random variable whose range is the interval $[1, 4]$ and whose PDF is

$$f(x) = \frac{2}{9}(x - 1) \text{ when } 1 \le x \le 4 \text{ and 0 otherwise,}$$

and let $g(X) = X^2$ be the square of X. Then

$$E(X) = \int_1^4 x \, f(x) \, dx = \int_1^4 x \, \frac{2}{9}(x - 1) \, dx = 3$$

and

$$E(g(X)) = \int_1^4 x^2 \, f(x) \, dx = \int_1^4 x^2 \, \frac{2}{9}(x - 1) \, dx = 19/2.$$

Note that $E(g(X)) \neq g(E(X))$.

4.1.3 Properties

Properties of sums and integrals imply the following properties of expectation:

1. If a is a constant, then $E(a) = a$.
2. If a and b are constants, then $E(a + bX) = a + bE(X)$.
3. If c_i is a constant and $g_i(X)$ is a real-valued function for $i = 1, 2, \ldots, k$, then

$$E(c_1 g_1(X) + c_2 g_2(X) + \cdots + c_k g_k(X)) = c_1 E(g_1(X)) + c_2 E(g_2(X)) + \cdots + c_k E(g_k(X)).$$

The first property says that the mean of a constant function is the constant itself. The second property says that if $g(X) = a + bX$, then $E(g(X)) = g(E(X))$. The third property generalizes the first two.

Note that if $g(X) \neq a + bX$, then $E(g(X))$ and $g(E(X))$ may be different.

4.2 Mean, variance, standard deviation

Let X be a random variable, and let $\mu = E(X)$ be its mean. The *variance* of X, $Var(X)$, is defined as follows:

$$Var(X) = E\left((X - \mu)^2\right).$$

The notation $\sigma^2 = Var(X)$ is used to denote the variance. The *standard deviation* of X, $\sigma = SD(X)$, is the positive square root of the variance.

The symbols used for mean (μ) and standard deviation (σ) are the same as the symbols used for the parameters of the normal distribution.

The mean is a measure of the *center* (or *location*) of a distribution. The variance and standard deviation are measures of the *scale* (or *spread*) of a distribution. If X is the height of an individual in inches, say, then the values of $E(X)$ and $SD(X)$ are in inches, while the value of $Var(X)$ is in square inches.

Table 4.1 lists the values of the mean and variance for the univariate families of distributions from Chapters 2 and 3. Note, in particular, that the mean and variance of the Cauchy distribution do not exist.

4.2.1 Properties

Properties of sums and integrals can be used to prove the following properties of the variance and standard deviation:

1. $Var(X) = E(X^2) - (E(X))^2$.
2. If $Y = a + bX$, then $Var(Y) = b^2 Var(X)$ and $SD(Y) = |b| SD(X)$.

The first property provides a quick by-hand method for computing the variance. For example, the variance of the triangular distribution discussed on page 47 is $19/2 - 3^2 = 1/2$.

The second property relates the spread of the distribution of a linear transformation of X to the spread of the distribution of X. In particular, if $Y = a + X$ (the values are shifted), then the standard deviation remains the same; if $Y = bX$ with $b > 0$ (the values are rescaled), then the standard deviation of Y is b times the standard deviation of X.

Table 4.1. *Model summaries for the univariate families of distributions.*

Distribution	$E(X)$	$Var(X)$
Discrete Uniform n	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Hypergeometric n, M, N	$n\frac{M}{N}$	$n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(\frac{N-n}{N-1}\right)$
Bernoulli p	p	$p(1-p)$
Binomial n, p	np	$np(1-p)$
Geometric p	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Negative Binomial r, p	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$
Poisson λ	λ	λ
Uniform a, b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma α, β	$\alpha\beta$	$\alpha\beta^2$
Cauchy a, b	indeterminate	indeterminate
Normal μ, σ	μ	σ^2
Laplace μ, β	μ	$2\beta^2$

Standardization

If X is a random variable with mean μ and standard deviation σ, then the random variable $Z = (X - \mu)/\sigma$ is called the *standardization* of X. By the properties above, the standardization of X has mean 0 and standard deviation 1.

Note that if X is a member of a location-scale family, then the standardization of X is a member of the same family. In particular, if X is a normal random variable, then $Z = (X - \mu)/\sigma$ is the standard normal random variable.

4.2.2 Chebyshev inequality

The Chebyshev inequality illustrates the importance of the concepts of mean, variance, and standard deviation.

Theorem 4.1 (Chebyshev Inequality). *Let X be a random variable with finite mean μ and standard deviation σ, and let k be a positive constant. Then*

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

For example, if $k = 2.5$, then the Chebyshev inequality states that at least 84% of the distribution of X is concentrated in the interval $|x - \mu| < 2.5\sigma$ and that at most 16% of the distribution is in the complementary interval $|x - \mu| \geq 2.5\sigma$.

4.2.3 Markov inequality

For random variables with nonnegative values and finite positive mean, the Markov inequality can be used to give a bound on certain probabilities.

Theorem 4.2 (Markov Inequality). *Let X be a random variable with values in the nonnegative reals and finite positive mean μ, and let k be a positive constant. Then*

$$P(X \geq k\mu) \leq \frac{1}{k}.$$

For example, if $k = 4$, then the Markov inequality states that at most 25% of the distribution of X is in the interval $x \geq 4\mu$ and that at least 75% is in the complementary interval $x < 4\mu$.

4.3 Functions of two or more random variables

Let $X_1, X_2, \ldots, X_k$ be discrete random variables with joint PDF $f(x_1, x_2, \ldots, x_k)$, and let $g(X_1, X_2, \ldots, X_k)$ be a real-valued function. The *mean* or *expected value* or *expectation* of $g(X_1, X_2, \ldots, X_k)$ is

$$E(g(X_1, X_2, \ldots, X_k)) = \sum_{x_k} \sum_{x_{k-1}} \cdots \sum_{x_1} g(x_1, x_2, \ldots, x_k) f(x_1, x_2, \ldots, x_k),$$

provided that the sum converges absolutely. The multiple sum is assumed to include all k-tuples with nonzero joint PDF.

Similarly, let $X_1, X_2, \ldots, X_k$ be continuous random variables with joint PDF $f(x_1, x_2, \ldots, x_k)$, and let $g(X_1, X_2, \ldots, X_k)$ be a real-valued function. The *mean* or *expected value* or *expectation* of $g(X_1, X_2, \ldots, X_k)$ is

$$\begin{aligned} &E(g(X_1, X_2, \ldots, X_k)) \\ &\quad = \int_{x_k} \int_{x_{k-1}} \cdots \int_{x_1} g(x_1, x_2, \ldots, x_k)\, f(x_1, x_2, \ldots, x_k) dx_1 dx_2 \cdots dx_k, \end{aligned}$$

provided that the integral converges absolutely. The multiple integral is assumed to include all k-tuples with nonzero joint PDF.

Example: Finite joint distribution

For example, if (X, Y) has the discrete bivariate distribution displayed in Table 2.1 and $g(X, Y) = |X - Y|$ is the absolute difference in the variables, then

$$E(g(X, Y)) = \sum_{y=0}^{3} \sum_{x=0}^{3} g(x, y) f(x, y) = 0(0.47) + 1(0.40) + 2(0.13) = 0.66.$$

Example: Bivariate uniform distribution on a triangular region

Similarly, if (X, Y) has a bivariate uniform distribution on the triangle with vertices $(0, 0)$, $(3, 0)$, $(3, 5)$ and $g(X, Y) = XY$ is the product of the variables, then

$$E(g(X, Y)) = \int_{x=0}^{3} \int_{y=0}^{5x/3} g(x, y) f(x, y) dy dx = \int_{x=0}^{3} \int_{y=0}^{5x/3} xy \frac{2}{15} dy dx = \frac{15}{4}.$$

4.3.1 Properties

The following properties of expectation can be proven using properties of sums and integrals and the fact that the joint PDF of mutually independent random variables equals the product of the marginal PDFs:

1. If a and $b_1, b_2, \ldots, b_n$ are constants and $g_i(X_1, X_2, \ldots, X_k)$ are real-valued functions for $i = 1, 2, \ldots, n$, then

$$E\left(a + \sum_{i=1}^{n} b_i g_i(X_1, X_2, \ldots, X_k)\right) = a + \sum_{i=1}^{n} b_i E\left(g_i(X_1, X_2, \ldots, X_k)\right).$$

2. If $X_1, X_2, \ldots, X_k$ are mutually independent random variables and $g_i(X_i)$ are real-valued functions for $i = 1, 2, \ldots, k$, then

$$E\left(g_1(X_1) g_2(X_2) \cdots g_k(X_k)\right) = E\left(g_1(X_1)\right) E\left(g_2(X_2)\right) \cdots E\left(g_k(X_k)\right).$$

The first property generalizes the properties given in Section 4.2.1. The second property, for mutually independent random variables, is useful when studying associations among random variables.

4.3.2 Covariance, correlation

Let X and Y be random variables with finite means (μ_x, μ_y) and finite standard deviations (σ_x, σ_y). The *covariance* of X and Y, $Cov(X, Y)$, is defined as follows:

$$Cov(X, Y) = E((X - \mu_x)(Y - \mu_y)).$$

The notation $\sigma_{xy} = Cov(X, Y)$ is used to denote the covariance. The *correlation* of X and Y, $Corr(X, Y)$, is defined as follows:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

The notation $\rho = Corr(X, Y)$ is used to denote the correlation of X and Y; ρ is called the *correlation coefficient*.

The symbol used for the correlation coefficient (ρ) is the same as the symbol used for the correlation parameter of the bivariate normal distribution.

Covariance and correlation are measures of the association between two random variables. Specifically, the following hold:

Table 4.2. *Correlations for the bivariate families of distributions.*

Distribution	$Corr(X, Y)$
Trinomial n, (p_1, p_2, p_3)	$-\sqrt{\frac{p_1}{1-p_1}\,\frac{p_2}{1-p_2}}$
Bivariate Hypergeometric n, (M_1, M_2, M_3)	$-\sqrt{\frac{M_1}{M_2+M_3}\,\frac{M_2}{M_1+M_3}}$
Bivariate Normal μ_x, μ_y, σ_x, σ_y, ρ	ρ

1. The random variables X and Y are said to be *positively associated* if as X increases, Y tends to increase. If X and Y are positively associated, then $Cov(X, Y)$ and $Corr(X, Y)$ will be positive.
2. The random variables X and Y are said to be *negatively associated* if as X increases, Y tends to decrease. If X and Y are negatively associated, then $Cov(X, Y)$ and $Corr(X, Y)$ will be negative.

For example, the height and weight of individuals in a given population are positively associated. Educational level and indices of poor health are often negatively associated.

Table 4.2 lists correlations for the trinomial, bivariate hypergeometric, and bivariate normal distributions. Note that if the random pair (X, Y) has a trinomial or bivariate hypergeometric distribution, then X and Y are negatively associated.

Properties of covariance

The following properties of covariance are important:

1. $Cov(X, X) = Var(X)$.
2. $Cov(X, Y) = Cov(Y, X)$.
3. $Cov(a + bX, c + dY) = bdCov(X, Y)$, where a, b, c, and d are constants.
4. $Cov(X, Y) = E(XY) - E(X)E(Y)$.
5. If X and Y are independent, then $Cov(X, Y) = 0$.

The first two properties follow immediately from the definition of covariance. The third property relates the covariance of linearly transformed random variables to the covariance of the original random variables. In particular, if $b = d = 1$ (values are shifted only), then the covariance is unchanged; if $a = c = 0$ and $b, d > 0$ (values are rescaled), then the covariance is multiplied by bd.

The fourth property gives an alternative method for calculating the covariance; the method is particularly well suited for by-hand computations. Since $E(XY) = E(X)E(Y)$ for independent random variables, the fourth property can be used to prove that the covariance of independent random variables is zero (property 5).

Properties of correlation

The following properties of correlation are important:

1. $-1 \le Corr(X, Y) \le 1$.

2. If a, b, c, and d are constants, then

$$Corr(a + bX, c + dY) = \begin{cases} Corr(X, Y) & \text{when } bd > 0, \\ -Corr(X, Y) & \text{when } bd < 0. \end{cases}$$

   In particular, $Corr(X, a + bX)$ equals 1 if $b > 0$ and equals -1 if $b < 0$.

3. If X and Y are independent, then $Corr(X, Y) = 0$.

Correlation is often called standardized covariance since, by the first property, its values always lie in the $[-1, 1]$ interval. If the correlation is close to -1, then there is a strong negative association between the variables; if the correlation is close to 1, then there is a strong positive association between the variables.

The second property relates the correlation of linearly transformed variables to the correlation of the original variables. Note, in particular, that the correlation is unchanged if the random variables are shifted ($b = d = 1$) or if the random variables are rescaled ($a = c = 0$ and $b, d > 0$). For example, the correlation between the height and weight of individuals in a population is the same no matter which measurement scale is used for height (e.g., inches, feet) and which measurement scale is used for weight (e.g., pounds, kilograms).

If $Corr(X, Y) = 0$, then X and Y are said to be *uncorrelated*; otherwise, they are said to be *correlated*. The third property says that independent random variables are uncorrelated. Note that uncorrelated random variables are *not* necessarily independent. For example, if (X, Y) has a bivariate uniform distribution on the diamond-shaped region with vertices $(-1, 0)$, $(0, 1)$, $(1, 0)$, $(0, -1)$, then X and Y are uncorrelated but not independent.

Example: Bivariate uniform distribution on a triangular region

Assume that (X, Y) has a bivariate uniform distribution on the triangle with vertices $(0, 0)$, $(3, 0)$, $(3, 5)$. Then

$$E(X) = 2, \quad E(Y) = \frac{5}{3}, \quad E(X^2) = \frac{9}{2}, \quad E(Y^2) = \frac{25}{6}, \quad E(XY) = \frac{15}{4}.$$

(See page 51 for the computation of $E(XY)$.) Using properties of variance and covariance,

$$Var(X) = \frac{1}{2}, \quad Var(Y) = \frac{25}{18}, \quad Cov(X, Y) = \frac{5}{12}.$$

Finally, $\rho = Corr(X, Y) = 1/2$.

4.3.3 Sample summaries

Recall that a random sample of size n from the X distribution is a list of n mutually independent random variables, each with the same distribution as X.

If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with mean μ and standard deviation σ, then the *sample mean*, $\overline{X}$, is the random variable

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n),$$

the *sample variance*, S^2, is the random variable

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

and the *sample standard deviation*, S, is the positive square root of the sample variance. The following theorem can be proven using properties of expectation.

Theorem 4.3 (Sample Summaries). *If $\overline{X}$ is the sample mean and S^2 is the sample variance of a random sample of size n from a distribution with mean μ and standard deviation σ, then the following hold:*

1. $E(\overline{X}) = \mu$ *and* $Var(\overline{X}) = \sigma^2/n$.
2. $E(S^2) = \sigma^2$.

In statistical applications, the observed value of the sample mean is used to estimate an unknown mean μ, and the observed value of the sample variance is used to estimate an unknown variance σ^2.

Sample correlation

A *random sample* of size n from the joint (X, Y) distribution is a list of n mutually independent random pairs, each with the same distribution as (X, Y).

If $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is a random sample of size n from a bivariate distribution with correlation $\rho = Corr(X, Y)$, then the *sample correlation*, R, is the random variable

$$R = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}},$$

where $\overline{X}$ and $\overline{Y}$ are the sample means of the X and Y samples, respectively.

In statistical applications, the observed value of the sample correlation is used to estimate an unknown correlation ρ.

4.3.4 Conditional expectation; regression

Let X and Y be discrete random variables with joint PDF $f(x, y)$. If $f_X(x) \neq 0$, then the *conditional expectation* (or the *conditional mean*) of Y given $X = x$, $E(Y|X = x)$, is defined as follows:

$$E(Y|X = x) = \sum_y y\, f_{Y|X=x}(y|x),$$

where the sum is over all y with nonzero conditional PDF ($f_{Y|X=x}(y|x) \neq 0$), provided that the series converges absolutely.

Let X and Y be continuous random variables with joint PDF $f(x, y)$. If $f_X(x) \neq 0$, then the *conditional expectation* (or the *conditional mean*) of Y given $X = x$, $E(Y|X = x)$, is defined as

$$E(Y|X = x) = \int_y y\, f_{Y|X=x}(y|x)\, dy,$$

where the integral is over all y with nonzero conditional PDF ($f_{Y|X=x}(y|x) \neq 0$), provided that the integral converges absolutely.

Definitions for the conditional expectation of X given $Y = y$, $E(X|Y = y)$, in the discrete and continuous cases are similar to those given above.

Regression of Y on X

The formula for the conditional expectation $E(Y|X = x)$ as a function of x is often called the *regression equation* of Y on X.

An important problem in statistical applications is to determine the formula for the conditional mean $E(Y|X = x)$ as a function of x. See Chapter 14.

Example: Finite joint distribution

For example, the following table gives the values of $E(Y|X = x)$ for the discrete bivariate distribution given in Table 2.1:

x	0	1	2	3
$E(Y\|X = x)$	0.60	1.65	1.975	2.50

Example: Joint distribution on a quarter plane

Assume the continuous pair (X, Y) takes values in the quarter plane $[1, \infty) \times [0, \infty)$ with joint PDF

$$f(x, y) = \frac{e^{-yx}}{2\sqrt{x}} \text{ when } x \geq 1,\ y \geq 0 \text{ and 0 otherwise.}$$

For $x \geq 1$, the marginal PDF of X is $f_X(x) = 1/(2x^{3/2})$, and the conditional PDF of Y given $X = x$ is

$$f_{Y|X=x}(y|x) = xe^{-yx} \text{ when } y \geq 0 \text{ and 0 otherwise.}$$

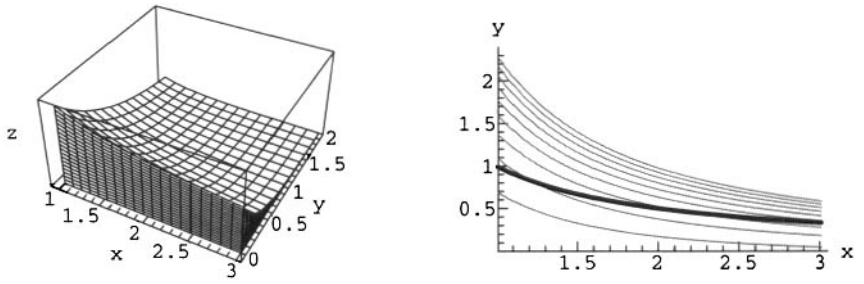


Figure 4.2. *Graph of $z = f(x, y)$ for a continuous bivariate distribution (left plot) and contour plot with $z = 1/2, 1/4, 1/6, \ldots, 1/20$ and conditional expectation of Y given $X = x$ superimposed (right plot).*

Since the conditional distribution is exponential with parameter $\lambda = x$, the formula for the conditional mean is $E(Y|X = x) = 1/x$ for $x \geq 1$.

The left part of Figure 4.2 is a graph of $z = f(x, y)$, and the right part is a contour plot with $z = 1/2, 1/4, 1/6, \ldots, 1/20$ (in gray). The formula for the conditional mean, $y = 1/x$, is superimposed on the contour plot (in black).

Linear conditional expectation

Let X and Y be random variables with finite means (μ_x, μ_y), standard deviations (σ_x, σ_y), and correlation (ρ). If the conditional expectation $E(Y|X = x)$ is a linear function of x, then the formula is of the form

$$E(Y|X = x) = \mu_y + \frac{\rho\sigma_y}{\sigma_x}(x - \mu_x).$$

If the conditional expectation $E(X|Y = y)$ is a linear function of y, then the formula is of the form

$$E(X|Y = y) = \mu_x + \frac{\rho\sigma_x}{\sigma_y}(y - \mu_y).$$

The trinomial, bivariate hypergeometric, and bivariate normal distributions have linear conditional means.

4.4 Linear functions of random variables

Let $X_1, X_2, \ldots, X_n$ be random variables with

1. $\mu_i = E(X_i)$ and $\sigma_i = SD(X_i)$ for $i = 1, 2, \ldots, n$ and
2. $\sigma_{i,j} = Cov(X_i, X_j)$ for $i \neq j$.

This section considers properties of linear functions of $X_1, X_2, \ldots, X_n$.

Theorem 4.4 (Mean and Variance). *Let $X_1, X_2, \ldots, X_n$ be random variables with summary measures given above and $Y = a + \sum_{i=1}^{n} b_i X_i$, where a and $b_1, b_2, \ldots, b_n$ are constants. Then*

$$E(Y) = a + \sum_{i=1}^{n} b_i \mu_i \text{ and } Var(Y) = \sum_{i=1}^{n} b_i^2 \sigma_i^2 + \sum_{i \neq j} b_i b_j \sigma_{i,j}.$$

If, in addition, $X_1, X_2, \ldots, X_n$ are mutually independent, then

$$E(Y) = a + \sum_{i=1}^{n} b_i \mu_i \quad \text{and} \quad Var(Y) = \sum_{i=1}^{n} b_i^2 \sigma_i^2.$$

For example, if $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with mean μ and standard deviation σ and $Y = \sum_{i=1}^{n} X_i$ is the *sample sum*, then the mean of Y is $E(Y) = n\mu$ and the variance of Y is $Var(Y) = n\sigma^2$.

Theorem 4.5 (Covariance). *Let $X_1, X_2, \ldots, X_n$ be random variables with summary measures given above, $V = a + \sum_{i=1}^{n} b_i X_i$ and $W = c + \sum_{i=1}^{n} d_i X_i$, where a, c, $b_1, b_2, \ldots, b_n$ and $d_1, d_2, \ldots, d_n$ are constants. Then*

$$Cov(V, W) = \sum_{i=1}^{n} b_i d_i \sigma_i^2 + \sum_{i \neq j} b_i d_j \sigma_{i,j}.$$

If, in addition, $X_1, X_2, \ldots, X_n$ are mutually independent, then

$$Cov(V, W) = \sum_{i=1}^{n} b_i d_i \sigma_i^2.$$

For example, if X_1, X_2, X_3, X_4 is a random sample of size 4 from a distribution with mean 3 and standard deviation 5,

$$V = X_1 + X_2 + X_3 + X_4 \text{ and } W = X_1 - 2X_2 + 3X_3 - 4X_4,$$

then $Cov(V, W) = -50$ and $Corr(V, W) = \frac{-1}{\sqrt{30}} \approx -0.18$.

4.4.1 Independent normal random variables

Theorem 4.4 can be used to determine summary measures of a linear function of random variables, but it says nothing about the distribution of the linear function. If $X_1, X_2, \ldots, X_n$ are independent normal variables, then the distribution is known.

Theorem 4.6 (Independent Normal Random Variables). *Let $X_1, X_2, \ldots, X_n$ be mutually independent normal random variables, and let $Y = a + \sum_{i=1}^{n} b_i X_i$. Then Y is a normal random variable.*

In particular, if $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution, then the sample sum and the sample mean are normal random variables.

Theorem 4.6 can be proven using the method of moment generating functions. Moment generating functions are discussed in Chapter 5.

4.5 Laboratory problems

The laboratory problems for this chapter introduce *Mathematica* commands for summarizing distributions and summarizing samples from distributions. The problems are designed to reinforce mathematical expectation concepts.

4.5.1 Laboratory: Mathematical expectation

In the main laboratory notebook (Problems 1 to 6), you are asked to compute model summaries and probabilities; apply the Chebyshev inequality; compute and summarize simulated random samples from distributions; use the sample mean and sample standard deviation to estimate the parameters in a model for percent body fat in men [59] and to estimate model probabilities; and work with conditional expectation. Exponential, uniform, negative binomial, gamma, Poisson, normal, and trinomial models are used.

4.5.2 Additional problem notebooks

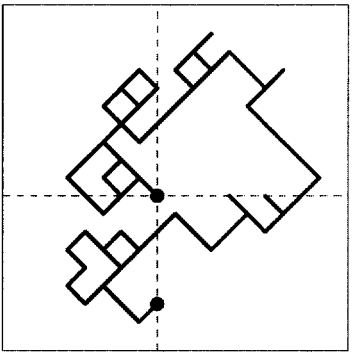
Problem 7 considers the Markov inequality and applies the Markov and Chebyshev inequalities in an advertising setting.

Problem 8 uses expectation to study the potential advantages of *pooled blood testing* when screening for a particular disease. Different implementation strategies are considered.

In Problems 9 and 10, sample summaries are used to estimate unknown parameters in models; the estimated models are then used to compute probabilities. Problem 9 uses a Poisson distribution to model bombing patterns during the second world war [27]. Problem 10 uses an exponential distribution to model the time between successive coal mining disasters [71], [57].

Problems 11, 12, and 13 consider bivariate distributions with linear conditional means. Bivariate uniform distributions on parallelograms are used in Problem 11. The distribution of outcomes in a dice and coin experiment is used in Problem 12. In Problem 13, bivariate normal distributions are applied to height-weight data for athletes [28].

Problems 14 and 15 consider bivariate discrete distributions, their corresponding marginal and conditional distributions, and the *law of total expectation*. Problem 14 uses data from a marriage study [22], [80]. Problem 15 uses data from an eyesight study [102].



Chapter 5
Limit Theorems

This chapter considers properties of sample sums and sample means as the sample size n approaches infinity. Definitions are given in Section 1. The law of large numbers is stated in Section 2, and an outline of its proof is given. The central limit theorem is stated in Section 3. Section 4 introduces the concept of moment generating functions and shows how moment generating functions can be used to prove the central limit theorem. Section 5 outlines the laboratory problems for this chapter.

5.1 Definitions

Let $X_1, X_2, X_3, \ldots$ be a sequence of mutually independent random variables, each with the same distribution as X. Two related sequences are of interest:

1. The sequence of *running sums*:

$$S_m = \sum_{i=1}^{m} X_i \text{ for } m = 1, 2, 3, \ldots.$$

2. The sequence of *running averages*:

$$\overline{X}_m = \frac{1}{m} \sum_{i=1}^{m} X_i \text{ for } m = 1, 2, 3, \ldots.$$

Example: Independent Bernoulli trials

For example, if X is a Bernoulli random variable with parameter p, then S_m is the total number of successes, and $\overline{X}_m$ is the average number of successes, in m repeated trials of the experiment. S_m is a binomial random variable with mean mp and variance $mp(1-p)$. The distribution of $\overline{X}_m$ has mean p and variance $p(1-p)/m$. Further,

$$\lim_{m \to \infty} Var(S_m) = \infty \text{ and } \lim_{m \to \infty} Var(\overline{X}_m) = 0.$$

5.2 Law of large numbers

For distributions with finite mean and standard deviation, the law of large numbers says that the sample mean, $\overline{X}$, is unlikely to be far from the true mean, μ, when the sample size is large enough. Formally, the following theorem holds.

Theorem 5.1 (Law of Large Numbers). *Let X be a random variable with finite mean μ and standard deviation σ, and let $\overline{X}_m$, $m = 1, 2, 3, \ldots$, be the sequence of running averages. For every positive real number ϵ,*

$$\lim_{m\to\infty} P\left(|\overline{X}_m - \mu| \geq \epsilon\right) = 0.$$

The law of large numbers can be proven using Chebyshev's inequality (Theorem 4.1). An outline of the proof is as follows:

(i) Using complements, it is sufficient to demonstrate that

$$\lim_{m\to\infty} P\left(\mu - \epsilon < \overline{X}_m < \mu + \epsilon\right) = 1.$$

(ii) Since $SD(\overline{X}_m) = \sigma/\sqrt{m}$, if $\epsilon = k\sigma/\sqrt{m}$ (correspondingly, $k = \epsilon\sqrt{m}/\sigma$) is substituted into the Chebyshev inequality, then we get the following lower bound on probability:

$$P\left(\mu - \epsilon < \overline{X}_m < \mu + \epsilon\right) \geq 1 - \frac{1}{k^2} = 1 - \frac{\sigma^2}{\epsilon^2 m}.$$

(iii) As $m \to \infty$, the quantity on the right in (ii) approaches 1, implying that the limit in (i) is greater than or equal to 1. Since probabilities must be between 0 and 1, the limit must be exactly 1.

5.2.1 Example: Monte Carlo evaluation of integrals

An interesting application of the law of large numbers is to the approximation of multiple integrals. For example, consider evaluating the double integral

$$\int_{x=0}^{1}\int_{y=0}^{2} e^{-xy}\, dy\, dx,$$

and let R be the region of integration (the $[0, 1] \times [0, 2]$ rectangle). Assume that (X, Y) has a bivariate uniform distribution on R, and let $W = e^{-XY}$. Since the area of R is 2, the expected value of W,

$$\mu = E(W) = E(e^{-XY}) = \int_{x=0}^{1}\int_{y=0}^{2} e^{-xy}\,\frac{1}{2}\, dy\, dx,$$

is exactly one-half the value of the integral above. For a given sample size n, the sample mean $\overline{W} = \frac{1}{n}\sum_{i=1}^{n} e^{-X_i Y_i}$ is an estimate of μ, and twice the sample mean is an estimate of the integral of interest.

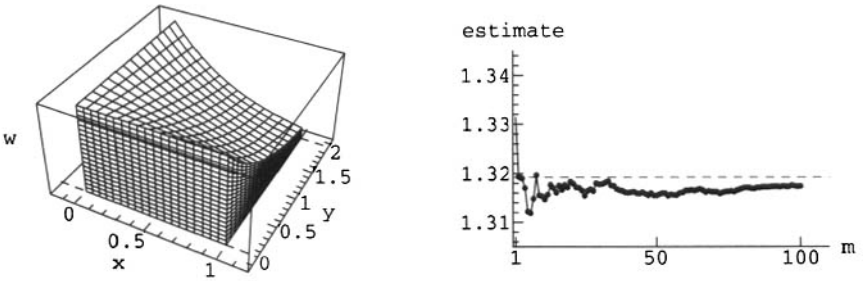


Figure 5.1. *The surface* $w = e^{-xy}$ *over the* $[0, 1] \times [0, 2]$ *rectangle (left plot) and Monte Carlo estimates of the volume under the surface using samples of size* $1000m$ *for* $m = 1, 2, \ldots, 100$ *(right plot). A horizontal dashed line is drawn in the right plot at the numerical value of the volume.*

The left plot in Figure 5.1 shows the surface $w = e^{-xy}$ over R. The right plot shows an observed sequence of running estimates

$$2\overline{w}_{1000m} \text{ for } m = 1, 2, \ldots, 100$$

based on simulations from the joint (X, Y) distribution. The final estimate of 1.31743 is based on a sample of size 100,000. Each estimate is called a Monte Carlo estimate of the double integral. (In a *Monte Carlo* analysis, simulation is used to estimate a quantity of interest.)

Monte Carlo methods are used often in statistical applications. Chapters 11 through 15 contain many examples of these methods.

5.3 Central limit theorem

The most important theorem in a probability course is the central limit theorem. Its proof is attributed to P. Laplace and A. de Moivre.

Theorem 5.2 (Central Limit Theorem). *Let* X *be a random variable with finite mean* μ *and standard deviation* σ. *Let* S_m *and* $\overline{X}_m$, $m = 1, 2, 3, \ldots$, *be the sequence of running sums and averages, respectively, and let*

$$Z_m = \frac{S_m - m\mu}{\sqrt{m\sigma^2}} = \frac{\overline{X}_m - \mu}{\sqrt{\sigma^2/m}}, \quad m = 1, 2, 3, \ldots,$$

be the sequence of standardized sums (or averages). Then for each real number x,

$$\lim_{m\to\infty} P(Z_m \le x) = \Phi(x),$$

where $\Phi(\cdot)$ *is the CDF of the standard normal random variable.*

For distributions with finite mean and standard deviation, the central limit theorem implies that the distributions of the sample sum and the sample mean are approximately normal when the sample size n is large enough.

In statistical applications, the central limit theorem is used to answer questions about an unknown mean μ. See Chapters 7 and 8.

5.3.1 Continuity correction

Let X be a discrete random variable with values in the integers and with finite mean μ and standard deviation σ, and let S be the sample sum of a random sample of size n from the X distribution. The normal approximation to the distribution of the sample sum can be improved using the *correction for continuity*,

$$P(S = x) \approx P(x - 0.50 \leq N \leq x + 0.50),$$

where N is the normal random variable with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.

For example, let X be the discrete random variable with PDF

x	1	2	3	4
$f(x)$	0.10	0.20	0.30	0.40

Then $E(X) = 3$ and $Var(X) = 1$. If S is the sample sum of a random sample of size 50 from the X distribution, then S takes integer values in the range $50, 51, \ldots, 200$, and its summary measures are

$$E(S) = 50E(X) = 150 \text{ and } Var(S) = 50Var(X) = 50.$$

By the central limit theorem (with continuity correction),

$$P(120 \leq S \leq 160) \approx P(119.5 \leq N \leq 160.5) \approx 0.9306,$$

where N is a normal random variable with mean 150 and standard deviation $\sqrt{50}$.

5.3.2 Special cases

The central limit theorem implies that the distributions of certain binomial, negative binomial, Poisson, and gamma random variables can be approximated by normal distributions. Specifically, the following hold:

1. *Binomial distribution.* Let X be a binomial random variable with parameters n and p. If n is large, then the distribution of X is approximately normal.

2. *Negative binomial distribution.* Let X be a negative binomial random variable with parameters r and p. If r is large, then the distribution of X is approximately normal.

3. *Poisson distribution.* Let X be a Poisson random variable with parameter λ. If λ is large, then the distribution of X is approximately normal.

4. *Gamma distribution.* Let X be a gamma random variable with parameters α and β. If α is large, then the distribution of X is approximately normal.

In each case, the random variable X can be written as a sample sum. In the first three cases, the correction for continuity can be used to improve the normal approximation.

5.4 Moment generating functions

If X is a random variable and k is a nonnegative integer, then $E(X^k)$ is known as the k^{th} *moment* of the random variable. The notation $\mu_k = E(X^k)$ is used to denote the k^{th} moment.

For a given X, $\mu_0 = 1$, μ_1 is the mean of X, and $\mu_2 - \mu_1^2$ is the variance of X. The value of μ_3 is related to the *skewness* (lack of symmetry) of the X distribution; the value of μ_4 is related to the *kurtosis* (peakedness) of the X distribution; etc.

The higher-order moments of X are similar in application to the higher-order derivatives of a function $y = f(x)$ at $x = a$.

The sequence of summary measures μ_k, $k = 0, 1, 2, \ldots$, can often be obtained quickly using moment generating functions. The *moment generating function* of X, where it exists, is defined as follows:

$$\text{MGF}(t) = E(e^{tX}).$$

For example, the moment generating function of a binomial random variable is

$$\text{MGF}(t) = \sum_{i=0}^{n} \binom{n}{i} e^{ti} p^i (1-p)^{n-i} = \left(pe^t + (1-p)\right)^n.$$

Similarly, the moment generating function of a normal random variable is

$$\text{MGF}(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{(x-\mu)^2}{\sigma^2}} dx = e^{\mu t + \frac{1}{2}\sigma^2 t^2}.$$

Note that, in general, MGF(0) equals 1. For binomial and normal random variables MGF(t) exists (has a finite value) for all real numbers t.

Theorem 5.3 (Moment Generating Function). *Let* MGF(t) *be the moment generating function of the random variable X. If* MGF(t) *exists for all t in an open interval containing $t = 0$, then the k^{th} moment of X is equal to the k^{th} derivative of* MGF(t) *at $t = 0$:*

$$\text{MGF}^{(k)}(0) = E(X^k) = \mu_k$$

for $k = 0, 1, 2, \ldots$.

An outline of the proof of Theorem 5.3 when X takes values in the nonnegative integers is as follows:

(i) Let $p_i = P(X = i)$. Then

$$\text{MGF}(t) = p_0 + e^t p_1 + e^{2t} p_2 + \cdots = \sum_{i=0}^{\infty} e^{it} p_i.$$

(ii) If $k = 0$, then $\text{MGF}^{(0)}(t) = \text{MGF}(t)$ and

$$\text{MGF}(0) = p_0 + p_1 + p_2 + \cdots = 1 = \mu_0.$$

(iii) If $k > 0$, then $\text{MGF}^{(k)}(t) = \sum_{i=1}^{\infty} i^k e^{it} p_i$ and

$$\text{MGF}^{(k)}(0) = 1p_1 + 2^k p_2 + 3^k p_3 + \cdots = E(X^k) = \mu_k.$$

Note that the existence of the moment generating function in an open interval containing $t = 0$ ensures the existence of the sequence of summary measures.

For example, consider the moment generating function of the standard normal random variable: $\text{MGF}(t) = e^{t^2/2}$. The first two derivatives of $\text{MGF}(t)$ are

$$\text{MGF}'(t) = e^{t^2/2}\, t \quad \text{and} \quad \text{MGF}''(t) = e^{t^2/2} + e^{t^2/2}\, t^2,$$

and the evaluations when $t = 0$ are

$$\text{MGF}'(0) = 0 = E(X) \text{ and } \text{MGF}''(0) = 1 = E(X^2).$$

The following theorems can be proven using properties of expectation.

Theorem 5.4 (Linear Functions). *Let* $\text{MGF}_1(t)$ *be the moment generating function of* X *and* $\text{MGF}_2(t)$ *be the moment generating function of* Y. *If* $Y = aX + b$, *where* $a \neq 0$ *and* b *are constants, then*

$$\text{MGF}_2(t) = e^{bt}\text{MGF}_1(at).$$

Theorem 5.5 (Convolution Theorem). *Let* $\text{MGF}_1(t)$ *be the moment generating function of* X_1 *and* $\text{MGF}_2(t)$ *be the moment generating function of* X_2. *If* X_1 *and* X_2 *are independent, then the moment generating function of* $W = X_1 + X_2$ *is*

$$\text{MGF}(t) = \text{MGF}_1(t)\text{MGF}_2(t).$$

Corollary 5.6. *More generally, if* $X_1, X_2, \ldots, X_n$ *are mutually independent random variables and* W *is their sum, then the moment generating function of* W *is*

$$\text{MGF}(t) = \text{MGF}_1(t)\text{MGF}_2(t)\cdots\text{MGF}_n(t),$$

where $\text{MGF}_i(t)$ *is the moment generating function of* X_i.

Finally, when the moment generating function exists in an open interval containing $t = 0$, it uniquely defines the distribution of the random variable.

Theorem 5.7 (Uniqueness Theorem). *Let* $\text{MGF}_1(t)$ *be the moment generating function of* X_1 *and* $\text{MGF}_2(t)$ *be the moment generating function of* X_2. *If* $\text{MGF}_1(t) = \text{MGF}_2(t)$ *in an open interval containing* $t = 0$, *then* X_1 *and* X_2 *have the same probability distribution.*

5.4.1 Method of moment generating functions

Moment generating functions can be used to determine the distribution of a random variable. For example, the moment generating function of a Poisson random variable with parameter λ is

$$\mathrm{MGF}(t) = \sum_{x=0}^{\infty} e^{tx} e^{-\lambda} \frac{\lambda^x}{x!} = \exp\left(\lambda\left(e^t - 1\right)\right),$$

where exp() is the exponential function. If X and Y are independent Poisson random variables with parameters λ_1 and λ_2, respectively, and $W = X + Y$ is their sum, then the convolution theorem (Theorem 5.5) implies that the moment generating function of W is

$$\mathrm{MGF}(t) = \exp\left(\lambda_1\left(e^t - 1\right)\right)\exp\left(\lambda_2\left(e^t - 1\right)\right) = \exp\left(\left(\lambda_1 + \lambda_2\right)\left(e^t - 1\right)\right).$$

Since the form of $\mathrm{MGF}(t)$ is the same as the form of the moment generating function of a Poisson random variable with parameter $\lambda_1 + \lambda_2$ and $\mathrm{MGF}(t)$ exists for all real numbers t, the uniqueness theorem (Theorem 5.7) implies that W has a Poisson distribution with parameter $\lambda_1 + \lambda_2$.

The method of moment generating functions can be used to prove the last two special cases in Section 5.3.2.

5.4.2 Relationship to the central limit theorem

It is possible to sketch a proof of the central limit theorem (Theorem 5.2) when the moment generating function of X exists in an open interval containing $t = 0$. Let Z_m be the standardized form of the sum S_m,

$$Z_m = \frac{S_m - m\mu}{\sigma\sqrt{m}} = \sum_{i=1}^{m} \frac{X_i - \mu}{\sigma\sqrt{m}},$$

and $W_i = (X_i - \mu)/(\sigma\sqrt{m})$ for $i = 1, 2, \ldots, m$. Then $E(Z_m) = 0$, $Var(Z_m) = 1$, and, for each i, $E(W_i) = 0$ and $Var(W_i) = E(W_i^2) = 1/m$.

If $\mathrm{MGF}_m(t)$ is the moment generating function of Z_m and $\mathrm{MGF}(t)$ is the moment generating function of each W_i, then by Corollary 5.6

$$\mathrm{MGF}_m(t) = (\mathrm{MGF}(t))^m = \left(1 + \frac{1}{2m}t^2 + \cdots\right)^m,$$

where the expression in parentheses on the right is the Maclaurin series expansion of $\mathrm{MGF}(t)$. For values of t near zero, it can be shown that

$$\lim_{m\to\infty} \mathrm{MGF}_m(t) = \lim_{m\to\infty}\left(1 + \frac{t^2/2}{m}\right)^m = e^{t^2/2}.$$

The formula on the right is the moment generating function of the standard normal random variable.

Finally, it can be shown that if the sequence of moment generating functions ($MGF_m(t)$) approaches the moment generating function of the standard normal random variable for values of t in an open interval containing 0, then the sequence of cumulative distribution functions must approach the cumulative distribution function of the standard normal random variable.

Recall that the Maclaurin series for $f(x)$ is the Taylor expansion around $a = 0$:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k = f(0) + f'(0)x + \frac{f''(0)}{2} x^2 + \cdots.$$

In the application above, $f(0) = 1$, $f'(0) = 0$, and $f''(0) = 1/m$.

Recall that for each constant a

$$\lim_{m \to \infty} \left(1 + \frac{a}{m}\right)^m = e^a.$$

In the application above, the terms of the Maclaurin expansion of degree 3 or more are small enough to be "ignored" in the limit process; the remaining part has limit $e^{t^2/2}$.

5.5 Laboratory problems

Laboratory problems for this chapter introduce *Mathematica* commands for visualizing sequences of running sums and averages, and moment generating functions of sums. The problems are designed to reinforce ideas related to the limit theorems.

5.5.1 Laboratory: Sums and averages

In the main laboratory notebook (Problems 1 to 6), you are asked to generate and graph simulated sequences of running sums and running averages; compute exact and approximate probabilities for sample sums and sample means; and study errors in using the normal approximation to a discrete random variable. Uniform, exponential, Poisson, normal, and Cauchy models are used.

5.5.2 Additional problem notebooks

Problems 7 and 8 relate running sums to *random walks*. In Problem 7, uniform steps are used to model the ups and downs of the stock market. In Problem 8, theoretical and sample random walks in the plane are considered.

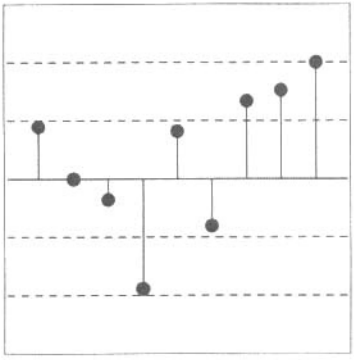
Problem 9 considers errors in using the normal approximation to the binomial distribution as n varies and as p varies.

Problems 10 and 11 use probability generating functions to find exact probabilities for sample sums and compare exact probabilities with normal approximations. Problem 10 uses a poker game setting. Problem 11 uses a coupon collecting setting.

In Problem 12, simulation is used to determine if the distributions of summary statistics other than the sample sum and sample mean are approximately normal when the sample size n is large. Exponential and uniform models are used.

Problems 13, 14, and 15 concern moment generating functions and sums. In each case, the functions are studied using both computational and graphical tools. Problem 13 uses a dice setting. Problem 14 uses a roulette game setting. Problem 15 considers the rate of convergence of the moment generating function of the standardized sum to the moment generating function of the standard normal random variable when X is Bernoulli, geometric, or exponential.

This page intentionally left blank



Chapter 6
Transition to Statistics

The normal distribution is the most widely used model in statistical applications. In the first two sections of this chapter, three families of distributions related to sampling from normal distributions are introduced. Applications of these three families will appear throughout the rest of the book. Section 3 is an informal introduction to one of these applications: the problem of testing the goodness-of-fit of a probability model to sample data. Section 4 outlines the laboratory problems.

6.1 Distributions related to the normal distribution

This section introduces three families of distributions related to the normal distribution and states properties of these distributions.

6.1.1 Chi-square distribution

Let $Z_1, Z_2, \ldots, Z_m$ be independent standard normal random variables. Then

$$V = Z_1^2 + Z_2^2 + \cdots + Z_m^2$$

is said to be a *chi-square random variable*, or to have a *chi-square distribution*, with parameter m. The PDF of V is as follows:

$$f(x) = \frac{1}{2^{m/2}\,\Gamma(m/2)} x^{(m/2)-1} e^{-x/2} \quad \text{when } x > 0 \text{ and 0 otherwise.}$$

The number of independent summands, m, is called the *degrees of freedom* (df) of the chi-square distribution. The notation χ_p^2 is used to denote the p^{th} quantile of the distribution.

The chi-square distribution with m degrees of freedom is the same as the gamma distribution with parameters $\alpha = m/2$ and $\beta = 1/2$.

A table of selected quantiles of chi-square distributions, suitable for doing problems by hand without using the computer, is given in Appendix B on the CD.

Properties of the chi-square distribution

Properties of the chi-square distribution are as follows:

1. If V is a chi-square random variable with m degrees of freedom, then $E(V) = m$ and $Var(V) = 2m$.
2. If m is large, then, by the central limit theorem, the distribution of V is approximately normal.
3. If V_1 and V_2 are independent chi-square random variables with m_1 and m_2 degrees of freedom, respectively, then the sum $V_1 + V_2$ has a chi-square distribution with $m_1 + m_2$ degrees of freedom.
4. If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ and standard deviation σ, then

$$V = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$

is a chi-square random variable with n degrees of freedom.

Recall that if X is a normal random variable with mean μ and standard deviation σ, then $Z = (X - \mu)/\sigma$ is a standard normal random variable. Thus, the random variable V given in property 4 is the sum of squares of n independent standard normal random variables.

6.1.2 Student t distribution

Assume that Z is a standard normal random variable, V is a chi-square random variable with m degrees of freedom, and Z and V are independent. Then

$$T = \frac{Z}{\sqrt{V/m}}$$

is said to be a *Student t random variable*, or to have a *Student t distribution*, with parameter m. The PDF of T is as follows:

$$f(x) = \frac{\Gamma((m+1)/2)}{\sqrt{m\pi}\,\Gamma(m/2)} \left(\frac{m}{m + x^2} \right)^{(m+1)/2} \quad \text{for all real numbers } x.$$

The parameter m is called the *degrees of freedom* (df) of the Student t distribution. The notation t_p is used to denote the p^{th} quantile of the distribution.

If T is a Student t random variable with 1 degree of freedom, then T has a Cauchy distribution with center $a = 0$ and scale $b = 1$.

A table of selected quantiles of Student t distributions, suitable for doing problems by hand without using the computer, is given in Appendix B on the CD.

Properties of the Student t distribution

Let T be a Student t random variable with m degrees of freedom. Then the following hold:

1. The distribution of T is symmetric around $x = 0$.
2. If $m > 1$, then $E(T) = 0$. If $m > 2$, then $Var(T) = m/(m-2)$.
3. If m is large, then the distribution of T is approximately standard normal.

6.1.3 F ratio distribution

Let U and V be independent chi-square random variables with n_1 and n_2 degrees of freedom, respectively. Then

$$F = \frac{U/n_1}{V/n_2}$$

is said to be an *f ratio random variable*, or to have an *f ratio distribution*, with parameters n_1 and n_2. The PDF of F is as follows:

$$f(x) = \frac{\Gamma((n_1+n_2)/2)}{\Gamma(n_1/2)\,\Gamma(n_2/2)} \left(\frac{n_1}{n_2}\right)^{n_1/2} \left(\frac{n_2}{n_2+n_1x}\right)^{(n_1+n_2)/2}$$

when $x > 0$ and 0 otherwise.

The parameters n_1 and n_2 are called the *degrees of freedom* (df) of the f ratio distribution. The notation f_p is used to denote the p^{th} quantile of the distribution.

A table of selected quantiles of f ratio distributions, suitable for doing problems by hand without using the computer, is given in Appendix B on the CD.

Properties of f ratio distributions

Let F be an f ratio random variable with n_1 and n_2 degrees of freedom. Then the following hold:

1. If $n_2 > 2$, then $E(F) = n_2/(n_2-2)$. If $n_2 > 4$, then
$$Var(F) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-4)(n_2-2)^2}.$$
2. The reciprocal of F, $1/F$, is an f ratio random variable with n_2 and n_1 degrees of freedom.

6.2 Random samples from normal distributions

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and standard deviation σ. Recall that the sample mean, $\overline{X}$, and sample variance, S^2, are the following random variables:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2.$$

In addition, the sample standard deviation, S, is the positive square root of the sample variance.

6.2.1 Sample mean, sample variance

The following distribution theorem states important properties of the sample mean and variance.

Theorem 6.1 (Distribution Theorem). *Let $\overline{X}$ be the sample mean and S^2 be the sample variance of a random sample of size n from a normal distribution with mean μ and standard deviation σ. Then the following hold:*

1. *$\overline{X}$ is a normal random variable with mean μ and standard deviation $\sqrt{\sigma^2/n}$.*
2. *$V = (n-1)S^2/\sigma^2$ is a chi-square random variable with $(n-1)$ degrees of freedom.*
3. *$\overline{X}$ and S^2 are independent random variables.*

The distribution theorem can be proven using moment generating functions. The first part of the theorem is a special case of Theorem 4.6.

Note that since $V = (n-1)S^2/\sigma^2$ is a chi-square random variable with $(n-1)$ degrees of freedom,

$$E(S^2) = \frac{\sigma^2}{(n-1)}\,E(V) = \sigma^2 \quad \text{and} \quad Var(S^2) = \frac{\sigma^4}{(n-1)^2}\,Var(V) = \frac{2\sigma^4}{(n-1)}.$$

Application: Interval estimation

Knowledge of the distribution of a sample summary is important in statistical applications. For example, suppose that the sample mean and sample variance of a random sample of size n from a normal distribution are used to estimate the unknown μ and σ^2. Let χ^2_p and χ^2_{1-p} be the p^{th} and $(1-p)^{th}$ quantiles of the chi-square distribution with $(n-1)$ degrees of freedom, respectively. Then

$$1 - 2p = P\left(\chi^2_p \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{1-p}\right) = P\left(\frac{(n-1)S^2}{\chi^2_{1-p}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_p}\right).$$

If the observed value of the sample variance is $s^2 = 8.72$, $n = 12$ and $p = 0.05$, then the interval

$$\left[\frac{11(8.72)}{19.68}, \frac{11(8.72)}{4.57}\right] = [4.87, 20.99]$$

is an estimate of an interval containing σ^2 with probability 0.90.

The estimated interval above is an example of a confidence interval for the variance. Confidence intervals for variances are introduced in Chapter 7.

6.2.2 Approximate standardization of the sample mean

Since the sample mean, $\overline{X}$, is a normal random variable with mean μ and standard deviation $\sqrt{\sigma^2/n}$, the standardized sample mean

$$Z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}}$$

is a standard normal random variable. An *approximation* is obtained by substituting the sample variance S^2 for the true variance σ^2.

Theorem 6.2 (Approximate Standardization). *Let $\overline{X}$ be the sample mean and S^2 be the sample variance of a random sample of size n from a normal distribution with mean μ and standard deviation σ. Then*

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}}$$

has a Student t distribution with $(n-1)$ degrees of freedom.

Note that, by Theorem 6.1, $Z = (\overline{X} - \mu)/\sqrt{\sigma^2/n}$ is a standard normal random variable, $V = (n-1)S^2/\sigma^2$ is a chi-square random variable with $(n-1)$ degrees of freedom, and Z and V are independent. Thus,

$$T = \frac{Z}{\sqrt{V/(n-1)}} = \frac{(\overline{X} - \mu)}{\sqrt{\sigma^2/n}} \frac{1}{\sqrt{S^2/\sigma^2}} = \frac{(\overline{X} - \mu)}{\sqrt{S^2/n}}$$

has a Student t distribution with $n-1$ degrees of freedom.

Application: Interval estimation

The distribution of the approximate standardization of the sample mean is important in statistical applications. For example, suppose that the sample mean and sample variance of a random sample of size n from a normal distribution are used to estimate the unknown μ and σ^2. Let t_p and t_{1-p} be the p^{th} and $(1-p)^{th}$ quantiles of the Student t distribution with $(n-1)$ degrees of freedom, respectively. Then

$$1 - 2p = P\left(t_p \le \frac{(\overline{X} - \mu)}{\sqrt{S^2/n}} \le t_{1-p}\right) = P\left(\overline{X} - t_{1-p}\sqrt{\frac{S^2}{n}} \le \mu \le \overline{X} - t_p\sqrt{\frac{S^2}{n}}\right).$$

If the observed values of the sample summaries are $\overline{x} = 36.5$ and $s^2 = 3.75$, $n = 15$, and $p = 0.05$, then the interval

$$\left[36.5 - 1.761\sqrt{\frac{3.75}{15}}, 36.5 + 1.761\sqrt{\frac{3.75}{15}}\right] = [35.62, 37.38]$$

is an estimate of an interval containing μ with probability 0.90.

The estimated interval above is an example of a confidence interval for the mean. Confidence intervals for means are introduced in Chapter 7.

6.2.3 Ratio of sample variances

Assume that

$$X_1, X_2, \ldots, X_n \text{ and } Y_1, Y_2, \ldots, Y_m$$

are independent random samples from normal distributions with parameters μ_x and σ_x, and μ_y and σ_y, respectively. Let $\overline{X}$ and S_x^2 be the sample mean and variance of the X sample, and let $\overline{Y}$ and S_y^2 be the sample mean and variance of the Y sample.

The ratio of sample variances, S_x^2/S_y^2, can be used to estimate σ_x^2/σ_y^2. Further, the following theorem holds.

Theorem 6.3 (Distribution Theorem). *Let S_x^2 and S_y^2 be the sample variances of independent random samples of sizes n and m, respectively, from normal distributions. Then*

$$F = \frac{S_x^2/S_y^2}{\sigma_x^2/\sigma_y^2}$$

has an f ratio distribution with $(n-1)$ and $(m-1)$ degrees of freedom.

Note that, by Theorem 6.1, $U = (n-1)S_x^2/\sigma_x^2$ and $V = (m-1)S_y^2/\sigma_y^2$ are independent chi-square random variables with $(n-1)$ and $(m-1)$ degrees of freedom, respectively. Thus,

$$F = \frac{U/(n-1)}{V/(m-1)} = \frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} = \frac{S_x^2/S_y^2}{\sigma_x^2/\sigma_y^2}$$

is an f ratio random variable with $(n-1)$ and $(m-1)$ degrees of freedom.

Application: Interval estimation

The distribution of the ratio of S_x^2/S_y^2 to σ_x^2/σ_y^2 is important in statistical applications. For example, suppose that all four parameters $(\mu_x, \sigma_x, \mu_y, \sigma_y)$ are unknown. Let f_p and f_{1-p} be the p^{th} and $(1-p)^{th}$ quantiles of the f ratio distribution with $(n-1)$ and $(m-1)$ degrees of freedom, respectively. Then

$$1 - 2p = P\left(f_p \le \frac{S_x^2/S_y^2}{\sigma_x^2/\sigma_y^2} \le f_{1-p}\right) = P\left(\frac{S_x^2/S_y^2}{f_{1-p}} \le \frac{\sigma_x^2}{\sigma_y^2} \le \frac{S_x^2/S_y^2}{f_p}\right).$$

If the observed sample variances are $s_x^2 = 18.75$ and $s_y^2 = 3.45$, $n = 8$, $m = 10$, and $p = 0.05$, then the interval

$$\left[\frac{18.75/3.45}{3.29}, \frac{18.75/3.45}{0.27}\right] = [1.65, 20.13]$$

is an estimate of an interval containing σ_x^2/σ_y^2 with probability 0.90.

The estimated interval above is an example of a confidence interval for the variance ratio. Confidence intervals for variance ratios are introduced in Chapter 10.

6.3 Multinomial experiments

A *multinomial experiment* is an experiment with exactly k outcomes. The probability of the i^{th} outcome is p_i, $i = 1, 2, \ldots, k$. The outcomes of a multinomial experiment are often referred to as *categories* or *groups*.

6.3.1 Multinomial distribution

Let X_i be the number of occurrences of the i^{th} outcome in n independent trials of a multinomial experiment, $i = 1, 2, \ldots, k$. Then the random k-tuple $(X_1, X_2, \ldots, X_k)$ is said to have a *multinomial distribution* with parameters n and $(p_1, p_2, \ldots, p_k)$. The joint PDF for the k-tuple is

$$f(x_1, x_2, \ldots, x_k) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

when $x_1, x_2, \ldots, x_k = 0, 1, \ldots, n$ and $\sum_i x_i = n$ (and zero otherwise).

The multinomial distribution generalizes the binomial and trinomial distributions. Specifically, the following hold:

1. If X is a binomial random variable with parameters n and p, then $(X, n - X)$ has a multinomial distribution with parameters n and $(p, 1 - p)$.
2. If (X, Y) has a trinomial distribution with parameters n and (p_1, p_2, p_3), then $(X, Y, n - X - Y)$ has a multinomial distribution with parameters n and (p_1, p_2, p_3).

Properties of the multinomial distribution

If $(X_1, X_2, \ldots, X_k)$ has a multinomial distribution, then the following hold:

1. For each i, X_i is a binomial random variable with parameters n and p_i.
2. For each $i \neq j$, (X_i, X_j) has a trinomial distribution with parameters n and $(p_i, p_j, 1 - p_i - p_j)$. In particular, X_i and X_j are negatively associated.

6.3.2 Goodness-of-fit: Known model

In 1900, K. Pearson developed a quantitative method to determine if observed data are consistent with a given multinomial model.

If $(X_1, X_2, \ldots, X_k)$ has a multinomial distribution with parameters n and $(p_1, p_2, \ldots, p_k)$, then *Pearson's statistic* is the following random variable:

$$\mathbf{X^2} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}.$$

For each i, the observed frequency, X_i, is compared to the expected frequency, $E(X_i) = np_i$, under the multinomial model. If each observed frequency is close to expected, then the value of $\mathbf{X^2}$ will be close to zero. If at least one observed frequency

is far from expected, then the value of $\mathbf{X}^2$ will be large and the appropriateness of the given multinomial model will be called into question. A test can be set up using the following distribution theorem.

Theorem 6.4 (Pearson's Theorem). *Under the assumptions above, if n is large, the distribution of* $\mathbf{X}^2$ *is approximately chi-square with* $(k-1)$ *degrees of freedom.*

Note that the chi-square approximation is adequate when $E(X_i) = np_i \geq 5$ for $i = 1, 2, \ldots, k$.

Pearson's goodness-of-fit test

For a given k-tuple, $(x_1, x_2, \ldots, x_k)$, let

$$\mathbf{x}^2_{\text{obs}} = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i}$$

be the observed value of Pearson's statistic. Use the chi-square approximation to the distribution of $\mathbf{X}^2$ to compute $P(\mathbf{X}^2 \geq \mathbf{x}^2_{\text{obs}})$. Then the following hold:

(i) If $P(\mathbf{X}^2 \geq \mathbf{x}^2_{\text{obs}}) > 0.10$, the fit is judged to be good (the observed data are judged to be consistent with the multinomial model).

(ii) If $0.05 < P(\mathbf{X}^2 \geq \mathbf{x}^2_{\text{obs}}) < 0.10$, the fit is judged to be fair (the observed data are judged to be marginally consistent with the model).

(iii) If $P(\mathbf{X}^2 \geq \mathbf{x}^2_{\text{obs}}) < 0.05$, the fit is judged to be poor (the observed data are judged to be not consistent with the model).

The probability $P(\mathbf{X}^2 \geq \mathbf{x}^2_{\text{obs}})$ is called the *p value* of the test. The p value measures the strength of the evidence against the given multinomial model.

Analysis of standardized residuals

For a given k-tuple, $(x_1, x_2, \ldots, x_k)$, the list of *standardized residuals*,

$$r_i = \frac{(x_i - np_i)}{\sqrt{np_i}}, \quad i = 1, 2, \ldots, k,$$

serve as diagnostic values for the goodness-of-fit test.

When n is large, the r_i's are approximate values from a standard normal distribution. Values outside the interval $[-2, +2]$ are considered to be unusual and deserve comment in a statistical analysis.

Example: Survey analysis

For example, assume that the table below gives age ranges for adults and approximate proportions in each age range according to the 1980 census.

Age Group	18–24	25–34	35–44	45–64	65+
1980 Proportion	0.18	0.23	0.16	0.27	0.16

Assume also that in a recent survey of 250 adults, there were 40, 52, 43, 59, and 56 individuals in ranges 18–24, 25–34, 35–44, 45–64, and 65+, respectively. Of interest is whether the recent survey results are consistent with the 1980 census model.

Observed and expected frequencies, and standardized residuals are as follows:

Observed Frequency	40	52	43	59	56
Expected Frequency	45	57.5	40	67.5	40
Standardized Residual	−0.745	−0.725	0.474	−1.035	2.530

The observed value of Pearson's statistic is 8.778 (the sum of squares of the standardized residuals) and the p value, based on the chi-square distribution with 4 degrees of freedom, is 0.067. The recent survey data are judged to be only marginally consistent with the 1980 census model. In particular, the observed number of adults in the 65+ group was much larger than expected.

Note that the analysis above assumes that the 250 individuals chosen for the survey are a simple random sample of adults in the United States. Since the total number of adults is quite large, a multinomial distribution can be used to analyze the results of the survey (generalizing ideas introduced in Sections 2.2.4 and 2.3.5).

Pearson's goodness-of-fit procedure is an example of a hypothesis test. Hypothesis tests are studied in detail in Chapter 8.

6.3.3 Goodness-of-fit: Estimated model

In many practical situations, certain parameters of the multinomial model need to be estimated from the sample data. R. A. Fisher proved a generalization of Theorem 6.4 to handle this case.

Theorem 6.5 (Fisher's Theorem). *Assume that* $(X_1, X_2, \ldots, X_k)$ *has a multinomial distribution with parameters n and* $(p_1, p_2, \ldots, p_k)$*, and that the list of probabilities has e free parameters. Then, under smoothness conditions and when n is large, the distribution of the statistic*

$$\mathbf{X}^2 = \sum_{i=1}^{k} \frac{(X_i - n\widehat{p}_i)^2}{n\widehat{p}_i}$$

is approximately chi-square with $(k - 1 - e)$ *degrees of freedom, where* $\widehat{p}_i$ *is an appropriate estimate of* p_i*, for* $i = 1, 2, \ldots, k$.

The smoothness conditions mentioned in the theorem, and methods for estimating free parameters in models, are studied in detail in Chapter 7. The method of *minimum chi-square* is often used to estimate the free parameters in goodness-of-fit problems (see laboratory Problems 10 and 11).

Pearson's goodness-of-fit test is conducted in the same way as before, with estimated expected frequencies taking the place of expected frequencies and with $(k - 1 - e)$ degrees of freedom taking the place of $(k - 1)$ degrees of freedom.

Table 6.1. *Goodness-of-fit analysis of IQ scores data.*

Event	Observed Frequency	Expected Frequency	Standardized Residual	Component of $\mathbf{X^2}$
$IQ < 77.63$	5	5.6	−0.25	0.06
$77.63 \leq IQ < 83.55$	7	5.6	0.59	0.35
$83.55 \leq IQ < 87.55$	6	5.6	0.17	0.03
$87.55 \leq IQ < 90.73$	5	5.6	−0.25	0.06
$90.73 \leq IQ < 93.45$	6	5.6	0.17	0.03
$93.45 \leq IQ < 95.90$	2	5.6	−1.52	2.31
$95.90 \leq IQ < 98.17$	9	5.6	1.44	2.06
$98.17 \leq IQ < 100.32$	4	5.6	−0.68	0.46
$100.32 \leq IQ < 102.41$	7	5.6	0.59	0.35
$102.41 \leq IQ < 104.46$	3	5.6	−1.10	1.21
$104.46 \leq IQ < 106.50$	4	5.6	−0.68	0.46
$106.50 \leq IQ < 108.59$	6	5.6	0.17	0.03
$108.59 \leq IQ < 110.74$	9	5.6	1.44	2.06
$110.74 \leq IQ < 113.01$	8	5.6	1.01	1.03
$113.01 \leq IQ < 115.46$	6	5.6	0.17	0.03
$115.46 \leq IQ < 118.18$	3	5.6	−1.10	1.21
$118.18 \leq IQ < 121.36$	8	5.6	1.01	1.03
$121.36 \leq IQ < 125.36$	5	5.6	−0.25	0.06
$125.36 \leq IQ < 131.28$	5	5.6	−0.25	0.06
$IQ \geq 131.28$	4	5.6	−0.68	0.46

Example: Analysis of IQ scores

A study was conducted using the Stanford–Binet intelligence scale to determine the intelligence quotients (IQ scores) of children in five kindergarten classes in San Jose and San Mateo, California [106], [80, p. 387]. There were 112 children (64 boys and 48 girls), ranging in age from 3.5 to 7 years old. The majority of the kindergarteners were from the middle class, and all were native born. A sample mean of $\bar{x} = 104.455$ and a sample standard deviation of $s = 16.3105$ were observed. Of interest was whether a normal distribution could be used to model IQ scores.

Let X be the IQ score of a randomly chosen kindergarten student. To obtain a multinomial model, the observations are grouped as follows:

$$X < x_{0.05},\ x_{0.05} \leq X < x_{0.10},\ \ldots,\ x_{0.90} \leq X < x_{0.95},\ X \geq x_{0.95},$$

where x_p is the p^{th} quantile of the normal distribution with mean 104.455 and standard deviation 16.3105. The multinomial model has 20 equally likely outcomes ($\widehat{p}_i = 0.05$ for $i = 1, 2, \ldots, 20$); two free parameters have been estimated.

Table 6.1 summarizes the important information needed in the analysis. The observed value of Pearson's statistic is 13.3571, and the p value, based on the chi-square distribution with 17 degrees of freedom, is 0.712. The IQ scores data

are judged to be consistent with a normal distribution. There are no unusual standardized residuals.

Note that the analysis above assumes that the 112 children chosen for the study are a simple random sample of kindergarten children in the United States and that the total number of kindergarten children is large enough that a multinomial model can be used to analyze the results of the experiment.

A standard rule of thumb for using a multinomial approximation is that the number of individuals in the simple random sample is less than 5% of the total population size.

6.4 Laboratory problems

Laboratory problems for this chapter introduce *Mathematica* commands for working with chi-square and multinomial models, and for conducting goodness-of-fit tests. The problems are designed to reinforce the ideas of this chapter.

6.4.1 Laboratory: Transition to statistics

In the main laboratory notebook (Problems 1 to 6), you will use simulation to study chi-square distributions and multinomial distributions; compute probabilities in multinomial models; use simulation to study Pearson's goodness-of-fit procedure; and apply the goodness-of-fit method to data on major coal mining accidents in Great Britain in the nineteenth century [71], [57].

6.4.2 Additional problem notebooks

Problems 7 through 13 are applications of Pearson's goodness-of-fit method. Problem 7 uses data on computer-shuffled and hand-shuffled bridge hands [13], [33]; of interest is whether the data are consistent with a model for well-shuffled decks. Problem 8 uses data on the numbers of boys and girls in German families with exactly 12 children [46], [90]; of interest is whether the data are consistent with a binomial model. Problem 9 uses data on radioactive decay [93], [50]; of interest is whether the data are consistent with a Poisson model.

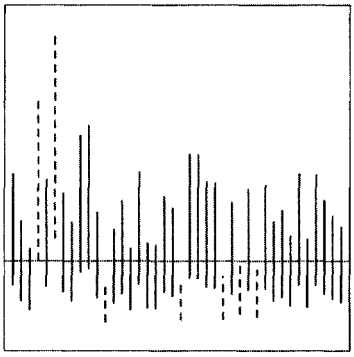
Problems 10 and 11 introduce the method of minimum chi-square. Problem 10 applies the method to data from a memory study [109], [48]. Problem 11 applies the method to data from a genetics study [24], [65].

Applications of gamma distributions are presented in Problems 12 and 13. In each case, sample values are used to estimate the parameters of the gamma distribution. Problem 12 uses data on the lifetimes of machine components subjected to repeated alternating stress [15]. Problem 13 uses data on rainfall amounts [66], [90].

In Problem 14, simulation is used to determine if the distribution of the random variable $(n-1)S^2/\sigma^2$ is approximately chi-square when sampling from distributions other than the normal distribution. Exponential, gamma, and uniform models are used.

In Problem 15, simulation is used to study the quality of the chi-square approximation to the distribution of Pearson's statistic when some expected frequencies are small.

This page intentionally left blank



Chapter 7
Estimation Theory

Statistical inference, which includes estimation theory and hypothesis testing theory, refers to a broad collection of methods for analyzing random samples from probability distributions. If the family of distributions from which the data were drawn is known except for the values of one or more parameters, then estimation theory can be used to make probabilistic statements about the unknown parameters.

This chapter introduces estimation theory. The first three sections give important definitions and examples. Method of moments estimation and maximum likelihood estimation are introduced in Sections 4 and 5, respectively. Section 6 outlines the laboratory problems.

7.1 Definitions

Recall that a random sample of size n is a list of n mutually independent random variables, each with the same probability distribution.

A *statistic* is a function of one or more random samples. The probability distribution of a statistic is known as its *sampling distribution*.

An *estimator* (or *point estimator*) is a statistic used to estimate an unknown parameter. An *estimate* is the value of an estimator for a given set of data.

Example: Sampling from a normal distribution

For example, let $\overline{X}$ and S^2 be the sample mean and sample variance of a random sample of size n from a normal distribution.

Then $\overline{X}$ is an estimator of μ and S^2 is an estimator of σ^2. By Theorem 6.1, the sampling distribution of $\overline{X}$ is normal with parameters μ and $\sigma/\sqrt{n}$, and the sampling distribution of $(n-1)S^2/\sigma^2$ is chi-square with $(n-1)$ degrees of freedom.

Further, if the numbers 90.8, 98.0, 113.0, 134.7, 80.5, 97.6, 117.6, 119.9 are observed, then an estimate of μ is $\overline{x} = 106.513$ and an estimate of σ^2 is $s^2 = 316.316$.

7.2 Properties of point estimators

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with parameter θ. The notation

$$\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$$

is often used to denote an estimator of θ. ($\widehat{\theta}$ is a function of the random sample, although the arguments are often suppressed.)

7.2.1 Bias; unbiased estimator

The *bias* of the estimator $\widehat{\theta}$ is the difference between the expected value of the estimator and the true parameter:

$$BIAS(\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$

If $E(\widehat{\theta}) = \theta$, then $\widehat{\theta}$ is said to be an *unbiased* estimator of θ; otherwise, $\widehat{\theta}$ is said to be a *biased* estimator of θ.

For example, let $\overline{X}$, S^2, and S be the sample mean, sample variance, and sample standard deviation of a random sample of size n from a normal distribution. Since $E(\overline{X}) = \mu$, $\overline{X}$ is an unbiased estimator of μ. Since $E(S^2) = \sigma^2$, S^2 is an unbiased estimator of σ^2. Since

$$E(S) = \sigma \sqrt{\frac{2}{n-1}} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \neq \sigma,$$

S is a biased estimator of σ.

Asymptotically unbiased estimator

The estimator $\widehat{\theta}$ is said to be *asymptotically unbiased* if $\lim_{n\to\infty} E(\widehat{\theta}) = \theta$. For example, the sample standard deviation of a random sample from a normal distribution is an asymptotically unbiased estimator of σ.

7.2.2 Efficiency for unbiased estimators

Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two unbiased estimators of θ, each based on a random sample of size n from the X distribution. $\widehat{\theta}_1$ is said to be *more efficient* than $\widehat{\theta}_2$ if

$$Var(\widehat{\theta}_1) < Var(\widehat{\theta}_2).$$

Given two unbiased estimators, we would prefer to use the more efficient one.

For example, let X_1, X_2, X_3, X_4 be a random sample of size 4 from a distribution with mean μ and standard deviation σ. Consider two estimators of μ:

$$\widehat{\mu}_1 = \frac{1}{2}X_1 + \frac{1}{6}X_2 + \frac{1}{6}X_3 + \frac{1}{6}X_4 \quad \text{and} \quad \widehat{\mu}_2 = \frac{1}{4}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_3 + \frac{1}{4}X_4.$$

Each statistic is an unbiased estimator of μ. Since $Var(\widehat{\mu}_1) = \sigma^2/3$ and $Var(\widehat{\mu}_2) = \sigma^2/4$, $\widehat{\mu}_2$ is more efficient then $\widehat{\mu}_1$. Note that $\widehat{\mu}_2$ is the sample mean.

MVUE estimator

The unbiased estimator $\widehat{\theta}$ is called a *minimum variance unbiased estimator* (MVUE) of θ if it has the minimum variance among all unbiased estimators of θ.

An interesting and difficult problem in the field of statistics is that of determining when an MVUE exists. Criteria for the existence of an MVUE are given in Section 7.5.2.

7.2.3 Mean squared error

The *mean squared error* (MSE) of an estimator is the expected value of the square of the difference between the estimator and the true parameter:

$$MSE(\widehat{\theta}) = E\left((\widehat{\theta} - \theta)^2\right).$$

Properties of expectation can be used to show that

$$MSE(\widehat{\theta}) = Var(\widehat{\theta}) + \left(BIAS(\widehat{\theta})\right)^2.$$

Thus, in particular, if $\widehat{\theta}$ is an unbiased estimator of θ, then $MSE(\widehat{\theta}) = Var(\widehat{\theta})$.

Efficiency

Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two (not necessarily unbiased) estimators of θ, each based on a random sample of size n from the X distribution. $\widehat{\theta}_1$ is said to be *more efficient* than $\widehat{\theta}_2$ if

$$MSE(\widehat{\theta}_1) < MSE(\widehat{\theta}_2).$$

MSE is the average squared distance between $\widehat{\theta}$ and θ. It is possible for a biased estimator of θ to have a smaller MSE than an unbiased estimator. Based on the mean squared error criterion, the biased estimator would be preferred to the unbiased estimator.

7.2.4 Consistency

A consistent estimator of θ is one that is unlikely to be far from θ when the sample size n is large. The formal definition is as follows.

The estimator $\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$ is said to be a *consistent* estimator of θ if, for every positive number ϵ,

$$\lim_{n\to\infty} P\left(|\widehat{\theta} - \theta| \geq \epsilon\right) = 0.$$

Consistent estimators are not necessarily unbiased but are generally asymptotically unbiased. For unbiased estimators, the following theorem gives a criterion for consistency.

Theorem 7.1 (Consistency Theorem). *If $\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$ is an unbiased estimator of θ based on a random sample of size n from the X distribution and*

$$\lim_{n\to\infty} Var(\widehat{\theta}) = 0,$$

then $\widehat{\theta}$ is a consistent estimator of θ.

For example, the sample mean and sample variance of a random sample from a normal distribution are consistent estimators of μ and σ^2, respectively.

Note that the law of large numbers (Theorem 5.1) is a special case of the consistency theorem. Like the law of large numbers, the consistency theorem can be proven using Chebyshev's inequality (Theorem 4.1). Further, the consistency theorem can be extended to include asymptotically unbiased estimators.

7.3 Interval estimation

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with parameter θ. The goal in interval estimation is to find two statistics

$$L = L(X_1, X_2, \ldots, X_n) \quad \text{and} \quad U = U(X_1, X_2, \ldots, X_n)$$

with the property that θ lies in the interval $[L, U]$ with high probability.

It is customary to let α (the *error probability*) denote the probability that θ is not in the interval and to find statistics L and U satisfying

$$P(\theta < L) = P(\theta > U) = \frac{\alpha}{2} \text{ and } P(L \leq \theta \leq U) = 1 - \alpha.$$

The probability $(1 - \alpha)$ is called the *confidence coefficient*, and the interval $[L, U]$ is called a $100(1 - \alpha)\%$ *confidence interval* for θ.

Sample data are used to estimate the lower (L) and upper (U) endpoints of the confidence interval $[L, U]$. $100(1 - \alpha)\%$ of the estimated intervals will contain θ.

7.3.1 Example: Normal distribution

Let $\overline{X}$ be the sample mean and S^2 be the sample variance of a random sample of size n from a normal distribution with mean μ and standard deviation σ.

Confidence intervals for μ

If the value of σ^2 is known, then Theorem 6.1 can be used to demonstrate that

$$\overline{X} \pm z(\alpha/2)\sqrt{\frac{\sigma^2}{n}}$$

is a $100(1 - \alpha)\%$ confidence interval for μ, where $z(\alpha/2)$ is the $100(1 - \alpha/2)\%$ point of the standard normal distribution.

If the value of σ^2 is estimated from the data, then Theorem 6.2 can be used to demonstrate that

$$\overline{X} \pm t_{n-1}(\alpha/2)\sqrt{\frac{S^2}{n}}$$

is a $100(1 - \alpha)\%$ confidence interval for μ, where $t_{n-1}(\alpha/2)$ is the $100(1 - \alpha/2)\%$ point of the Student t distribution with $(n - 1)$ degrees of freedom.

For example, if $n = 12$, $\alpha = 0.10$, $\bar{x} = 81.282$ is used to estimate μ and $s^2 = 11.833$ is used to estimate σ^2, then

$$81.282 \pm 1.796\sqrt{\frac{11.833}{12}} \quad \Longrightarrow \quad [79.499, 83.066]$$

is a 90% confidence interval for μ.

Confidence intervals for σ^2

If the value of μ is known, then

$$\left[\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi_n^2(\alpha/2)}, \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi_n^2(1-\alpha/2)}\right]$$

is a $100(1-\alpha)\%$ confidence interval for σ^2, where $\chi_n^2(p)$ is the $100(1-p)\%$ point of the chi-square distribution with n degrees of freedom.

If the value of μ is estimated from the data, then Theorem 6.1 can be used to demonstrate that

$$\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\chi_{n-1}^2(\alpha/2)}, \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\chi_{n-1}^2(1-\alpha/2)}\right]$$

is a $100(1-\alpha)\%$ confidence interval for σ^2, where $\chi_{n-1}^2(p)$ is the $100(1-p)\%$ point of the chi-square distribution with $(n-1)$ degrees of freedom.

Confidence intervals for σ

If $[L, U]$ is a $100(1-\alpha)\%$ confidence interval for σ^2, then $[\sqrt{L}, \sqrt{U}]$ is a $100(1-\alpha)\%$ confidence interval for σ.

For example, if $n = 12$, $\alpha = 0.10$, $\bar{x} = 81.282$ is used to estimate μ and $s^2 = 11.833$ is used to estimate σ^2, then

$$\left[\frac{11(11.833)}{19.68}, \frac{11(11.833)}{4.57}\right] = [6.614, 28.482]$$

and $[2.572, 5.337]$ are 90% confidence intervals for σ^2 and σ, respectively.

7.3.2 Approximate intervals for means

Let $\overline{X}$ be the sample mean and S^2 be the sample variance of a random sample of size n from a distribution with unknown (but finite) mean and variance. If n is large, then the central limit theorem (Theorem 5.2) can be used to demonstrate that

$$\overline{X} \pm z(\alpha/2)\sqrt{\frac{S^2}{n}}$$

is an *approximate* $100(1-\alpha)\%$ confidence interval for μ, where $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

This large sample method for μ is useful because it does not require precise knowledge of the X distribution. Approximate methods for other parameters are given in Section 7.5.3.

7.4 Method of moments estimation

The *method of moments* (MOM), introduced by K. Pearson in the 1880's, is a general method for estimating one or more unknown parameters. In general, MOM estimators are consistent but are not necessarily unbiased.

Recall that $\mu_k = E\left(X^k\right)$ is called the k^{th} *moment* of the X distribution for $k = 1, 2, \ldots$. The k^{th} *sample moment* is the random variable

$$\widehat{\mu_k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k \quad \text{for } k = 1, 2, 3, \ldots,$$

where $X_1, X_2, \ldots, X_n$ is a random sample of size n from the X distribution. For example, if the numbers 1.57, −2.41, 2.42, 0.80, 4.20, −2.97 are observed, then 0.602 is the first sample moment (the sample mean), 6.872 is the second sample moment, and 8.741 is the third sample moment.

Note that for each k, the k^{th} sample moment is an unbiased estimator of μ_k.

7.4.1 Single parameter estimation

If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with parameter θ, and $\mu_k = E\left(X^k\right)$ is a function of θ for some k, then a *method of moments* estimator (or MOM estimator) of θ is obtained using the following procedure:

$$\text{Solve } \mu_k = \widehat{\mu_k} \quad \text{for the parameter } \theta.$$

For example, let X be a uniform random variable on the interval $[0, b]$, and assume that $b > 0$ is unknown. Since $E(X) = b/2$, a MOM estimator is obtained as follows:

$$\mu_1 = \widehat{\mu_1} \implies E(X) = \frac{1}{n}\sum_{i=1}^{n} X_i \implies \frac{b}{2} = \overline{X} \implies \widehat{b} = 2\overline{X}.$$

Suppose instead that X is a uniform random variable on the interval $[-b, b]$ and that $b > 0$ is unknown. Since $E(X) = 0$ and $E(X^2) = Var(X) = b^2/3$, a MOM estimator is obtained as follows:

$$\mu_2 = \widehat{\mu_2} \implies E\left(X^2\right) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \implies \frac{b^2}{3} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \implies \widehat{b} = \sqrt{\frac{3}{n}\sum_{i=1}^{n} X_i^2}.$$

7.4.2 Multiple parameter estimation

The procedure can be generalized to any number of unknown parameters. For example, if the X distribution has two unknown parameters (say θ_1 and θ_2), then MOM estimators are obtained using the procedure

$$\text{Solve } \mu_{k_1} = \widehat{\mu_{k_1}} \quad \text{and} \quad \mu_{k_2} = \widehat{\mu_{k_2}} \text{ simultaneously for } \theta_1 \text{ and } \theta_2$$

for appropriately chosen k_1 and k_2.

Example: Normal distribution

For example, let X be a normal random variable with unknown mean μ and variance σ^2. Since $E(X) = \mu$ and $E(X^2) = Var(X) + (E(X))^2 = \sigma^2 + \mu^2$, MOM estimators are obtained by solving two equations in two unknowns:

$$\mu_1 = \widehat{\mu_1},\ \mu_2 = \widehat{\mu_2} \implies \widehat{\mu} = \overline{X},\ \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2.$$

Note that $\widehat{\mu}$ is an unbiased estimator of μ, but $\widehat{\sigma^2}$ is a biased estimator of σ^2.

7.5 Maximum likelihood estimation

The method of *maximum likelihood* (ML), introduced by R.A. Fisher in the 1920's, is a general method for estimating one or more unknown parameters. In general, ML estimators are consistent but are not necessarily unbiased.

7.5.1 Single parameter estimation

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with parameter θ and PDF $f(x)$. The *likelihood function* is the joint PDF of the random sample thought of as a function of θ:

$$Lik(\theta) = \prod_{i=1}^{n} f(X_i).$$

The *log-likelihood function* is the natural logarithm of the likelihood function:

$$\ell(\theta) = \log(Lik(\theta)).$$

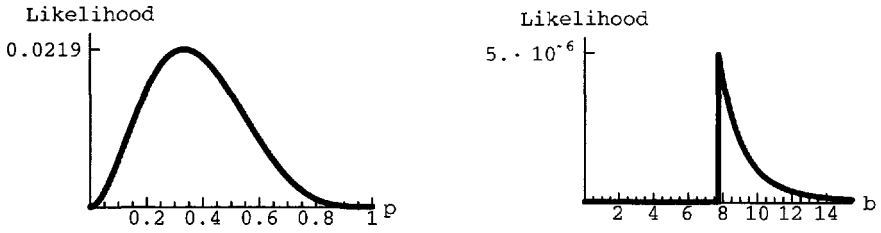


Figure 7.1. *Likelihood functions for samples from a Bernoulli distribution (left plot) and from a uniform distribution (right plot).*

Note that the symbol log is used to represent the natural logarithm function, not the common logarithm function.

The *maximum likelihood* estimator (or ML estimator) of θ is the value that maximizes the likelihood function (or the log-likelihood function).

The ML estimator is the value of θ that maximizes the likelihood of the observed sample.

Example: Bernoulli distribution

Let X be a Bernoulli random variable with parameter p. Since the Bernoulli random variable is a binomial random variable with $n = 1$, its PDF can be written as follows:

$$f(x) = p^x(1-p)^{1-x} \text{ when } x = 0, 1 \text{ and } 0 \text{ otherwise.}$$

The likelihood function is

$$Lik(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\Sigma_i X_i}(1-p)^{\Sigma_i(1-X_i)} = p^Y(1-p)^{n-Y},$$

where $Y = \sum_i X_i$ is the sample sum, and the log-likelihood function is

$$\ell(p) = \log(Lik(p)) = Y\log(p) + (n-Y)\log(1-p).$$

Assume that $0 < Y < n$. The ML estimator can be obtained by solving the derivative equation $\ell'(p) = 0$ for p:

$$\ell'(p) = \frac{Y}{p} - \frac{(n-Y)}{(1-p)} = 0 \Longrightarrow \frac{Y}{p} = \frac{(n-Y)}{(1-p)} \Longrightarrow \widehat{p} = \frac{Y}{n}.$$

Further, the second derivative test can be used to demonstrate that $\widehat{p} = Y/n$ maximizes the likelihood function on $0 < p < 1$.

The left part of Figure 7.1 shows the Bernoulli likelihood function when two successes are observed in six trials. The function is maximized at the ML estimate of the success probability, $\widehat{p} = 1/3$.

Note that if the observed value of Y is 0, then the ML estimate is $\widehat{p} = 0$, and if the observed value of Y is n, then the ML estimate is $\widehat{p} = 1$.

Example: Uniform distribution

Let X be a uniform random variable on the interval $[0, b]$. Since the PDF of X is $f(x) = 1/b$ for $0 \le x \le b$ and 0 otherwise, the likelihood function is

$$Lik(b) = \prod_{i=1}^{n} \frac{1}{b} = \frac{1}{b^n} \quad \text{if each } X_i \text{ is between 0 and } b$$

(and 0 otherwise). Likelihood is maximized at $\widehat{b} = \max(X_1, X_2, \ldots, X_n)$.

The right part of Figure 7.1 shows the uniform likelihood function when the numbers 1.80, 1.89, 2.14, 3.26, 4.85, 7.74 are observed. The function is maximized at the ML estimate of the upper endpoint, $\widehat{b} = 7.74$.

7.5.2 Cramer–Rao lower bound

The theorem below gives a formula for the lower bound on the variance of an unbiased estimator of θ. The formula is valid under what are called *smoothness conditions* on the X distribution.

If the three conditions

1. the PDF of X has continuous second partial derivatives (except, possibly, at a finite number of points),
2. the parameter θ is not at the boundary of possible parameter values, and
3. the range of X does not depend on θ

hold, then X satisfies the smoothness conditions of the theorem. The theorem excludes, for example, the Bernoulli distribution with $p = 1$ (condition 2 is violated) and the uniform distribution on $[0, b]$ (condition 3 is violated).

Many important models are excluded from the theorem. Additional tools, not covered here, are needed to study these models.

Theorem 7.2 (Cramer–Rao Lower Bound). *Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution with parameter θ, and let $\widehat{\theta}$ be an unbiased estimator of θ based on this sample. Under smoothness conditions on the X distribution,*

$$Var(\widehat{\theta}) \ge \frac{1}{nI(\theta)},$$

where $nI(\theta)$ can be computed as follows:

$$nI(\theta) = -E(\ell''(\theta)).$$

In this formula, $\ell''(\theta)$ is the second derivative of the log-likelihood function, and the expectation is computed using the joint distribution of the X_i's for fixed θ.

The quantity $nI(\theta)$ is called the *information* in a sample of size n. The *Cramer–Rao lower bound* is the reciprocal of the information, $1/(nI(\theta))$.

Efficient estimator

The estimator $\widehat{\theta}$ is said to be *efficient* if $E(\widehat{\theta}) = \theta$ and $Var(\widehat{\theta}) = 1/(nI(\theta))$.

If the X distribution satisfies the smoothness conditions and $\widehat{\theta}$ is an efficient estimator, then Theorem 7.2 implies that $\widehat{\theta}$ is a minimum variance unbiased estimator of θ.

For the Bernoulli example from page 88, the information is

$$nI(p) = -E(\ell''(p)) = -E\left(-\frac{Y}{p^2} - \frac{(n-Y)}{(1-p)^2}\right) = \frac{n}{p(1-p)}$$

($E(Y) = nE(X) = np$), and the Cramer–Rao lower bound is $p(1-p)/n$. Since

$$E(\widehat{p}) = p \text{ and } Var(\widehat{p}) = \frac{p(1-p)}{n},$$

the ML estimator $\widehat{p}$ is an efficient estimator of p.

7.5.3 Approximate sampling distribution

R. A. Fisher proved a generalization of the central limit theorem (Theorem 5.2) for ML estimators.

Theorem 7.3 (Fisher's Theorem). *Let $\widehat{\theta}_n$ be the ML estimator of θ based on a random sample of size n from the X distribution, let $nI(\theta)$ be the information, and*

$$Z_n = \frac{\widehat{\theta}_n - \theta}{\sqrt{1/(nI(\theta))}}.$$

Under smoothness conditions on the X distribution,

$$\lim_{n\to\infty} P(Z_n \le x) = \Phi(x) \quad \textit{for every real number } x,$$

where $\Phi(\cdot)$ is the CDF of the standard normal random variable.

Fisher's theorem says that if the X distribution satisfies smoothness conditions and the sample size is large, then the sampling distribution of the ML estimator is approximately normal with mean θ and variance equal to the Cramer–Rao lower bound. Thus, under smoothness conditions, the ML estimator is asymptotically efficient.

Approximate confidence intervals

Under the conditions of Theorem 7.3, an approximate $100(1-\alpha)\%$ confidence interval for θ has the form

$$\widehat{\theta} \pm z(\alpha/2)\sqrt{\frac{1}{nI(\widehat{\theta})}},$$

where $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution, and $nI(\widehat{\theta})$ is the estimate of $nI(\theta)$ obtained by substituting the ML estimate for θ.

Example: Bernoulli/binomial distribution

Let Y be the sample sum of a random sample of size n from a Bernoulli distribution with parameter p. If n is large and $0 < Y < n$, then an approximate $100(1-\alpha)\%$ confidence interval for p has the following form:

$$\widehat{p} \pm z(\alpha/2)\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \qquad \text{where } \widehat{p} = Y/n,$$

and $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

The sample sum Y is a binomial random variable with parameters n and p. The estimator $\widehat{p} = Y/n$ is called the *sample proportion*.

For example, suppose that in a recent national survey of 1570 adults in the United States, 30% (471/1570) said they considered the most serious problem facing the nation's public schools to be drugs. An approximate 95% confidence interval for the proportion p of all adults in the United States who consider the most serious problem facing public schools to be drugs is

$$0.30 \pm 1.960\sqrt{\frac{(0.30)(0.70)}{1570}} \implies [0.2773, 0.3227].$$

Note that the analysis above assumes that the 1570 individuals chosen for the survey are a simple random sample of adults in the United States. Since the total number of adults is quite large, the results can be assumed to summarize a random sample from a Bernoulli distribution.

In survey applications, the halfwidth of the confidence interval

$$z(\alpha/2)\sqrt{\widehat{p}(1-\widehat{p})/n}$$

is often called the *margin of error*. Unless otherwise specified, the confidence level is assumed to be 0.95 (that is, $\alpha = 0.05$).

Example: Poisson distribution

Let Y be the sample sum of a random sample of size n from a Poisson distribution with parameter λ. If n is large and $Y > 0$, then an approximate $100(1-\alpha)\%$ confidence interval for λ has the form

$$\widehat{\lambda} \pm z(\alpha/2)\sqrt{\frac{\widehat{\lambda}}{n}}, \qquad \text{where } \widehat{\lambda} = Y/n,$$

and $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

The sample sum Y is a Poisson random variable with parameter $n\lambda$. The statistic $\widehat{\lambda} = Y/n$ is the ML estimator of λ. Further, Theorem 7.2 can be used to show that $\widehat{\lambda}$ is an efficient estimator of λ.

For example, suppose that traffic accidents occurring during the workday in a large metropolitan area follow an approximate Poisson process and that an average

of 7.1625 (573/80) accidents per day were observed in 80 workdays. If this information summarizes the values of a random sample from a Poisson distribution with parameter λ (the average number of traffic accidents per workday in the area), then an approximate 95% confidence interval for λ is

$$7.1625 \pm 1.960\sqrt{\frac{7.1625}{80}} \implies [6.576, 7.749].$$

Example: Multinomial distribution

Assume that $(X_1, X_2, \ldots, X_k)$ has a multinomial distribution with parameters n and

$$p_i = p_i(\theta) \text{ for } i = 1, 2, \ldots, k.$$

Then the likelihood function can be written as follows:

$$Lik(\theta) = \binom{n}{X_1, X_2, \ldots, X_k} p_1(\theta)^{X_1}\, p_2(\theta)^{X_2} \cdots p_k(\theta)^{X_k}.$$

Since the random k-tuple summarizes the results of n independent trials of a multinomial experiment, the results of Theorem 7.3 can be applied in many situations.

For example, let $k = 4$ and

$$(p_1, p_2, p_3, p_4) = ((1-\theta), \theta(1-\theta), \theta^2(1-\theta), \theta^3),$$

where θ is a proportion. If n is large and each $X_i > 0$, Theorem 7.3 can be used to demonstrate that an approximate $100(1-\alpha)\%$ confidence interval for θ has the form

$$\widehat{\theta} \pm z(\alpha/2)\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n(1+\widehat{\theta}+\widehat{\theta^2})}}, \qquad \text{where } \widehat{\theta} = \frac{X_2 + 2X_3 + 3X_4}{X_1 + 2X_2 + 3X_3 + 3X_4},$$

and $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

In particular, if $n = 1000$ and $(x_1, x_2, x_3, x_4) = (363, 205, 136, 296)$ is observed, then the ML estimate of θ is 0.660 and an approximate 95% confidence interval for θ is

$$0.660 \pm 1.96\sqrt{\frac{0.660(1-0.660)}{1000(1+0.660+0.660^2)}} \implies [0.639, 0.680].$$

The model above can be used in *survival analysis*. Let θ equal the probability of surviving one unit of time, and assume that survival is independent of time period. Then p_1 is the probability of dying in the first period, p_2 is the probability of surviving the first period but dying in the second, p_3 is the probability of surviving the first two periods but dying in the third, and p_4 is the probability of surviving three or more time periods.

7.5.4 Multiple parameter estimation

If the X distribution has two or more unknown parameters, then ML estimators are computed using the techniques of multivariable calculus.

Example: Normal distribution

Let X be a normal random variable with mean μ and variance σ^2. The likelihood function is

$$Lik(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{2\sigma^2}\right),$$

and the log-likelihood function is

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Partial derivatives with respect to μ ($\ell_1(\mu, \sigma^2)$) and σ^2 ($\ell_2(\mu, \sigma^2)$) are

$$\ell_1(\mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu), \quad \ell_2(\mu, \sigma^2) = \frac{1}{2\sigma^2}\left[-n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right].$$

The ML estimators can be obtained by solving the partial derivative equations $\ell_1(\mu, \sigma^2) = 0$ and $\ell_2(\mu, \sigma^2) = 0$ simultaneously for μ and σ^2:

$$\ell_1(\mu, \sigma^2) = 0, \ \ell_2(\mu, \sigma^2) = 0 \implies \widehat{\mu} = \overline{X}, \ \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

Note that ML and MOM estimators are the same in normal distributions.

Example: Gamma distribution

Let X be a gamma random variable with parameters α and β. The likelihood function is

$$Lik(\alpha, \beta) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} X_i^{\alpha-1} e^{-X_i/\beta} = \left(\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\right)^n (\Pi_i X_i)^{\alpha-1} e^{-\Sigma_i X_i/\beta},$$

and the log-likelihood function is

$$\ell(\alpha, \beta) = -n\log(\Gamma(\alpha)) - n\alpha\log(\beta) + (\alpha - 1)\log(\Pi_i X_i) - \frac{\Sigma_i X_i}{\beta}.$$

Partial derivatives with respect to α ($\ell_1(\alpha, \beta)$) and β ($\ell_2(\alpha, \beta)$) are

$$\ell_1(\alpha, \beta) = \frac{-n\Gamma'(\alpha)}{\Gamma(\alpha)} - n\log(\beta) + \log(\Pi_i X_i), \quad \ell_2(\alpha, \beta) = \frac{-n\alpha}{\beta} + \frac{\Sigma_i X_i}{\beta^2}.$$

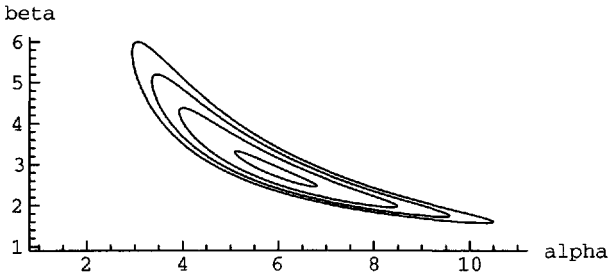


Figure 7.2. *Contours* $z = -39.6, -40.0, -40.4, -40.8$ *for a gamma log-likelihood function* $z = \ell(\alpha, \beta)$. *ML estimates are* $\widehat{\alpha} = 5.91075$ *and* $\widehat{\beta} = 2.84609$.

Since the system of equations $\ell_1(\alpha, \beta) = 0$ and $\ell_2(\alpha, \beta) = 0$ cannot be solved exactly, the computer is used to analyze specific samples.

For example, Figure 7.2 is a contour plot of the gamma log-likelihood function when the observations are 8.68, 8.91, 11.42, 12.04, 12.47, 14.61, 14.82, 15.77, 17.85, 23.44, 29.60, 32.26. Log-likelihood is maximized when α equals 5.91075 and β equals 2.84609. Thus, the ML estimates are $\widehat{\alpha} = 5.91075$ and $\widehat{\beta} = 2.84609$.

7.6 Laboratory problems

Laboratory problems for this chapter introduce *Mathematica* commands for working with the Student t distribution, for computing confidence intervals for normal means and variances, and for constructing plots of likelihood functions and simulated confidence intervals. The problems are designed to reinforce ideas related to estimation theory.

7.6.1 Laboratory: Estimation theory

In the main laboratory notebook (Problems 1 to 5), you are asked to use graphics and simulation to study Student t distributions; use simulation from normal distributions to study confidence interval procedures; apply confidence interval procedures to data from an IQ study [106], [80]; use simulation and graphics to study the concepts of confidence interval and ML estimation; and apply large sample methods to a hospital infection setting and to data from a spatial distribution study [85], [75]. Normal, binomial, and Poisson models are used.

7.6.2 Additional problem notebooks

Problems 6 and 7 are applications of estimation methods for samples from normal distributions. Problem 6 uses data from a study of forearm lengths [83] and includes a goodness-of-fit analysis. Problem 7 uses data from a physical anthropology study [107] and includes comparisons of three estimated models.

Problem 8 considers ML estimation and confidence interval procedures for subfamilies of the normal distribution where either the mean is known or the variance is known.

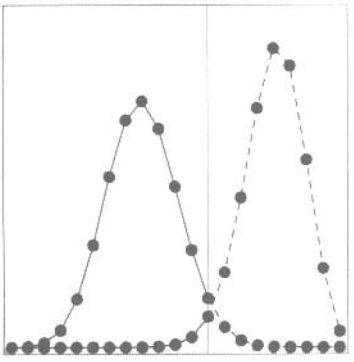
Problems 9 through 12 consider ML estimation methods in models with a single unknown parameter. Problem 9 applies the exponential distribution to data on major earthquakes [50]. Problem 10 applies the exponential distribution to data on major coal mining accidents in Great Britain in the nineteenth century [71], [57]. Problem 11 applies the negative binomial distribution in a political campaign setting. Problem 12 applies the gamma distribution with known shape parameter in a backup systems setting.

Problems 13 and 14 consider MOM and ML estimation for gamma distributions. Problem 13 uses data from a fuel leakage study [67]. Problem 14 uses data on cholesterol levels in two groups of male patients [95], [50]. Both problems include goodness-of-fit analyses and comparisons of different estimated models.

In Problem 15, simulation is used to determine if the sample mean and sample variance are uncorrelated when sampling from distributions other than the normal. Exponential, gamma, and uniform models are used.

In Problem 16, simulation is used to determine if confidence interval procedures designed for samples from normal distributions remain valid when sampling from distributions other than the normal. Exponential, gamma, and uniform models are used.

This page intentionally left blank



Chapter 8
Hypothesis Testing Theory

Statistical inference, which includes estimation theory and hypothesis testing theory, refers to a broad collection of methods for analyzing random samples from probability distributions. If the family of distributions from which the data were drawn is known except for the values of one or more parameters, then hypothesis testing theory can be used to determine if the unknown parameters lie in one subset of the set of possible parameters or in its complement.

This chapter introduces hypothesis testing theory. The first six sections give important definitions and examples. Likelihood ratio tests are introduced in Section 7, and Section 8 discusses the relationship between hypothesis tests and confidence intervals. Section 9 outlines the laboratory problems.

8.1 Definitions

An *hypothesis* is an assertion about the distribution of a random variable or a random k-tuple. A *simple hypothesis* specifies the distribution completely. A *compound hypothesis* does not specify the distribution completely. For example, the hypothesis

$$H\text{: } X \text{ is an exponential random variable with parameter } \lambda = 1/5$$

is simple, and the hypothesis

$$H\text{: } X \text{ is an exponential random variable with parameter } \lambda \geq 1/5$$

is compound.

A *hypothesis test* is a decision rule allowing the user to choose between competing assertions.

8.1.1 Neyman–Pearson framework

In the *Neyman–Pearson framework* of hypothesis testing, there are two competing assertions: the *null* hypothesis, denoted by H_o, and the *alternative* hypothesis,

denoted by H_a. The null hypothesis is accepted as true unless sufficient evidence is provided to the contrary; then the null hypothesis is rejected in favor of the alternative hypothesis.

For example, suppose that the standard treatment for a given medical condition is effective in 45% of patients. A new treatment promises to be effective in more than 45% of patients. In testing the efficacy of the new treatment, the hypotheses could be set up as follows:

H_o: The new treatment is no more effective than the standard treatment.

H_a: The new treatment is more effective than the standard treatment.

If p is the proportion of patients for whom the new treatment would be effective, then the hypotheses above could be written as follows:

$$H_o\colon\ p = 0.45 \text{ versus } H_a\colon\ p > 0.45.$$

Similarly, in testing whether an exponential distribution is a reasonable model for sample data, the hypotheses would be set up as follows:

H_o: The distribution of X is exponential.

H_a: The distribution of X is not exponential.

If Pearson's goodness-of-fit test is used, then the data would be grouped (using k ranges of values for some k) and the hypotheses would be set up as follows:

H_o: The data are consistent with the grouped exponential model.

H_a: The data are not consistent with the grouped exponential model.

Test setup

Let $X_1, X_2, \ldots, X_n$ be a random sample from the X distribution. To set up a test, the following is done:

1. A *test statistic*, $T = T(X_1, \ldots, X_n)$, is chosen.
2. The range of T is subdivided into the *rejection region* and the complementary *acceptance region*.
3. If the observed value of T is in the acceptance region, then the null hypothesis is accepted. Otherwise, the null hypothesis is rejected in favor of the alternative.

The test statistic and acceptance and rejection regions are chosen so that the probability that T is in the rejection region is small (close to 0) when the null hypothesis is true. Hopefully, although it is not guaranteed, the probability that T is in the rejection region is large (close to 1) when the alternative hypothesis is true.

Upper tail test example

Let Y be the sample sum of a random sample of size 25 from a Bernoulli distribution with parameter p. Consider the following decision rule for a test of the null hypothesis $p = 0.45$ versus the alternative hypothesis $p > 0.45$:

$$\text{Reject } p = 0.45 \text{ in favor of } p > 0.45 \text{ when } Y \geq 16.$$

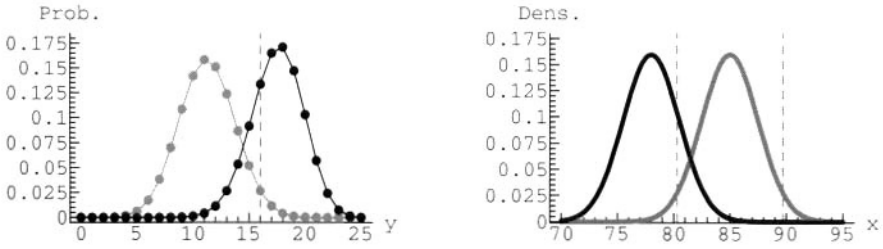


Figure 8.1. *Graphs of distributions under null and alternative hypotheses for an upper tail test (left plot) and a two tailed test (right plot).*

If the null hypothesis is true, then $P(Y \geq 16 \text{ when } p = 0.45) = 0.044$. If the actual success probability is 0.70, then $P(Y \geq 16 \text{ when } p = 0.70) = 0.811$.

The test statistic Y is a binomial random variable with $n = 25$. The left part of Figure 8.1 shows the distribution of Y under the null hypothesis when $p = 0.45$ (in gray) and under the alternative hypothesis when $p = 0.70$ (in black). A vertical dashed line is drawn at $y = 16$.

This is an example of an upper tail test. In an *upper tail* test, the null hypothesis is rejected if the test statistic is in the upper tail of distributions satisfying the null hypothesis.

A test that rejects the null hypothesis when the test statistic is in the lower tail of distributions satisfying the null hypothesis is called a *lower tail* test. Upper tail and lower tail tests are also called *one sided* tests.

Two tailed test example

Let $\overline{X}$ be the sample mean of a random sample of size 16 from a normal distribution with standard deviation 10. Consider the following decision rule for a test of the null hypothesis $\mu = 85$ versus the alternative hypothesis $\mu \neq 85$:

$$\text{Reject } \mu = 85 \text{ in favor of } \mu \neq 85 \text{ when } \overline{X} \leq 80.3 \text{ or } \overline{X} \geq 89.7.$$

If the null hypothesis is true, then $P(\overline{X} \leq 80.3 \text{ or } \overline{X} \geq 89.7 \text{ when } \mu = 85) = 0.06$. If the actual mean is 78, then $P(\overline{X} \leq 80.3 \text{ or } \overline{X} \geq 89.7 \text{ when } \mu = 78) = 0.821$.

The test statistic $\overline{X}$ is a normal random variable with standard deviation $10/\sqrt{16} = 2.5$. The right part of Figure 8.1 shows the distribution of $\overline{X}$ under the null hypothesis when $\mu = 85$ (in gray) and under the alternative hypothesis when $\mu = 78$ (in black). Vertical dashed lines are drawn at $x = 80.3$ and $x = 89.7$.

This is an example of a two tailed test. In a *two tailed* (or *two sided*) test, the null hypothesis is rejected if the test statistic is either in the lower tail or in the upper tail of distributions satisfying the null hypothesis.

8.1.2 Equivalent tests

Consider two tests, each based on a random sample of size n: (1) a test based on statistic T with rejection region RR_T and (2) a test based on statistic W with rejection

region RR_W. The tests are said to be *equivalent* if

$$T \in RR_T \quad \Longleftrightarrow \quad W \in RR_W.$$

That is, given the same information, either both tests accept the null hypothesis or both reject the null hypothesis. Equivalent tests have the same properties.

For example, let $\overline{X}$ be the sample mean of a random sample of size 16 from a normal distribution with standard deviation 10, and let $Z = (\overline{X} - 85)/2.5$ be the standardized mean when $\mu = 85$. Then the two tailed test given in the last example is equivalent to the test with decision rule:

$$\text{Reject } \mu = 85 \text{ in favor of } \mu \neq 85 \text{ when } |Z| \geq 1.88.$$

8.2 Properties of tests

Let X be a distribution with parameter θ. Assume that the null and alternative hypotheses can be stated in terms of values of θ as follows:

$$H_o\colon \theta \in \omega_o \qquad \text{versus} \qquad H_a\colon \theta \in \Omega - \omega_o,$$

where Ω represents the full set of parameter values under consideration, and ω_o is a subset of Ω. For example, if X is a Bernoulli random variable with parameter p, the null hypothesis is $p \leq 0.30$, and the alternative hypothesis is $p > 0.30$, then

$$\Omega = \{p \mid 0 \leq p \leq 1\} \text{ and } \omega_o = \{p \mid 0 \leq p \leq 0.30\}.$$

8.2.1 Errors, size, significance level

When carrying out a test, two types of errors can occur:

1. An error of *type* I occurs when a true null hypothesis is rejected.
2. An error of *type* II occurs when a false null hypothesis is accepted.

The *size* or *significance level* of the test with decision rule

$$\text{Reject } \theta \in \omega_o \text{ in favor of } \theta \in \Omega - \omega_o \text{ when } T \in RR$$

is defined as follows:

$$\alpha = \sup_{\theta \in \omega_o} P(T \in RR \text{ when the true parameter is } \theta).$$

A test of size α is often called a "100α% test."

The size or significance level is the maximum type I error (or the least upper bound of type I errors, if a maximum does not exist).

For example, the upper tail test on page 98 is a 4.4% test of the null hypothesis $p = 0.45$ versus the alternative hypothesis $p > 0.45$. If the actual success probability is 0.55, then the type II error is $P(Y < 16 \text{ when } p = 0.55) = 0.758$.

Statistical significance

If the significance level is α and the observed data lead to rejecting the null hypothesis, then the result is said to be *statistically significant* at level α. If the observed data do not lead to rejecting the null hypothesis, then the result is not statistically significant at level α.

Observed significance level, p value

The *observed significance level* or *p value* is the minimum significance level for which the observed data indicate that the null hypothesis should be rejected.

Note that the p value measures the strength of evidence against the null hypothesis. For a size α test, if the p value is greater than α, then the null hypothesis is accepted; if the p value is less than or equal to α, then the null hypothesis is rejected in favor of the alternative hypothesis.

For example, in the upper tail test on page 98, if $y = 14$ is observed, then the p value is $P(Y \geq 14 \text{ when } p = 0.45) = 0.183$ and the result is not statistically significant at level 0.044.

In the two tailed test on page 99, if $\bar{x} = 90.5$ is observed, then the p value is $2P(\overline{X} \geq 90.5 \text{ when } \mu = 85) = 0.0278$ and the result is statistically significant at level 0.06. If $\bar{x} = 80.6$ is observed, then the p value is $2P(\overline{X} \leq 80.6 \text{ when } \mu = 85) = 0.0784$ and the result is not statistically significant at level 0.06.

8.2.2 Power, power function

The *power* of the test with decision rule

$$\text{Reject } \theta \in \omega_o \text{ in favor of } \theta \in \Omega - \omega_o \text{ when } T \in RR$$

at $\theta \in \Omega$ is the probability $P(T \in RR \text{ when the true parameter is } \theta)$. The *power function* of the test is the function

$$\text{Power}(\theta) = P(T \in RR \text{ when the true parameter is } \theta) \qquad \text{for } \theta \in \Omega.$$

If $\theta \in \omega_o$, then the power at θ is the same as the type I error. If $\theta \in \Omega - \omega_o$, then the power corresponds to the test's ability to correctly reject the null hypothesis in favor of the alternative hypothesis.

For example, in the upper tail test on page 98, the power when $p = 0.70$ is 0.811; in the two tailed test on page 99, the power when $\mu = 78$ is 0.821.

Figure 8.2 shows power functions for the upper tail (left plot) and two tailed (right plot) examples. In the left plot, which has been extended to include all p between 0 and 1, power increases as p increases. In the right plot, power increases as μ gets further from 85 in either direction.

Uniformly more powerful test

Consider two $100\alpha\%$ tests of the null hypothesis $\theta \in \omega_o$ versus the alternative hypothesis $\theta \in \Omega - \omega_o$, each based on a random sample of size n: (1) a test based

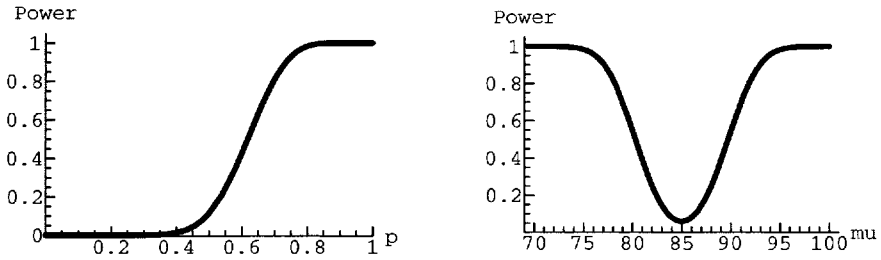


Figure 8.2. *Power curves for an upper tail test (left plot) and for a two tailed test (right plot).*

on statistic T with power function $\text{Power}_T(\theta)$ and (2) a test based on statistic W with power function $\text{Power}_W(\theta)$. The test based on T is said to be *uniformly more powerful* than the test based on W if

$$\text{Power}_T(\theta) \geq \text{Power}_W(\theta) \qquad \text{for all } \theta \in \Omega - \omega_o$$

with strict inequality ($>$) for at least one $\theta \in \Omega - \omega_o$.

It is possible that the test based on T is more powerful than the one based on W for some values of $\theta \in \Omega - \omega_o$ and that the test based on W is more powerful than the one based on T for other values of $\theta \in \Omega - \omega_o$. If the test based on T is uniformly more powerful than the one based on W, then T has a greater (or equal) chance of rejecting the false null hypothesis for each model satisfying the alternative hypothesis. Thus, we would prefer to use the test based on T.

Uniformly most powerful test

The test based on T is a *uniformly most powerful test* (UMPT) if it is uniformly more powerful than all other (nonequivalent) tests.

An interesting and difficult problem in the field of statistics is that of determining when a UMPT exists. This problem is considered in Section 8.7.1.

8.3 Example: Normal distribution

Let $\overline{X}$ be the sample mean and S^2 be the sample variance of a random sample of size n from a normal distribution with mean μ and standard deviation σ.

8.3.1 Tests of $\mu = \mu_o$

If the value of σ^2 is known, then the standardized mean when $\mu = \mu_o$,

$$Z = \frac{\overline{X} - \mu_o}{\sqrt{\sigma^2/n}},$$

can be used as test statistic. The following table gives the rejection regions for one sided and two sided $100\alpha\%$ tests:

Alternative Hypothesis	Rejection Region
$\mu < \mu_o$	$Z \leq -z(\alpha)$
$\mu > \mu_o$	$Z \geq z(\alpha)$
$\mu \neq \mu_o$	$\lvert Z\rvert \geq z(\alpha/2)$

where $z(p)$ is the $100(1-p)\%$ point of the standard normal distribution.

These are examples of z tests. A *z test* is a test based on a statistic with a standard normal distribution under the null hypothesis.

If the value of σ^2 is estimated from the data, then the approximate standardization when $\mu = \mu_o$,

$$T = \frac{\overline{X} - \mu_o}{\sqrt{S^2/n}},$$

can be used as test statistic. The following table gives the rejection regions for one sided and two sided $100\alpha\%$ tests:

Alternative Hypothesis	Rejection Region
$\mu < \mu_o$	$T \leq -t_{n-1}(\alpha)$
$\mu > \mu_o$	$T \geq t_{n-1}(\alpha)$
$\mu \neq \mu_o$	$\lvert T\rvert \geq t_{n-1}(\alpha/2)$

where $t_{n-1}(p)$ is the $100(1-p)\%$ point of the Student t distribution with $(n-1)$ degrees of freedom.

These are examples of t tests. A *t test* is a test based on a statistic with a Student t distribution under the null hypothesis.

For example, consider testing the null hypothesis $\mu = 120$ versus the alternative hypothesis $\mu < 120$ at the 5% significance level using a sample of size 16.

(i) If the distribution has standard deviation 5, then the test statistic is $Z = (\overline{X} - 120)/1.25$ and the rejection region is $Z \leq -1.645$.

(ii) If the standard deviation is not known, then the test statistic is $T = (\overline{X} - 120)/(S/4)$ and the rejection region is $T \leq -1.753$.

8.3.2 Tests of $\sigma^2 = \sigma_o^2$

If the value of μ is known, then the sum of squared deviations from μ divided by the hypothesized variance,

$$V = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma_o^2},$$

can be used as test statistic. The following table gives the rejection regions for one sided and two sided $100\alpha\%$ tests:

Alternative Hypothesis	Rejection Region
$\sigma^2 < \sigma_o^2$	$V \leq \chi_n^2(1-\alpha)$
$\sigma^2 > \sigma_o^2$	$V \geq \chi_n^2(\alpha)$
$\sigma^2 \neq \sigma_o^2$	$V \leq \chi_n^2(1-\alpha/2)$ or $V \geq \chi_n^2(\alpha/2)$

where $\chi_n^2(p)$ is the $100(1-p)\%$ point of the chi-square distribution with n degrees of freedom.

If the value of μ is estimated from the data, then the sum of squared deviations from the sample mean divided by the hypothesized variance,

$$V = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma_o^2},$$

can be used as test statistic. The following table gives the rejection regions for one sided and two sided $100\alpha\%$ tests:

Alternative Hypothesis	Rejection Region
$\sigma^2 < \sigma_o^2$	$V \leq \chi_{n-1}^2(1-\alpha)$
$\sigma^2 > \sigma_o^2$	$V \geq \chi_{n-1}^2(\alpha)$
$\sigma^2 \neq \sigma_o^2$	$V \leq \chi_{n-1}^2(1-\alpha/2)$ or $V \geq \chi_{n-1}^2(\alpha/2)$

where $\chi_{n-1}^2(p)$ is the $100(1-p)\%$ point of the chi-square distribution with $n-1$ degrees of freedom.

These are examples of chi-square tests. A *chi-square test* is a test based on a statistic with a chi-square distribution under the null hypothesis.

For example, consider testing the null hypothesis $\sigma^2 = 16$ versus the alternative hypothesis $\sigma^2 \neq 16$ at the 5% significance level using a sample of size 20.

(i) If the distribution has mean 80, then the test statistic is $V = \sum_i (X_i - 80)^2/16$ and the rejection region is $V \leq 9.59$ or $V \geq 34.17$.

(ii) If the mean is not known, then the test statistic is $V = \sum_i (X_i - \overline{X})^2/16$ and the rejection region is $V \leq 8.91$ or $V \geq 32.85$.

8.4 Example: Bernoulli/binomial distribution

Let Y be the sample sum of a random sample of size n from a Bernoulli distribution with parameter p. Y is a binomial random variable with parameters n and p.

Small sample tests of $p = p_o$

Rejection regions for one sided and two sided $100\alpha\%$ tests are as follows:

Alternative Hypothesis	Rejection Region
$p < p_o$	$Y \leq c$, where c is chosen so that $\alpha = P(Y \leq c \text{ when } p = p_o)$
$p > p_o$	$Y \geq c$, where c is chosen so that $\alpha = P(Y \geq c \text{ when } p = p_o)$
$p \neq p_o$	$Y \leq c_1$ or $Y \geq c_2$, where c_1 and c_2 are chosen so that $\alpha = P(Y \leq c_1 \text{ when } p = p_o) + P(Y \geq c_2 \text{ when } p = p_o)$ (and the two probabilities are approximately equal)

For example, consider testing the null hypothesis $p = 0.30$ versus the alternative hypothesis $p \neq 0.30$ at a significance level close to 5% using a sample of size 18. Since $P(Y \leq 1 \text{ when } p = 0.30) = 0.014$ and $P(Y \geq 10 \text{ when } p = 0.30) = 0.021$, the rejection region for a 3.5% test is $Y \leq 1$ or $Y \geq 10$.

Large sample tests of $p = p_o$

The standardized sample sum when $p = p_o$,

$$Z = \frac{Y - np_o}{\sqrt{np_o(1 - p_o)}},$$

can be used as test statistic. Since, by the central limit theorem (Theorem 5.2), the distribution of Z is approximately standard normal when n is large, rejection regions for approximate one sided and two sided $100\alpha\%$ tests are as follows:

Alternative Hypothesis	Rejection Region
$p < p_o$	$Z \leq -z(\alpha)$
$p > p_o$	$Z \geq z(\alpha)$
$p \neq p_o$	$\lvert Z \rvert \geq z(\alpha/2)$

where $z(p)$ is the $100(1 - p)\%$ point of the standard normal distribution.

For example, consider testing the null hypothesis $p = 0.45$ versus the alternative hypothesis $p > 0.45$ at the 1% significance level using a sample of size 250. The test statistic is $Z = (Y - 112.5)/\sqrt{61.875}$, and the rejection region is $Z \geq 2.326$.

8.5 Example: Poisson distribution

Let Y be the sample sum of a random sample of size n from a Poisson distribution with parameter λ. Y is a Poisson random variable with parameter $n\lambda$.

Small sample tests of $\lambda = \lambda_o$

Rejection regions for one sided and two sided $100\alpha\%$ tests are as follows:

Alternative Hypothesis	Rejection Region
$\lambda < \lambda_o$	$Y \leq c$, where c is chosen so that $\alpha = P(Y \leq c \text{ when } \lambda = \lambda_o)$
$\lambda > \lambda_o$	$Y \geq c$, where c is chosen so that $\alpha = P(Y \geq c \text{ when } \lambda = \lambda_o)$
$\lambda \neq \lambda_o$	$Y \leq c_1$ or $Y \geq c_2$, where c_1 and c_2 are chosen so that $\alpha = P(Y \leq c_1 \text{ when } \lambda = \lambda_o) + P(Y \geq c_2 \text{ when } \lambda = \lambda_o)$ (and the two probabilities are approximately equal)

For example, consider testing the null hypothesis $\lambda = 2$ versus the alternative hypothesis $\lambda < 2$ at a significance level close to 2% using a sample of size 5. Since $P(Y \leq 4 \text{ when } \lambda = 2) = 0.0294$, the rejection region for a 2.94% test is $Y \leq 4$.

Large sample tests of $\lambda = \lambda_o$

The standardized sample sum when $\lambda = \lambda_o$,

$$Z = \frac{Y - n\lambda_o}{\sqrt{n\lambda_o}},$$

can be used as test statistic. Since, by the central limit theorem (Theorem 5.2), the distribution of Z is approximately standard normal when n is large, rejection regions for approximate one sided and two sided $100\alpha\%$ tests are as follows:

Alternative Hypothesis	Rejection Region
$\lambda < \lambda_o$	$Z \leq -z(\alpha)$
$\lambda > \lambda_o$	$Z \geq z(\alpha)$
$\lambda \neq \lambda_o$	$\lvert Z \rvert \geq z(\alpha/2)$

where $z(p)$ is the $100(1-p)\%$ point of the standard normal distribution.

For example, consider testing the null hypothesis $\lambda = 2$ versus the alternative hypothesis $\lambda \neq 2$ at the 5% significance level using a sample of size 80. The test statistic is $Z = (Y - 160)/\sqrt{160}$, and the rejection region is $|Z| \geq 1.960$.

8.6 Approximate tests of $\mu = \mu_o$

Let $\overline{X}$ be the sample mean and S^2 be the sample variance of a random sample of size n from a distribution with unknown (but finite) mean and variance. The approximate standardization of the sample mean when $\mu = \mu_o$,

$$Z = \frac{\overline{X} - \mu_o}{\sqrt{S^2/n}},$$

can be used as test statistic. Since, by the central limit theorem (Theorem 5.2), the distribution of Z is approximately standard normal when n is large, rejection regions for approximate one sided and two sided $100\alpha\%$ tests are as follows:

Alternative Hypothesis	Rejection Region
$\mu < \mu_o$	$Z \leq -z(\alpha)$
$\mu > \mu_o$	$Z \geq z(\alpha)$
$\mu \neq \mu_o$	$\lvert Z \rvert \geq z(\alpha/2)$

where $z(p)$ is the $100(1-p)\%$ point of the standard normal distribution.

This large sample method for $\mu = \mu_o$ is useful because it does not require precise knowledge of the X distribution. Additional approximate methods are given in Section 8.7.3.

8.7 Likelihood ratio tests

The *likelihood ratio* method, introduced by J. Neyman and E. Pearson in the 1930's, is a general method for constructing tests.

In many practical situations, likelihood ratio tests are uniformly most powerful. In situations where no uniformly most powerful test (UMPT) exists, likelihood ratio tests are popular choices because they have good statistical properties.

8.7.1 Likelihood ratio statistic; Neyman–Pearson lemma

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with parameter θ, and let $Lik(\theta)$ be the likelihood function based on this sample. Consider testing the null hypothesis $\theta = \theta_o$ versus the alternative hypothesis $\theta = \theta_a$, where θ_o and θ_a are constants. Then the *likelihood ratio statistic*, Λ, is the ratio of the likelihood functions,

$$\Lambda = \frac{Lik(\theta_o)}{Lik(\theta_a)},$$

and a *likelihood ratio test* based on this statistic is a test whose decision rule has the following form:

$$\text{Reject } \theta = \theta_o \text{ in favor of } \theta = \theta_a \text{ when } \Lambda \leq c.$$

Note that if the null hypothesis is true, then the value of the likelihood function in the numerator will tend to be larger than the value in the denominator. If the alternative

hypothesis is true, then the value of the likelihood function in the denominator will tend to be larger than the value in the numerator. Thus, it is reasonable to "reject when Λ is small."

The following theorem, proven by Neyman and Pearson, states that the likelihood ratio test is a UMPT for a simple null hypothesis versus a simple alternative hypothesis.

Theorem 8.1 (Neyman–Pearson Lemma). *Given the situation above, if c is chosen so that* $P(\Lambda \leq c \text{ when } \theta = \theta_o) = \alpha$, *then the test with decision rule*

$$\textit{Reject } \theta = \theta_o \textit{ in favor of } \theta = \theta_a \textit{ when } \Lambda \leq c$$

is a UMPT of size α.

In general, a likelihood ratio test is not implemented as shown above. Instead, an equivalent test (with a simpler statistic and rejection region) is used.

Example: Bernoulli distribution

Let Y be the sample sum of a random sample of size n from a Bernoulli distribution, and consider testing the null hypothesis $p = p_o$ versus the alternative hypothesis $p = p_a$, where $p_a > p_o$. Since $Lik(p) = p^Y(1-p)^{n-Y}$ (see page 88), the likelihood ratio statistic is

$$\Lambda = \frac{Lik(p_o)}{Lik(p_a)} = \frac{(p_o)^Y(1-p_o)^{n-Y}}{(p_a)^Y(1-p_a)^{n-Y}} = \left(\frac{p_o}{p_a}\right)^Y \left(\frac{1-p_o}{1-p_a}\right)^{n-Y}.$$

Assume that $P(\Lambda \leq c \text{ when } p = p_o) = \alpha$ for some α. Since

$$\begin{aligned}
\Lambda \leq c &\iff \log(\Lambda) \leq \log(c) \\
&\iff Y \log\left(\tfrac{p_o}{p_a}\right) + (n-Y)\log\left(\tfrac{1-p_o}{1-p_a}\right) \leq \log(c) \\
&\iff Y\left(\log\left(\tfrac{p_o}{p_a}\right) - \log\left(\tfrac{1-p_o}{1-p_a}\right)\right) \leq \log(c) - n\log\left(\tfrac{1-p_o}{1-p_a}\right) \\
&\iff Y \log\left(\tfrac{p_o(1-p_a)}{p_a(1-p_0)}\right) \leq \log(c) - n\log\left(\tfrac{1-p_o}{1-p_a}\right) \\
&\iff Y \geq k, \qquad \text{where } k = \left(\log(c) - n\log\left(\tfrac{1-p_o}{1-p_a}\right)\right) \Big/ \log\left(\tfrac{p_o(1-p_a)}{p_a(1-p_0)}\right),
\end{aligned}$$

the likelihood ratio test is equivalent to the test with decision rule

$$\text{Reject } p = p_o \text{ in favor of } p = p_a \text{ when } Y \geq k,$$

where k is chosen so that $P(Y \geq k \text{ when } p = p_o) = \alpha$. Thus, by Theorem 8.1, the test based on the sample sum is a uniformly most powerful test of size α.

The inequality switches in the last equivalence above since $p_a > p_o$ implies that the ratio $(p_o(1-p_a))/(p_a(1-p_0))$ is less than 1, and its logarithm is a negative number.

Application: One sided tests in single parameter families

The Neyman–Pearson lemma can sometimes be used to derive UMPTs for composite hypotheses in single parameter families of distributions.

Continuing with the Bernoulli example above, suppose that k is chosen so that $P(Y \geq k \text{ when } p = p_o) = \alpha$, and consider the test with decision rule

$$\text{Reject } p = p_o \text{ in favor of } p > p_o \text{ when } Y \geq k.$$

Since Theorem 8.1 implies that $P(Y \geq k \text{ when } p = p_a)$ is maximum possible for each $p_a > p_o$, the test based on Y is a uniformly most powerful size α test for the simple null hypothesis $p = p_o$ versus the composite alternative hypothesis $p > p_o$.

8.7.2 Generalized likelihood ratio tests

The methods in this section generalize the approach of Theorem 8.1 to compound hypotheses and to multiple parameter families. Generalized likelihood ratio tests are not guaranteed to be uniformly most powerful. In fact, in many situations (e.g., two tailed tests) UMPTs do not exist.

Let X be a distribution with parameter θ, where θ is a single parameter or a k-tuple of parameters. Assume that the null and alternative hypotheses can be stated in terms of values of θ as follows:

$$H_o\colon \theta \in \omega_o \qquad \text{versus} \qquad H_a\colon \theta \in \Omega - \omega_o,$$

where Ω represents the full set of parameter values under consideration, and ω_o is a subset of Ω. For example, if X is a normal random variable with unknown mean μ and unknown variance σ^2, the null hypothesis is $\mu = 120$, and the alternative hypothesis is $\mu \neq 120$, then $\theta = (\mu, \sigma^2)$,

$$\Omega = \{(\mu, \sigma^2) \mid -\infty < \mu < \infty, \sigma^2 > 0\} \text{ and } \omega_o = \{(120, \sigma^2) \mid \sigma^2 > 0\}.$$

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with parameter θ, and let $Lik(\theta)$ be the likelihood function based on this sample. The generalized *likelihood ratio statistic*, Λ, is the ratio of the maximum value of the likelihood function for models satisfying the null hypothesis to the maximum value of the likelihood function for all models under consideration,

$$\Lambda = \frac{\max_{\theta \in \omega_o} Lik(\theta)}{\max_{\theta \in \Omega} Lik(\theta)},$$

and a generalized *likelihood ratio test* based on this statistic is a test whose decision rule has the following form:

$$\text{Reject } \theta \in \omega_o \text{ in favor of } \theta \in \Omega - \omega_o \text{ when } \Lambda \leq c.$$

Note that the value in the denominator of Λ is the value of the likelihood at the ML estimator, and the value in the numerator is less than or equal to the value in the denominator. Thus, $\Lambda \leq 1$. Further, if the null hypothesis is true, then the

numerator and denominator values will be close (and Λ will be close to 1); otherwise, the numerator is likely to be much smaller than the denominator (and Λ will be close to 0). Thus, it is reasonable to "reject when Λ is small."

In general, a likelihood ratio test is not implemented as shown above. Instead, an equivalent test (with a simpler statistic and rejection region) is used.

Example: Normal distribution

Let $\overline{X}$ be the sample mean and S^2 be the sample variance of a random sample of size n from a normal distribution with unknown mean μ and variance σ^2. Consider testing the null hypothesis $\mu = \mu_o$ versus the alternative hypothesis $\mu \neq \mu_o$, where μ_o is a constant.

(i) The numerator in the likelihood ratio statistic

$$\max_{(\mu,\sigma^2)\in\omega_o} Lik(\mu, \sigma^2), \text{ where } \omega_o = \{(\mu_o, \sigma^2) \mid \sigma^2 > 0\},$$

is the value of the likelihood when (μ, σ^2) is $\left(\mu_o, \frac{1}{n}\sum_{i=1}^n (X_i - \mu_o)^2\right)$. After cancellations, the numerator becomes

$$\left(\frac{2\pi}{n}\sum_{i=1}^n (X_i - \mu_o)^2\right)^{-n/2} e^{-n/2}.$$

(See page 93 for the normal likelihood function.)

(ii) The denominator in the likelihood ratio statistic

$$\max_{(\mu,\sigma^2)\in\Omega} Lik(\mu, \sigma^2), \text{ where } \Omega = \{(\mu, \sigma^2) \mid -\infty < \mu < \infty, \sigma^2 > 0\},$$

is the value of the likelihood when (μ, σ^2) is $\left(\overline{X}, \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2\right)$. After cancellations, the denominator becomes

$$\left(\frac{2\pi}{n}\sum_{i=1}^n (X_i - \overline{X})^2\right)^{-n/2} e^{-n/2},$$

and the ratio simplifies to

$$\Lambda = \left(\frac{\sum_{i=1}^n (X_i - \mu_o)^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)^{-n/2}.$$

(iii) Since $\sum_{i=1}^n (X_i - \mu_o)^2$

$$= \sum_{i=1}^n (X_i - \overline{X} + \overline{X} - \mu_o)^2$$

$$= \sum_{i=1}^n (X_i - \overline{X})^2 + 2(\overline{X} - \mu_o)\sum_{i=1}^n (X_i - \overline{X}) + n(\overline{X} - \mu_o)^2$$

$$= \sum_{i=1}^n (X_i - \overline{X})^2 + n(\overline{X} - \mu_o)^2 \qquad \text{(the middle sum is zero)}$$

and since $\sum_{i=1}^{n}(X_i - \overline{X})^2 = (n-1)S^2$, the likelihood ratio statistic can be further simplified as follows:

$$\Lambda = \left(1 + \frac{n(\overline{X} - \mu_o)^2}{(n-1)S^2}\right)^{-n/2}.$$

(iv) Assume that $P(\Lambda \leq c \text{ when } \mu = \mu_o) = \alpha$ for some α. Then

$$\begin{aligned}
\Lambda \leq c &\iff \frac{n(\overline{X}-\mu_o)^2}{(n-1)S^2} \geq \sqrt[n/2]{1/c} - 1 \\
&\iff \frac{(\overline{X}-\mu_o)^2}{S^2/n} \geq (n-1)\left(\sqrt[n/2]{1/c} - 1\right) \\
&\iff \left|\frac{\overline{X}-\mu_o}{\sqrt{S^2/n}}\right| \geq k, \quad \text{where } k = \sqrt{(n-1)\left(\sqrt[n/2]{1/c} - 1\right)}.
\end{aligned}$$

Since $P(|\overline{X} - \mu_o/\sqrt{S^2/n}| \geq k \text{ when } \mu = \mu_o) = \alpha$, the test with decision rule

$$\text{Reject } \mu = \mu_o \text{ in favor of } \mu \neq \mu_o \text{ when } \left|\frac{\overline{X}-\mu_o}{\sqrt{S^2/n}}\right| \geq k$$

is a (generalized) likelihood ratio test of size α. Thus, the two sided t test is equivalent to a likelihood ratio test.

The tests given in Section 8.3 for samples from normal distributions are examples of likelihood ratio tests (or approximate likelihood ratio tests).

8.7.3 Approximate sampling distribution

In many situations, the exact distribution of Λ (or an equivalent form) is not known. The theorem below, proven by S.S. Wilks in the 1930's, provides a useful large sample approximation to the distribution of $-2\log(\Lambda)$ (where log is the natural logarithm function) under the smoothness conditions of Theorems 7.2 and 7.3.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with parameter θ, and let $Lik(\theta)$ be the likelihood function based on this sample. Consider testing the null hypothesis $\theta \in \omega_o$ versus the alternative hypothesis $\theta \in \Omega - \omega_o$ using the likelihood ratio test.

Theorem 8.2 (Wilks Theorem). *Given the situation above, under smoothness conditions on the X distribution and when n is large, the distribution of* $-2\log(\Lambda)$ *is approximately chi-square with* $r - r_o$ *degrees of freedom, where r is the number of free parameters in* Ω, r_o *is the number of free parameters in* ω_o, *and* log() *is the natural logarithm function.*

Approximate tests

Under the conditions of Theorem 8.2, an approximate $100\alpha\%$ test of $\theta \in \omega_o$ versus $\theta \in \Omega - \omega_o$ has the following decision rule:

$$\text{Reject } \theta \in \omega_o \text{ in favor of } \theta \in \Omega - \omega_o \text{ when } -2\log(\Lambda) \geq \chi^2_{r-r_o}(\alpha),$$

where $\chi^2_{r-r_o}(\alpha)$ is the $100(1-\alpha)\%$ point on the chi-square distribution with $r - r_o$ degrees of freedom, r is the number of free parameters in Ω, and r_o is the number of free parameters in ω_o.

Rejecting when Λ is small is equivalent to rejecting when $-2\log(\Lambda)$ is large.

Example: Comparing Bernoulli parameters

Let Y_i be the sample sum of a random sample of size n_i from a Bernoulli distribution with parameter p_i for $i = 1, 2, \ldots, k$.

Consider testing the null hypothesis that the k Bernoulli parameters are equal versus the alternative that not all parameters are equal. Under the null hypothesis, the combined sample is a random sample from a Bernoulli distribution with parameter $p = p_1 = p_2 = \cdots = p_k$. The parameter sets for this test are

$$\Omega = \{(p_1, p_2, \ldots, p_k) : 0 \le p_1, p_2, \ldots, p_k \le 1\} \text{ and}$$
$$\omega_o = \{(p, p, \ldots, p) : 0 \le p \le 1\}.$$

There are k free parameters in Ω and 1 free parameter in ω_o.

The statistic $-2\log(\Lambda)$ simplifies to

$$-2\log(\Lambda) = \sum_{i=1}^{k}\left[2Y_i \log\left(\frac{Y_i}{n_i\widehat{p}}\right) + 2(n_i - Y_i)\log\left(\frac{n_i - Y_i}{n_i(1-\widehat{p})}\right)\right],$$

where $\widehat{p}$ is the estimate of the common parameter under the null hypothesis,

$$\widehat{p} = \frac{Y_1 + Y_2 + \cdots + Y_k}{n_1 + n_2 + \cdots + n_k},$$

and log() is the natural logarithm function. If each n_i is large, then $-2\log(\Lambda)$ has an approximate chi-square distribution with $(k-1)$ degrees of freedom.

For example, assume the table below summarizes the values of independent random samples from four Bernoulli distributions, and consider testing the null hypothesis $p_1 = p_2 = p_3 = p_4$ at the 5% significance level.

y_i	41	26	13	17
n_i	219	102	95	49
y_i/n_i	0.1872	0.2549	0.1368	0.3469

The estimated common proportion is $\widehat{p} = 97/465 = 0.2086$, and the observed value of $-2\log(\Lambda)$ is 10.156. The observed significance level, based on the chi-square distribution with 3 degrees of freedom, is $P(-2\log(\Lambda) \ge 10.156) = 0.0173$. Since the p value is less than 0.05, the null hypothesis that the parameters are equal is rejected. Notice, in particular, that the largest sample proportion is more than twice the smallest proportion.

Example: Comparing Poisson parameters

Let Y_i be the sample sum of a random sample of size n_i from a Poisson distribution with parameter λ_i for $i = 1, 2, \ldots, k$.

Consider testing the null hypothesis that the k Poisson parameters are equal versus the alternative that not all parameters are equal. Under the null hypothesis, the combined sample is a random sample from a Poisson distribution with parameter $\lambda = \lambda_1 = \lambda_2 = \cdots = \lambda_k$. The parameter sets for this test are

$$\Omega = \{(\lambda_1, \lambda_2, \ldots, \lambda_k) : \lambda_1, \lambda_2, \ldots, \lambda_k \geq 0\} \text{ and } \omega_o = \{(\lambda, \lambda, \ldots, \lambda) : \lambda \geq 0\}.$$

There are k free parameters in Ω and 1 free parameter in ω_o.

The statistic $-2\log(\Lambda)$ simplifies to

$$-2\log(\Lambda) = \sum_{i=1}^{k} 2Y_i \log\left(\frac{Y_i}{n_i\widehat{\lambda}}\right),$$

where $\widehat{\lambda}$ is the estimate of the common parameter under the null hypothesis,

$$\widehat{\lambda} = \frac{Y_1 + Y_2 + \cdots + Y_k}{n_1 + n_2 + \cdots + n_k},$$

and log() is the natural logarithm function. If each mean ($E(Y_i) = n_i\lambda$) is large, then $-2\log(\Lambda)$ has an approximate chi-square distribution with $(k-1)$ degrees of freedom.

For example, as part of a study of incidence of childhood leukemia in upstate New York, data were collected on the number of children contracting the disease in the 5-year period from 1978 to 1982 [111]. The table below summarizes results using geographic regions, running from west to east, of equal total population.

	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6
#Cases	89	86	97	96	120	102

Let λ_i be the average number of new cases in Region i for a 5-year period and $n_i = 1$ for $i = 1, 2, \ldots, 6$. Assume the information above are the values of independent Poisson random variables with parameters λ_i, and consider testing the null hypothesis $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = \lambda_6$ at the 5% significance level.

The estimated common 5-year rate is $\widehat{\lambda} = 98.33$ cases, and the observed value of $-2\log(\Lambda)$ is 7.20. The observed significance level, based on the chi-square distribution with 5 degrees of freedom, is $P(-2\log(\Lambda) \geq 7.20) = 0.2062$. Since the p value is greater than 0.05, the null hypothesis that the 5-year rates for geographic regions running from west to east are equal is accepted.

Example: Multinomial goodness-of-fit

Assume that $(X_1, X_2, \ldots, X_k)$ has a multinomial distribution with parameters n and $(p_1, p_2, \ldots, p_k)$. Consider testing the null hypothesis $p_i = p_{i_o}$ for $i = 1, 2, \ldots, k$ versus the alternative that the given model for probabilities does not hold. Then

$$\Omega = \left\{(p_1, p_2, \ldots, p_k) : 0 \leq p_1, p_2, \ldots, p_k \leq 1, \sum_i p_i = 1\right\}$$

and $r = k - 1$. There are two cases to consider.

Case 1. If the model for probabilities is known, then ω_o contains a single k-tuple and has 0 free parameters. The statistic $-2\log(\Lambda)$ simplifies to

$$-2\log(\Lambda) = \sum_{i=1}^{k} 2X_i \log\left(\frac{X_i}{np_{i_o}}\right),$$

where log() is the natural logarithm function. If n is large, then $-2\log(\Lambda)$ has an approximate chi-square distribution with $(k-1)$ degrees of freedom.

Case 2. If e parameters need to be estimated (ω_o has e free parameters), then the statistic $-2\log(\Lambda)$ simplifies to

$$-2\log(\Lambda) = \sum_{i=1}^{k} 2X_i \log\left(\frac{X_i}{n\widehat{p_{i_o}}}\right),$$

where log() is the natural logarithm function and $\widehat{p_{i_o}}$ is the estimated value of p_{i_o} for $i = 1, 2, \ldots, k$. If n is large, then $-2\log(\Lambda)$ has an approximate chi-square distribution with $(k-1-e)$ degrees of freedom.

For example, consider testing the goodness-of-fit of the survival model

$$(p_1, p_2, p_3, p_4) = ((1-\theta), \theta(1-\theta), \theta^2(1-\theta), \theta^3)$$

when $n = 1000$ and $(x_1, x_2, x_3, x_4) = (363, 205, 136, 296)$ is observed. (See page 92.) The ML estimate of θ is 0.660, and the observed value of $-2\log(\Lambda)$ is 4.515. The observed significance level, based on the chi-square distribution with 2 degrees of freedom, is $P(-2\log(\Lambda) \geq 4.515) = 0.105$. Since the p value is greater than 0.05, the null hypothesis that the probabilities have the form above is accepted.

Note that the value of Pearson's statistic for these data is 4.472, which is quite close to the value of $-2\log(\Lambda)$.

Multivariable calculus can be used to demonstrate that Pearson's statistic is a second-order Taylor approximation of $-2\log(\Lambda)$. Thus, Pearson's goodness-of-fit test is an approximate likelihood ratio test when ML estimates are used for free parameters.

8.8 Relationship with confidence intervals

Confidence intervals can sometimes be used as an alternative method to report the results of a hypothesis test. For example, consider testing the null hypothesis that $\mu = 120$ versus the alternative hypothesis that $\mu \neq 120$ at the 5% significance level using a random sample of size n from a normal distribution with standard deviation 5. Then, the null hypothesis would be accepted if and only if the value of 120 is in the 95% confidence interval for μ constructed using the sample data.

Similarly, hypothesis tests can sometimes be *inverted* to produce confidence interval procedures. Some examples are given in the laboratory problems, and others examples will be discussed in later chapters.

8.9 Laboratory problems

Laboratory problems for this chapter introduce *Mathematica* commands for conducting tests for normal means and variances, for designing tests and computing power in single parameter families, and for constructing plots of simulated test statistics. The problems are designed to reinforce ideas related to hypothesis testing theory.

8.9.1 Laboratory: Hypothesis testing

In the main laboratory notebook (Problems 1 to 5), you are asked to use simulation to study test procedures for normal samples; apply test procedures for normal samples to data from a physical anthropology study [11] and answer questions about the estimated model; construct tests for single parameter models and display power at fixed alternatives; construct power curves; and find sample sizes for a proposed cholesterol-reduction study and a proposed traffic-pattern study. Normal and Poisson models are used.

8.9.2 Additional problem notebooks

Problem 6 applies test and confidence interval methods for normal samples to data from a study on treatments for anorexia [50].

Problem 7 is a study design question for samples from the subfamily of normal distributions with known mean. An industrial setting is used.

Problems 8, 9, and 10 concern constructing confidence intervals by inverting hypothesis tests. Problem 8 considers binomial distributions and the anorexia treatment data from Problem 6. Problem 9 considers hypergeometric distributions and data from an EPA study on mercury contamination in Maine lakes [54]. Problem 10 considers hypergeometric distributions and data from a study of incidence of the disease spina bifida [89].

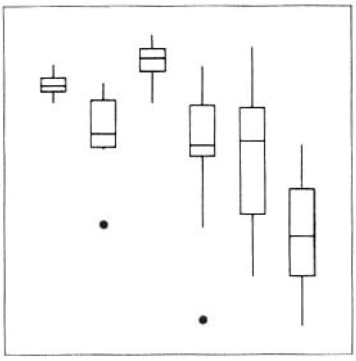
Problems 11 and 12 apply large sample test and confidence interval methods for means to differences data. Problem 11 uses differences in calcium content using two different measurement techniques [52], [90]. Problem 12 uses differences in daily maximum ozone levels in two different cities in the northeast [25].

In Problem 13, simulation is used to study the power of the t test for normal means and the chi-square test for normal variances.

In Problem 14, simulation is used to determine if t tests and chi-square tests remain valid when samples are drawn from distributions other than the normal. Exponential, gamma, and uniform models are used.

Problems 15 and 16 are applications of the likelihood ratio goodness-of-fit test method. Problem 15 uses data on memory and stressful events [109], [48]. Problem 16 uses data from a genetics study [24], [65].

This page intentionally left blank



Chapter 9
Order Statistics and Quantiles

In many statistical applications, interest focuses on estimating the quantiles of a continuous distribution or conducting hypothesis tests about the quantiles. For example, a medical researcher might be interested in determining the median lifetime of patients with a serious illness, or a geophysicist might be interested in determining the 90th percentile of the distribution of earthquake magnitudes in a region.

This chapter introduces methods for estimating quantiles of continuous distributions. Order statistics are defined and studied in the first section. Procedures for estimating quantiles and for constructing confidence intervals for quantiles based on order statistics are given in the next two sections. Two graphical methods are also introduced. Section 4 outlines the laboratory problems.

9.1 Order statistics

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a continuous distribution with PDF $f(x)$ and CDF $F(x)$, and let k be an integer between 1 and n. The k^{th} *order statistic*, $X_{(k)}$, is the k^{th} observation in order:

$$X_{(k)} \text{ is the } k^{th} \text{ smallest of } X_1, X_2, \ldots, X_n.$$

The largest observation, $X_{(n)}$, is called the *sample maximum* and the smallest observation, $X_{(1)}$, is called the *sample minimum*.

Sample median

The *sample median* is the middle observation when n is odd and the average of the two middle observations when n is even:

$$\text{Sample median} = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{when } n \text{ is odd,} \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right) & \text{when } n \text{ is even.} \end{cases}$$

Distribution of the sample maximum

Let $X_{(n)}$ be the sample maximum. The CDF and PDF of $X_{(n)}$ are as follows:

$$F_{(n)}(x) = (F(x))^n \text{ for all real numbers } x,$$
$$f_{(n)}(x) = n(F(x))^{n-1} f(x) \text{ for all real numbers } x.$$

To demonstrate that the formula for $F_{(n)}(x)$ is correct, observe that

$$\begin{aligned} F_{(n)}(x) &= P(X_{(n)} \le x) \\ &= P(X_1 \le x, X_2 \le x, \ldots, X_n \le x) \\ &= \textstyle\prod_i P(X_i \le x) \quad \text{by independence} \\ &= (F(x))^n. \end{aligned}$$

(The maximum value is x or less if and only if all n values are x or less.) The formula for the PDF is obtained by applying the chain rule.

For example, if X is a uniform random variable on the interval $[a, b]$, then

$$F_{(n)}(x) = \begin{cases} 0 & \text{when } x \le a, \\ \left(\frac{x-a}{b-a}\right)^n & \text{when } a < x < b, \\ 1 & \text{when } x \ge b, \end{cases}$$

and $f_{(n)}(x) = n(x-a)^{n-1}/(b-a)^n$ when $a < x < b$ and 0 otherwise.

Distribution of the sample minimum

Let $X_{(1)}$ be the sample minimum. The CDF and PDF of $X_{(1)}$ are as follows:

$$F_{(1)}(x) = 1 - (1 - F(x))^n \text{ for all real numbers } x,$$
$$f_{(1)}(x) = n(1 - F(x))^{n-1} f(x) \text{ for all real numbers } x.$$

To demonstrate that the formula for $F_{(1)}(x)$ is correct, observe that

$$\begin{aligned} F_{(1)}(x) &= 1 - P(X_{(1)} > x) \\ &= 1 - P(X_1 > x, X_2 > x, \ldots, X_n > x) \\ &= 1 - \textstyle\prod_i P(X_i > x) \quad \text{by independence} \\ &= 1 - (1 - F(x))^n. \end{aligned}$$

(The minimum is greater than x if and only if all n values are greater than x.) The formula for the PDF is obtained by applying the chain rule.

For example, if X is an exponential random variable with parameter λ, then

$$F_{(1)}(x) = 1 - \left(e^{-\lambda x}\right)^n = 1 - e^{-n\lambda x} \text{ when } x > 0 \text{ and 0 otherwise,}$$

and $f_{(1)}(x) = n\lambda e^{-n\lambda x}$ when $x > 0$ and 0 otherwise. Note that the sample minimum is an exponential random variable with parameter $n\lambda$.

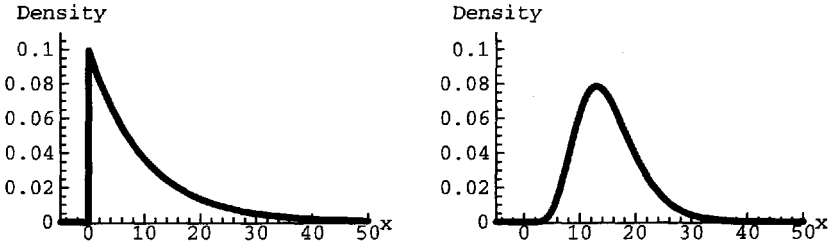


Figure 9.1. *PDF of an exponential random variable with parameter* $1/10$ *(left plot) and the* 9^{th} *order statistic of a random sample of size* 11 *from the exponential distribution (right plot).*

Distribution in the general case

Let $X_{(k)}$ be the k^{th} order statistic with $1 < k < n$. The CDF and PDF of $X_{(k)}$ are as follows:

$$F_{(k)}(x) = \sum_{j=k}^{n} \binom{n}{j} (F(x))^j (1 - F(x))^{n-j} \text{ for all real numbers } x,$$

$$f_{(k)}(x) = \binom{n}{k-1,1,n-k} (F(x))^{k-1} f(x) (1 - F(x))^{n-k} \text{ for all real numbers } x.$$

To demonstrate that the formula for $F_{(k)}(x)$ is correct, first note that

$$F_{(k)}(x) = \mathrm{P}(X_{(k)} \leq x) = \mathrm{P}(k \text{ or more } X_i\text{'s are} \leq x).$$

The probability that exactly j observations are $\leq x$ is a binomial probability,

$$\binom{n}{j} p^j (1-p)^{n-j}, \text{ where } p = P(X \leq x) = F(x),$$

and the formula for $F_{(k)}(x)$ is the sum of binomial probabilities. The formula for $f_{(k)}(x)$ is obtained using the product and chain rules for derivatives.

For example, if X is an exponential random variable with parameter $\frac{1}{10}$, and $X_{(9)}$ is the 9^{th} order statistic of a random sample of size 11 from the X distribution, then the CDF of $X_{(9)}$ is

$$F_{(9)}(x) = \sum_{j=9}^{11} \binom{11}{j} \left(1 - e^{-x/10}\right)^j \left(e^{-x/10}\right)^{11-j} \text{ when } x > 0 \text{ and 0 otherwise,}$$

and the PDF is

$$f_{(9)}(x) = \tfrac{99}{2} \left(1 - e^{-x/10}\right)^8 e^{-3x/10} \text{ when } x > 0 \text{ and 0 otherwise.}$$

The left part of Figure 9.1 is a graph of the PDF of X, and the right part is a graph of the PDF of $X_{(9)}$. Further, $P(X \leq 12) = 0.699$ and $P(X_{(9)} \leq 12) = 0.310$.

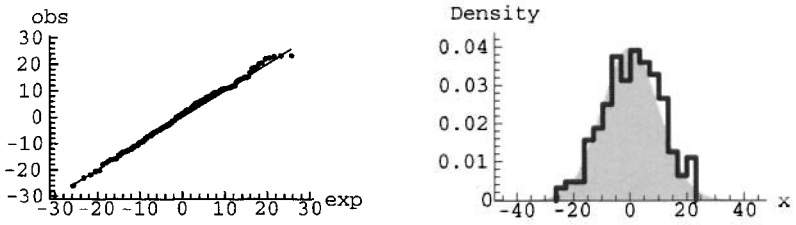


Figure 9.2. *Normal probability plot (left plot) and empirical histogram (right plot) for a simulated sample from a normal distribution. The empirical histogram is superimposed on a plot of the normal density function.*

9.1.1 Approximate mean and variance

Finding the mean and variance of a k^{th} order statistic can sometimes be difficult. The following theorem gives useful approximate formulas for these summary measures.

Theorem 9.1 (Approximate Summaries). *Let X be a continuous random variable with PDF $f(x)$, $X_{(k)}$ be the k^{th} order statistic of a random sample of size n from the X distribution, and θ be the p^{th} quantile of the X distribution, where $p = \frac{k}{n+1}$. If $f(\theta) \neq 0$, then*

$$E(X_{(k)}) \approx \theta \text{ and } Var(X_{(k)}) \approx \frac{p(1-p)}{(n+2)(f(\theta))^2}.$$

The formulas given in the theorem are exact for uniform distributions. For example, let X be a uniform random variable on the interval $[0, 10]$, $n = 4$, and $k = 3$. Then $p = 0.6$, $\theta = x_{0.6} = 6$, $E(X_{(3)}) = 6$, and $Var(X_{(3)}) = 4$.

Probability plots

Theorem 9.1 implies that order statistics can be used as estimators of quantiles. In particular, the k^{th} order statistic can be used to estimate the $(k/(n+1))^{st}$ quantile for $k = 1, 2, \ldots, n$. An interesting graphical comparison of a model with sample data uses this result.

A *probability plot* is a plot of pairs of the form

$$\left(\left(\tfrac{k}{n+1}\right)^{st} \text{quantile},\ x_{(k)}\right) \text{ for } k = 1, 2, \ldots, n,$$

where $x_{(k)}$ is the observed k^{th} order statistic for each k.

For example, let X be a normal random variable with mean 0 and standard deviation 10. The left part of Figure 9.2 is a normal probability plot of a simulated sample of size 195 from the X distribution. Observed order statistics (vertical axis) are paired with approximate expected values (horizontal axis). The gray line is the line $y = x$. Since the sample was drawn from the X distribution, the points are close to the line.

The right part of Figure 9.2 is another graphical comparison: an empirical histogram of the simulated data is superimposed on a graph of the density function of X (filled plot). Once again, the comparison is good.

If n is large, then both plots give good graphical comparisons of model and data. If n is small to moderate, then the probability plot may be better since the shapes of the empirical histogram and density curve may be quite different.

9.2 Confidence intervals for quantiles

This section presents an approximate confidence interval method for the median of a continuous distribution and exact confidence interval methods for p^{th} quantiles.

9.2.1 Approximate distribution of the sample median

Let X be a continuous random variable with median θ. The following theorem says that, under certain conditions, the sampling distribution of the sample median is approximately normal with mean θ.

Theorem 9.2 (Approximate Distribution). *Let X be a continuous random variable with median θ, and let*

$$\widehat{\theta} = X_{(\frac{n+1}{2})}$$

be the sample median of a random sample of size n, where n is odd. If n is large and $f(\theta) \neq 0$, then the distribution of the sample median is approximately normal with mean θ and variance $1/(4n(f(\theta))^2)$.

Approximate confidence interval for the median

Under the conditions of Theorem 9.2, an approximate $100(1-\alpha)\%$ confidence interval for the median θ has the form

$$\widehat{\theta} \pm z(\alpha/2)\sqrt{\frac{1}{4n\left(\widehat{f(\theta)}\right)^2}}, \qquad \text{where } \widehat{\theta} = X_{(\frac{n+1}{2})},$$

and $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution. In this formula, $\widehat{f(\theta)}$ is the estimate of $f(\theta)$ obtained by substituting the sample median for θ.

Example: Cauchy distribution

Let X be a Cauchy random variable with center (and median) θ and scale 1. The PDF of X is

$$f(x) = \frac{1}{\pi(1+(x-\theta)^2)} \quad \text{for all real numbers } x$$

and $f(\theta) = 1/\pi$.

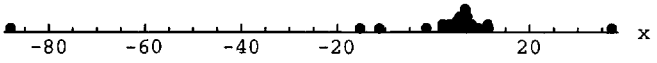


Figure 9.3. *Simulated sample from a Cauchy distribution.*

Figure 9.3 is a dot plot of a simulated sample of size 75 from a Cauchy distribution with scale 1. The observed sample median is 5.985, and an approximate 90% confidence interval for θ is computed as follows:

$$5.985 \pm 1.645\sqrt{\pi^2/(4(75))} \implies [5.687, 6.283].$$

9.2.2 Exact confidence interval procedure

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a continuous distribution, and let θ be the p^{th} quantile of the distribution.

The order statistics, $X_{(k)}$, divide the real line into $n+1$ intervals

$$(-\infty, X_{(1)}),\ (X_{(1)}, X_{(2)}), \ldots, (X_{(n-1)}, X_{(n)}),\ (X_{(n)}, \infty)$$

(ignoring the endpoints). The probability that θ lies in a given interval follows a binomial distribution with parameters n and p. Specifically,

$$\begin{aligned} P(\theta < X_{(1)}) &= (1-p)^n, \\ P(X_{(k)} < \theta < X_{(k+1)}) &= \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 1, 2, \ldots, n-1, \\ P(\theta > X_{(n)}) &= p^n. \end{aligned}$$

These facts can be used to prove the following theorem.

Theorem 9.3 (Quantile Confidence Interval). *Under the assumptions above, if k_1 and k_2 are chosen so that*

$$\begin{aligned} P(\theta < X_{(k_1)}) &= \sum_{j=0}^{k_1-1} \binom{n}{j} p^j (1-p)^{n-j} = \alpha/2, \\ P(X_{(k_1)} < \theta < X_{(k_2)}) &= \sum_{j=k_1}^{k_2-1} \binom{n}{j} p^j (1-p)^{n-j} = 1-\alpha, \\ P(\theta > X_{(k_2)}) &= \sum_{j=k_2}^{n} \binom{n}{j} p^j (1-p)^{n-j} = \alpha/2, \end{aligned}$$

then the interval $[X_{(k_1)}, X_{(k_2)}]$ is a $100(1-\alpha)\%$ confidence interval for θ.

In applications of Theorem 9.3, k_1 and k_2 are chosen so that

$$P(\theta < X_{(k_1)}) \approx P(\theta > X_{(k_2)}) \approx \frac{\alpha}{2}.$$

Table 9.1. *Confidence intervals for quartiles of a continuous distribution.*

Quantile	Confidence Interval	Confidence Level
0.25	$[x_{(13)}, x_{(25)}] = [4.610, 5.380]$	$\sum_{j=13}^{24} \binom{75}{j} 0.25^j 0.75^{75-j} = 0.8912$
0.50	$[x_{(31)}, x_{(45)}] = [5.652, 6.301]$	$\sum_{j=31}^{44} \binom{75}{j} 0.50^j 0.50^{75-j} = 0.8945$
0.75	$[x_{(51)}, x_{(63)}] = [6.485, 7.302]$	$\sum_{j=51}^{62} \binom{75}{j} 0.75^j 0.25^{75-j} = 0.8912$

For moderate to large data sets, it is best to let the computer do the work.

The method is valid for any continuous distribution with p^{th} quantile θ and any sample size. It does not require precise knowledge of the X distribution.

Table 9.1 displays confidence intervals for the quartiles of a continuous distribution using the simulated sample of size 75 displayed in Figure 9.3 and the methods of this section. The intervals were constructed to have confidence level as close to 0.90 as possible. Note that the confidence interval for the median is close to the one computed in the Cauchy example on page 121.

9.3 Sample quantiles

Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution, and let θ be the p^{th} quantile of the distribution, where $\frac{1}{n+1} \le p \le \frac{n}{n+1}$. Then the p^{th} *sample quantile*, $\widehat{\theta}$, is defined as follows:

$$\widehat{\theta} = \begin{cases} X_{(k)} & \text{when } p = \frac{k}{n+1}, \\ X_{(k)} + ((n+1)p - k)\left(X_{(k+1)} - X_{(k)}\right) & \text{when } \frac{k}{n+1} < p < \frac{k+1}{n+1}, \end{cases}$$

where $X_{(k)}$ is the k^{th} order statistic for $k = 1, 2, \ldots, n$.

When $p = k/(n+1)$, the p^{th} sample quantile is the k^{th} order statistic; otherwise, it is defined so that the following three points lie on a single line:

$$\left(X_{(k)}, \frac{k}{n+1}\right), \quad (\widehat{\theta}, p), \quad \left(X_{(k+1)}, \frac{k+1}{n+1}\right).$$

Note that when $p = 0.50$, the definition given above reduces to the definition of the sample median given earlier.

9.3.1 Sample quartiles, sample IQR

Estimates of p^{th} quantiles, where $p = 0.25, 0.50, 0.75$, are called the *sample quartiles* and are denoted by q_1, q_2, and q_3, respectively (q_2 is also the sample median). The difference $q_3 - q_1$ is called the *sample interquartile range* (sample IQR).

Table 9.2. *Lifetimes (in days) of guinea pigs exposed to an infectious disease.*

Low Exposure:														
33	44	56	59	74	77	93	100	102	105	107	107	108	108	109
115	120	122	124	136	139	144	153	159	160	163	163	168	171	172
195	202	215	216	222	230	231	240	245	251	253	254	278	458	555
Medium Exposure:														
10	45	53	56	56	58	66	67	73	81	81	81	82	83	88
91	91	92	92	97	99	99	102	102	103	104	107	109	118	121
128	138	139	144	156	162	178	179	191	198	214	243	249	380	522
High Exposure:														
15	22	24	32	33	34	38	38	43	44	54	55	59	60	60
60	61	63	65	65	67	68	70	70	76	76	81	83	87	91
96	98	99	109	127	129	131	143	146	175	258	263	341	341	376

For example, as part of a study on the effects of an infectious disease on the lifetimes of guinea pigs, more than 400 animals were infected [16], [90, p. 349]. Table 9.2 gives the lifetimes (in days) of 45 animals in each of three exposure groups. In the low exposure group, the sample median is 153 days and the sample IQR is 112 days. In the medium exposure group, the sample median is 102 days and the sample IQR is 69 days. In the high exposure group, the sample median is 70 days and the sample IQR is 63.5 days.

9.3.2 Box plots

A *box plot* is a graphical display of a data set that shows the sample median, the sample IQR, and the presence of possible outliers (numbers that are far from the center). Box plots were introduced by J. Tukey in the 1970's.

Box plot construction

To construct a box plot, the following is done:

1. A *box* is drawn from the first to the third sample quartiles (q_1 to q_3).
2. A *bar* is drawn through the box at the sample median (q_2).
3. A *whisker* is drawn from q_3 to the largest observation that is less than or equal to $q_3 + 1.50(q_3 - q_1)$. Another whisker is drawn from q_1 to the smallest observation that is greater than or equal to $q_1 - 1.50(q_3 - q_1)$.

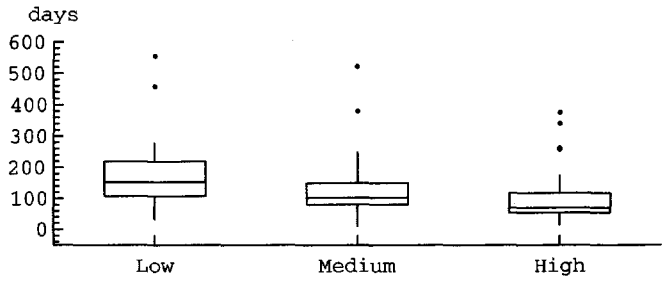


Figure 9.4. *Side-by-side box plots of the lifetimes data.*

4. Observations outside the interval

$$[q_1 - 1.50(q_3 - q_1), q_3 + 1.50(q_3 - q_1)]$$

are drawn as separate points. These observations are called the *outliers*.

Figure 9.4 shows side-by-side box plots of the data on lifetimes of guinea pigs from Table 9.2. The plot suggests a strong relationship between level of exposure and lifetime. For the low, medium, and high exposure groups, the estimated median lifetimes are 153 days, 102 days, and 70 days, respectively. In each case, there are large outliers. In addition, as exposure increases, the sample distributions become more skewed. (In each case, the distance between the first and second sample quartiles is smaller than the distance between the second and third sample quartiles. As the exposure increases, the differences are more pronounced.)

9.4 Laboratory problems

Laboratory problems for this chapter introduce *Mathematica* commands for plotting the distributions of order statistics, for computing quantile confidence intervals and sample quantiles, and for constructing probability plots and box plots. The problems are designed to reinforce ideas related to order statistics and quantiles.

9.4.1 Laboratory: Order statistics and quantiles

In the main laboratory notebook (Problems 1 to 5), you will use graphics, simulation, and probability computations to study order statistic distributions; apply quantile estimation methods to data on daily maximum ozone levels in two cities in the northeast [25]; use simulation to study the components of box plots; and apply quantile estimation methods and box plots to data from a cholesterol-reduction study [36]. Gamma, normal, and exponential models are used.

9.4.2 Additional problem notebooks

Problems 6, 7, and 8 are on uniform distributions. Problem 6 considers the mean and *mode* (point of maximum density) of order statistics of a random sample from a

uniform distribution. Problem 7 uses simulation to study ML estimation and confidence procedures in subfamilies with known lower endpoint or with known upper endpoint. Problem 8 is a study design question for samples from the subfamily of uniform distributions with known lower endpoint. A traffic-pattern setting is used.

Problem 9 compares several methods for constructing median confidence intervals. Data from a study on plasma retinol levels in women is used [104]. The problem includes a goodness-of-fit analysis.

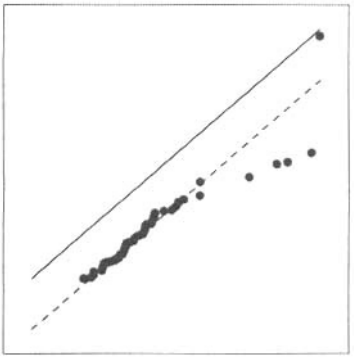
The center of the normal distribution is both the mean and the median of the distribution. Problem 10 uses simulation to determine how confidence interval procedures for the center change as one sample value changes. Normal probability plots are used to visualize the changes.

Problem 11 applies several computational and graphical methods, including box plots and gamma probability plots, to two data sets from a study on lifetimes of components under sustained pressure [10], [6].

Problem 12 uses simulation to determine if normal probability plots can be used as diagnostic tools. Exponential, gamma, uniform, and Laplace models are considered.

Problem 13 uses quantile methods and box plots to study factors related to mercury contamination in Maine lakes [54]. Problem 14 uses quantile methods and box plots to study factors related to plasma levels of beta-carotene in women [104].

Problems 15 and 16 consider properties of the *sign test* for medians. In Problem 15, upper tail tests of the null hypothesis $a = a_o$ of a Cauchy distribution when the scale parameter is known are studied. In problem 16, two tailed tests of the null hypothesis $\mu = \mu_o$ of a normal distribution when the standard deviation is known are studied.



Chapter 10
Two Sample Analysis

In many statistical applications, interest focuses on comparing two probability distributions. For example, an education researcher might be interested in determining if the distributions of standardized test scores for students in public and private schools are equal, or a medical researcher might be interested in determining if mean blood pressure levels are the same in patients on two different treatment protocols.

This chapter considers statistical methods for comparing independent random samples from two continuous distributions. Methods for samples from normal distributions are given in the first two sections. Large sample methods for the difference in means are given in Section 3. Methods applicable to a broad range of distributions are given in Section 4. Section 5 considers sampling and study design questions, and Section 6 outlines the laboratory problems for this chapter. A general reference for the material in Section 4 is [68].

10.1 Normal distributions: Difference in means

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be independent random samples, of sizes n and m, from normal distributions with parameters

$$\mu_x = E(X), \quad \sigma_x = SD(X), \quad \mu_y = E(Y), \quad \sigma_y = SD(Y).$$

This section focuses on answering statistical questions about the difference in means, $\mu_x - \mu_y$. Note that the difference in sample means, $\overline{X} - \overline{Y}$, can be used to estimate the difference in means. By Theorem 4.6, $\overline{X} - \overline{Y}$ is a normal random variable with

$$E\left(\overline{X} - \overline{Y}\right) = \mu_x - \mu_y \quad \text{and} \quad Var\left(\overline{X} - \overline{Y}\right) = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}.$$

10.1.1 Known variances

Assume that σ_x and σ_y are known.

Confidence intervals for $\mu_x - \mu_y$

If σ_x and σ_y are known, then

$$(\overline{X} - \overline{Y}) \pm z(\alpha/2)\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$$

is a $100(1-\alpha)\%$ confidence interval for $\mu_x - \mu_y$, where $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

Tests of $\mu_x - \mu_y = \delta_o$

If σ_x and σ_y are known, then the standardized difference when $\mu_x - \mu_y = \delta_o$,

$$Z = \frac{(\overline{X} - \overline{Y}) - \delta_o}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}},$$

can be used as test statistic. The following table gives the rejection regions for one sided and two sided $100\alpha\%$ tests:

Alternative Hypothesis	Rejection Region
$\mu_x - \mu_y < \delta_o$	$Z \le -z(\alpha)$
$\mu_x - \mu_y > \delta_o$	$Z \ge z(\alpha)$
$\mu_x - \mu_y \neq \delta_o$	$\lvert Z\rvert \ge z(\alpha/2)$

where $z(p)$ is the $100(1-p)\%$ point of the standard normal distribution.

For example, consider testing the null hypothesis $\mu_x - \mu_y = 4$ versus the alternative hypothesis $\mu_x - \mu_y \neq 4$ at the 5% significance level using samples of sizes $n = 8$ and $m = 12$. The rejection region is $|Z| \ge 1.960$. If $\sigma_x = \sigma_y = 2$ and the observed difference in means is $\overline{x} - \overline{y} = 3.27$, then the observed value of Z is

$$z_{\text{obs}} = \frac{(\overline{x} - \overline{y}) - 4}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} = \frac{3.27 - 4}{\sqrt{\frac{4}{8} + \frac{4}{12}}} = -0.80.$$

Since $|z_{\text{obs}}| < 1.960$, the null hypothesis is accepted. Further, a 95% confidence interval for $\mu_x - \mu_y$ is

$$3.27 \pm 1.960\sqrt{\frac{4}{8} + \frac{4}{12}} \implies [1.48, 5.06].$$

Note that the confidence interval contains 4.

10.1.2 Pooled t methods

Pooled t methods are used when X and Y have a common unknown variance. Let $\sigma^2 = \sigma_x^2 = \sigma_y^2$ be the common variance. The *pooled estimate* of σ^2, S_p^2, is defined as follows:

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2},$$

where S_x^2 and S_y^2 are the sample variances of the X and Y samples, respectively.

The statistic S_p^2 is a weighted average of the separate estimates of σ^2. If $n = m$, then the weights are equal; otherwise, the estimate based on the larger sample is given the larger weight.

S_p^2 is an unbiased estimator of σ^2. The following theorem says that the approximate standardization of the difference in sample means has a Student t distribution.

Theorem 10.1 (Approximate Standardization). *Under the assumptions of this section, the statistic*

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

has a Student t distribution with $(n+m-2)$ degrees of freedom.

Confidence intervals for $\mu_x - \mu_y$

If the value of $\sigma^2 = \sigma_x^2 = \sigma_y^2$ is estimated from the data, then Theorem 10.1 can be used to demonstrate that

$$(\overline{X} - \overline{Y}) \pm t_{n+m-2}(\alpha/2)\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}$$

is a $100(1-\alpha)\%$ confidence interval for $\mu_x - \mu_y$, where $t_{n+m-2}(\alpha/2)$ is the $100(1-\alpha/2)\%$ point on the Student t distribution with $(n+m-2)$ degrees of freedom.

Tests of $\mu_x - \mu_y = \delta_o$

If the value of $\sigma^2 = \sigma_x^2 = \sigma_y^2$ is estimated from the data, then the approximate standardization when $\mu_x - \mu_y = \delta_o$,

$$T = \frac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}},$$

can be used as test statistic. The following table gives the rejection regions for one sided and two sided $100\alpha\%$ tests:

Alternative Hypothesis	Rejection Region
$\mu_x - \mu_y < \delta_o$	$T \le -t_{n+m-2}(\alpha)$
$\mu_x - \mu_y > \delta_o$	$T \ge t_{n+m-2}(\alpha)$
$\mu_x - \mu_y \neq \delta_o$	$\lvert T\rvert \ge t_{n+m-2}(\alpha/2)$

where $t_{n+m-2}(p)$ is the $100(1-p)\%$ point of the Student t distribution with $(n+m-2)$ degrees of freedom.

10.1.3 Welch t methods

Welch t methods are used when X and Y have distinct unknown variances. The following theorem, proven by B. Welch in the 1930's, says that the approximate standardization of the difference in sample means has an approximate Student t distribution.

Theorem 10.2 (Welch Theorem). *Under the assumptions of this section, the statistic*

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

has an approximate Student t distribution with degrees of freedom as follows:

$$df = \frac{\left((S_x^2/n) + (S_y^2/m)\right)^2}{\left((S_x^2/n)^2/n + (S_y^2/m)^2/m\right)} - 2.$$

To apply the formula for *df*, you would round the expression to the closest whole number. The computed *df* satisfies the following inequality:

$$\min(n, m) - 1 \leq df \leq n + m - 2.$$

A quick by-hand method is to use the lower bound for *df* instead of Welch's formula.

Approximate confidence intervals for $\mu_x - \mu_y$

If the values of σ_x^2 and σ_y^2 are estimated from the data, then Theorem 10.2 can be used to demonstrate that

$$(\overline{X} - \overline{Y}) \pm t_{df}(\alpha/2)\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}$$

is an *approximate* $100(1-\alpha)\%$ confidence interval for $\mu_x - \mu_y$, where $t_{df}(\alpha/2)$ is the $100(1-\alpha/2)\%$ point on the Student t distribution with degrees of freedom computed using Welch's formula.

Approximate tests of $\mu_x - \mu_y = \delta_o$

If the values of σ_x^2 and σ_y^2 are estimated from the data, then the approximate standardization when $\mu_x - \mu_y = \delta_o$,

$$T = \frac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}},$$

can be used as test statistic. The following table gives the rejection regions for *approximate* one sided and two sided $100\alpha\%$ tests:

Alternative Hypothesis	Rejection Region
$\mu_x - \mu_y < \delta_o$	$T \leq -t_{df}(\alpha)$
$\mu_x - \mu_y > \delta_o$	$T \geq t_{df}(\alpha)$
$\mu_x - \mu_y \neq \delta_o$	$\lvert T\rvert \geq t_{df}(\alpha/2)$

where $t_{df}(p)$ is the $100(1-p)\%$ point of the Student t distribution with degrees of freedom computed using Welch's formula.

Comparison of pooled t and Welch t methods

To illustrate the differences in using pooled t and Welch t methods, consider constructing a 95% confidence interval for $\mu_x - \mu_y$ using samples of sizes $n = 8$ and $m = 12$. Assume that the observed difference in means is $\bar{x} - \bar{y} = 3.27$ and that the observed sample variances are $s_x^2 = 4.672$ and $s_y^2 = 2.435$.

(i) If the distributions have a common unknown variance, then the pooled estimate of the common variance is 3.305 and the confidence interval is

$$3.27 \pm t_{18}(0.025)\sqrt{3.305\left(\frac{1}{8}+\frac{1}{12}\right)} \Rightarrow 3.27 \pm 2.101\sqrt{0.689} \Rightarrow [1.527, 5.013].$$

(ii) If the distributions have distinct unknown variances, then the degrees of freedom formula yields $df = 11$ and the approximate confidence interval is

$$3.27 \pm t_{11}(0.025)\sqrt{\frac{4.672}{8}+\frac{2.435}{12}} \Rightarrow 3.27 \pm 2.201\sqrt{0.787} \Rightarrow [1.318, 5.222].$$

The interval produced using Welch t methods is slightly wider than the interval produced using pooled t methods.

10.2 Normal distributions: Ratio of variances

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be independent random samples, of sizes n and m, from normal distributions with parameters

$$\mu_x = E(X), \quad \sigma_x = SD(X), \quad \mu_y = E(Y), \quad \sigma_y = SD(Y).$$

The ratio of sample variances, S_x^2/S_y^2, is used to answer statistical questions about the ratio of model variances σ_x^2/σ_y^2 when the means are estimated from the data. By Theorem 6.3, the statistic

$$F = \frac{S_x^2/S_y^2}{\sigma_x^2/\sigma_y^2}$$

has an f ratio distribution with $(n-1)$ and $(m-1)$ degrees of freedom.

Confidence intervals for σ_x^2/σ_y^2

If μ_x and μ_y are estimated from the data, then Theorem 6.3 can be used to demonstrate that

$$\left[\frac{S_x^2/S_y^2}{f_{n-1,m-1}(\alpha/2)}, \frac{S_x^2/S_y^2}{f_{n-1,m-1}(1-\alpha/2)}\right]$$

is a $100(1-\alpha)\%$ confidence interval for σ_x^2/σ_y^2, where $f_{n-1,m-1}(p)$ is the $100(1-p)\%$ point of the f ratio distribution with $(n-1)$ and $(m-1)$ degrees of freedom.

Tests of $\sigma_x^2/\sigma_y^2 = r_o$

If μ_x and μ_y are estimated from the data, then the ratio when $\sigma_x^2/\sigma_y^2 = r_o$,

$$F = \frac{S_x^2/S_y^2}{r_o},$$

can be used as test statistic. The following table gives the rejection regions for one sided and two sided $100\alpha\%$ tests:

Alternative Hypothesis	Rejection Region
$\sigma_x^2/\sigma_y^2 < r_o$	$F \le f_{n-1,m-1}(1-\alpha)$
$\sigma_x^2/\sigma_y^2 > r_o$	$F \ge f_{n-1,m-1}(\alpha)$
$\sigma_x^2/\sigma_y^2 \ne r_o$	$F \le f_{n-1,m-1}(1-\alpha/2)$ or $F \ge f_{n-1,m-1}(\alpha/2)$

where $f_{n-1,m-1}(p)$ is the $100(1-p)\%$ point of the f ratio distribution with $(n-1)$ and $(m-1)$ degrees of freedom.

These tests are examples of f tests. An *f test* is a test based on a statistic with an f ratio distribution under the null hypothesis.

For example, consider testing the null hypothesis that the variances are equal ($r_o = 1$) versus the alternative hypothesis that the variances are not equal ($r_o \ne 1$) at the 5% significance level using samples of sizes $n = 8$ and $m = 12$. The rejection region is $F \le 0.212$ or $F \ge 3.759$. If the observed sample variances are $s_x^2 = 4.672$ and $s_y^2 = 2.435$, then the observed value of the f statistic is $f_{\text{obs}} = 1.919$. Since $0.212 < f_{\text{obs}} < 3.759$, the null hypothesis is accepted. Further, a 95% confidence interval for σ_x^2/σ_y^2 is

$$\left[\frac{1.919}{3.759}, \frac{1.919}{0.212}\right] = [0.511, 9.052].$$

Note that the confidence interval contains 1.

Example: Comparison of plasma beta-carotene distributions

Several studies have suggested that low plasma concentrations of beta-carotene (a precursor of vitamin A) may be associated with increased risk of certain types of

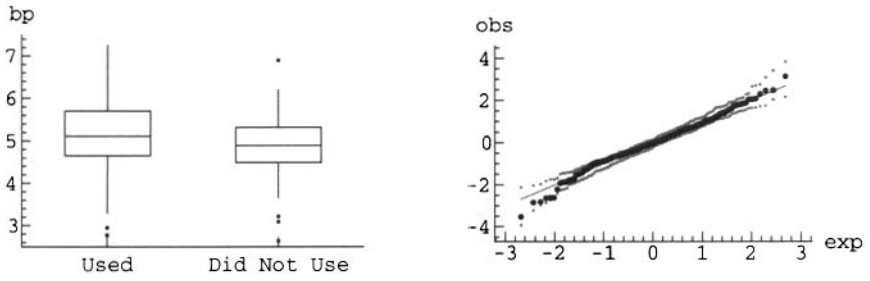


Figure 10.1. *Side-by-side box plots of plasma beta-carotene levels (in log-ng/ml) for women who used vitamin supplements regularly and for women who did not use supplements regularly (left plot) and an enhanced normal probability plot of standardized residuals (right plot).*

cancer. As part of a study to investigate the relationship between personal characteristics (including diet) and levels of beta-carotene in the blood, measurements were made on over 300 subjects [104]. This example uses a subset of their data.

The left plot in Figure 10.1 shows side-by-side box plots of plasma levels of beta-carotene for 108 women who regularly used vitamin supplements and 163 women who did not use supplements regularly. The scale is the natural logarithm of nanograms per milliliter (log-ng/ml). The right plot is an *enhanced* normal probability plot of *standardized residuals*. Construct the plot as follows:

(i) Each observation in the first group, x, is replaced by its standardized value, $(x - \overline{x})/s_x$, where $\overline{x}$ is the observed value of the sample mean and s_x is the observed value of the sample standard deviation.

(ii) Similarly, each observation in the second group, y, is replaced by $(y - \overline{y})/s_y$.

(iii) The combined list of 271 standardized values is compared to the standard normal distribution using a normal probability plot (black dots).

(iv) The normal probability plot is enhanced using the results of 100 simulations from the standard normal distribution (gray dots). Specifically, 100 random samples of size 271 are generated and the points

$$(k/272 \text{ quantile, minimum of } 100\ k^{th} \text{ order statistics}) \text{ and}$$
$$(k/272 \text{ quantile, maximum of } 100\ k^{th} \text{ order statistics})$$

for $k = 1, 2, \ldots, 271$ are drawn.

If these data are the values of independent random samples, then the right plot suggests that methods for samples from normal distributions can be used to analyze plasma levels of beta-carotene on the log scale.

Let X be the plasma beta-carotene level (in log-ng/ml) for women who use vitamin supplements regularly, and let Y be the corresponding level for women who do not regularly use supplements. Assume that X and Y are normal random variables and that the data shown in Figure 10.1 are the values of independent random samples from the X and Y distributions.

The observed ratio of sample variances is $s_x^2/s_y^2 = 0.710/0.409 = 1.736$, and a 95% confidence interval for σ_x^2/σ_y^2 is

$$\left[\frac{s_x^2/s_y^2}{f_{107,162}(0.025)}, \frac{s_x^2/s_y^2}{f_{107,162}(0.975)}\right] = \left[\frac{1.736}{1.406}, \frac{1.736}{0.703}\right] = [1.235, 2.469].$$

Since both endpoints are greater than 1, there is evidence that the variance of plasma levels of beta-carotene for women who use supplements is greater than the variance for women who do not use supplements.

The observed difference in sample means is $\bar{x} - \bar{y} = 0.275$, and an approximate 95% confidence interval for $\mu_x - \mu_y$, using Welch t methods, is

$$(\bar{x} - \bar{y}) \pm t_{186}(0.025)\sqrt{\frac{s_x^2}{108} + \frac{s_y^2}{163}} \implies 0.275 \pm 1.973\sqrt{\frac{7.10}{108} + \frac{0.409}{163}}$$

$$\implies [0.087, 0.463].$$

Since both endpoints are positive, there is evidence that the mean plasma level of beta-carotene for women who use supplements is greater than the mean for women who do not use supplements.

Welch t methods are used to construct the confidence interval for the difference in means since there is evidence that the variances are not equal.

10.3 Large sample: Difference in means

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be independent random samples, of sizes n and m, from continuous distributions with unknown (but finite) means and variances. Let $\overline{X}$ and $\overline{Y}$ be the sample means, and let S_x^2 and S_y^2 be the sample variances computed from these samples.

Approximate confidence intervals for $\mu_x - \mu_y$

If n and m are large, then Theorem 5.2 can be used to demonstrate that

$$(\overline{X} - \overline{Y}) \pm z(\alpha/2)\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}$$

is an *approximate* $100(1-\alpha)\%$ confidence interval for $\mu_x - \mu_y$, where $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point on the standard normal distribution.

Approximate tests of $\mu_x - \mu_y = \delta_o$

If n and m are large, then the approximate standardization when $\mu_x - \mu_y = \delta_o$,

$$Z = \frac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}},$$

can be used as test statistic. The following table gives the rejection regions for *approximate* one sided and two sided $100\alpha\%$ tests:

Alternative Hypothesis	Rejection Region
$\mu_x - \mu_y < \delta_o$	$Z \leq -z(\alpha)$
$\mu_x - \mu_y > \delta_o$	$Z \geq z(\alpha)$
$\mu_x - \mu_y \neq \delta_o$	$\lvert Z\rvert \geq z(\alpha/2)$

where $z(p)$ is the $100(1-p)\%$ point of the standard normal distribution.

Consider comparing the plasma beta-carotene distributions from page 132, using measurements in ng/ml instead of log-ng/ml. An approximate 5% test of the null hypothesis that the means are equal versus the alternative that they are not equal has rejection region $|Z| \geq 1.960$. The observed difference in sample means is $\bar{x} - \bar{y} = 88.334$, the observed sample variances are $s_x^2 = 65420.1$ and $s_y^2 = 13250.4$, and the observed value of the test statistic is $z_{\text{obs}} = 3.37$. Since $|z_{\text{obs}}| > 1.960$, the null hypothesis is rejected. Further, an approximate 95% confidence interval for the difference in means is

$$88.334 \pm 1.960\sqrt{\frac{65420.1}{108} + \frac{13250.4}{163}} \implies [36.960, 139.708].$$

The mean level of plasma beta-carotene for women who use vitamin supplements regularly is estimated to be between 36.960 ng/ml and 139.708 ng/ml higher than for women who do not use supplements regularly.

10.4 Rank sum test

A *distribution-free* (or *nonparametric*) method is a statistical procedure applicable to a broad range of distributions.

In the 1940's, F. Wilcoxon, H. Mann, and D. Whitney developed equivalent nonparametric methods for testing the null hypothesis that the X and Y distributions are equal versus alternatives that one distribution is stochastically larger than the other (see below). In some situations, confidence procedures for the difference in medians can be developed.

Stochastically larger; stochastically smaller

Let V and W be continuous random variables. V is *stochastically larger* than W (correspondingly, W is *stochastically smaller* than V) if

$$P(V \geq x) \geq P(W \geq x) \text{ for all real numbers } x$$

with strict inequality ($P(V \geq x) > P(W \geq x)$) for at least one x.

The definition is illustrated in Figure 10.2, where the V distribution is shown in gray and the W distribution in black. Note, in particular, that if F_v and F_w are the CDFs of V and W, respectively, then

$$F_v(x) \leq F_w(x) \text{ for all real numbers } x.$$

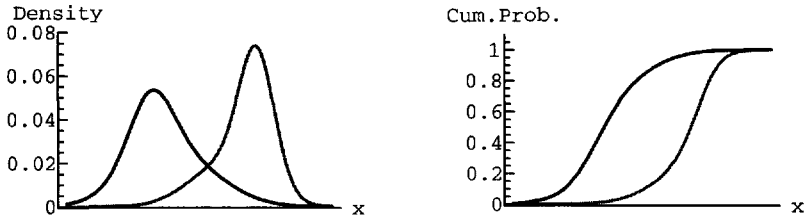


Figure 10.2. *PDFs (left plot) and CDFs (right plot) of two distributions. The distribution pictured in gray is stochastically larger than the distribution pictured in black.*

10.4.1 Rank sum statistic

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be independent random samples, of sizes n and m, from continuous distributions. The Wilcoxon *rank sum statistics* for the X sample (R_1) and for the Y sample (R_2) are computed as follows:

1. Pool and sort the $n + m$ observations.
2. Replace each observation by its *rank* (or position) in the sorted list.
3. Let R_1 equal the sum of the ranks for observations in the X sample, and let R_2 equal the sum of the ranks for observations in the Y sample.

For example, let $n = 9$ and $m = 5$. If the observations in the first sample are 8.3, 8.8, 10.8, 12.3, 13.5, 14.4, 27.6, 31.4, 35.0, and the observations in the second sample are 17.2, 18.1, 21.6, 35.5, 39.9, then the sorted combined list of 14 observations is

8.3, 8.8, 10.8, 12.3, 13.5, 14.4, 17.2, 18.1, 21.6, 27.6, 31.4, 35.0, 35.5, 39.9,

the observed value of R_1 is 54, and the observed value of R_2 is 51.

Note that the sum of the statistics is $R_1 + R_2 = (n + m)(n + m + 1)/2$ and that tests based on R_1 and R_2 are equivalent. We will use R_1.

The following theorem gives information about the sampling distribution of R_1 when the distributions of X and Y are equal.

Theorem 10.3 (Rank Sum Distribution). *Assume that the X and Y distributions are equal, and let R_1 be the Wilcoxon rank sum statistic for the first sample. Then the following hold:*

1. *The range of R_1 is $n(n+1)/2, 1 + n(n+1)/2, \ldots, nm + n(n+1)/2$.*
2. *$E(R_1) = n(n + m + 1)/2$ and $Var(R_1) = nm(n + m + 1)/12$.*
3. *The distribution of R_1 is symmetric around its mean. In particular,*

$$P(R_1 = x) = P(R_1 = n(n + m + 1) - x) \text{ for each } x.$$

4. *If n and m are large, then the distribution of R_1 is approximately normal.*

Theorem 10.3 can be proven, in part, using counting methods. The ideas are as follows. If the X and Y distributions are equal, then each ordering of the $n + m$ random variables is equally likely. This fact implies that each choice of the n ranks used to compute a value of R_1 is equally likely. There are a total of $\binom{n+m}{n}$ such choices to consider when tabulating the distribution of rank sums.

Rank sum test: Observed significance level

Let r_{obs} be the observed value of R_1. Large values of R_1 support the alternative hypothesis that X is stochastically larger than Y. Small values of R_1 support the alternative hypothesis that X is stochastically smaller than Y. Thus, the following hold:

(i) The observed significance level (p value) for a test of the null hypothesis that the X and Y distributions are equal versus the alternative hypothesis that X is stochastically larger than Y is $P(R_1 \geq r_{obs})$.

(ii) The observed significance level (p value) for a test of the null hypothesis that the X and Y distributions are equal versus the alternative hypothesis that X is stochastically smaller than Y is $P(R_1 \leq r_{obs})$.

The p value for a two tailed test is twice the p value for a one tailed test.

If $n > 20$ and $m > 20$, then the normal approximation to the R_1 distribution can be used to compute p values. Otherwise, the exact sampling distribution should be used. It is best to let the computer do the work.

Example: $n = 9, m = 5$

If $n = 9$ and $m = 5$, then R_1 takes integer values between 45 and 90. The R_1 distribution has mean 67.5 and variance 56.25.

Consider the test of the null hypothesis that the X and Y distributions are equal versus the alternative hypothesis that one of the distributions (either the X or the Y distribution) is stochastically larger than the other using the 5% significance level. If the observed value of R_1 is 54, then the observed significance level, obtained using the computer, is $2P(R_1 \leq 54) = 0.083$. Since the p value is greater than 0.05, the null hypothesis that the distributions are equal is accepted.

Example: $n = 45, m = 27$

If $n = 45$ and $m = 27$, then R_1 takes integer values between 1035 and 2250. The R_1 distribution has mean 1642.5 and variance 7391.25.

Consider the test of the null hypothesis that the X and Y distributions are equal versus the alternative hypothesis that one of the distributions (either the X or the Y distribution) is stochastically larger than the other using the 5% significance level. If the observed value of R_1 is 1859, then the observed significance level, based on the normal approximation to the R_1 distribution, is $2P(R_1 \geq 1859) = 0.012$. Since the p value is less than 0.05, the null hypothesis that the distributions are equal is rejected. There is evidence that the X distribution is stochastically larger than the Y distribution.

Table 10.1. *Midranks for samples of sizes* 12 *and* 8.

	Observation	Midrank		Observation	Midrank
1	10.3	1.0	11	22.9	10.5
2	11.4	2.0	12	22.9	10.5
3	17.5	4.5	13	24.4	14.5
4	17.5	4.5	14	24.4	14.5
5	17.5	4.5	15	24.4	14.5
6	17.5	4.5	16	24.4	14.5
7	20.8	7.5	17	27.5	18.0
8	20.8	7.5	18	27.5	18.0
9	22.9	10.5	19	27.5	18.0
10	22.9	10.5	20	29.9	20.0

10.4.2 Tied observations; midranks

Continuous data are often rounded to a fixed number of decimal places, causing two or more observations to be equal. Equal observations are said to be *tied* at a given value. If two or more observations are tied at a given value, then their average rank (or *midrank*) is used in computing the rank sum statistic.

For example, let $n = 12$ and $m = 8$. Suppose that the observations in the first sample are 17.5, 20.8, 22.9, 22.9, 22.9, 24.4, 24.4, 24.4, 27.5, 27.5, 27.5, 29.9 and that the observations in the second sample are 10.3, 11.4, 17.5, 17.5, 17.5, 20.8, 22.9, 24.4. Table 10.1 shows the combined sorted list of 20 observations and their midranks. (The midrank for 17.5, for example, is the average of the 4 positions with values of 17.5: $(3 + 4 + 5 + 6)/4 = 4.5$.) The observed value of R_1 is 161.

Rank sum distribution and test

If the X and Y distributions are equal, then counting methods can be used to compute the sampling distribution of R_1 for a given list of midranks. Observed significance levels can be computed as described earlier.

For example, for the midranks in Table 10.1 and $n = 12$, R_1 takes integer and half-integer values between 78.0 and 174.0. The R_1 distribution has mean 143.0 and variance 186.333. For the example above, the p value for a two tailed test of the null hypothesis of equality distributions is $2P(R_1 \geq 161) = 0.005$.

To compute the exact distribution, imagine writing the $n + m$ midranks on separate slips of paper and placing the slips in an urn. A subset of n slips is chosen, and the sum of midranks is recorded. If each choice of subset is equally likely, then the R_1 distribution is the resulting distribution of midrank sums. It is best to let the computer do the work.

10.4.3 Mann–Whitney U statistic

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be independent random samples, of sizes n and m, from continuous distributions. The *Mann–Whitney U statistic* for the first sample, U_1, is the number of times an X observation is greater than a Y observation. Similarly, the Mann–Whitney U statistic for the second sample, U_2, is the number of times a Y observation is greater than an X observation. Note that the sum of the statistics is $U_1 + U_2 = nm$.

For example, let $n = 4$ and $m = 6$. Suppose that the observations in the first sample are 1.1, 2.5, 3.2, 4.1 and the observations in the second sample are 2.8, 3.6, 4.0, 5.2, 5.8, 7.2. Then the sorted combined list of observations (with the x-values underlined) is

$$\underline{1.1},\ \underline{2.5},\ 2.8,\ \underline{3.2},\ 3.6,\ 4.0,\ \underline{4.1},\ 5.2,\ 5.8,\ 7.2.$$

Since $x_{(1)}$ and $x_{(2)}$ are each greater than 0 y-values, $x_{(3)}$ is greater than 1 y-value, and $x_{(4)}$ is greater than 3 y-values, the observed value of U_1 is $0 + 0 + 1 + 3 = 4$. Similarly, since $y_{(1)}$ is greater than 2 x-values, $y_{(2)}$ and $y_{(3)}$ are each greater than 3 x-values, and $y_{(4)}$, $y_{(5)}$, and $y_{(6)}$ are each greater than 4 x-values, the observed value of U_2 is $2+3+3+4+4+4 = 20$. The sum of the two statistics is the total number of comparisons, $24 = 4 \times 6$.

Sum of Bernoulli random variables

The Mann–Whitney U statistic for the first sample can be written as the sum of nm dependent Bernoulli random variables:

$$U_1 = \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}, \text{ where } U_{ij} \text{ equals 1 if } X_i > Y_j \text{ and 0 if } X_i < Y_j.$$

Since $E(U_{ij}) = P(X > Y)$ for each i and j, $E(U_1) = nmP(X > Y)$. Thus, the ratio $U_1/(nm)$ is an unbiased estimator of $P(X > Y)$.

Similarly, U_2 can be written as the sum of nm dependent Bernoulli random variables, and the ratio $U_2/(nm)$ is an unbiased estimator of $P(Y > X)$.

For the example above, the estimates of $P(X > Y)$ and $P(Y > X)$ are 1/6 and 5/6, respectively.

The collection $\{U_{i,j}\}$ are dependent random variables because each X_i is used in m comparisons (correspondingly, each Y_j is used in n comparisons).

Relationship to Wilcoxon rank sum statistic

The U_1 and R_1 statistics are related. Specifically, $U_1 = R_1 - n(n+1)/2$. The following distribution theorem can be proven using Theorem 10.3.

Theorem 10.4 (U Statistic Distribution). *Assume that the X and Y distributions are equal, and let U_1 be the Mann–Whitney U statistic for the first sample. Then the*

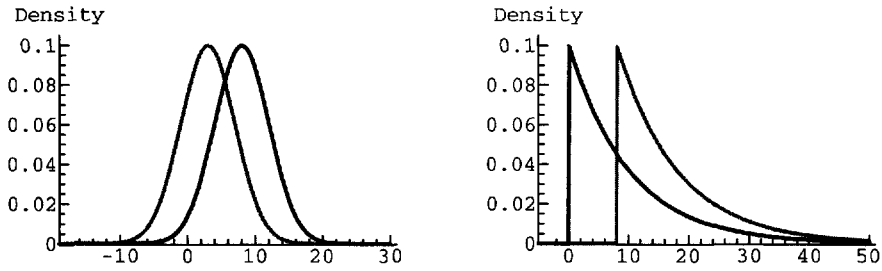


Figure 10.3. *Normal distributions satisfying a shift model (left plot) and shifted exponential distributions (right plot).*

following hold:

1. *The range of U_1 is $0, 1, 2, \ldots, nm$.*
2. *$E(U_1) = nm/2$ and $Var(U_1) = nm(n+m+1)/12$.*
3. *The distribution of U_1 is symmetric around its mean. In particular,*
$$P(U_1 = x) = P(U_1 = nm - x) \text{ for each } x.$$
4. *If n and m are large, then the distribution of U_1 is approximately normal.*

Since $U_1 = R_1 - n(n+1)/2$, tests based on U_1 are equivalent to tests based on R_1.

10.4.4 Shift models

The random variables X and Y are said to satisfy a *shift model* if

$$X - \Delta \text{ and } Y \text{ have the same distribution,}$$

where Δ is the difference in medians, $\Delta = \text{Median}(X) - \text{Median}(Y)$. The parameter Δ is called the *shift parameter*.

Assume that X and Y satisfy a shift model and $\Delta \neq 0$. If $\Delta > 0$, then X is stochastically larger than Y; otherwise, X is stochastically smaller than Y.

For example, if X is a normal random variable with mean 3 and standard deviation 4, and Y is a normal random variable with mean 8 and standard deviation 4, then X and Y satisfy a shift model with shift parameter $\Delta = -5$. The left part of Figure 10.3 shows the distribution of X in gray and the distribution of Y in black.

If X has the *shifted exponential distribution* with PDF

$$f(x) = \frac{1}{10} e^{-(x-8)/10} \text{ when } x > 8 \text{ and 0 otherwise,}$$

and Y is an exponential random variable with parameter 1/10, then X and Y satisfy a shift model with $\Delta = 8$. The right part of Figure 10.3 shows the distribution of X in gray and the distribution of Y in black.

If X and Y satisfy a shift model, then their distributions differ in location only. In studies comparing a treatment group to a no treatment group, where the effect of the treatment is additive, the shift parameter is referred to as the *treatment effect*. Note that if X and Y have finite means, then the shift parameter Δ is also the difference in means, $\Delta = E(X) - E(Y)$.

Hodges–Lehmann estimator

If X and Y satisfy a shift model with shift parameter Δ, then the *Hodges–Lehmann* (HL) estimator of Δ is the median of the list of nm differences

$$X_i - Y_j \text{ for } i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, m.$$

The differences are often referred to as the *Walsh differences*.

For example, let $n = 5$ and $m = 7$. Suppose that the observations in the first sample are 4.9, 7.3, 9.2, 11.0, 17.3 and that the observations in the second sample are 0.5, 0.7, 1.5, 2.7, 5.6, 8.7, 13.4. The following 5-by-7 table gives the 35 Walsh differences (row value minus column value):

	0.5	0.7	1.5	2.7	5.6	8.7	13.4
4.9	4.4	4.2	3.4	2.2	−0.7	−3.8	−8.5
7.3	6.8	6.6	5.8	4.6	1.7	−1.4	−6.1
9.2	8.7	8.5	7.7	6.5	3.6	0.5	−4.2
11.0	10.5	10.3	9.5	8.3	5.4	2.3	−2.4
17.3	16.8	16.6	15.8	14.6	11.7	8.6	3.9

For these data, the HL estimate is 5.4.

Confidence interval procedure for shift parameter

The ordered Walsh differences

$$D_{(k)} \text{ for } k = 1, 2, \ldots, nm$$

divide the real line into $nm + 1$ intervals

$$(-\infty, D_{(1)}), \; (D_{(1)}, D_{(2)}), \ldots, \; (D_{(nm-1)}, D_{(nm)}), \; (D_{(nm)}, \infty)$$

(ignoring the endpoints). The following theorem relates the probability that Δ is in one of these intervals (or in a union of these intervals) to the null distribution of the Mann–Whitney U statistic for the first sample, U_1.

Theorem 10.5 (Shift Confidence Intervals). *Under the assumptions of this section, if k is chosen so that the null probability $P(U_1 < k) = \frac{\alpha}{2}$, then the interval*

$$[\, D_{(k)}, \; D_{(nm-k+1)} \,]$$

is a $100(1 - \alpha)\%$ confidence interval for the shift parameter, Δ.

An outline of the proof is as follows:

(i) Since X and Y satisfy a shift model, the samples

$$\begin{aligned} &\text{Sample 1:} \quad X_1 - \Delta, X_2 - \Delta, \ldots, X_n - \Delta, \\ &\text{Sample 2:} \quad Y_1, Y_2, \ldots, Y_m \end{aligned}$$

are independent random samples from the same distribution. Thus, the distribution of

$$U_1 = \#(X_i - \Delta > Y_j) = \#(X_i - Y_j > \Delta)$$

can be tabulated, assuming that each assignment of n values to the first sample is equally likely, where U_1 is the number of times a shifted X observation is greater than a Y observation; equivalently, U_1 is the number of times a Walsh difference is greater than Δ.

(ii) The following statements are equivalent:

- $D_{(k)} < \Delta < D_{(k+1)}$.
- Exactly k differences of the form $X_i - Y_j$ are less than Δ, and exactly $nm - k$ differences are greater than Δ.
- $U_1 = nm - k$.

And, by symmetry of the U_1 distribution,

$$P(D_{(k)} < \Delta < D_{(k+1)}) = P(U_1 = nm - k) = P(U_1 = k).$$

(iii) The statement in the theorem corresponds to choosing k so that

- $P(\Delta < D_{(k)}) = P(U_1 < k) = \alpha/2$,
- $P(D_{(k)} < \Delta < D_{(nm-k+1)}) = P(k \leq U_1 \leq nm - k) = 1 - \alpha$, and
- $P(\Delta > D_{(nm-k+1)}) = P(U_1 > nm - k) = \alpha/2$.

The procedure given in Theorem 10.5 is an example of *inverting* a hypothesis test: A value δ_o is in a $100(1-\alpha)\%$ confidence interval if the two sided rank sum test of

$$H_o: \text{ The distributions of } X - \delta_o \text{ and } Y \text{ are equal}$$

is accepted at the α significance level.

For example, assume the data in the previous example are the values of independent random samples from distributions satisfying a shift model. Since $P(U_1 < 6) = 0.024$, a 95.2% confidence interval for the shift parameter is

$$[d_{(6)}, d_{(30)}] = [-1.4, 10.5].$$

Since this interval contains 0, the possibility that the X and Y distributions are equal cannot be ruled out.

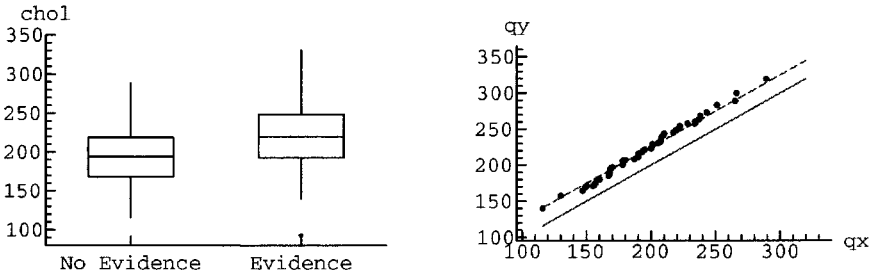


Figure 10.4. *Side-by-side box plots of cholesterol levels (in mg/dl) for men with no evidence of heart disease and for men with evidence of disease (left plot) and a QQ plot of the data (right plot). The dashed line in the right plot is* $y = x + 25$.

Example: Comparison of cholesterol distributions

As part of a study to identify risk factors for coronary artery disease, cholesterol levels in milligrams per deciliter (mg/dl) were measured in more than 300 male patients complaining of chest pain [95], [50, p. 221]. This example uses a subset of their data.

The left part of Figure 10.4 shows side-by-side box plots for 51 men with no evidence of heart disease and 150 men with evidence of disease. The right plot is a *quantile-quantile* (QQ) plot of the data. A QQ plot is constructed by pairing the $k/(N+1)$ sample quantile of the first sample with the $k/(N+1)$ sample quantile of the second sample, for $k = 1, 2, \ldots, N$, where $N = \min(n, m)$. The solid line in the plot is $y = x$, and the dashed line is $y = x + 25$.

Assume these data are the values of independent random samples. Since the points are close to a line parallel to $y = x$, the QQ plot suggests that the cholesterol distributions satisfy a shift model.

Let X be the serum cholesterol level (in mg/dl) for male patients experiencing chest pain but with no evidence of heart disease and Y be the corresponding level for men with evidence of disease. Assume that the X and Y distributions satisfy a shift model and that the data shown in Figure 10.4 are the values of independent random samples from these distributions. The HL estimate of $\Delta = \text{Median}(X) - \text{Median}(Y)$ is -25.0, and an approximate 95% confidence interval is $[-38.0, -13.0]$. The median serum cholesterol level for men with no evidence of disease is estimated to be between 13 mg/dl and 38 mg/dl lower than for men with evidence of disease.

10.5 Sampling models

The methods of this chapter assume that the measurements under study are the values of independent random samples from continuous distributions.

In most applications, simple random samples of individuals are drawn from finite populations, and measurements are made on these individuals. If population sizes are large enough, then the resulting measurements can be treated as if they were the values of independent random samples.

10.5.1 Population model

If simple random samples are drawn from sufficiently large populations of individuals, then sampling is said to be done under a *population model*. Under a population model, measurements can be treated as if they were the values of independent random samples.

When comparing two distributions, sampling can be done separately from two subpopulations or from a total population. For example, a researcher interested in comparing achievement test scores of girls and boys in the fifth grade might sample separately from the subpopulations of fifth-grade girls and fifth-grade boys or might sample from the population of all fifth-graders and then split the sample into subsamples of girls and boys.

A third possibility in the two sample setting is sampling from a total population followed by *randomization* to one of two treatments under study. For example, a medical researcher interested in determining if a new treatment to reduce serum cholesterol levels is more effective than the standard treatment in a population of women with very high levels of cholesterol might do the following:

1. Choose a simple random sample of $n + m$ subjects from the population of women with very high levels of serum cholesterol.
2. Partition the $n + m$ subjects into distinguishable subsets (or groups) of sizes n and m.
3. Administer the standard treatment to each subject in the first group for a fixed period of time and the new treatment to each subject in the second group for the same fixed period of time.

By randomly assigning subjects to treatment groups, the effect is as if sampling was done from two subpopulations: the subpopulation of women with high cholesterol who have been treated with the standard treatment for a fixed period of time and the subpopulation of women with high cholesterol who have been treated with the new treatment for a fixed period of time. Note that, by design, the subpopulations differ in treatment only.

10.5.2 Randomization model

The following is a common research scenario:

> A researcher is interested in comparing two treatments and has $n + m$ subjects willing to participate in a study. The researcher randomly assigns n subjects to receive the first treatment; the remaining m subjects will receive the second treatment.

Treatments could be competing drugs for reducing cholesterol (as above) or competing methods for teaching multivariable calculus.

If the $n+m$ subjects are *not* a simple random sample from the study population, but the assignment of subjects to treatments is one of $\binom{n+m}{n}$ equally likely assignments, then sampling is said to be done under a *randomization model*.

Under a randomization model for the comparison of treatments, chance enters into the experiment only through the assignment of subjects to treatments. The results

of experiments conducted under a randomization model cannot be generalized to a larger population of interest but may still be of interest to researchers.

The Wilcoxon rank sum test is an example of a method that can be used to analyze data sampled under either the population model or the randomization model. Additional methods will be discussed in Chapter 11 and in later chapters.

10.6 Laboratory problems

Laboratory problems for this chapter introduce *Mathematica* commands for working with the f ratio distribution, for analyzing samples from normal distributions, for analyzing samples using rank sum methods, and for constructing QQ plots. The problems are designed to reinforce ideas related to the analysis of two samples.

10.6.1 Laboratory: Two sample analysis

In the main laboratory notebook (Problems 1 to 5), you will use simulation to study methods for comparing samples from normal distributions and apply graphical and formal inference methods to data from four studies: (1) stamina in lizards with and without disease [94], [44]; (2) calories and sodium levels in beef and poultry franks [56], [78]; (3) diets of two types of lizards [86], [73]; and (4) Olympic marathon finishing times for men and women [81].

10.6.2 Additional problem notebooks

Problems 6 and 7 are applications of methods for samples from normal distributions. Problem 6 uses data on body temperatures of healthy men and women [97]. Problem 7 uses data from a physical anthropology study [11].

Problems 8 and 9 focus on shift models. Problem 8 uses the Olympic marathon finishing times from the main laboratory notebook. Problem 9 uses data from a cloud-seeding experiment [99], [25].

In Problem 10, a variety of graphical and computational methods are applied to several data sets from an ecology study [87]. In Problem 11, a variety of graphical and computational methods are applied to several data sets from a study of factors related to mercury contamination in Maine lakes [54].

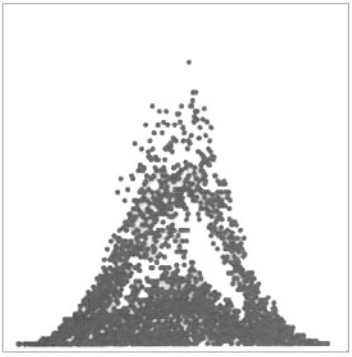
Problem 12 considers data generated under the randomization model. Wilcoxon rank sum methods are applied to study military drafting procedures during the Vietnam War [38].

In Problems 13 and 14, simulation is used to determine if there is an advantage to using a balanced study design ($n = m$), assuming that the total number of observations ($n+m$) is fixed. Problem 13 considers pooled t tests. Problem 14 considers f ratio tests.

In Problem 15, simulation is used to compare the power of two tailed pooled t tests and rank sum tests when sample sizes are equal and when sample sizes are not equal.

In Problem 16, simulation is used to determine if f ratio methods remain valid when sampling is done from distributions other than the normal. Exponential, gamma, and uniform models are considered.

This page intentionally left blank



Chapter 11
Permutation Analysis

In many statistical applications, the null and alternative hypotheses of interest can be paraphrased in the following simple terms:

H_o: Any patterns appearing in the data are due to chance alone.

H_a: There is a tendency for a certain type of pattern to appear.

Permutation methods allow researchers to determine whether to accept or reject a null hypothesis of randomness and, in some cases, to construct confidence intervals for unknown parameters. The methods are applicable in many settings since they require few mathematical assumptions.

This chapter introduces permutation analysis. The first two sections give the important definitions and applications in the two sample and paired sample settings. The third section is on correlation analysis. Additional applications are given in Section 4. Section 5 outlines the laboratory problems. General references for this chapter are [68], [74].

11.1 Introduction

As an introduction to permutation methods, consider the analysis of two samples, introduced in Chapter 10. Let $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$ be the observed samples, where each is a list of numbers with repetitions.

The data could have been generated under one of two sampling models.

Population model

Data generated under a population model can be treated as the values of independent random samples from continuous distributions. For example, consider testing the null hypothesis that the X and Y distributions are equal versus the alternative hypothesis that X is stochastically smaller than Y. If $n = 3$, $m = 5$, and the observed lists are

$$\{x_1, x_2, x_3\} = \{1.4, 1.2, 2.8\} \text{ and } \{y_1, y_2, y_3, y_4, y_5\} = \{1.7, 2.9, 3.7, 2.3, 1.3\},$$

then the event

$$X_2 < Y_5 < X_1 < Y_1 < Y_4 < X_3 < Y_2 < Y_3$$

has occurred. Under the null hypothesis, each permutation of the $n+m=8$ random variables is equally likely; thus, the observed event is one of $(n+m)! = 40,320$ equally likely choices. Patterns of interest under the alternative hypothesis are events where the X_i's tend to be smaller than the Y_j's.

Randomization model

Data generated under a *randomization model* cannot be treated as the values of independent random samples from continuous distributions but can be thought of as one of N equally likely choices under the null hypothesis.

In the example above, the null hypothesis of randomness is that the observed data is one of $N = 40,320$ equally likely choices, where a choice corresponds to a matching of the 8 numbers to the labels $x_1, x_2, x_3, y_1, y_2, y_3, y_4, y_5$.

11.1.1 Permutation tests

Conduct a permutation test using a test statistic T as follows:

1. The sampling distribution of T is obtained by computing the value of the statistic for each reordering of the data.
2. The observed significance level (or p value) is computed by comparing the observed value of T to the sampling distribution from step 1.

The sampling distribution from the first step is called the *permutation distribution* of the statistic T, and the p value is called a *permutation p value*.

> In some books, the term *permutation test* is used when sampling is done under a population model, and the term *randomization test* is used when sampling is done under a randomization model.

Continuing with the example above, let S be the sum of numbers in the x sample. Since the value of S depends only on which observations are labeled x's, and not on the relative ordering of all 8 observations, the sampling distribution of S under the null hypothesis is obtained by computing its value for each partition of the 8 numbers into subsets of sizes 3 and 5, respectively.

Table 11.1 shows the values of S for each of the $\binom{8}{3} = 56$ choices of observations for the first sample. The observed value of S if 5.4. Since small values of S support the alternative hypothesis that x values tend to be smaller than y values, the observed significance level is $P(S \leq 5.4) = 13/56$.

Conditional test, nonparametric test

A permutation test is an example of a *conditional test*, since the sampling distribution of T is computed *conditional* on the observations.

For example, the Wilcoxon rank sum test is a permutation test where the observations have been replaced by ranks. The test is conditional on the pattern of ties in

Table 11.1. *Sampling distribution of the sum of observations in the first sample.*

Sum	x sample	Sum	x sample	Sum	x sample
3.9	{1.2, 1.3, 1.4}	5.9	{1.4, 1.7, 2.8}	7.1	{1.4, 2.8, 2.9}
4.2	{1.2, 1.3, 1.7}	5.9	{1.3, 1.7, 2.9}	7.2	{1.2, 2.3, 3.7}
4.3	{1.2, 1.4, 1.7}	6.0	{1.4, 1.7, 2.9}	7.3	{1.3, 2.3, 3.7}
4.4	{1.3, 1.4, 1.7}	6.2	{1.2, 1.3, 3.7}	7.4	{1.4, 2.3, 3.7}
4.8	{1.2, 1.3, 2.3}	6.3	{1.2, 1.4, 3.7}	7.4	{1.7, 2.8, 2.9}
4.9	{1.2, 1.4, 2.3}	6.3	{1.2, 2.3, 2.8}	7.7	{1.2, 2.8, 3.7}
5.0	{1.3, 1.4, 2.3}	6.4	{1.3, 2.3, 2.8}	7.7	{1.7, 2.3, 3.7}
5.2	{1.2, 1.7, 2.3}	6.4	{1.2, 2.3, 2.9}	7.8	{1.2, 2.9, 3.7}
5.3	{1.2, 1.3, 2.8}	6.4	{1.3, 1.4, 3.7}	7.8	{1.3, 2.8, 3.7}
5.3	{1.3, 1.7, 2.3}	6.5	{1.3, 2.3, 2.9}	7.9	{1.4, 2.8, 3.7}
5.4	{1.2, 1.4, 2.8}	6.5	{1.4, 2.3, 2.8}	7.9	{1.3, 2.9, 3.7}
5.4	{1.4, 1.7, 2.3}	6.6	{1.2, 1.7, 3.7}	8.0	{1.4, 2.9, 3.7}
5.4	{1.2, 1.3, 2.9}	6.6	{1.4, 2.3, 2.9}	8.0	{2.3, 2.8, 2.9}
5.5	{1.2, 1.4, 2.9}	6.7	{1.3, 1.7, 3.7}	8.2	{1.7, 2.8, 3.7}
5.5	{1.3, 1.4, 2.8}	6.8	{1.4, 1.7, 3.7}	8.3	{1.7, 2.9, 3.7}
5.6	{1.3, 1.4, 2.9}	6.8	{1.7, 2.3, 2.8}	8.8	{2.3, 2.8, 3.7}
5.7	{1.2, 1.7, 2.8}	6.9	{1.2, 2.8, 2.9}	8.9	{2.3, 2.9, 3.7}
5.8	{1.2, 1.7, 2.9}	6.9	{1.7, 2.3, 2.9}	9.4	{2.8, 2.9, 3.7}
5.8	{1.3, 1.7, 2.8}	7.0	{1.3, 2.8, 2.9}		

the observations. If there are no ties in the observations, then integers between 1 and $n + m$ are used to construct the permutation distribution of R_1; if some values are tied (see, for example, Table 10.1), then the observed midranks are used to construct the permutation distribution of R_1.

Note that the term *nonparametric test* is often used to describe permutation tests where observations have been replaced by ranks.

Monte Carlo test

If the number of reorderings of the data is very large, then the computer can be used to approximate the sampling distribution of T by computing its value for a fixed number of random reorderings of the data. The approximate sampling distribution can then be used to estimate the p value.

A test conducted in this way is an example of a Monte Carlo test. (In a *Monte Carlo* analysis, simulation is used to estimate a quantity of interest. Here the quantity of interest is a p value.)

11.1.2 Example: Difference in means test

This section considers permutation tests based on the difference in means statistic

$$D = \text{Mean of } x \text{ sample} - \text{Mean of } y \text{ sample.}$$

Tests based on D are appropriate in the following situations:

1. *Population model.* The observed data are the values of independent random samples from distributions differing in mean only. The null hypothesis is that the distributions are equal, equivalently that $\mu_x = \mu_y$. Alternatives of interest are that the mean of one distribution is larger than the mean of the other.
2. *Randomization model.* The data are measurements taken on $n + m$ individuals in distinguishable groups of sizes n and m. The null hypothesis is that the observed difference in means is due to chance alone. Alternatives of interest are that values in one group tend to be larger (but not more variable) than values in the other group.

The sampling distribution of D under the null hypothesis of randomness is obtained by computing the difference in means for each partition of the $n + m$ observations into subsets of sizes n and m, respectively. The following theorem gives summary measures of the resulting distribution.

Theorem 11.1 (Difference in Means). *Conditional on the observed values in the two samples, the permutation distribution of D has the summary measures*

$$E(D) = 0 \text{ and } Var(D) = \frac{n+m}{n\,m\,(n+m-1)} \sum_{i=1}^{n+m} (z_i - \bar{z})^2,$$

where $z_1, z_2, \ldots, z_{n+m}$ is the combined list (with repetitions) of the $n + m$ observations, and $\bar{z}$ is the mean of the $n + m$ observations.

For example, let $n = 8$ and $m = 10$. If the observations in the first sample are 9.4, 22.9, 14.6, 7.9, 0.7, 19.2, 16.9, 5.6 and the observations in the second sample are 18.7, 19.5, 15.0, 17.4, 22.6, 26.0, 31.5, 8.8, 8.5, 10.6, then the permutation distribution of D has mean 0 and variance 13.8617. Consider testing the null hypothesis of randomness using a two sided alternative and the 5% significance level. The observed difference in means is -5.71, and the observed significance level is

$$P(|D| \geq |-5.71|) = \frac{5632}{43758} \approx 0.1287.$$

(There are $\binom{18}{8} = 43,758$ partitions to consider.) Since the p value is greater than 0.05, the null hypothesis that the observed difference in means is due to chance alone is accepted.

Comparison of tests

Permutation tests based on the difference in means (D) and on the sum of observations in the first sample (S) are equivalent.

In situations where both the pooled t test of the null hypothesis of equality of means and the difference in means tests are appropriate, the pooled t test is preferred. However, it is interesting to note that the tests give similar results.

In situations where both the difference in means test and the rank sum test are appropriate, if the samples are highly skewed or have extreme outliers, then the rank

sum test is preferred; otherwise, the difference in means test is preferred. In practice, the rank sum test is used almost exclusively since it is easy to implement.

11.1.3 Example: Smirnov two sample test

The *empirical cumulative distribution* function (or empirical CDF) of a sample of n numbers, $\{x_1, x_2, \ldots, x_n\}$, is defined as follows:

$$\text{ECDF}(x) = \text{Proportion of } x_i\text{'s} \leq x \qquad \text{for all real numbers } x.$$

For example, if $n = 10$ and the ordered observations are

$$2.3,\ 3.6,\ 4.4,\ 4.7,\ 6.3,\ 6.7,\ 8.1,\ 8.7,\ 9.4,\ 10.7,$$

then ECDF(x) is a step function with values

$$0 \text{ when } x < 2.3, \quad \frac{1}{10} \text{ when } 2.3 \leq x < 3.6, \quad \frac{2}{10} \text{ when } 3.6 \leq x < 4.4, \quad \ldots.$$

In the 1930's, Smirnov proposed a two sample test based on a comparison of empirical CDFs. Let ECDF_1 and ECDF_2 be the empirical CDFs of the x and y samples, respectively. The Smirnov statistic, S, is the maximum absolute difference in the empirical CDFs:

$$S = \max_x \left| \text{ECDF}_1(x) - \text{ECDF}_2(x) \right|.$$

Tests based on S are appropriate in the following situations:

1. *Population model.* The observed data are the values of independent random samples. The null hypothesis is that the distributions from which the data were drawn are equal versus the general alternative that the distributions are not equal.

2. *Randomization model.* The data are measurements taken on $n + m$ individuals in distinguishable groups of sizes n and m. The null hypothesis is that observed differences in the empirical CDFs are due to chance alone. Alternatives of interest are that the samples differ in some way.

Values of S lie in the interval $[0, 1]$. Observed values near 0 support the null hypothesis; large observed values support the alternative.

For example, suppose that 45 Australian consumers and 48 Japanese consumers were asked to rate a particular brand of cholocate on a 10-point scale (where a score of 10 indicates the consumer liked the sweetness, while a score of 1 indicates a consumer did not like the sweetness at all), with results summarized in the following table:

Score	1	2	3	4	5	6	7	8	9	10	Sum
Australian	0	0	11	7	15	4	2	2	2	2	45
Japanese	3	3	12	0	3	3	7	13	1	3	48

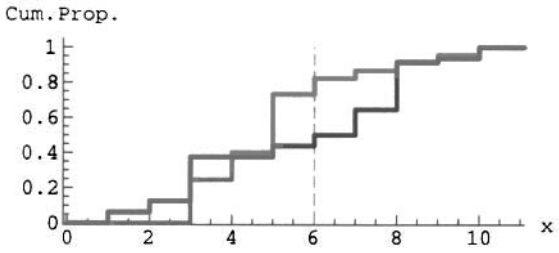


Figure 11.1. *Empirical CDFs for the taste test example.*

Empirical CDFs for the Australian (in gray) and Japanese (in black) samples are shown in Figure 11.1. The observed value of the Smirnov statistic is 0.32, occurring when the score is 6.

Consider testing the null hypothesis of randomness using the 5% significance level. In a Monte Carlo analysis using 2000 random partitions (including the observed partition of the 93 scores), 0.55% (11/2000) of S values were greater than or equal to the observed value. Thus, there is evidence that observed differences in responses by Australian and Japanese consumers were not due to chance alone. Although the mean scores in both groups were close (5.07 for the Australians versus 5.63 for the Japanese), most Australians gave scores in the 3–4–5 range, while most Japanese gave scores of 3 and 8.

The sampling distribution of S depends on the ranks (or midranks, in case of ties) of the observations and not on the observations themselves. The relationship is quite complicated. Using simulation to estimate p values is a good approach.

11.2 Paired sample analysis

This section considers several approaches to analyzing lists of pairs of numbers,

$$\{\{x_1, y_1\}, \{x_2, y_2\}, \ldots, \{x_n, y_n\}\},$$

or corresponding lists of differences,

$$\{d_1, d_2, \ldots, d_n\} = \{x_1 - y_1, x_2 - y_2, \ldots, x_n - y_n\}.$$

Paired samples arise in many experimental settings. Examples include the following:

(i) *Before-and-after experiments.* For each of n individuals, the x value is a measurement made before a treatment begins, and the y value is the corresponding measurement after a fixed treatment period.

(ii) *Randomized pairs experiments.* For each of n pairs of individuals, where each pair is matched on important factors (for example, age, sex, severity of disease), one member of the pair is randomly assigned to treatment 1 and the other to treatment 2. After a fixed period of time, measurements are taken on each individual.

Table 11.2. *Proportions of women in the labor force in* 1972 *(x) and* 1968 *(y), and the difference in proportions (x − y), for women living in* 19 *U.S. cities.*

City	x	y	$x-y$	City	x	y	$x-y$
Baltimore	0.57	0.49	0.08	Minneapolis/St. Paul	0.59	0.50	0.09
Boston	0.60	0.45	0.15	Newark	0.53	0.54	−0.01
Buffalo	0.64	0.58	0.06	New York	0.45	0.42	0.03
Chicago	0.52	0.52	0.00	Patterson	0.57	0.56	0.01
Cincinnati	0.53	0.51	0.02	Philadelphia	0.45	0.45	0.00
Hartford	0.55	0.54	0.01	Pittsburgh	0.49	0.34	0.15
Dallas	0.64	0.63	0.01	San Francisco	0.55	0.55	0.00
Detroit	0.46	0.43	0.03	St. Louis	0.35	0.45	−0.10
Houston	0.50	0.49	0.01	Washington, D.C.	0.52	0.42	0.10
Los Angeles	0.50	0.50	0.00				

Researchers use paired designs to reduce the variability of the results. In paired designs, individuals within each pair are expected to respond similarly to treatment, while individuals in different pairs are expected to respond differently to treatment.

Paired t methods

If $D = X - Y$ is a normal random variable, the differences data are the values of a random sample from the D distribution, and questions about $E(D) = E(X) - E(Y)$ are of interest, then methods discussed in Chapters 7 and 8 can be applied.

For example, consider the data in Table 11.2 on the proportions of women in the labor force in 19 cities in the United States in 1968 and 1972 [30]. Of interest is whether the mean labor force participation rate for women in 1972 has changed from the mean rate in 1968. Assuming these data are the values of a random sample from a normal distribution, a 95% confidence interval for the mean difference is

$$\overline{d} \pm t_{18}(0.025)\sqrt{\frac{s_d^2}{19}} \implies 0.03368 \pm 2.101\sqrt{\frac{0.003569}{19}} \implies [0.0049, 0.0625].$$

Since both endpoints are positive, there is evidence that the mean rate increased over the 4-year period.

11.2.1 Example: Signed rank test

In the 1940's Wilcoxon developed a nonparametric test for the analysis of paired data. The test is appropriate in the following situations:

1. *Population model.* The paired data are the values of a random sample from a bivariate continuous distribution. The null hypothesis of interest is that the distribution of differences $D = X - Y$ is symmetric around zero versus alternatives that the values of D tend to be positive or tend to be negative.

2. *Randomization model.* The paired data are measurements taken on n individuals (or n pairs of individuals). The null hypothesis is that the signs (positive or negative) of the observed differences are due to chance alone versus alternatives that observed differences tend to be positive or tend to be negative.

Consider the population model. Under the null hypothesis,

$$P(D_i > 0) = P(D_i < 0) = 1/2 \text{ for } i = 1, 2, \ldots, n,$$

where $D_i = X_i - Y_i$. By independence, each of the 2^n events of the form

$$\begin{Bmatrix} D_1 > 0 \\ D_1 < 0 \end{Bmatrix} \text{ and } \begin{Bmatrix} D_2 > 0 \\ D_2 < 0 \end{Bmatrix} \text{ and } \cdots \text{ and } \begin{Bmatrix} D_n > 0 \\ D_n < 0 \end{Bmatrix}$$

is equally likely. (Choose either $D_i > 0$ or $D_i < 0$ in each bracket.)

Signed rank statistic

The Wilcoxon signed rank statistics for positive differences (W_+) and for negative differences (W_-) are computed as follows:

1. Sort the list of absolute differences.
2. Replace each observed difference by its rank (or position) in the sorted list. Use midranks in case of ties in the absolute differences.
3. Let W_+ equal the sum of the ranks for positive differences, and let W_- equal the sum of the ranks for negative differences.

For example, let $n = 10$, and assume that the differences are -3.54, -3.05, -0.66, 0.65, 1.66, 2.16, 2.75, 3.23, 4.24, 5.15. The ordered list of absolute differences is

$$0.65,\ 0.66,\ 1.66,\ 2.16,\ 2.75,\ 3.05,\ 3.23,\ 3.54,\ 4.24,\ 5.15,$$

the observed value of W_+ is 39, and the observed value of W_- is 16.

Tests based on W_+ and W_- are equivalent. We will use W_+.

Permutation distribution

The permutation distribution of W_+ is computed as follows. For each assignment of signs to absolute differences, the sum of the ranks for positive differences is computed. Absolute differences of zero ($|d_i| = 0$) drop out of the analysis of signed ranks. Thus, the total number of assignments of signs is 2^m, where m is the number of nonzero differences.

The following theorem gives information about the sampling distribution of W_+ when there are no ties and no zeros in the list of differences.

Theorem 11.2 (Signed Rank Distribution). *Consider the permutation distribution of W_+ under the null hypothesis. If there are no ties in the absolute differences, and all differences are nonzero, then the following hold:*

Table 11.3. *Differences in labor force participation and midranks.*

	d_i	$\|d_i\|$	Midrank		d_i	$\|d_i\|$	Midrank
1	0.00	0.00	2.5	11	0.03	0.03	11.5
2	0.00	0.00	2.5	12	0.03	0.03	11.5
3	0.00	0.00	2.5	13	0.06	0.06	13.0
4	0.00	0.00	2.5	14	0.08	0.08	14.0
5	0.01	0.01	5.0	15	0.09	0.09	15.0
6	0.01	0.01	7.5	16	−0.10	0.10	16.5
7	0.01	0.01	7.5	17	0.10	0.10	16.5
8	0.01	0.01	7.5	18	0.15	0.15	18.5
9	−0.01	0.01	7.5	19	0.15	0.15	18.5
10	0.02	0.02	10.0				

1. *The range of* W_+ *is* $0, 1, 2, \ldots, n(n+1)/2$.
2. $E(W_+) = n(n+1)/4$ *and* $Var(W_+) = n(n+1)(2n+1)/24$.
3. *The distribution of* W_+ *is symmetric around its mean. In particular,*

$$P(W_+ = x) = P\left(W_+ = \frac{n(n+1)}{2} - x\right) \text{ for all } x.$$

4. *If n is large, then the distribution of* W_+ *is approximately normal.*

For example, if $n = 10$ and there are no ties and no zero differences, then W_+ takes integer values between 0 and 55. The W_+ distribution has mean 27.5 and variance 192.5.

Signed rank test: Observed significance level

Large values of W_+ support the alternative that differences tend to be positive, and small values support the alternative that differences tend to be negative.

If $n > 20$ and there are no tied observations and no zeros, then the normal approximation to the W_+ distribution can be used to estimate p values.

Table 11.3 shows the differences in labor force participation data from Table 11.2 and corresponding midranks. To determine if the observed differences between the 1968 and 1972 proportions of women participating in the labor force are due to chance alone versus an alternative that the values in one of these years tend to be higher than the other, a two sided test will be conducted at the 5% significance level. There are 15 nonzero differences. The W_+ distribution has mean 90 and variance 608.375. The observed value of W_+ is 156, and the observed significance level is

$$2P(W_+ \geq 156) = \frac{190}{32768} \approx 0.005798$$

(there are $2^{15} = 32,768$ assignments of signs). Since the p value is less than 0.05, the null hypothesis that differences in sign are due to chance alone is rejected. In fact, there is evidence that the participation of women in the labor force has increased over time.

11.2.2 Shift models

Assume the paired data are the values of a random sample from a bivariate continuous distribution. The difference $D = X - Y$ is said to satisfy a *shift model* if the distribution of D is symmetric around Δ, where $\Delta = \text{Median}(D) = \text{Median}(X - Y)$. The parameter Δ is called the *shift parameter*.

Assume that D satisfies a shift model and $\Delta \neq 0$. If $\Delta > 0$, then the values of D tend to be positive; otherwise, the values of D tend to be negative.

Note that in experiments comparing a treatment group to a no treatment group, Δ corresponds to an *additive treatment effect*.

Hodges–Lehmann estimator

If D satisfies a shift model with shift parameter Δ, then the *Hodges–Lehmann* (HL) estimator of Δ is the median of the list of $n(n+1)/2$ averages

$$\frac{1}{2}(D_i + D_j), \quad \text{where } i, j = 1, 2, \ldots, n \text{ and } j \leq i.$$

The averages are often referred to as the *Walsh averages*.

The list of Walsh averages includes the original differences (when $j = i$) and averages of each pair of differences (when $j \neq i$).

For example, let $n = 6$. Suppose that the observed differences are -8.09, -7.7, -7.4, -5.7, 2.13, 9.3. The following 6-by-6 table gives the 21 Walsh averages:

	−8.090	−7.700	−7.400	−5.700	2.130	9.300
−8.090	−8.090					
−7.700	−7.895	−7.700				
−7.400	−7.745	−7.550	−7.400			
−5.700	−6.895	−6.700	−6.550	−5.700		
2.130	−2.980	−2.785	−2.635	−1.785	2.130	
9.300	0.605	0.800	0.950	1.800	5.715	9.300

For these data, the HL estimate is -2.980.

Confidence interval procedure for shift parameter

The ordered Walsh averages

$$A_{(k)} \text{ for } k = 1, 2, \ldots, \tfrac{n(n+1)}{2}$$

divide the real line into $\frac{n(n+1)}{2} + 1$ intervals

$$\left(-\infty, A_{(1)}\right), \ \left(A_{(1)}, A_{(2)}\right), \ \ldots, \ \left(A_{\left(\frac{n(n+1)}{2}-1\right)}, A_{\left(\frac{n(n+1)}{2}\right)}\right), \ \left(A_{\left(\frac{n(n+1)}{2}\right)}, \infty\right)$$

(ignoring endpoints). The following theorem relates the probability that Δ is in one of these intervals (or a union of these intervals) to the null distribution of W_+.

Theorem 11.3 (Shift Confidence Intervals). *Under the assumptions of this section, if k is chosen so that the null probability $P(W_+ < k) = \frac{\alpha}{2}$, then the interval*

$$\left[A_{(k)}, \ A_{\left(\frac{n(n+1)}{2}-k+1\right)}\right]$$

is a $100(1-\alpha)\%$ confidence interval for Δ.

The procedure given in Theorem 11.3 is an example of *inverting* a hypothesis test: A value δ_o is in a $100(1-\alpha)\%$ confidence interval if the two sided signed rank test of

$$H_o: \text{ The distribution of } (X-Y)-\delta_o \text{ is symmetric around } 0$$

is accepted at the α significance level. The proof is similar to the proof of Theorem 10.5.

For example, assume the differences data in the previous example are the values of a random sample from a distribution satisfying a shift model. Since $P(W_+ < 3) = 0.047$, a 90.6% confidence interval for the shift parameter is

$$[a_{(3)}, a_{(19)}] = [-7.745, 2.130].$$

Since this interval contains 0, the possibility that the median of the differences distribution is zero cannot be ruled out.

11.2.3 Example: Fisher symmetry test

R. A. Fisher, who pioneered the use of permutation methods, proposed a test for paired samples using the sum of differences statistic, $S = \sum_{i=1}^{n} d_i$. Tests based on S are appropriate in the following situations:

1. *Population model.* The paired data are the values of a random sample from a bivariate continuous distribution. $D = X - Y$ satisfies a shift model with $\Delta = E(D) = E(X-Y)$. The null hypothesis is that the mean is zero. Alternatives of interest are that the mean is positive or negative.

2. *Randomization model.* The paired data are measurements taken on n individuals (or n pairs of individuals). The null hypothesis is that the signs (positive or negative) of observed differences are due to chance alone. Alternatives of interest are that the observed differences tend to be positive or negative.

Permutation distribution

The sampling distribution of S under the null hypothesis of randomness is obtained by computing the sum for each assignment of signs to the observed differences. The following theorem gives information about the resulting distribution.

Theorem 11.4 (Sum of Differences). *Conditional on the observed differences, the permutation distribution of the sum of differences statistic, S, has the following summary measures:*

$$E(S) = 0 \text{ and } Var(S) = \sum_{i=1}^{n} d_i^2.$$

If n is large enough, then the S distribution is approximately normal.

For example, if $n = 6$ and the observed differences are -8.09, -7.7, -7.4, -5.7, 2.13, 9.3, then $2^6 = 64$ sums of the form

$$\pm 8.09 \pm 7.7 \pm 7.4 \pm 5.7 \pm 2.13 \pm 9.3$$

would be computed. (Choose either $+$ or $-$ in each summand.) The resulting distribution has mean 0 and variance 303.015.

Symmetry test: Observed significance level

Large values of S support the alternative that differences tend to be positive, and small values support the alternative that differences tend to be negative.

For example, in a classic experiment on plant growth [41], [73], Charles Darwin took 15 pairs of the plant *Zea mays*, where the two plants in each pair were

> of exactly the same age, were subjected from the first to last to the same conditions, were descended from the same parents.

One individual was cross-fertilized (CF), and the other was self-fertilized (SF). Darwin hypothesized that cross-fertilized plants produced taller offspring than self-fertilized plants. The heights of offspring of the 15 pairs were then measured to the nearest eighth of an inch; the table below gives the results in eighths of an inch over 12 inches. The first row is the value for the cross-fertilized plant, the second row gives the value for the self-fertilized plant, and the last row is the difference (CF−SF).

CF	92	0	72	80	57	76	81	67	50	77	90	72	81	88	0
SF	43	67	64	64	51	53	53	26	36	48	34	48	6	28	48
CF−SF	49	−67	8	16	6	23	28	41	14	29	56	24	75	60	−48

For these data, the observed sum is 314.

To determine if observed differences between cross-fertilized and self-fertilized plants are due to chance alone, versus the alternative that cross-fertilized plants produce taller offspring, a one sided test will be conducted at the 5% significance level. The observed significance level is

$$P(S \geq 314) = \frac{863}{32768} \approx 0.0263.$$

Since the p value is less than 0.05, there is evidence supporting Darwin's hypothesis that cross-fertilized plants produce taller offspring than self-fertilized plants.

Comparison of tests

In situations where both the paired t test of the null hypothesis that the mean is zero and the sum of differences test are appropriate, the paired t test is preferred. However, it is interesting to note that the tests give similar results.

In situations where both the sum of differences test and the signed rank test are appropriate, if the differences data are highly skewed or have extreme outliers, then the signed rank test is preferred; otherwise, the sum of differences test is preferred. In practice, the signed rank test is used almost exclusively since it is easy to implement.

11.3 Correlation analysis

This section considers permutation methods for analyzing lists of pairs of numbers

$$\{\{x_1, y_1\}, \{x_2, y_2\}, \ldots, \{x_n, y_n\}\}$$

in one of the following situations:

1. *Population model.* The paired data are the values of a random sample from a bivariate continuous distribution. The null hypothesis of interest is that X and Y are independent versus alternatives that there is a positive or negative association between the variables.

2. *Randomization model.* The paired data are measurements of two characteristics in each of n individuals. The null hypothesis of interest is that there is no relationship between the characteristics. Alternatives of interest are that the characteristics are positively or negatively associated.

Consider the population model. If the observations are indexed so that

$$X_1 < X_2 < \cdots < X_n$$

and if X and Y are independent (the null hypothesis of interest), then each of the $n!$ orderings of the Y_i's is equally likely. Thus, an observed matching of Y values to ordered X values can be thought of as one of $n!$ equally likely choices. This fact forms the basis of the permutation methods below.

11.3.1 Example: Correlation test

The sample correlation statistic

$$R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

can be used to test the null hypothesis of randomness.

Table 11.4. *Permutation distribution of sample correlations.*

r	Permutation of y's	r	Permutation of y's
−0.90	{ 0.7, −0.2, −0.8, −1.1}	0.14	{−0.2, −0.8, 0.7, −1.1}
−0.88	{ 0.7, −0.2, −1.1, −0.8}	0.16	{−0.8, −0.2, 0.7, −1.1}
−0.86	{−0.2, 0.7, −0.8, −1.1}	0.27	{−0.2, −0.8, −1.1, 0.7}
−0.84	{−0.2, 0.7, −1.1, −0.8}	0.30	{−0.8, −0.2, −1.1, 0.7}
−0.50	{ 0.7, −0.8, −0.2, −1.1}	0.36	{−0.2, −1.1, 0.7, −0.8}
−0.44	{−0.8, 0.7, −0.2, −1.1}	0.40	{−1.1, −0.2, 0.7, −0.8}
−0.43	{ 0.7, −0.8, −1.1, −0.2}	0.47	{−0.2, −1.1, −0.8, 0.7}
−0.37	{−0.8, 0.7, −1.1, −0.2}	0.51	{−1.1, −0.2, −0.8, 0.7}
−0.27	{ 0.7, −1.1, −0.2, −0.8}	0.83	{−0.8, −1.1, 0.7, −0.2}
−0.23	{ 0.7, −1.1, −0.8, −0.2}	0.84	{−1.1, −0.8, 0.7, −0.2}
−0.21	{−1.1, 0.7, −0.2, −0.8}	0.90	{−0.8, −1.1, −0.2, 0.7}
−0.16	{−1.1, 0.7, −0.8, −0.2}	0.91	{−1.1, −0.8, −0.2, 0.7}

Values of R lie in the interval $[-1, 1]$. Positive values favor the alternative that the characteristics under study are positively associated. Negative values favor the alternative that the characteristics under study are negatively associated.

Note that under the population model, R is an estimate of the correlation coefficient $\rho = Corr(X, Y)$. Technically, tests based on R are tests of the null hypothesis that $\rho = 0$ and *not* tests of the null hypothesis that X and Y are independent.

Permutation distribution

The permutation distribution of R is obtained by computing the sample correlation for each matching of a permutation of the y values to the ordered x values. The following theorem gives information about the resulting distribution.

Theorem 11.5 (Sample Correlations). *Conditional on the observed pairs, the permutation distribution of R has the following summary measures:*

$$E(R) = 0 \text{ and } Var(R) = \frac{1}{n-1}.$$

If n is large enough, then the R distribution is approximately normal.

For example, let $n = 4$. Suppose the observed pairs are as follows:

$$\{\{-1.3, -0.2\}, \{-1.2, -1.1\}, \{0.6, 0.7\}, \{0.8, -0.8\}\}.$$

Table 11.4 shows the value of R (with two decimal places of accuracy) for each of the $4! = 24$ permutations of y values. Consider testing the null hypothesis of randomness using a two sided alternative. The observed value of R is 0.36, and the observed significance level is $P(|R| \geq |0.36|) = 16/24$.

Table 11.5. *Cholesterol (x) and triglycerides (y) for the 51 men with no evidence of heart disease.*

x	116	130	147	149	150	155	156	157	158	160	167	168	168
y	87	64	95	146	167	48	126	134	87	116	177	71	100
x	168	169	170	178	178	178	180	187	190	190	190	193	194
y	227	86	90	116	157	166	82	109	85	108	132	210	121
x	195	200	201	201	205	206	207	207	208	209	210	217	219
y	348	154	72	171	158	99	160	195	139	97	91	114	98
x	221	222	228	234	234	237	238	243	251	265	266	289	
y	156	284	119	116	143	174	172	101	211	73	486	120	

Example: Analysis of cholesterol-triglycerides pairs

Cholesterol and triglycerides belong to the class of chemicals known as lipids (fats). As part of a study to determine the relationship between high levels of lipids and coronary artery disease, researchers measured plasma levels of cholesterol and triglycerides in milligrams per deciliter (mg/dl) in more than 300 men complaining of chest pain [95], [50, p. 221].

Table 11.5 gives the cholesterol (x) and triglycerides (y) measurements for the 51 men with no evidence of disease. The observed correlation is 0.325. To determine if the observed association is due to chance alone, a permutation test will be conducted using a two sided alternative and a 5% significance level.

In a Monte Carlo analysis using 5000 random permutations (including the observed permutation of the y values), 1.96% (98/5000) of $|R|$ values were greater than or equal to $|0.325|$. Thus, there is evidence that the observed association between cholesterol and triglycerides in men complaining of chest pain but with no evidence of disease is not due to chance alone.

11.3.2 Example: Rank correlation test

In the early 1900's, Spearman proposed a test based on the ranks of the x and y values. Spearman's rank correlation statistic, R_s, is computed as follows:

1. Replace each x by its rank (or midrank) in the ordered x values.
2. Replace each y by its rank (or midrank) in the ordered y values.
3. Let R_s equal the sample correlation of the paired ranks.

For example, if $n = 6$ and the list of paired data is

{{10.42, 13.18}, {11.43, 14.03}, {11.79, 13.24}, {13.17, 12.03}, {13.4, 11.75}, {13.53, 11.83}},

then the list of paired ranks is

$$\{\{1, 4\}, \{2, 6\}, \{3, 5\}, \{4, 3\}, \{5, 1\}, \{6, 2\}\}$$

and the observed value of Spearman's statistic is -0.771429.

Permutation tests using R_s are conducted in the same way as permutation tests using the sample correlation. Unless there are many ties in the data or n is very small, the large sample normal approximation to the R_s distribution can be used to estimate p values.

For example, consider the cholesterol-triglycerides data in Table 11.5. The observed value of Spearman's statistic is 0.288, and the observed significance level for a two sided test is

$$P(|R_s| \geq |0.288|) \approx P(|N| \geq |0.288|) \approx 0.042,$$

where N is the normal random variable with mean 0 and variance 1/50. Once again, the null hypothesis that the observed association between cholesterol and triglycerides is due to chance alone is rejected at the 5% significance level.

Comparison of tests

Permutation tests based on the sample correlation and rank correlation statistics are valid in the same situations. If the data are highly skewed or have extreme outliers, then the rank correlation test is preferred; otherwise, the sample correlation statistic should be used.

It is interesting to note that the rank correlation statistic is unchanged if either the x values or the y values are transformed using an increasing function (such as square or square root).

11.4 Additional tests and extensions

This section introduces additional permutation tests and ways in which the ideas of this chapter can be extended. Other methods will be presented in the laboratory problems and in later chapters.

11.4.1 Example: One sample trend test

Consider the analysis of a single sample, $\{x_1, x_2, \ldots, x_n\}$, where the index corresponds to time. That is, x_1 is the first observed measurement, x_2 is the second observed measurement, etc.

Of interest is whether there is a linear trend over time. For example, a manufacturer may be interested in determining whether a critical measurement has increased or decreased over time. If a systematic change has occurred, then the manufacturing process would need to be adjusted.

In the 1940's, Mann proposed a simple approach to testing for trend. Let

$$S = \sum_{i<j} U_{ij}, \quad \text{where } U_{ij} = \begin{cases} +1 & \text{when } x_i < x_j, \\ -1 & \text{when } x_i > x_j, \\ 0 & \text{when } x_i = x_j. \end{cases}$$

A total of $\binom{n}{2}$ comparisons are made. If the n observations are strictly increasing, then the value of S is $\binom{n}{2}$; if the n observations are strictly decreasing, then the value

is $-\binom{n}{2}$. If there is no trend over time, then the $+1$'s and -1's will roughly balance and the value of S will be near zero.

Tests using the S statistic are appropriate in the following situations:

1. *Population model.* The data are the values of independent continuous random variables. The null hypothesis is that the X_i's have the same distribution:

$$F_1 = F_2 = \cdots = F_n,$$

where F_i is the CDF of X_i, for $i = 1, 2, \ldots, n$. Alternatives of interest are

$$F_1(x) < F_2(x) < \cdots < F_n(x) \text{ for all real numbers } x$$

(values tend to decrease with time) or

$$F_1(x) > F_2(x) > \cdots > F_n(x) \text{ for all real numbers } x$$

(values tend to increase with time).

2. *Randomization model.* The data are measurements taken at n time points. The null hypothesis is that there is no relationship between the measurements and time. Alternatives of interest are that measurements tend to increase or decrease with time.

Consider the population model. Under the null hypothesis, the list $X_1, X_2, \ldots, X_n$ is a random sample from a continuous distribution. Thus, each of $n!$ orderings of the n observations is equally likely. This fact forms the basis of the permutation method.

Permutation distribution

The sampling distribution of S is obtained by computing the value of S for each permutation of the x values. The distribution depends on the ranks (or midranks, in case of ties) of the observations and not on the observations themselves. The following theorem gives information about the distribution when there are no ties in the data.

Theorem 11.6 (Trend Statistic). *Consider the permutation distribution of S under the null hypothesis. If there are no ties in the observed data, then the following hold:*

1. *S takes integer values between $-n(n-1)/2$ and $n(n-1)/2$.*
2. *$E(S) = 0$ and $Var(S) = n(n-1)(2n+5)/18$.*
3. *The distribution of S is symmetric around 0:*

$$P(S = x) = P(S = -x) \text{ for all } x.$$

4. *If n is large, then the distribution of S is approximately normal.*

For example, if $n = 8$ and there are no ties in the data, then S takes integer values between -28 and 28. The S distribution has mean 0 and variance 65.33. If the observations (in time order) are

$$24.11,\ 23.23,\ 28.07,\ 15.08,\ 16.81,\ 21.26,\ 9.61,\ 10.57,$$

then the observed value of S is -16 and the observed significance level for a two sided trend test is

$$P(|S| \geq |-16|) = \frac{2460}{40320} \approx 0.0610$$

(there are $8! = 40,320$ permutations to consider).

11.4.2 Example: Two sample scale test

Consider again the analysis of two samples, $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$, where the data are the values of independent random samples from continuous distributions.

Let X and Y be continuous random variables with common median $\theta = \text{Median}(X) = \text{Median}(Y)$, and let Δ be a positive constant. Assume that

$$(X - \theta) \text{ and } \Delta(Y - \theta) \text{ have the same distribution.}$$

That is, assume that the distributions differ in scale only.

Note that if $\Delta > 1$, then X is more variable than Y; if $\Delta < 1$, then Y is more variable than X; and if $\Delta = 1$, the distributions are equal.

Tests of the null hypothesis that the distributions are equal ($\Delta = 1$) versus alternatives that one distribution is more variable than the other can be conducted using a simple test proposed by Siegel and Tukey (and modified by Ansari and Bradley) in 1960.

Specifically, the (symmetrized) Siegel–Tukey statistics for the X sample (W_1) and the Y sample (W_2) are computed as follows:

1. Pool and sort the $n + m$ observations.
2. Assign rank 1 to the smallest and largest observations, rank 2 to the second smallest and second largest observations, rank 3 to the third smallest and third largest observations, and so forth. Use midranks in case of ties.
3. Let W_1 equal the sum of the ranks for observations in the X sample, and let W_2 equal the sum of the ranks for observations in the Y sample.

For example, let $n = 6$ and $m = 8$. If the observations in the first sample are 5.71, 5.94, 5.95, 6.05, 6.38, 6.63 and the observations in the second sample are 1.17, 2.42, 4.18, 4.72, 4.78, 5.07, 11.39, 12.32, then the ordered combined list of 14 observations is

1.17, 2.42, 4.18, 4.72, 4.78, 5.07, 5.71, 5.94, 5.95, 6.05, 6.38, 6.63, 11.39, 12.32.

1.17 and 12.32 would each receive rank 1, 2.42 and 11.39 would each receive rank 2, etc. The observed value of W_1 is 32, and the observed value of W_2 is 24.

Tests based on W_1 and W_2 are equivalent. For tests based on W_1, large values of W_1 support the alternative that the X distribution is less variable than the Y distribution ($\Delta < 1$); small values support the opposite alternative.

Permutation distribution

The sampling distribution of W_1 is obtained by computing the value of the statistic for each partition of the combined list of $n + m$ observations (with repetitions) into sublists of lengths n and m.

For the example above, W_1 takes integer values between 12 and 36. The distribution of W_1 is symmetric, with mean 24 and variance 14.7692. The observed significance level for a two sided test of the null hypothesis is

$$2P(W_1 \geq 32) = \frac{144}{3003} \approx 0.047952$$

(there are $\binom{14}{6} = 3003$ partitions to consider).

11.4.3 Stratified analyses

Researchers use stratified samples when they expect individuals in different subpopulations (or *strata*) to respond differently to treatment. This section presents two general examples of how permutation methods can be applied to analyze data from stratified studies.

Example: Stratified two sample analysis

Suppose that a researcher is interested in comparing two methods for teaching multivariable calculus and expects students at different schools to respond differently to the teaching methods.

If the researcher is interested in comparing the two methods at each of four schools using 20 students at each school, for example, then a simple design would be as follows:

> At each school, randomly assign 10 of the 20 students to a class using the first teaching method, with the remaining 10 assigned to a class using the second teaching method.

Assume that the measurement of interest is a score on a standardized final exam, and let

$$\{\{s_{1,1}, s_{1,2}\}, \{s_{2,1}, s_{2,2}\}, \{s_{3,1}, s_{3,2}\}, \{s_{4,1}, s_{4,2}\}\}$$

be the nested list of scores. For school i, $s_{i,1}$ is the list of scores for students using the first teaching method, and $s_{i,2}$ is the list of scores for the second method.

Under the null hypothesis that the teaching methods are equivalent, the scores for each school may be randomly partitioned into subsets of sizes 10 (for the first

method) and 10 (for the second method). Since there are four schools, there are a total of $\binom{20}{10}^4$ choices to consider.

Example: Paired sample analysis

Paired sample analyses are examples of stratified analyses. In addition to the methods of Section 11.2 for analyses of the differences list, the list of pairs can be analyzed directly.

Suppose that a researcher is interested in comparing different treatments for a serious medical condition and expects individuals of different age, sex, and disease status to respond differently to the proposed treatments.

If the researcher is interested in comparing the two treatments using 30 pairs of individuals matched on sex, age group, and disease status score, for example, then a simple design is as follows:

> For each pair of individuals, randomly assign one individual to receive the first treatment and the other individual to receive the second treatment.

Assume that the measurement of interest is the disease status score after a fixed period of time, and let

$$\{\{x_1, y_1\}, \{x_2, y_2\}, \ldots, \{x_{30}, y_{30}\}\}$$

be the nested list of scores. For pair i, x_i is the score for the individual assigned to the first treatment, and y_i is the score for the individual assigned to the second treatment.

Under the null hypothesis that the treatments are equivalent, the scores for each pair may be randomly permuted. Since there are 30 pairs, there are a total of 2^{30} choices to consider.

11.5 Laboratory problems

Laboratory problems for this chapter introduce *Mathematica* commands for using simulation to estimate p values; for analyzing samples using signed rank, trend, and rank correlation methods; and for constructing plots of empirical cumulative distribution functions. The problems are designed to reinforce ideas about permutation analysis.

11.5.1 Laboratory: Permutation analysis

In the main laboratory notebook (Problems 1 to 5), you will use simulation and graphics to study two sample rank sum and Smirnov tests and apply the Smirnov test to data from a study of earthquake locations from historical and current records [60]; use simulation and graphics to study the correlation test and apply the test to data from a study comparing overseas and domestic stock market returns [78]; and use simulation and graphics to study the signed rank test and apply the test to data from two studies: (1) a study to determine if the labeling of so-called health foods is accurate [2], [30] and (2) a study to determine if a diet containing oat bran can help reduce serum cholesterol levels [4], [82].

11.5.2 Additional problem notebooks

Problems 6, 7, and 8 are applications of permutation methods for two samples. In Problem 6, two woodlands areas are compared using a variety of biodiversity indices [72], [101]. In Problem 7, treated and control multiple sclerosis patients are compared using stratified two sample methods [63], [8]. In Problem 8, consumption rates for male and female Asian shore crabs are compared using stratified two sample methods [19].

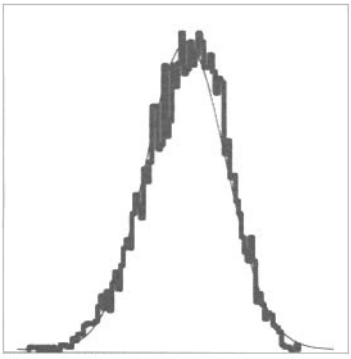
Problem 9 applies the rank correlation test to data on factors related to death rates in large metropolitan areas [49]. Problem 10 applies the trend test to data on manganese content of iron ore [90].

Problems 11 and 12 are applications of permutations methods for paired samples. In Problem 11, a cloud-seeding experiment is analyzed using a double-ratio statistic [76]. In Problem 12, spatial distributions of cod are studied using two-dimensional Cramer–von Mises statistics [105].

Problems 13 and 14 concern tests for frequency data. In Problem 13, the goal is to study the association between genetics and longevity using data on female twins and the number of twin pairs alive at certain ages as test statistic [114]. In Problem 14, the goal is to study the association between incidence of childhood leukemia and distance to a hazardous waste site using data from an area in upstate New York and the Stone statistic [111].

Problem 15 uses frequency generating functions to study properties of the Wilcoxon signed rank statistic. Problem 16 demonstrates how frequency generating functions can be used to construct the sampling distribution for the Fisher symmetry test quickly; the method is applied to the data on the accuracy of health-food labeling from the main laboratory notebook.

This page intentionally left blank



Chapter 12
Bootstrap Analysis

In many statistical applications, interest focuses on estimating a quantity using a random sample from a probability distribution, the distribution from which the data were drawn is not known exactly, and the sampling distribution of the statistic used to estimate the quantity is not known exactly (or approximately). Bootstrap methods allow researchers to make approximate probability calculations in these situations by using the computer to simulate the original experiment many times.

This chapter introduces bootstrap analysis. The first three sections introduce bootstrap estimation methods and give many applications. Section 4 considers bootstrap hypothesis testing methods. Section 5 outlines the laboratory problems. General references for this chapter are [31], [36].

12.1 Introduction

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with parameter θ, and let $T = T(X_1, X_2, \ldots, X_n)$ be a statistic used to estimate θ. The computer can be used to approximate the sampling distribution of T and to estimate the mean and standard deviation of the T distribution.

Two types of computer analysis will be discussed.

Nonparametric bootstrap analysis

If the n observations are $x_1, x_2, \ldots, x_n$, then the *observed distribution* of the sample data is the discrete distribution with PDF

$$f(x) = \frac{\#(x_i\text{'s equal to } x)}{n} \quad \text{for all } x.$$

For each observed x, $f(x)$ is the proportion of times x appears in the sample; otherwise, $f(x)$ equals 0.

In nonparametric bootstrap analysis, the sampling distribution of T is approximated using replicate data sets of size n created by sampling from the observed

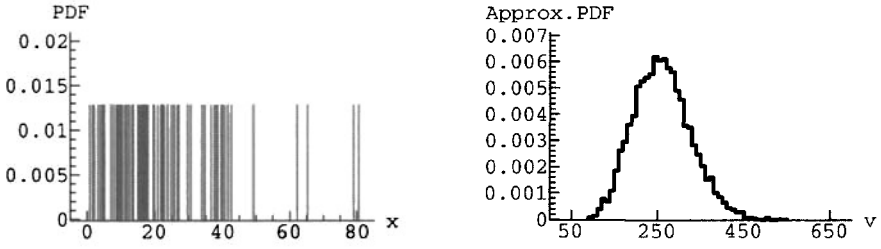


Figure 12.1. *Observed distribution of a random sample of size* 78 *(left plot) and bootstrap approximate sampling distribution of the sample variance using* 5000 *resamples from the observed distribution (right plot).*

distribution of the sample data. Each replicate data set is called a random *resample* of the original data.

A useful way to think about sampling from the observed distribution is as follows. Imagine writing the n observations on n slips of paper and placing the slips in an urn. The following experiment is repeated n times: Thoroughly mix the urn, choose a slip, record the value, return the slip to the urn. Thus, each replicate data set is the result of sampling *with* replacement n times from the original list of n observations.

For example, let X be a continuous random variable with unknown variance, and let S^2 be the sample variance of a random sample of size n from the X distribution. Assume that the left part of Figure 12.1 is a line plot representing the observed distribution of a random sample of size 78 from the X distribution. (In a *line plot* of an observed distribution, a segment from $(x, 0)$ to $(x, f(x))$ is used to represent each observation and its probability.) The observed sample variance is 264.54.

The right part of Figure 12.1 is a histogram of 5000 sample variances, where each sample variance is based on a random sample of size 78 from the observed distribution. The histogram is an approximation of the sampling distribution of S^2 when $n = 78$. The mean of the approximate sampling distribution is 260.51, and the standard deviation is 65.18.

Parametric bootstrap analysis

In parametric bootstrap analysis, the observations are used to fit a model to the X distribution. The sampling distribution of T is approximated using replicate data sets of size n created by sampling from the estimated model.

Continuing with the example above, if X is a gamma random variable, then the parameters of the gamma model can be fit using maximum likelihood. For the data in Figure 12.1, the ML estimate of α is 1.84 and the ML estimate of β is 11.91. The left part of Figure 12.2 shows the estimated gamma model superimposed on an empirical histogram of the data; the right part is a histogram of 5000 sample variances, where each sample variance is based on a different random sample of size 78 from the estimated gamma distribution. The right part is an approximation of the sampling distribution of S^2 for samples of size 78 from a gamma distribution. The mean of the approximate sampling distribution is 263.19, and the standard deviation is 69.15.

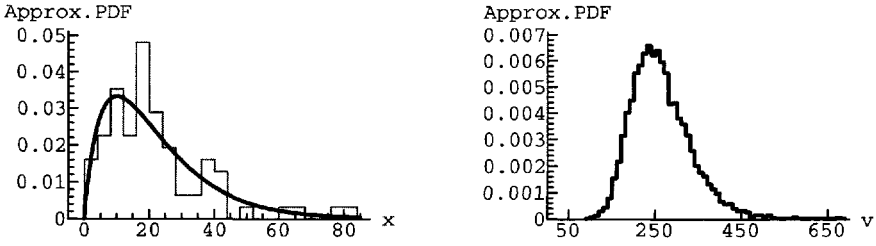


Figure 12.2. *Empirical histogram of a random sample of size* 78 *with estimated gamma model superimposed (left plot) and bootstrap approximate sampling distribution of the sample variance using* 5000 *resamples from the estimated gamma distribution (right plot).*

Note that the two methods (nonparametric and parametric bootstrap) produced roughly equivalent results in this case because the gamma model fits the sample data well.

12.1.1 Approximate conditional estimation

A bootstrap analysis is an example of a *conditional* analysis, since the distribution used to create replicate data sets is constructed *conditional* on the observed data.

Since simulation is used to approximate the sampling distribution of T and to estimate its summary measures, a bootstrap analysis is also an example of a Monte Carlo analysis. (In a *Monte Carlo* analysis, simulation is used to estimate quantities of interest.)

Sources of error

There are two sources of error in a bootstrap analysis:

1. the error in using the observed distribution or an estimated model instead of the X distribution itself and

2. the error in using a fixed number of replicate data sets to approximate the sampling distribution of T and to estimate its summary measures.

If the sample size n is large, the sample data approximate the X distribution well, and the resampling scheme does not rely strongly on a small subset of the observed data, then the results of bootstrap analyses are generally good.

To illustrate a situation where the approximate T distribution is not close to its true distribution, let X be a continuous random variable, let θ be the 90^{th} percentile of the X distribution, and let $X_{(72)}$ be the 72^{nd} order statistic in a random sample of size 79 from the X distribution. $X_{(72)}$ can be used to estimate θ. (See Theorem 9.1.) Assume that the left part of Figure 12.3 is a line plot representing the observed distribution of a random sample of size 79 from the X distribution. The observed 72^{nd} order statistic is 9.14.

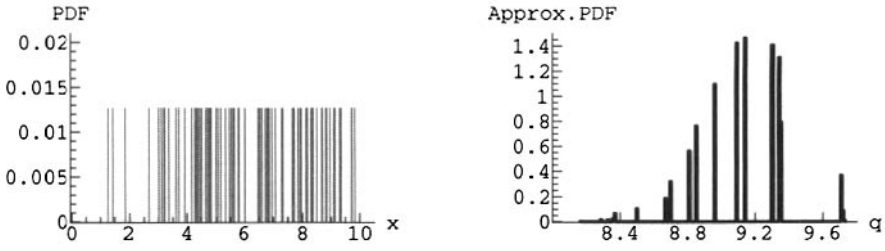


Figure 12.3. *Observed distribution of a random sample of size* 79 *(left plot) and bootstrap approximate sampling distribution of the* 72nd *order statistic using* 5000 *resamples from the observed distribution (right plot).*

The right part of Figure 12.3 is a histogram of 5000 sample 72nd order statistics, where each order statistic is based on a random sample of size 79 from the observed distribution. For the data on the right, the mean is 9.13 and the standard deviation is 0.26.

Although 9.13 and 0.26 are reasonable estimates of $E(X_{(72)})$ and $SD(X_{(72)})$, respectively, the shape of the distribution on the right is not close to the shape of the distribution of an order statistic from a continuous distribution.

In the resampling step of the example above, the values of the 72nd order statistic are restricted to the 79 observed values of X. 94.9% (4746/5000) of the simulated values were equal to one of 10 numbers from the original list of 79 numbers. Thus, the histogram does not approximate a continuous curve very well. This problem would not be alleviated by using a larger number of resampled values.

Number of bootstrap resamples

The idea behind the bootstrap is very simple and attractive: the computer is used to estimate properties of a sampling distribution. However, it is difficult to assess the reliability of results from a bootstrap analysis. Thus, bootstrap results should be used with caution.

One way to reduce the second type of error above is to use a large number of resamples. In the examples below, 5000 resamples are used. In general, 5000 resamples are sufficient to give reasonable estimates of the quantities of interest in this chapter.

12.2 Bootstrap estimation

Let θ be a parameter of interest, T be a statistic used to estimate θ from sample data, and t_{obs} be the observed value of T. Assume the sample data are the values of a random sample or of independent random samples.

In the resampling step of a bootstrap analysis, either the observed distribution or an estimated model is used to produce simulated data and simulated values of T, say t^*, as illustrated in the right column of Figure 12.4. This process is repeated a large number of times, say B, to produce an approximate sample from the sampling

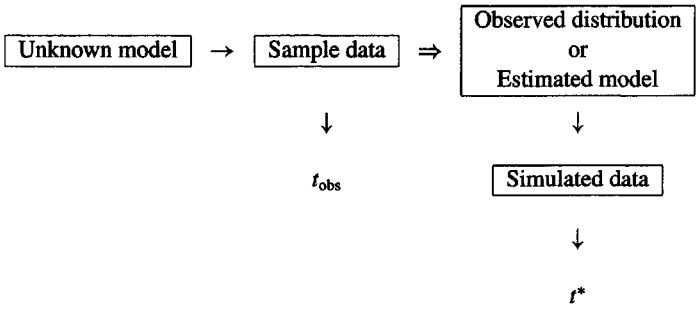


Figure 12.4. *Illustration of bootstrap resampling.*

distribution of T:

$$t_1^*,\ t_2^*,\ t_3^*, \ldots,\ t_B^*.$$

12.2.1 Error distribution

Recall that the bias of an estimator T is the difference between its expected value and the parameter of interest, $BIAS(T) = E(T) - \theta$. The *standard error* (SE) of an estimator T is the same as the standard deviation of T, $SE(T) = SD(T)$.

The *error distribution* is the distribution of $T - \theta$. Its mean and standard deviation are equal to the bias and standard error of T, respectively:

$$E(T - \theta) = E(T) - \theta = BIAS(T) \text{ and } SD(T - \theta) = SD(T) = SE(T).$$

(The error distribution is a shift of the T distribution.)

Bias and standard error are estimated using the mean and standard deviation of the B approximate errors, where t_{obs} takes the place of θ:

$$t_1^* - t_{obs},\ t_2^* - t_{obs},\ t_3^* - t_{obs}, \ldots,\ t_B^* - t_{obs}.$$

12.2.2 Simple approximate confidence interval procedures

This section presents two simple approximate confidence interval procedures for θ.

Standard bootstrap confidence intervals

If the error distribution is approximately normal when n is large, then an approximate $100(1 - \alpha)\%$ confidence interval for θ has the form

$$t_{obs} - b \pm z(\alpha/2)\, se,$$

where b is the estimated bias, se is the estimated standard error, and $z(\alpha/2)$ is the $100(1 - \alpha/2)\%$ point of the standard normal distribution.

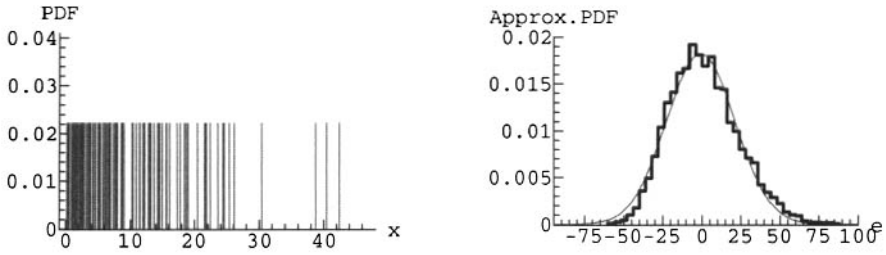


Figure 12.5. *Observed distribution of a random sample of size* 110 *(left plot) and bootstrap approximate error distribution based on* 5000 *resamples from the observed distribution (right plot). The approximate error distribution is superimposed on a normal distribution with the same mean and standard deviation.*

Basic bootstrap confidence intervals

If the error distribution is not approximately normal, then approximate confidence intervals can be based on sample quantiles. Specifically, an approximate $100(1-\alpha)\%$ confidence interval for θ has the form

$$\left[t_{\text{obs}} - t^*_{1-\alpha/2},\ t_{\text{obs}} - t^*_{\alpha/2}\right],$$

where t^*_p is the sample p^{th} quantile of the list of B estimated errors.

Demonstrations

To demonstrate the approximate confidence procedures, note that

$$1-\alpha = P\left(t_{\alpha/2} \le T-\theta \le t_{1-\alpha/2}\right) = P\left(T - t_{1-\alpha/2} \le \theta \le T - t_{\alpha/2}\right),$$

where $t_{\alpha/2}$ and $t_{1-\alpha/2}$ are quantiles of the error distribution.

(i) If sample quantiles are used to estimate $t_{\alpha/2}$ and $t_{1-\alpha/2}$, then the estimated lower endpoint is $t_{\text{obs}} - t^*_{1-\alpha/2}$ and the estimated upper endpoint is $t_{\text{obs}} - t^*_{\alpha/2}$.

(ii) If the normal approximation is used, then

$$t_{\alpha/2} \approx b - z(\alpha/2)se \quad \text{and} \quad t_{1-\alpha/2} \approx b + z(\alpha/2)se.$$

The estimated lower endpoint is $t_{\text{obs}} - b - z(\alpha/2)se$, and the estimated upper endpoint is $t_{\text{obs}} - b + z(\alpha/2)se$.

Example: Simple confidence intervals for variance

To illustrate the approximate procedures, assume that the left part of Figure 12.5 is a line plot representing the observed distribution of a random sample of size 110 from a distribution with unknown variance σ^2. For these data the sample variance is 92.51.

The right part of Figure 12.5 is a histogram of 5000 estimated errors, $s^2 - 92.51$, superimposed on a normal density curve. Each error is based on a random sample of size 110 from the observed distribution. The normal density has mean equal to the estimated bias (-0.74) and standard deviation equal to the estimated standard error (21.92). An approximate 95% confidence interval for σ^2 based on a normal approximation to the error distribution is

$$(92.51 + 0.74) \pm 1.96(21.92) \implies [50.30, 136.21].$$

The 0.025 and 0.975 sample quantiles of the estimated error list are -38.71 and 46.43, respectively, and an approximate 95% confidence interval for σ^2 using the basic bootstrap procedure is

$$[92.51 - 46.43, 92.51 + 38.71] = [46.08, 131.22].$$

The graphic suggests that the second interval is the better choice in this case.

12.2.3 Improved intervals: Nonparametric case

Recall that the interval $[L, U]$ is a $100(1-\alpha)\%$ confidence interval for θ if

$$P(\theta < L) = P(\theta > U) = \frac{\alpha}{2} \text{ and } P(L \le \theta \le U) = 1 - \alpha.$$

A method for computing intervals is said to be *approximately accurate* when

$$P(\theta < L) \approx \frac{\alpha}{2} \text{ and } P(\theta > U) \approx \frac{\alpha}{2} \text{ and } P(L \le \theta \le U) \approx 1 - \alpha.$$

Confidence intervals are *transformation-preserving*. That is, if $[L, U]$ is a $100(1-\alpha)\%$ confidence interval for θ and

(i) g is an increasing function, then $[g(L), g(U)]$ is a $100(1-\alpha)\%$ CI for $g(\theta)$;
(ii) g is a decreasing function, then $[g(U), g(L)]$ is a $100(1-\alpha)\%$ CI for $g(\theta)$.

The approximate procedures described in the last section are *not* true approximate confidence procedures. The methods are not transforming-preserving, and the endpoints are not approximately accurate.

B. Efron, who pioneered the use of the bootstrap, proposed an improved procedure known as the *bias-corrected and adjusted* (or BC_a) percentile method (see, for example, [36, p. 184]). In general, Efron's improved method produces intervals whose endpoints are approximately accurate and which are transformation-preserving. The details for the improved method will not be discussed here. The algorithms have been implemented in the nonparametric case.

For example, Efron's improved method (with 5000 resamples) applied to the data in Figure 12.5 produced the following approximate 95% confidence interval for the variance σ^2: [57.27, 163.47].

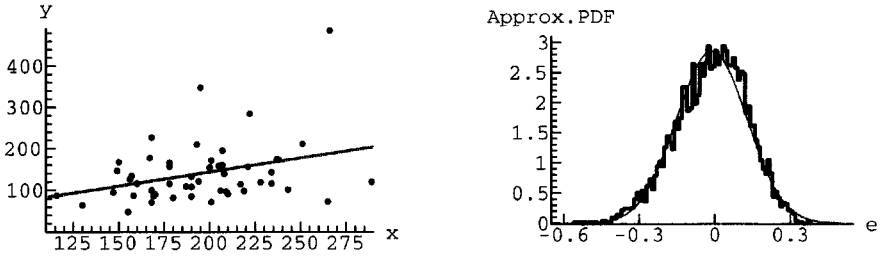


Figure 12.6. *Scatter plot of cholesterol-triglycerides pairs (left plot) and bootstrap approximate error distribution based on* 5000 *resamples from the observed distribution (right plot). The approximate error distribution is superimposed on a normal distribution with the same mean and standard deviation.*

12.3 Applications of bootstrap estimation

This section considers applications of bootstrap estimation when the observed data are the values of a single random sample and when they are the values of two or more independent random samples. Three example analyses are presented. Other examples will be considered in the laboratory problems and in later chapters.

12.3.1 Single random sample

In addition to using the bootstrap to study the sampling distribution of the sample variance, bootstrap methods can be used to study the sampling distributions of the following:

1. MOM and ML estimators. In particular, the bootstrap can be used to check if the large sample approximate normal distribution of an ML estimator is adequate in a particular situation.
2. HL estimators of shift in the paired sample setting.
3. Correlation estimators (see the example below).
4. Estimators of location, such as the trimmed mean (see the example below).

Example: Correlation analysis

Consider again the data on cholesterol and triglycerides levels in 51 men complaining of chest pain but with no evidence of heart disease (see Table 11.5). Assume these data are the values of a random sample from a joint cholesterol-triglycerides distribution with correlation coefficient ρ, and let R be the sample correlation.

The left part of Figure 12.6 is a scatter plot of the data. The observed sample correlation is 0.325. The right part is a histogram of 5000 estimated errors, $r - 0.235$, superimposed on a normal density curve. Each error is based on a random sample of size 51 from the observed distribution. The normal density has mean equal to the estimated bias (-0.012) and standard deviation equal to the estimated standard error (0.140). Using Efron's method (with 5000 resamples), an approximate 95%

confidence interval for ρ is [0.043, 0.641]. Since both endpoints are positive, this analysis suggests a positive association between levels of cholesterol and triglycerides in the population from which these men were sampled.

Example: Trimmed mean analysis

Let X be a continuous random variable with mean μ and PDF $f(x)$, and let α be a proportion in the interval $0 < \alpha < \frac{1}{2}$. The $100\alpha\%$ *trimmed mean* of X is the expected value of the middle $100(1-2\alpha)\%$ of the X distribution:

$$100\alpha\% \text{ Trimmed Mean} = \frac{1}{1-2\alpha} \int_{x_\alpha}^{x_{1-\alpha}} x\, f(x)dx,$$

where x_p is the p^{th} quantile of the X distribution.

As α approaches 0, the $100\alpha\%$ trimmed mean approaches μ. As α approaches 1/2, the $100\alpha\%$ trimmed mean approaches the median of the distribution.

The sample $100\alpha\%$ trimmed mean is the sample mean of the middle $100(1-2\alpha)\%$ of the sample data. For example, if $n = 10$ and the ordered data are

$$2.16,\ 6.26,\ 8.64,\ 8.82,\ 11.82,\ 13.61,\ 17.39,\ 27.84,\ 29.40,\ 58.42,$$

then the sample 20% trimmed mean is the mean of the middle six numbers, 14.6867.

If the X distribution is symmetric around $x = \mu$, then μ is the mean, median, and $100\alpha\%$ trimmed mean for each α. Researchers often use sample $100\alpha\%$ trimmed means to estimate the mean in the symmetric case when there are outliers in the data, since outliers do not affect these estimators as much as they do sample means. (Sample trimmed means are examples of robust estimators of location. A *robust* estimator is one that is not sensitive to outliers.)

If the X distribution is not symmetric, then the mean, the median, and the $100\alpha\%$ trimmed mean (for each α) are different measures of the center of the X distribution. When distributions are extremely skewed, a $100\alpha\%$ trimmed mean may be a better measure of center than either the mean or the median.

For example, the left part of Figure 12.7 is a line plot representing the observed distribution of 227 rainfall measurements (in inches). The measurements were made at a series of rain gauges in southern Illinois in the summers of 1960 through 1964 [66], [90, p. 249]. For these data, the mean is 0.224 inches, the median is 0.07 inches, and the 20% trimmed mean is 0.106 inches.

Assume these data are the values of a random sample from a distribution with 20% trimmed mean θ, and let $\widehat{\theta}$ be the sample 20% trimmed mean. The right part of Figure 12.7 is a histogram of 5000 estimated errors, $\widehat{\theta} - 0.106$, superimposed on a normal density curve. Each error is based on a random sample of size 227 from the observed distribution. The normal density has mean equal to the estimated bias (0.0008) and standard deviation equal to the estimated standard error (0.015). Using Efron's method (with 5000 resamples), an approximate 95% confidence interval for the 20% trimmed mean θ is [0.078, 0.144].

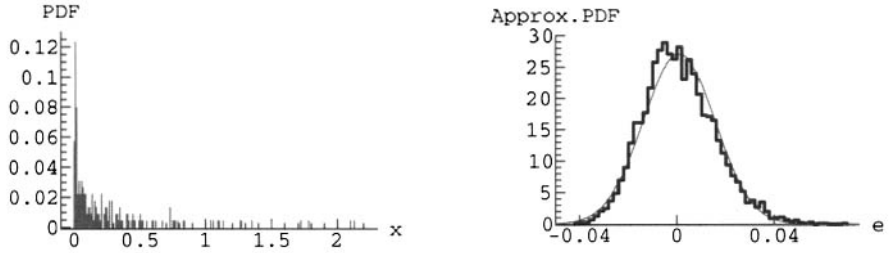


Figure 12.7. *Observed distribution of rainfall data (left plot) and bootstrap approximate error distribution based on* 5000 *resamples from the observed distribution (right plot). The approximate error distribution is superimposed on a normal distribution with the same mean and standard deviation.*

12.3.2 Independent random samples

If the observed data are the values of two or more independent random samples, then resampling is done separately from each estimated distribution. In the two sample setting, applications of bootstrap methods include studying the sampling distributions of

1. variance ratio estimators when X and Y are not normal random variables,
2. IQR ratio estimators (see the example below),
3. mean or median ratio estimators,
4. HL estimators of shift in the two sample setting,
5. estimators of $P(X < Y)$ or $P(X > Y)$.

Example: Ratio of IQRs analysis

Let X and Y be continuous random variables with interquartile ranges IQR_x and IQR_y, respectively, and let $\theta = \text{IQR}_x/\text{IQR}_y$ be the ratio of IQRs. Let $\widehat{\theta}$ be the ratio of sample IQRs, based on independent random samples of sizes n and m from the X and Y distributions. $\widehat{\theta}$ can be used to estimate θ.

In the resampling step of a nonparametric bootstrap analysis, resamples are taken separately from the observed X and Y distributions. For example, the left part of Figure 12.8 shows side-by-side box plots of the finishing times (in hours) of the 65 women and 111 men who completed the 1996 Olympic marathon competition in Atlanta, Georgia [81]. For these data, the sample IQR for women is 0.163 and for men is 0.209. The sample IQR ratio is $0.163/0.209 = 0.779$.

Assume the marathon data are the values of independent random samples. The right part of Figure 12.8 is a histogram of 5000 estimated errors, $\widehat{\theta} - 0.779$, superimposed on a normal density curve. Each error is based on independent random samples of sizes 65 and 111 from the observed distributions. The normal density has mean equal to the estimated bias (0.065) and standard deviation equal to the estimated standard error (0.208). Using Efron's method (with 5000 resamples), an

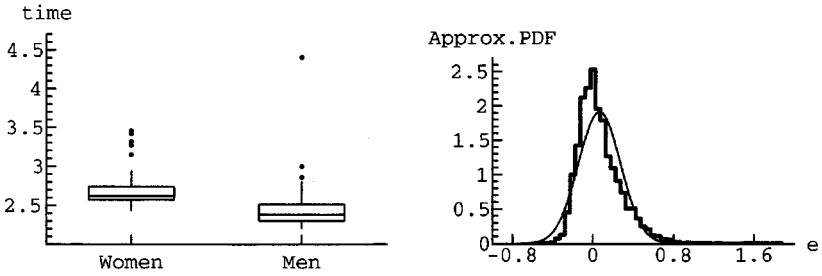


Figure 12.8. *Side-by-side box plots of the finishing times (in hours) of Olympic finishing times data (left plot) and bootstrap approximate error distribution based on* 5000 *resamples from the observed distributions (right plot). The approximate error distribution is superimposed on a normal distribution with the same mean and standard deviation.*

approximate 95% confidence interval for the IQR ratio θ is [0.477, 1.349]. Since 1 is in the interval, the results suggest that the IQRs of the two distributions are equal.

12.4 Bootstrap hypothesis testing

Bootstrap resampling methods can be adapted to conduct approximate hypothesis tests. This section introduces two examples.

Example: Difference in means test

Let X and Y be continuous nonnormal random variables with unknown (but finite) means and variances. Consider testing the null hypothesis

$$H_o\text{: The means of the } X \text{ and } Y \text{ distributions are equal}$$

versus alternatives that the mean of one distribution is larger than that of the other distribution using independent random samples of sizes n and m, respectively, from the X and Y distributions, and Welch's t statistic,

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}.$$

Large values of T favor the alternative that the mean of X is larger than the mean of Y; small values of T favor the opposite alternative.

> A test of the equality of means is not the same as a test of equality of distributions. For example, X and Y may be members of different families of distributions.

Nonparametric bootstrap analysis can be used to approximate the sampling distribution of T under the null hypothesis of equality of means and to estimate p values [36, p. 212]. To carry out the analysis, the observed data need to be adjusted to satisfy the null hypothesis.

Specifically, let $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$ be the observed samples and t_{obs} be the observed value of the T statistic.

(i) The samples used in the resampling step are

$$\begin{aligned} &\text{Adjusted } x \text{ sample:} \quad x_1 - \bar{x} + \bar{z},\ x_2 - \bar{x} + \bar{z}, \ldots,\ x_n - \bar{x} + \bar{z}, \\ &\text{Adjusted } y \text{ sample:} \quad y_1 - \bar{y} + \bar{z},\ y_2 - \bar{y} + \bar{z}, \ldots,\ y_m - \bar{y} + \bar{z}, \end{aligned}$$

where $\bar{z}$ is the mean of all $n + m$ observations. The adjusted samples have a common mean of $\bar{z}$ (as required by the null hypothesis). In addition, they retain the approximate shapes of the X and Y distributions.

(ii) Replicate data sets are constructed by resampling separately from the adjusted x and y samples. For each replicate data set, the value of T is computed.

(iii) If the alternative hypothesis is $\mu_x > \mu_y$, then the estimated p value is the proportion of resampled T statistics greater than or equal to t_{obs}. If the alternative hypothesis is $\mu_x < \mu_y$, then the estimated p value is the proportion of T statistics less than or equal to t_{obs}. For a two tailed test, the estimated p value is the proportion of $|T|$ statistics greater than or equal to $|t_{obs}|$.

Example: Separate families test

Let X be a continuous random variable. Consider testing

$$\begin{aligned} &H_o\text{: The PDF of the } X \text{ distribution has form } f_o(x) \quad \text{versus} \\ &H_a\text{: The PDF of the } X \text{ distribution has form } f_a(x) \end{aligned}$$

using a random sample of size n from the X distribution and a log-likelihood ratio statistic of the form

$$T = \log(\Lambda) = \log\left(L_a / L_o\right),$$

where L_o is the maximum value of the likelihood function under the null hypothesis, and L_a is the maximum value of the likelihood function under the alternative hypothesis. Large values of T favor the alternative hypothesis.

> The null family of distributions may be easier to work with than a "better" alternative family of distributions. Thus, the null family is preferred unless evidence is provided to say that the alternative family should be used.

Parametric bootstrap analysis can be used to approximate the sampling distribution of T under the null hypothesis and to estimate p values [31, p. 148].

Specifically, let $\{x_1, x_2, \ldots, x_n\}$ be the observed sample and t_{obs} be the observed value of the T statistic.

(i) Use the observed data to compute ML estimates under the null hypothesis.

(ii) Replicate data sets are sampled from the estimated model under the null hypothesis. For each replicate data set, the value of T is computed.

(iii) The estimated p value is the proportion of resampled T statistics greater than or equal to t_{obs}.

Note that the setup for a separate families test is different from the setup of the usual goodness-of-fit test. In the goodness-of-fit test, the alternative hypothesis is that the null family of models should not be used; in the separate families test, the alternative hypothesis is that a specific alternative family of models should be used. In general, a separate families test is applied when goodness-of-fit tests would not reject the use of either family.

12.5 Laboratory problems

Laboratory problems for this chapter introduce *Mathematica* commands for constructing and summarizing bootstrap approximate sampling distributions and for applying Efron's improved confidence procedure in the nonparametric setting. The problems are designed to reinforce ideas related to bootstrap analyses.

12.5.1 Laboratory: Bootstrap analysis

In the main laboratory notebook (Problems 1 to 5), you will use simulation from exponential distributions to study the performance of nonparametric bootstrap techniques applied to the reciprocal-mean estimator of λ; apply a variety of methods to study the rate of minor-to-light earthquakes in the northeastern United States and eastern Canada [35]; apply bootstrap methods to study the sampling distribution of the sample correlation using data from a study of blood fats [95], [50]; apply a variety of methods to a study on ozone levels in two cities in the northeast [25]; and apply bootstrap methods for independent random samples to two data sets: the first uses a weighted-mean statistic to estimate gravity [29], [31], and the second uses a Mann–Whitney estimator to estimate $P(X < Y)$ from a visual perception study [25].

12.5.2 Additional problem notebooks

The *delta method* is a commonly used technique for estimating bias and standard error. Problem 6 uses bootstrap and delta methods to study properties of the reciprocal-mean estimator of λ in exponential distributions. Problem 7 uses bootstrap and delta methods to study properties of the reciprocal log-mean estimator of the shape parameter in Pareto distributions.

Problem 8 is on correlation in bivariate normal distributions. A variety of techniques (including nonparametric bootstrap methods) are applied to height-weight data of athletes [28].

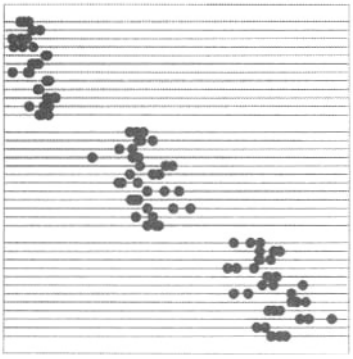
Problem 9 is on paired sample analysis. A variety of techniques (including nonparametric bootstrap methods) are applied to data from a study comparing methods for estimating the percent of calcium in animal feed [52], [90].

Problem 10 is on nonparametric bootstrap analysis of trimmed means. Data from a study of trends in catch-per-unit effort of Alaskan king crabs over a 14-year period is used [32], [58].

Problem 11 applies parametric bootstrap methods to data on earthquakes in southern California [35]. Problem 12 applies nonparametric bootstrap methods to data comparing Olympic marathon finishing times for men and women [81].

Problems 13 and 14 are on bootstrap hypothesis testing. Problem 13 applies the separate families test to data on levels of plasma retinol in women [104]. Problem 14 applies the mean difference test to data from a study comparing the spending patterns of single men and women in Hong Kong [50].

Problems 15 and 16 apply bootstrap methods to nonlinear least squares estimators. In Problem 15, nonparametric bootstrap methods are applied to study the effects of herbicides on the reproductive ability of microscopic animals [7]. In Problem 16, parametric bootstrap methods are applied to a whimsical comparison of world-class sprinters [62], [108].



Chapter 13
Multiple Sample Analysis

This chapter considers methods for comparing more than two samples, under both population and randomization sampling models. Methods introduced in the first two sections generalize methods from Chapters 10 and 11. Section 3 introduces methods for analyzing how two factors affect an outcome of interest (for example, how diet and exercise programs affect weight). Section 4 outlines the laboratory problems. References for this chapter are [17], [68].

13.1 One-way layout

This section considers methods for I samples

$$\{x_{1,1}, x_{1,2}, \ldots, x_{1,n_1}\},\ \{x_{2,1}, x_{2,2}, \ldots, x_{2,n_2}\},\ \ldots, \{x_{I,1}, x_{I,2}, \ldots, x_{I,n_I}\},$$

where n_i is the number of observations in the i^{th} sample, and $x_{i,j}$ is the j^{th} observation in the i^{th} sample. Let $N = \sum_i n_i$ be the total number of observations.

13.1.1 Example: Analysis of variance

The data are assumed to be the values of I independent random samples

$$X_{i,1}, X_{i,2}, \ldots, X_{i,n_i} \text{ for } i = 1, 2, \ldots, I$$

from normal distributions with a common unknown standard deviation σ. The samples are often referred to as *groups* and the mean of the i^{th} sample, μ_i, as the i^{th} *group mean.*

Of interest is a test of the null hypothesis that the I group means are equal versus the general alternative that at least two means differ.

Linear model

Let μ equal the expected value of the average of all N observations,

$$\mu = E\left(\frac{1}{N}\sum_{i,j} X_{i,j}\right) = \frac{1}{N}\sum_{i=1}^{I} n_i\mu_i,$$

and let $\alpha_i = \mu_i - \mu$ be the difference between the i^{th} group mean and μ. μ is called the *overall mean* and α_i is called the *differential effect* of the i^{th} group (or the i^{th} *group effect*) for $i = 1, 2, \ldots, I$. Then the general assumptions imply that $X_{i,j}$ can be written in the linear form

$$X_{i,j} = \mu + \alpha_i + \epsilon_{i,j} \text{ for } j = 1, 2, \ldots, n_i, i = 1, 2, \ldots, I,$$

where the collection $\{\epsilon_{i,j}\}$ is a random sample of size N from a normal distribution with mean 0 and standard deviation σ, and the weighted sum of the group effects is zero ($\sum_i n_i \alpha_i = 0$). Further, if the null hypothesis is true, then the I group effects are identically zero.

The random variable $X_{i,j}$ can be written as the sum of the overall mean, the differential effect of the i^{th} group, and an error term. Error terms have a common variance.

Theorem 13.1 (Parameter Estimation). *Given the assumptions and definitions above, the following are ML estimators of the parameters in the linear model:*

1. *Overall mean:*
$$\widehat{\mu} = \overline{X}_{..} = \frac{1}{N}\left(\sum_{i,j} X_{i,j}\right).$$

2. *Group effects:*
$$\widehat{\alpha_i} = \widehat{\mu_i} - \widehat{\mu} = \overline{X}_{i.} - \overline{X}_{..} = \frac{1}{n_i}\left(\sum_{j} X_{i,j}\right) - \overline{X}_{..} \text{ for each } i.$$

3. *Error terms:*
$$\widehat{\epsilon_{i,j}} = X_{i,j} - \widehat{\mu_i} = X_{i,j} - \overline{X}_{i.} \text{ for each } i, j.$$

Each is an unbiased estimator. Further, the pooled estimate of the common variance

$$S_p^2 = \frac{1}{N-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i} \widehat{\epsilon_{i,j}}^2 = \frac{1}{N-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i} \left(X_{i,j} - \overline{X}_{i.}\right)^2$$

is an unbiased estimator of σ^2.

To illustrate the computations, consider the data in Table 13.1 on lung cancer rates (per 100000 individuals per year for 1950–1969) for women in 26 counties in the northeastern United States [49, p. 93]. The data are grouped by each county's proximity to a bedrock area known as the Reading Prong. Group 1 corresponds to counties on the area, group 2 to counties on the fringe of the area, and group 3 to "control" counties (counties that are nearby but not on the prong). Bedrock areas such as the Reading Prong are suspected of emitting radon gas, a potential carcinogen.

Sample sizes, means, and standard deviations are given in the table. The mean of all 26 observations is 5.577, the estimated group effects are 1.423, −0.105, and −0.600, respectively, and the pooled estimate of the common variance is 1.319.

Table 13.1. *Lung cancer rates data.*

i	n_i	$x_{i,j}$	$\widehat{\mu_i} = \overline{x}_{i.}$	s_i
1	6	6.0, 10.5, 6.7, 6.0, 6.1, 6.7	7.000	1.746
2	7	5.2, 5.6, 5.8, 4.5, 5.5, 5.4, 6.3	5.471	0.553
3	13	6.3, 4.3, 4.0, 5.9, 4.7, 4.8, 5.8, 5.4, 5.2, 3.6, 4.3, 3.5, 6.9	4.977	1.051

Sources of variation

The formal analysis of the null hypothesis of equality of means is based on writing the sum of squared deviations of the observations from the estimated overall mean (known as the *total* sum of squares),

$$SS_t = \sum_{i,j} \left(X_{i,j} - \overline{X}_{..}\right)^2,$$

as the sum of squared deviations of the observations from the appropriate estimated group means (known as the *error* sum of squares),

$$SS_e = \sum_{i,j} \left(X_{i,j} - \overline{X}_{i.}\right)^2,$$

plus the weighted sum of squared deviations of estimated group means from the estimated overall mean (known as the *group* sum of squares),

$$SS_g = \sum_{i} n_i \left(\overline{X}_{i.} - \overline{X}_{..}\right)^2.$$

Error and group mean squares

The *error* mean square, MS_e, is defined as follows:

$$MS_e = \frac{1}{N-I} SS_e = \frac{1}{N-I} \sum_{i,j} \left(X_{i,j} - \overline{X}_{i.}\right)^2.$$

MS_e is equal to the pooled estimate of the common variance. Theorem 6.1 can be used to demonstrate that $(N-I)MS_e/\sigma^2$ is a chi-square random variable with $(N-I)$ degrees of freedom.

The *group* mean square, MS_g, is defined as follows:

$$MS_g = \frac{1}{I-1} SS_g = \frac{1}{I-1} \sum_{i} n_i \left(\overline{X}_{i.} - \overline{X}_{..}\right)^2.$$

Properties of expectation can be used to demonstrate that

$$E\left(MS_g\right) = \sigma^2 + \frac{1}{I-1} \sum_{i} n_i(\mu_i - \mu)^2 = \sigma^2 + \frac{1}{I-1} \sum_{i} n_i \alpha_i^2.$$

Table 13.2. *Analysis of variance table for the lung cancer rates data.*

Source	df	SS	MS	F	p value
Group	2	16.91	8.45	6.41	0.006
Error	23	30.34	1.32		
Total	25	47.25			

If the null hypothesis that the group means are equal is true, then the expected value of MS_g is σ^2; otherwise, values of MS_g will tend to be larger than σ^2.

The following theorem relates the error and group mean squares.

Theorem 13.2 (Distribution Theorem). *Under the general assumptions of this section and if the null hypothesis of equality of group means is true, then the ratio* $F = MS_g/MS_e$ *has an f ratio distribution with* $(I-1)$ *and* $(N-I)$ *degrees of freedom.*

Test of equality of means: Observed significance level

Large values of $F = MS_g/MS_e$ support the alternative hypothesis that some means differ. For an observed ratio, f_{obs}, the p value is $P(F \geq f_{\text{obs}})$.

For example, assume the data displayed in Table 13.1 are the values of independent random samples from normal distributions with a common variance. Table 13.2 shows the results of the test of equality of group means, organized into an *analysis of variance table*. The observed ratio of the group mean square to the error mean square is 6.41. The observed significance level, based on the f ratio distribution with 2 and 23 degrees of freedom, is 0.006. Since the p value is small, there is evidence that at least two means differ.

Informal model checking

The ratio of the maximum to the minimum sample standard deviation can be used to check the assumption of equality of variances. The usual rule of thumb is that ratios of 2 or less are fine. With small sample sizes, like the samples in the lung cancer rates data, ratios up to about 4 are considered reasonable.

Normal probability plots of the estimated errors (or *residuals*), $x_{i,j} - \bar{x}_{i\cdot}$, can be used to check if the error distribution is approximately normal.

Bonferroni method of multiple comparisons

If the null hypothesis of equality of means is rejected, then it is natural to try to determine which means differ. In the Bonferroni method of multiple comparisons, a total of $m = \binom{I}{2}$ two sided tests of equality of means of the form

$$\mu_i = \mu_k \text{ versus } \mu_i \neq \mu_k \text{ for } i < k$$

are conducted using pooled t statistics of the form

$$T_{i,k} = \frac{\overline{X}_{i\cdot} - \overline{X}_{k\cdot}}{\sqrt{S_p^2\left(\frac{1}{n_i} + \frac{1}{n_k}\right)}},$$

where S_p^2 is the pooled estimate of the common variance. In each case, $T_{i,k}$ has a Student t distribution with $(N - I)$ degrees of freedom under the null hypothesis.

If the significance level for each test is α/m, then the overall type I error for all m tests is at most α. That is, if all means are actually equal, then the probability of rejecting at least one of the m null hypotheses is at most α.

To demonstrate that the overall type I error is at most α, let $R_{i,k}$ be the event that the null hypothesis $\mu_i = \mu_k$ is rejected. Then

$$\begin{aligned} P(\text{at least 1 test is rejected}) &= P(R_{1,2} \cup R_{1,3} \cup \cdots \cup R_{I-1,I}) \\ &\leq P(R_{1,2}) + P(R_{1,3}) + \cdots + P(R_{I-1,I}) \\ &= m\,(\alpha/m) = \alpha. \end{aligned}$$

(The probability of the union is less than or equal to the sum of the probabilities.)

Continuing with the lung cancer rates example, a Bonferroni analysis with an overall 5% significance level uses decision rules of the form

$$\text{Reject } \mu_i = \mu_k \text{ in favor of } \mu_i \neq \mu_k \text{ when } \left|T_{i,k}\right| \leq t_{23}(.025/3) = 2.582$$

for $(i, k) = (1,2), (1,3), (2,3)$. Using these decision rules, only the hypothesis that $\mu_1 = \mu_3$ is rejected. In fact, there is evidence that $\mu_1 > \mu_3$, suggesting a link between radon and lung cancer.

13.1.2 Example: Kruskal–Wallis test

In the 1950's, Kruskal and Wallis developed a nonparametric version of the analysis of variance test appropriate in one of the following situations:

1. *Population model.* The data are the values of I independent random samples. The null hypothesis is that the distributions from which the data were drawn are equal.

2. *Randomization model.* The data are measurements on N individuals in distinguishable groups of sizes $n_1, n_2, \ldots, n_I$. The null hypothesis is that observed differences in the groups are due to chance alone.

The form of the test statistic is similar to the form of the group sum of squares.

Test statistic: No tied observations

Let $R_{i,j}$ be the rank of observation $x_{i,j}$ in the combined sample, and let $\overline{R}_{i\cdot}$ be the average rank of observations in the i^{th} sample. The Kruskal–Wallis statistic, K, is

defined as follows:

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{I} n_i \left(\overline{R}_{i\cdot} - \frac{N+1}{2} \right)^2 .$$

The average of all N ranks is $(N+1)/2$. The statistic is a weighted sum of squared deviations of average group ranks from the overall average rank.

The sampling distribution of K is obtained by computing its value for each partition of the N ranks into distinguishable groups of sizes $n_1, n_2, \ldots, n_I$. There are a total of $\binom{N}{n_1, n_2, \ldots, n_I}$ partitions to consider. The following theorem gives a large sample approximation to the distribution.

Theorem 13.3 (Kruskal–Wallis Statistic). *If there are no ties in the observations, then under the null hypothesis of randomness and when N is large, the distribution of K is approximately chi-square with $(I-1)$ degrees of freedom.*

Test statistic when some observations are tied

Midranks replace ranks when there are ties in the data. The Kruskal–Wallis statistic becomes

$$K = \sum_{i=1}^{I} w_i \left(\overline{R}_{i\cdot} - \frac{N+1}{2} \right)^2 ,$$

where the weights, w_i, in the weighted sum are chosen to make the approximate sampling distribution of K under the null hypothesis as close to the chi-square distribution as possible.

Most computer programs automatically use the appropriate weights.

Test of randomness: Observed significance level

Large values of K support the alternative hypothesis that the values in at least one sample tend to be larger or smaller than those in another. The observed significance level is $P(K \geq k_{\text{obs}})$, where k_{obs} is the observed value of the Kruskal–Wallis statistic. In most practical situations, the chi-square approximation to the sampling distribution of K is used to compute p values.

For example, Table 13.3 gives the midranks and average group ranks for the lung cancer rates data from Table 13.1. For these data, the observed value of the test statistic is 9.62. The observed significance level, based on the chi-square distribution with 2 degrees of freedom, is 0.00817. Since the p value is small, there is evidence that differences in the samples are not due to chance alone.

Multiple comparisons

If the Kruskal–Wallis test suggests that differences are not due to chance alone, then $m = \binom{I}{2}$ two sided Wilcoxon rank sum tests can be conducted to determine which samples differ. If each test is conducted at the α/m level, then the overall type I error will be at most α.

Table 13.3. *Midranks and average group ranks for the lung cancer rates data.*

i	$r_{i,j}$	$\bar{r}_{i.}$
1	18.5, 26.0, 23.5, 18.5, 20.0, 23.5	21.67
2	9.5, 14.0, 15.5, 6.0, 13.0, 11.5, 21.5	13.00
3	21.5, 4.5, 3.0, 17.0, 7.0, 8.0, 15.5, 11.5, 9.5, 2.0, 4.5, 1.0, 25.0	10.00

For the lung cancer rates data and $\alpha = 0.05$, there are significant differences between the first and second groups and between the first and third groups.

13.1.3 Example: Permutation f test

A permutation version of the analysis of variance f test is appropriate in one of the following situations:

1. *Population model.* The observed data are the values of independent random samples from distributions differing in mean only. The null hypothesis is that the distributions are equal (equivalently, all means are equal) versus the general alternative that at least two means differ.

2. *Randomization model.* The data are measurements taken on N individuals in distinguishable groups of sizes $n_1, n_2, \ldots, n_I$. The null hypothesis is that observed differences in means are due to chance alone versus the alternative that at least one sample has values that tend to be larger or smaller (but not more variable) than the values in another sample.

The sampling distribution of F is obtained by computing its value for each partition of the N observations into distinguishable groups of sizes $n_1, n_2, \ldots, n_I$. There are a total of $\binom{N}{n_1, n_2, \ldots, n_I}$ partitions to consider. Since this number can be quite large, Monte Carlo analysis is generally used to estimate a p value.

For example, consider testing the null hypothesis of randomness using the 5% significance level and the lung cancer rates data (Table 13.1). In a Monte Carlo analysis using 5000 random partitions (including the observed partition of the 26 rates), 0.3% (15/5000) of F values were greater than or equal to $f_{\text{obs}} = 6.41$. Thus, there is evidence that observed differences in mean rates are not due to chance alone.

Comparison of tests

In situations where both the analysis of variance f test and the permutation f test are appropriate, the analysis of variance f test is preferred. However, it is interesting to note that the tests give similar results.

In situations where both the permutation f test and the Kruskal–Wallis test are appropriate, if the samples are highly skewed or have extreme outliers, then the Kruskal–Wallis test is preferred; otherwise, the permutation f test is preferred.

In practice, the Kruskal–Wallis test is used almost exclusively since it is easy to implement.

13.2 Blocked design

This section considers methods for matrices of observations of the form

$$\{\{x_{1,1}, x_{1,2}, \ldots, x_{1,J}\}, \{x_{2,1}, x_{2,2}, \ldots, x_{2,J}\}, \ldots, \{x_{I,1}, x_{I,2}, \ldots, x_{I,J}\}\},$$

where the i^{th} *sample* is the list $\{x_{i,j} : j = 1, 2, \ldots, J\}$, and the j^{th} *block* is the list $\{x_{i,j} : i = 1, 2, \ldots, I\}$. Let $N = IJ$ be the total number of observations.

Blocked samples arise, for example, in *randomized block experiments*:

(i) There are I treatments under study and J blocks of I subjects each, where subjects in a given block are matched on important factors (such as age, sex, and general health measures).

(ii) The subjects within each block are randomly matched to the I treatments. The subject receiving the i^{th} treatment is in the i^{th} experimental *group*.

Blocked designs generalize paired designs. Researchers use blocked designs to reduce the variability of the results since individuals within each block are expected to respond similarly to treatment, while individuals in different blocks are expected to respond differently to treatment.

13.2.1 Example: Analysis of variance

The data are assumed to be the values of N independent normal random variables satisfying the linear model

$$X_{i,j} = \mu + \alpha_i + \beta_j + \epsilon_{i,j} \text{ for } i = 1, 2, \ldots, I, \; j = 1, 2, \ldots, J,$$

where

1. μ is the *overall mean*: $\mu = \frac{1}{N} \sum_{i,j} \mu_{i,j}$, where $\mu_{i,j} = E(X_{i,j})$;
2. α_i is the *differential effect* of the i^{th} group: $\alpha_i = \mu_{i\cdot} - \mu = \frac{1}{J} \sum_j \mu_{i,j} - \mu$;
3. β_j is the *differential effect* of the j^{th} block: $\beta_j = \mu_{\cdot j} - \mu = \frac{1}{I} \sum_i \mu_{i,j} - \mu$;
4. the collection $\{\epsilon_{i,j}\}$ is a random sample of size N from a normal distribution with mean 0 and standard deviation σ.

The random variable $X_{i,j}$ can be written as the sum of the overall mean, the differential effect of the i^{th} group, the differential effect of the j^{th} block, and an error term. The errors have a common variance.

The sum of the group effects is zero ($\sum_i \alpha_i = 0$), and the sum of the block effects is zero ($\sum_j \beta_j = 0$). The null hypothesis of primary interest is that the group effects are identically zero. The null hypothesis that the block effects are identically zero can also be tested.

Table 13.4. *Penicillin manufacturing data.*

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	
$i=1$	89	84	81	87	79	$\overline{x}_{1\cdot}=84$
$i=2$	88	77	87	92	81	$\overline{x}_{2\cdot}=85$
$i=3$	97	92	87	89	80	$\overline{x}_{3\cdot}=89$
$i=4$	94	79	85	84	88	$\overline{x}_{4\cdot}=86$
	$\overline{x}_{\cdot 1}=92$	$\overline{x}_{\cdot 2}=83$	$\overline{x}_{\cdot 3}=85$	$\overline{x}_{\cdot 4}=88$	$\overline{x}_{\cdot 5}=82$	

Theorem 13.4 (Parameter Estimation). *Given the assumptions and definitions above, the following are ML estimators of the parameters in the linear model:*

1. *Overall mean:*
$$\widehat{\mu}=\overline{X}_{\cdot\cdot}=\frac{1}{N}\left(\sum_{i,j}X_{i,j}\right).$$

2. *Group effects:*
$$\widehat{\alpha_i}=\widehat{\mu_{i\cdot}}-\widehat{\mu}=\overline{X}_{i\cdot}-\overline{X}_{\cdot\cdot}=\frac{1}{J}\left(\sum_{j}X_{i,j}\right)-\overline{X}_{\cdot\cdot}\ \textit{for all } i.$$

3. *Block effects:*
$$\widehat{\beta_j}=\widehat{\mu_{\cdot j}}-\widehat{\mu}=\overline{X}_{\cdot j}-\overline{X}_{\cdot\cdot}=\frac{1}{I}\left(\sum_{i}X_{i,j}\right)-\overline{X}_{\cdot\cdot}\ \textit{for all } j.$$

4. *Error terms:*
$$\widehat{\epsilon_{i,j}}=X_{i,j}-\widehat{\mu_{i,j}}=X_{i,j}-(\widehat{\mu}+\widehat{\alpha_i}+\widehat{\beta_j})=X_{i,j}-\overline{X}_{i\cdot}-\overline{X}_{\cdot j}+\overline{X}_{\cdot\cdot}\ \textit{for all } i,j.$$

Each is an unbiased estimator. Further, the pooled estimate of the common variance
$$S_p^2=\frac{1}{(I-1)(J-1)}\sum_{i,j}\widehat{\epsilon_{i,j}}^2=\frac{1}{(I-1)(J-1)}\sum_{i,j}\left(X_{i,j}-\overline{X}_{i\cdot}-\overline{X}_{\cdot j}+\overline{X}_{\cdot\cdot}\right)^2$$
is an unbiased estimator of σ^2.

To illustrate the computations, consider the data in Table 13.4 on the amount of penicillin produced using four different manufacturing processes (the groups) and five different blends of raw materials (the blocks) [17, p. 209]. Interest focuses on potential differences in manufacturing processes. A block design was used because different blends of raw materials could produce different results. For each blend of raw materials, the order in which the manufacturing processes were tested was randomized.

Sample group and block means are shown in the table. For these data, the overall mean is 86 (units of penicillin), the differential effects of the four manufacturing processes are $-2, -1, 3, 0$, the differential effects of the five blends are $6, -3, -1, 2, -4$, and the pooled estimate of the common variance is 18.83.

Sources of variation

Formal analyses are based on writing the sum of squared deviations of the observations from the estimated overall mean (known as the *total* sum of squares),

$$SS_t = \sum_{i,j} \left(X_{i,j} - \overline{X}_{..}\right)^2,$$

as the sum of squared deviations of the observations from their estimated means (known as the *error* sum of squares),

$$SS_e = \sum_{i,j} \left(X_{i,j} - \overline{X}_{i.} - \overline{X}_{.j} + \overline{X}_{..}\right)^2,$$

plus the weighted sum of squared deviations of estimated group means from the estimated overall mean (known as the *group* sum of squares),

$$SS_g = \sum_{i} J \left(\overline{X}_{i.} - \overline{X}_{..}\right)^2,$$

plus the weighted sum of squared deviations of estimated block means from the estimated overall mean (known as the *block* sum of squares),

$$SS_b = \sum_{j} I \left(\overline{X}_{.j} - \overline{X}_{..}\right)^2.$$

Error, group, and block mean squares

The *error* mean square, MS_e, is defined as follows:

$$MS_e = \frac{1}{(I-1)(J-1)} SS_e = \frac{1}{(I-1)(J-1)} \sum_{i,j} \left(X_{i,j} - \overline{X}_{i.} - \overline{X}_{.j} + \overline{X}_{..}\right)^2.$$

MS_e is equal to the pooled estimate of the common variance. Theorem 6.1 can be used to demonstrate that $(I-1)(J-1)MS_e/\sigma^2$ is a chi-square random variable with $(I-1)(J-1)$ degrees of freedom.

The *group* mean square, MS_g, is defined as follows:

$$MS_g = \frac{1}{I-1} SS_g = \frac{1}{I-1} \sum_{i} J \left(\overline{X}_{i.} - \overline{X}_{..}\right)^2.$$

Properties of expectation can be used to demonstrate that

$$E\left(MS_g\right) = \sigma^2 + \frac{J}{I-1} \sum_{i} (\mu_{i.} - \mu)^2 = \sigma^2 + \frac{J}{I-1} \sum_{i} \alpha_i^2.$$

Table 13.5. *Analysis of variance table for the penicillin manufacturing data.*

Source	df	SS	MS	F	p value
Group	3	70.0	23.33	1.24	0.339
Block	4	264.0	66.00	3.50	0.041
Error	12	226.0	18.83		
Total	19	560.0			

If the null hypothesis that the group effects are identically zero is true, then the expected value of MS_g is the common variance σ^2; otherwise, values of MS_g will tend to be larger than σ^2.

Similarly, the *block* mean square, MS_b, is defined as follows:

$$MS_b = \frac{1}{J-1} SS_b = \frac{1}{J-1} \sum_j I \left(\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot}\right)^2 .$$

Properties of expectation can be used to demonstrate that

$$E\left(MS_b\right) = \sigma^2 + \frac{I}{J-1} \sum_j \left(\mu_{\cdot j} - \mu\right)^2 = \sigma^2 + \frac{I}{J-1} \sum_j \beta_j^2 .$$

If the null hypothesis that the block effects are identically zero is true, then the expected value of MS_b is the common variance σ^2; otherwise, values of MS_b will tend to be larger than σ^2.

The following theorem relates the error, group, and block mean squares.

Theorem 13.5 (Distribution Theorem). *Under the general assumptions of this section, the following hold:*

(i) *If the null hypothesis that the group effects are identically zero is true, then the ratio $F = MS_g/MS_e$ has an f ratio distribution with $(I-1)$ and $(I-1)(J-1)$ degrees of freedom.*

(ii) *If the null hypothesis that the block effects are identically zero is true, then the ratio $F = MS_b/MS_e$ has an f ratio distribution with $(J-1)$ and $(I-1)(J-1)$ degrees of freedom.*

Observed significance levels

Large values of $F = MS_g/MS_e$ support the alternative hypothesis that some group effects are not zero, and large values of $F = MS_b/MS_e$ support the null hypothesis that some block effects are not zero.

For example, assume the data displayed in Table 13.4 are values of independent random variables satisfying the assumptions of this section. Table 13.5 shows the results of the two significance tests, organized into an *analysis of variance table*.

(i) The observed ratio of the group mean square to the error mean square is 1.24; the observed significance level, based on the f ratio distribution with 3 and 12 degrees of freedom, is 0.339.

(ii) The observed ratio of the block mean square to the error mean square is 3.50; the observed significance level, based on the f ratio distribution with 4 and 12 degrees of freedom, is 0.041.

The results of the analysis of variance suggest that once differences in raw material blends have been taken into account, mean levels of production for the four manufacturing processes are equal.

In the analysis of variance for blocked designs, potential group and block effects are separated. The effects can then be tested separately.

Bonferroni method of multiple comparisons

If the null hypothesis that the group effects are identically zero is rejected, then it is natural to try to determine which groups differ. In the Bonferroni method of multiple comparisons, a total of $m = \binom{I}{2}$ two sided tests of equality of group means of the form

$$\mu_{i\cdot} = \mu_{k\cdot} \text{ versus } \mu_{i\cdot} \neq \mu_{k\cdot} \text{ for } i < k$$

are conducted using paired t statistics of the form

$$T_{i,k} = \frac{\overline{X}_{i\cdot} - \overline{X}_{k\cdot}}{\sqrt{S_p^2/J}},$$

where S_p^2 is the pooled estimate of the common variance. In each case, $T_{i,k}$ has a Student t distribution with $(I-1)(J-1)$ degrees of freedom under the null hypothesis.

If the significance level for each test is α/m, then the overall type I error for all m tests is at most α. That is, if all means are actually equal, then the probability of rejecting at least one of the m null hypotheses is at most α.

Tests of the equality of group means in the blocked design setting are equivalent to tests of the equality of group effects. The Bonferroni analysis in the blocked design setting uses paired sample methods, while the analysis in the one-way layout setting uses two sample methods.

13.2.2 Example: Friedman test

In the 1930's, Friedman developed a nonparametric test for comparing groups in blocked designs appropriate in the following situations:

1. *Population model.* The data are the values of $N = IJ$ independent random variables. The null hypothesis is that for each j, the distributions of $X_{1,j}, X_{2,j}, \ldots, X_{I,j}$ are equal.

2. *Randomization model.* The blocked data are measurements on J individuals (or J blocks of I individuals each). The null hypothesis of interest is that observed differences in measurements within each block are due to chance alone.

The form of the test statistic is similar to the form of the group sum of squares.

Test statistic: No ties within blocks

Let $R_{i,j}$ be the rank of observation $x_{i,j}$ in the j^{th} block, and let $\overline{R}_{i\cdot}$ be the average rank of observations in the i^{th} sample. The Friedman statistic, Q, is defined as follows:

$$Q = \frac{12J}{I(I+1)} \sum_{i=1}^{I} \left(\overline{R}_{i\cdot} - \frac{I+1}{2} \right)^2 .$$

The average of all N ranks is $(I+1)/2$. (There are J copies of each integer between 1 and I.) The statistic is a weighted sum of squared deviations of average group ranks from the overall average rank.

The sampling distribution of Q is obtained by computing its value for each matching of the I ranks in each block to the I treatments. There are a total of $(I!)^J$ matchings to consider. The following theorem gives a large sample approximation to the distribution.

Theorem 13.6 (Friedman Statistic). *If there are no ties within blocks, then under the null hypothesis of randomness and when N is large, the distribution of Q is approximately chi-square with $(I-1)$ degrees of freedom.*

Test statistic when there are ties within blocks

Midranks replace ranks when there are ties in a given block. The Friedman statistic becomes

$$Q = \sum_{i=1}^{I} w_i \left(\overline{R}_{i\cdot} - \frac{I+1}{2} \right)^2 ,$$

where the weights, w_i, in the weighted sum are chosen to make the approximate sampling distribution of Q under the null hypothesis as close to the chi-square distribution as possible.

Most computer programs automatically use the appropriate weights.

Test of randomness: Observed significance level

Large values of Q support the alternative hypothesis that the values in at least one group tend to be larger or smaller than those in another. The observed significance level is $P(Q \geq q_{\text{obs}})$, where q_{obs} is the observed value of the Friedman statistic. In most practical situations, the chi-square approximation to the sampling distribution of Q is used to compute p values.

Table 13.6. *Midranks for the penicillin manufacturing data.*

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$\bar{r}_{i.}$
$i=1$	2.0	3.0	1.0	2.0	1.0	1.8
$i=2$	1.0	1.0	3.5	4.0	3.0	2.5
$i=3$	4.0	4.0	3.5	3.0	2.0	3.3
$i=4$	3.0	2.0	2.0	1.0	4.0	2.4

For example, Table 13.6 gives the midranks and average group ranks for the penicillin manufacturing data from Table 13.4. For these data, the observed value of the test statistic is 3.49. The observed significance level, based on the chi-square distribution with 3 degrees of freedom, is 0.322. Since the p value is large, the results suggest that once differences in raw material blends are taken into account, observed differences in the manufacturing processes are due to chance alone.

Multiple comparisons

If the Friedman test suggests that group differences are not due to chance alone, then $m = \binom{I}{2}$ two sided Wilcoxon signed rank tests can be conducted to determine which groups differ. If each test is conducted at the α/m level, then the overall type I error will be at most α.

13.3 Balanced two-way layout

This section considers methods for matrices of samples of the form

$$\{\{s_{1,1}, s_{1,2}, \ldots, s_{1,J}\}, \{s_{2,1}, s_{2,2}, \ldots, s_{2,J}\}, \ldots, \{s_{I,1}, x_{I,2}, \ldots, s_{I,J}\}\},$$

where each $s_{i,j}$ is a list of K observations:

$$s_{i,j} = \{x_{i,j,1}, x_{i,j,2}, \ldots, x_{i,j,K}\}.$$

i is used to index the I levels of the first factor of interest (the *row* factor), and j is used to index the J levels of the second factor of interest (the *column* factor). The total number of observations is $N = IJK$.

There are a total of IJ samples (or *groups*). The layout is called *balanced* because the sample sizes are equal.

13.3.1 Example: Analysis of variance

The data are assumed to be the values of IJ independent random samples

$$X_{i,j,1}, X_{i,j,2}, \ldots, X_{i,j,K} \text{ for } i = 1, 2, \ldots, I, j = 1, 2, \ldots, J$$

from normal distributions with a common unknown standard deviation σ. Let $\mu_{i,j}$ be the mean of observations in the (i, j) sample: $\mu_{i,j} = E(X_{i,j,k})$ for all k.

Linear model

The general assumptions imply that $X_{i,j,k}$ can be written in the linear form

$$X_{i,j,k} = \mu + \alpha_i + \beta_j + \delta_{i,j} + \epsilon_{i,j,k} \text{ for all } i, j, k,$$

where

1. μ is the *overall mean*: $\mu = \left(\sum_{i,j,k} \mu_{i,j}\right)/N = \left(\sum_{i,j} \mu_{i,j}\right)/(IJ)$;
2. α_i is the *differential effect* of the i^{th} level of the first factor,
$$\alpha_i = \mu_{i\cdot} - \mu = \frac{1}{JK}\sum_{j,k} \mu_{i,j} - \mu = \frac{1}{J}\sum_{j} \mu_{i,j} - \mu;$$
3. β_j is the *differential effect* of the j^{th} level of the second factor,
$$\beta_j = \mu_{\cdot j} - \mu = \frac{1}{IK}\sum_{i,k} \mu_{i,j} - \mu = \frac{1}{I}\sum_{i} \mu_{i,j} - \mu;$$
4. $\delta_{i,j}$ is the *interaction* between the i^{th} level of the first factor and the j^{th} level of the second factor,
$$\delta_{i,j} = \mu_{i,j} - \left(\mu + \alpha_i + \beta_j\right);$$
5. the collection $\{\epsilon_{i,j,k}\}$ is a random sample of size N from a normal distribution with mean 0 and standard deviation σ.

The random variable $X_{i,j,k}$ can be written as the sum of the overall mean, the differential effects of each factor, the interaction between factors, and an error term. The errors have a common variance.

The sum of the differential effects of each factor is zero ($\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$), and the sum of the interaction terms is zero at each fixed level of the first and second factors ($\sum_j \delta_{i,j} = 0$ for all i, $\sum_i \delta_{i,j} = 0$ for all j). Null hypotheses of interest are as follows: the differential effects of the first factor are identically zero; the differential effects of the second factor are identically zero; the interaction terms are identically zero.

Note that if the $\delta_{i,j} = 0$ for all i, j, then the means satisfy the simple additive model: $\mu_{i,j} = \mu + \alpha_i + \beta_j$.

Theorem 13.7 (Parameter Estimation). *Given the assumptions and definitions above, the following are ML estimators of the parameters in the linear model:*

1. *Overall mean:*

$$\widehat{\mu} = \overline{X}_{...} = \frac{1}{N}\left(\sum_{i,j,k} X_{i,j,k}\right).$$

2. *Row effects:*

$$\widehat{\alpha_i} = \widehat{\mu_{i.}} - \widehat{\mu} = \frac{1}{JK}\left(\sum_{j,k} X_{i,j,k}\right) - \widehat{\mu} = \overline{X}_{i..} - \overline{X}_{...} \quad \textit{for all } i.$$

3. *Column effects:*

$$\widehat{\beta_j} = \widehat{\mu_{.j}} - \widehat{\mu} = \frac{1}{IK}\left(\sum_{i,k} X_{i,j,k}\right) - \widehat{\mu} = \overline{X}_{.j.} - \overline{X}_{...} \quad \textit{for all } j.$$

4. *Interactions (row-by-column effects):*

$$\begin{aligned}\widehat{\delta_{i,j}} &= \widehat{\mu_{i,j}} - (\widehat{\mu} + \widehat{\alpha_i} + \widehat{\beta_j}) \\ &= \frac{1}{K}\left(\sum_{k} X_{i,j,k}\right) - (\widehat{\mu} + \widehat{\alpha_i} + \widehat{\beta_j}) \\ &= \overline{X}_{i,j.} - \overline{X}_{i..} - \overline{X}_{.j.} + \overline{X}_{...} \quad \textit{for all } i, j.\end{aligned}$$

5. *Error terms:*

$$\widehat{\epsilon_{i,j,k}} = X_{i,j,k} - \widehat{\mu_{i,j}} = X_{i,j,k} - \overline{X}_{i,j.} \quad \textit{for all } i, j, k.$$

Each is an unbiased estimator. Further, the pooled estimate of the common variance

$$S_p^2 = \frac{1}{N - IJ}\sum_{i,j,k} \widehat{\epsilon_{i,j,k}}^2 = \frac{1}{IJ(K-1)}\sum_{i,j,k}\left(X_{i,j,k} - \overline{X}_{i,j.}\right)^2$$

is an unbiased estimator of σ^2.

To illustrate the computations, consider the data in Table 13.7 on the times (in 1/100 minutes) needed to drill a hole through 5 feet of rock. Drilling was started at three different depths (row factor), using two different methods to drill each hole (column factor). Three holes were drilled at each combination of depth and drilling method [50, p. 204]. Row levels correspond to starting depths of 10, 30, and 50 feet; column levels correspond to dry drilling (where compressed air is used to flush cuttings) and wet drilling (where water is used to flush cuttings).

For these data, the overall mean is 773.5, and the pooled estimate of the common variance is 12612.6. The bottom table shows the estimated interactions, as well as the estimates of the differential effects of each factor.

Table 13.7. *Mining data (top table) and parameter estimates (bottom table).*

	$j=1$	$j=2$
$i=1$	816, 813, 771	855, 662, 507
$i=2$	827, 1022, 975	795, 634, 742
$i=3$	989, 814, 727	772, 599, 603

	$j=1$	$j=2$	
$i=1$	-25.39	25.39	$\widehat{\alpha_1} = -36.17$
$i=2$	20.78	-20.78	$\widehat{\alpha_2} = 59.00$
$i=3$	4.61	-4.61	$\widehat{\alpha_3} = -22.83$
	$\widehat{\beta_1} = 88.06$	$\widehat{\beta_2} = -88.06$	

Sources of variation

Formal analyses are based on writing the sum of squared deviations of the observations from the estimated overall mean (known as the *total* sum of squares),

$$SS_t = \sum_{i,j,k} \left(X_{i,j,k} - \overline{X}_{...}\right)^2,$$

as the sum of squared deviations of the observations from their estimated group means (known as the *error* sum of squares),

$$SS_e = \sum_{i,j,k} \left(X_{i,j,k} - \overline{X}_{ij.}\right)^2,$$

plus the weighted sum of squared deviations of estimated row means from the estimated overall mean (known as the *row* sum of squares),

$$SS_r = \sum_{i} JK \left(\overline{X}_{i..} - \overline{X}_{...}\right)^2,$$

plus the weighted sum of squared deviations of estimated column means from the estimated overall mean (known as the *column* sum of squares),

$$SS_c = \sum_{j} IK \left(\overline{X}_{.j.} - \overline{X}_{...}\right)^2,$$

plus the weighted sum of squared deviations of estimated group means from estimated means under the simple additive model (known as the *interaction* or *row-by-column* sum of squares),

$$SS_{r\times c} = \sum_{i,j} K \left(\overline{X}_{ij.} - \overline{X}_{i..} - \overline{X}_{.j.} + \overline{X}_{...}\right)^2.$$

Error, row, column, and interaction mean squares

The *error* mean square, MS_e, is defined as follows:

$$MS_e = \frac{1}{IJ(K-1)} SS_e = \frac{1}{IJ(K-1)} \sum_{i,j,k} \left(X_{i,j,k} - \overline{X}_{ij.}\right)^2.$$

MS_e is equal to the pooled estimate of the common variance. Theorem 6.1 can be used to demonstrate that $IJ(K-1)MS_e/\sigma^2$ is a chi-square random variable with $IJ(K-1)$ degrees of freedom.

The *row* mean square, MS_r, is defined as follows:

$$MS_r = \frac{1}{I-1} SS_r = \frac{1}{I-1} \sum_{i} JK \left(\overline{X}_{i..} - \overline{X}_{...}\right)^2.$$

Properties of expectation can be used to demonstrate that

$$E\left(MS_r\right) = \sigma^2 + \frac{JK}{I-1} \sum_{i} \left(\mu_{i.} - \mu\right)^2 = \sigma^2 + \frac{JK}{I-1} \sum_{i} \alpha_i^2.$$

If the null hypothesis that the row effects are identically zero is true, then the expected value of MS_r is the common variance σ^2; otherwise, values of MS_r will tend to be larger than σ^2.

The *column* mean square, MS_c, is defined as follows:

$$MS_c = \frac{1}{J-1} SS_c = \frac{1}{J-1} \sum_{j} IK \left(\overline{X}_{.j.} - \overline{X}_{...}\right)^2.$$

Properties of expectation can be used to demonstrate that

$$E\left(MS_c\right) = \sigma^2 + \frac{IK}{J-1} \sum_{j} \left(\mu_{.j} - \mu\right)^2 = \sigma^2 + \frac{IK}{J-1} \sum_{j} \beta_j^2.$$

If the null hypothesis that the column effects are identically zero is true, then the expected value of MS_c is the common variance σ^2; otherwise, values of MS_c will tend to be larger than σ^2.

Finally, the *interaction* or *row-by-column* mean square, $MS_{r\times c}$, is defined as follows:

$$MS_{r\times c} = \frac{1}{(I-1)(J-1)} SS_{r\times c} = \frac{1}{(I-1)(J-1)} \sum_{i,j} K \left(\overline{X}_{ij.} - \overline{X}_{i..} - \overline{X}_{.j.} + \overline{X}_{...}\right)^2.$$

Properties of expectation can be used to demonstrate that

$$\begin{aligned} E\left(MS_{r\times c}\right) &= \sigma^2 + \frac{K}{(I-1)(J-1)} \sum_{i,j} \left(\mu_{i,j} - \left(\mu + \alpha_i + \beta_j\right)\right)^2 \\ &= \sigma^2 + \frac{K}{(I-1)(J-1)} \sum_{i,j} \delta_{i,j}^2. \end{aligned}$$

Table 13.8. *Analysis of variance table for the mining data.*

Source	df	SS	MS	F	p value
Row	2	31862.3	15931.2	1.263	0.318
Col	1	139568.0	139568.0	11.066	0.006
Row-by-Col	2	6585.4	3292.7	0.261	0.774
Error	12	151351.0	12612.6		
Total	17	329367.0			

If the null hypothesis that the interaction effects are identically zero is true, then the expected value of $MS_{r\times c}$ is the common variance σ^2; otherwise, values of $MS_{r\times c}$ will tend to be larger than σ^2.

The following theorem relates the mean square random variables.

Theorem 13.8 (Distribution Theorem). *Under the general assumptions of this section, the following hold:*

(i) *If the null hypothesis that the row effects are identically zero is true, then the ratio* $F = MS_r/MS_e$ *has an f ratio distribution with* $(I-1)$ *and* $IJ(K-1)$ *degrees of freedom.*

(ii) *If the null hypothesis that the column effects are identically zero is true, then the ratio* $F = MS_c/MS_e$ *has an f ratio distribution with* $(J-1)$ *and* $IJ(K-1)$ *degrees of freedom.*

(iii) *If the null hypothesis that the interaction effects are identically zero is true, then the ratio* $F = MS_{r\times c}/MS_e$ *has an f ratio distribution with* $(I-1)(J-1)$ *and* $IJ(K-1)$ *degrees of freedom.*

For example, assume the data displayed in Table 13.7 are values of independent random variables satisfying the assumptions of this section. Table 13.8 shows the results of the three significance tests, organized into an *analysis of variance table.*

(i) The observed ratio of the row mean square to the error mean square is 1.263; the observed significance level, based on the f ratio distribution with 2 and 12 degrees of freedom, is 0.318.

(ii) The observed ratio of the column mean square to the error mean square is 11.066; the observed significance level, based on the f ratio distribution with 1 and 12 degrees of freedom, is 0.006.

(iii) The observed ratio of the row-by-column mean square to the error mean square is 0.261; the observed significance level, based on the f ratio distribution with 2 and 12 degrees of freedom, is 0.774.

The results of the analysis of variance suggest that starting depth (10, 30, or 50 feet) is not a significant factor in determining the time to drill a 5-foot hole, drilling method

(either dry or wet) is a significant factor, and the simple additive model holds. The wet method appears to be superior to the dry method, since the mean drilling times are smaller.

In the analysis of variance for balanced two-way layouts, potential row, column, and interaction effects are separated. The effects can then be tested separately.

Unbalanced two-way layouts can be analyzed using the more general linear regression methods. Linear regression is studied in Chapter 14.

13.3.2 Example: Permutation f tests

Permutation versions of the analysis of variance f tests for zero row or column effects can be developed. The analysis for column effects, for example, is appropriate in one of the following situations:

1. *Population model.* The observed data are the values of IJ independent random samples from distributions differing in mean only. The null hypothesis is that for each i
$$\mu_{i,1} = \mu_{i,2} = \cdots = \mu_{i,J}$$
(the means are equal within each level of the row factor) versus the general alternative that at least two means differ in some level of the row factor.

2. *Randomization model.* The data are measurements taken on N individuals in IJ distinguishable groups of size K each. The null hypothesis is that observed differences in means are due to chance alone versus the alternative that at least one sample has values that tend to be larger or smaller (but not more variable) than the values in another sample at the same level of the row factor.

The sampling distribution of $F = MS_c/MS_e$ is obtained by computing its value for each partition of the J samples at each level of the row factor into distinguishable groups of size K each. There are a total of $\binom{JK}{K,K,\ldots,K}^I$ partitions to consider. Since this number can be quite large, Monte Carlo analysis is generally used to estimate a p value.

For example, consider testing the null hypothesis of zero column effects using the 5% significance level and the mining data (Table 13.7). In a Monte Carlo analysis using 5000 random partitions (including the observed partition of the 18 drilling times), 0.62% (31/5000) of F values were greater than or equal to $f_{\text{obs}} = 11.066$. Thus, there is evidence that observed differences in mean drilling times using the dry and wet drilling methods are not due to chance alone.

Note that permutation analyses of zero row or column effects are examples of *stratified* analyses, where the strata correspond to the levels of the other factor.

13.4 Laboratory problems

Laboratory problems for this chapter introduce *Mathematica* commands for analysis of variance, Bonferroni analysis of multiple comparisons, and the Kruskal–Wallis

and Friedman tests. The problems are designed to reinforce ideas related to analyses of multiple samples.

13.4.1 Laboratory: Multiple sample analysis

In the main laboratory notebook (Problems 1 to 7), you will use simulation and graphics to study analysis of variance for one-way layouts and blocked designs and to study the Kruskal–Wallis and Friedman tests; apply methods for one-way layouts to an antibiotics study [65], [115]; and apply methods for blocked designs to data on tobacco yield [51], [110].

13.4.2 Additional problem notebooks

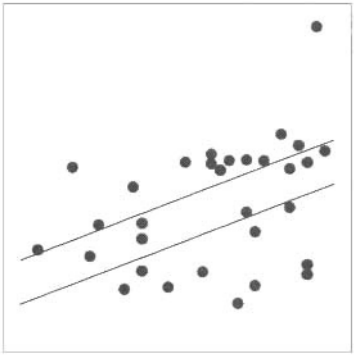
Problems 8, 9, and 10 are applications of methods for one-way layouts. Problem 8 uses data from a physical anthropology study [50]; Problem 9 uses data from a manufacturing study [64], [30]; and Problem 10 uses data from a study of factors related to plasma levels of beta-carotene in women [104].

Problems 11 and 12 are applications of methods for blocked designs. Problem 11 uses data from the 1996 Atlanta Olympics on running times for sprinters [81], and Problem 12 uses data from an ecology study [18], [42], [100].

Methods for balanced two-way layouts are applied in Problems 13 through 16. Problem 13 uses simulation to introduce analysis of variance for balanced two-way layouts and applies analysis of variance methods to data from a study of cardiovascular risk factors [78]. Problem 14 uses data from a survival study [17]. Problem 15 uses data from a marketing study [61], [78]. Problem 16 uses data from a niacin-enrichment study [23], [90]. In Problem 16, a method for adjusting for missing observations is introduced.

In Problem 17, simulation is used to study the validity of one-way analysis of variance when standard deviations are not all equal and to develop a rule of thumb for determining when the ratio of the maximum to the minimum sample standard deviation is small enough to use analysis of variance methods.

This page intentionally left blank



Chapter 14
Linear Least Squares Analysis

Linear least squares methods allow researchers to study how variables are related. For example, a researcher might be interested in determining the relationship between the weight of an individual and such variables as height, age, sex, and general body dimensions.

Sections 1 and 2 introduce methods used to analyze how one variable can be used to predict another (for example, how height can be used to predict weight). Section 3 introduces methods to analyze how several variables can be used to predict another (for example, how the combination of height, age, sex, and general body dimensions can be used to predict weight). Bootstrap applications are given in Section 4. Section 5 outlines the laboratory problems. References for regression diagnostic methods are [12], [28], [49].

14.1 Simple linear model

A *simple linear model* is a model of the form

$$Y = \alpha + \beta X + \epsilon,$$

where X and ϵ are independent random variables, and the distribution of ϵ has mean 0 and standard deviation σ. Y is called the *response* variable, and X is called the *predictor* variable. ϵ represents the measurement error.

The response variable Y can be written as a linear function of the predictor variable X plus an error term. The linear prediction function has slope β and intercept α.

The objective is to estimate the parameters in the conditional mean formula

$$E(Y|X = x) = \alpha + \beta x$$

using a list of paired observations. The observed pairs are assumed to be either the values of a random sample from the joint (X, Y) distribution or a collection of

independent responses made at predetermined levels of the predictor. Analysis is done conditional on the observed values of the predictor variable.

14.1.1 Least squares estimation

Assume that

$$Y_i = \alpha + \beta x_i + \epsilon_i \text{ for } i = 1, 2, \ldots, N$$

are independent random variables with means $E(Y_i) = \alpha + \beta x_i$, that the collection $\{\epsilon_i\}$ is a random sample from a distribution with mean 0 and standard deviation σ, and that all parameters (α, β, and σ) are unknown.

Least squares is a general estimation method introduced by A. Legendre in the early 1800's. In the simple linear case, the *least squares* (LS) estimators of α and β are obtained by minimizing the following sum of squared deviations of observed from expected responses:

$$S(\alpha, \beta) = \sum_{i=1}^{N} (Y_i - (\alpha + \beta x_i))^2 .$$

Multivariable calculus can be used to demonstrate that the LS estimators of slope and intercept can be written in the form

$$\widehat{\beta} = \sum_{i=1}^{N} \left[\frac{(x_i - \overline{x})}{S_{xx}} \right] Y_i \quad \text{and} \quad \widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{x} = \sum_{i=1}^{N} \left[\frac{1}{N} - \frac{(x_i - \overline{x})\overline{x}}{S_{xx}} \right] Y_i,$$

where $\overline{x}$ and $\overline{Y}$ are the mean values of predictor and response, respectively, and S_{xx} is the sum of squared deviations of observed predictors from their sample mean:

$$S_{xx} = \sum_{i=1}^{N} (x_i - \overline{x})^2.$$

Formulas for $\widehat{\alpha}$ and $\widehat{\beta}$ can be written in many different ways. The method used here emphasizes that each estimator is a linear combination of the response variables.

Example: Olympic winning times

To illustrate the computations, consider the following 20 data pairs, where x is the time in years since 1900 and y is the Olympic winning time in seconds for men in the final round of the 100-meter event [50, p. 248]:

x	0	4	8	12	20	24	28	32	36	48
y	10.8	11.0	10.8	10.8	10.8	10.6	10.8	10.3	10.3	10.3
x	52	56	60	64	68	72	76	80	84	88
y	10.4	10.5	10.2	10.0	9.95	10.14	10.06	10.25	9.99	9.92

The data set covers all Olympic events held between 1900 and 1988. (Olympic games were not held in 1916, 1940, and 1944.) For these data, $\overline{x} = 45.6$, $\overline{y} = 10.396$, and

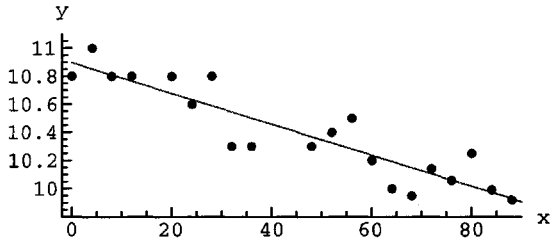


Figure 14.1. *Olympic winning time in seconds for men's 100-meter finals (vertical axis) versus year since* 1900 *(horizontal axis). The gray line is the linear least squares fit,* $y = 10.898 - 0.011x$.

the least squares estimates of slope and intercept are $\widehat{\beta} = -0.011$ and $\widehat{\alpha} = 10.898$, respectively. Figure 14.1 shows a scatter plot of the Olympic winning times data pairs superimposed on the least squares fitted line. The results suggest that the winning times have decreased at the rate of about 0.011 seconds per year during the 88 years of the study.

Properties of LS estimators

Theorem 4.4 can be used to demonstrate the following:

1. $E(\widehat{\beta}) = \beta$ and $Var(\widehat{\beta}) = \sigma^2/S_{xx}$.
2. $E(\widehat{\alpha}) = \alpha$ and $Var(\widehat{\alpha}) = \left(\sum_i x_i^2\right)\sigma^2/(N\,S_{xx})$.

In addition, the following theorem, proven by Gauss and Markov, states that LS estimators are best (minimum variance) among all linear unbiased estimators of intercept and slope.

Theorem 14.1 (Gauss–Markov Theorem). *Under the assumptions of this section, the least squares (LS) estimators are the best linear unbiased estimators of* α *and* β.

For example, consider estimating β using a linear function of the response variables, say $W = c + \sum_i d_i Y_i$ for some constants c and $d_1, d_2, \ldots, d_N$. If W is an unbiased estimator of β, then

$$Var(W) = Var\left(c + \sum_i d_i Y_i\right) = \sum_i d_i^2\, Var(Y_i) = \left(\sum_i d_i^2\right)\sigma^2$$

is minimized when $d_i = (x_i - \overline{x})/S_{xx}$ and $c = 0$. That is, the variance is minimized when W is the LS estimator of β.

Although LS estimators are best among linear unbiased estimators, they may not be ML estimators. Thus, there may be other more efficient methods of estimation.

14.1.2 Permutation confidence interval for slope

Permutation methods can be used to construct confidence intervals for the slope parameter β in the simple linear model. Let

$$(x_i, y_i) \text{ for } i = 1, 2, \ldots, N$$

be the observed pairs and π be a permutation of the indices $1, 2, \ldots, N$ other than the identity. Then the quantity

$$b(\pi) = \frac{\sum_i (x_i - \overline{x})(y_{\pi(i)} - y_i)}{\sum_i (x_i - \overline{x})(x_{\pi(i)} - x_i)}$$

is an estimate of β, and the collection

$$\{b(\pi) \ : \ \pi \text{ is a permutation other then the identity}\}$$

is a list of $N! - 1$ estimates. The ordered estimates

$$b_{(1)} < b_{(2)} < b_{(3)} < \cdots < b_{(N!-1)}$$

are used in constructing confidence intervals.

Theorem 14.2 (Slope Confidence Intervals). *Under the assumptions of this section, the interval*

$$[b_{(k)}, b_{(N!-k)}]$$

is a $100(1 - 2k/N!)\%$ *confidence interval for* β.

The procedure given in Theorem 14.2 is an example of *inverting* a hypothesis test: A value β_o is in a $100(1 - \gamma)\%$ confidence interval if the two sided permutation test of

$$H_o : \text{ The correlation between } Y - \beta_o X \text{ and } X \text{ is zero}$$

is accepted at the γ significance level. For a proof, see [74, p. 120].

Since the number of permutations can be quite large, Monte Carlo analysis is used to estimate endpoints. For example, assume the Olympic times data (page 206) are the values of random variables satisfying the assumptions of this section. An approximate 95% confidence interval for the slope parameter (based on 5000 random permutations) is $[-0.014, -0.008]$.

14.2 Simple linear regression

In simple *linear regression*, the error distribution is assumed to be normal, and, as above, analyses are done conditional on the observed values of the predictor variable. Specifically, assume that

$$Y_i = \alpha + \beta x_i + \epsilon_i \text{ for } i = 1, 2, \ldots, N$$

are independent random variables with means $E(Y_i) = \alpha + \beta x_i$, that the collection $\{\epsilon_i\}$ is a random sample from a normal distribution with mean 0 and standard deviation σ, and that all parameters are unknown.

In this setting, LS estimators are ML estimators.

Theorem 14.3 (Parameter Estimation). *Given the assumptions and definitions above, the LS estimators of α and β given on page 206 are ML estimators, and the statistics*

$$\begin{aligned}\widehat{\epsilon_i} = Y_i - (\widehat{\alpha} + \widehat{\beta} x_i) &= Y_i - (\overline{Y} + \widehat{\beta}(x_i - \overline{x})) \\ &= Y_i - \sum_j \left[\frac{1}{N} + \frac{(x_j - \overline{x})(x_i - \overline{x})}{S_{xx}} \right] Y_j\end{aligned}$$

are ML estimators of the error terms for $i = 1, 2, \ldots, N$. Each estimator is a normal random variable, and each is unbiased. Further, the statistic

$$S^2 = \frac{1}{N-2} \sum_i (Y_i - (\widehat{\alpha} + \widehat{\beta} x_i))^2$$

is an unbiased estimator of the common variance σ^2.

14.2.1 Confidence interval procedures

This section develops confidence interval procedures for the slope and intercept parameters, and for the mean response at a fixed value of the predictor variable.

Hypothesis tests can also be developed. Most computer programs automatically include both types of analyses.

Confidence intervals for β

Since the LS estimator $\widehat{\beta}$ is a normal random variable with mean β and variance σ^2/S_{xx}, Theorem 6.2 can be used to demonstrate that

$$\widehat{\beta} \pm t_{N-2}(\gamma/2) \sqrt{\frac{S^2}{S_{xx}}}$$

is a $100(1-\gamma)\%$ confidence interval for β, where S^2 is the estimate of the common variance given in Theorem 14.3 and $t_{N-2}(\gamma/2)$ is the $100(1-\gamma/2)\%$ point on the Student t distribution with $(N-2)$ degrees of freedom.

Confidence intervals for α

Since the LS estimator $\widehat{\alpha}$ is a normal random variable with mean α and variance $\sigma^2 \left(\sum_i x_i^2\right) / (N\, S_{xx})$, Theorem 6.2 can be used to demonstrate that

$$\widehat{\alpha} \pm t_{N-2}(\gamma/2) \sqrt{\frac{S^2 \left(\sum_i x_i^2\right)}{N\, S_{xx}}}$$

is a $100(1-\gamma)\%$ confidence interval for α, where S^2 is the estimate of the common variance given in Theorem 14.3 and $t_{N-2}(\gamma/2)$ is the $100(1-\gamma/2)\%$ point on the Student t distribution with $(N-2)$ degrees of freedom.

For example, if the Olympic times data (page 206) are the values of random variables satisfying the assumptions of this section, then a 95% confidence interval for the slope parameter is $[-0.013, -0.009]$, and a 95% confidence interval for the intercept parameter is $[10.765, 11.030]$.

Confidence intervals for mean response

The mean response $E(Y_o) = \alpha + \beta x_o$ at a new predictor-response pair, (x_o, Y_o), can be estimated using the statistic

$$\widehat{\alpha} + \widehat{\beta} x_o = \overline{Y} + \widehat{\beta}(x_o - \overline{x}) = \sum_{i=1}^{N} \left[\frac{1}{N} + \frac{(x_o - \overline{x})(x_i - \overline{x})}{S_{xx}} \right] Y_i.$$

This estimator is a normal random variable (by Theorem 4.6) with mean $\alpha + \beta x_o$ and

$$Var(\widehat{\alpha} + \widehat{\beta} x_o) = \sigma^2 \left(\frac{1}{N} + \frac{(x_o - \overline{x})^2}{S_{xx}} \right).$$

Thus, Theorem 6.2 can be used to demonstrate that

$$(\widehat{\alpha} + \widehat{\beta} x_o) \pm t_{N-2}(\gamma/2) \sqrt{S^2 \left(\frac{1}{N} + \frac{(x_o - \overline{x})^2}{S_{xx}} \right)}$$

is a $100(1-\gamma)\%$ confidence interval for $\alpha + \beta x_o$, where S^2 is the estimate of the common variance given in Theorem 14.3 and $t_{N-2}(\gamma/2)$ is the $100(1-\gamma/2)\%$ point on the Student t distribution with $(N-2)$ degrees of freedom.

Example: Percentage of dead or damaged spruce trees

For example, as part of a study on the relationship between environmental stresses and the decline of red spruce tree forests in the Appalachian Mountains, data were collected on the percentage of dead or damaged trees at various altitudes in forests in the northeast. The paired data were of interest because concentrations of airborne pollutants tend to be higher at higher altitudes [49, p. 102].

Figure 14.2 is based on information gathered in 53 areas. For these data, the least squares fitted line is $y = 8.24x - 33.66$, suggesting that the percentage of damaged or dead trees increases at the rate of 8.24 percentage points per 100 meters elevation.

An estimate of the mean response at 1000 meters ($x_o = 10$) is 48.76% damaged or dead. If these data are the values of independent random variables satisfying the assumptions of this section, then a 95% confidence interval for the mean response at 1000 meters is $[48.44, 49.07]$.

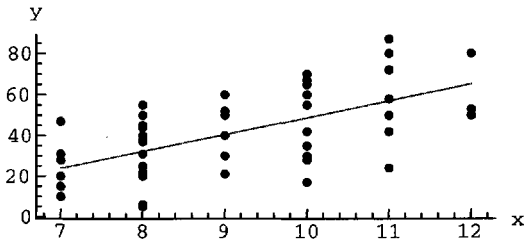


Figure 14.2. *Percentage dead or damaged red spruce trees (vertical axis) versus elevation in* 100 *meters (horizontal axis) at* 53 *locations in the northeast. The gray line is the linear least squares fit,* $y = 8.24x - 33.66$.

Comparison of procedures

The confidence interval procedure for β given in this section is valid when the error distribution is normal. When the error distribution is not normal, the permutation procedure given in Theorem 14.2 can be used.

The confidence interval procedures given in this section assume that the values of the predictor variable are known with certainty (the procedures are conditional on the observed values of the predictor) and assume that the error distributions are normal. Approximate bootstrap confidence interval procedures can also be developed under broader conditions; see Section 14.4.

14.2.2 Predicted responses and residuals

The i^{th} estimated mean (or *predicted response*) is the random variable

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta} x_i = \sum_j \left[\frac{1}{N} + \frac{(x_j - \overline{x})(x_i - \overline{x})}{S_{xx}} \right] Y_j \quad \text{for } i = 1, 2, \ldots, N,$$

and the i^{th} estimated error (or *residual*) is

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i \quad \text{for } i = 1, 2, \ldots, N.$$

Each random variable is a linear function of the response variables. Theorem 4.5 can be used to demonstrate that $Cov(\widehat{Y}_i, \widehat{\epsilon}_i) = 0$.

Although the error terms in the simple linear model have equal variances, the estimated errors do not. Specifically, the variance of the i^{th} residual is

$$Var(\widehat{\epsilon}_i) = \sigma^2 \left[\left(1 - \frac{1}{N} - \frac{(x_i - \overline{x})^2}{S_{xx}} \right)^2 + \sum_{j \neq i} \left(\frac{1}{N} + \frac{(x_j - \overline{x})(x_i - \overline{x})}{S_{xx}} \right)^2 \right] = \sigma^2 \, c_i.$$

The i^{th} estimated *standardized residual* is defined as follows:

$$r_i = \widehat{\epsilon}_i / \sqrt{S^2 \, c_i} \text{ for } i = 1, 2, \ldots, N,$$

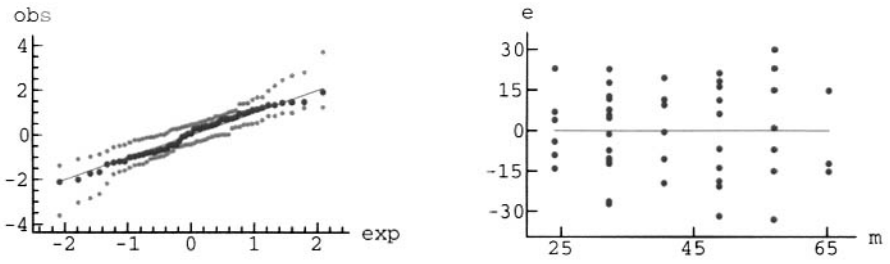


Figure 14.3. *Enhanced normal probability plot of standardized residuals (left plot) and scatter plot of residuals versus estimated means (right plot) for the spruce trees example.*

where S^2 is the estimate of the common variance given in Theorem 14.3 and c_i is the constant in brackets above.

Predicted responses, residuals, and estimated standardized residuals are used in diagnostic plots of model assumptions. For example, the left plot in Figure 14.3 is an enhanced normal probability of the estimated standardized residuals from the spruce trees example (page 210), and the right plot is a scatter plot of residuals (vertical axis) versus predicted responses (horizontal axis). The left plot suggests that the error distribution is approximately normally distributed; the right plot exhibits no relationship between the estimated errors and estimated means.

The scatter plot of residuals versus predicted responses should show no relationship between the variables. Of particular concern are the following:

1. If $\widehat{\epsilon_i} \approx h(\widehat{y_i})$ for some function h, then the assumption that the conditional mean is a linear function of the predictor may be wrong.
2. If $SD(\widehat{\epsilon_i}) \approx h(\widehat{y_i})$ for some function h, then the assumption of equal standard deviations may be wrong.

14.2.3 Goodness-of-fit

Suppose that the N predictor-response pairs can be written in the following form:

$$(x_i, Y_{i,j}) \text{ for } j = 1, 2, \ldots, n_i, \; i = 1, 2, \ldots, I.$$

(There are a total of n_i observed responses at the i^{th} level of the predictor variable for $i = 1, 2, \ldots, I$, and $N = \sum_i n_i$.) Then it is possible to use an analysis of variance technique to test the goodness-of-fit of the simple linear model.

Assumptions

The responses are assumed to be the values of I independent random samples

$$Y_{i,1}, Y_{i,2}, \ldots, Y_{i,n_i} \text{ for } i = 1, 2, \ldots, I$$

from normal distributions with a common unknown standard deviation σ.

Let μ_i be the mean of responses in the i^{th} sample: $\mu_i = E(Y_{i,j})$ for all j. Of interest is to test the null hypothesis that $\mu_i = \alpha + \beta x_i$ for $i = 1, 2, \ldots, I$.

Sources of variation

The formal goodness-of-fit analysis is based on writing the sum of squared deviations of the response variables from the predicted responses using the linear model (known as the *error* sum of squares),

$$SS_e = \sum_{i,j} \left(Y_{i,j} - (\widehat{\alpha} + \widehat{\beta} x_i)\right)^2,$$

as the sum of squared deviations of the response variables from the estimated group means (known as the *pure error* sum of squares),

$$SS_p = \sum_{i,j} \left(Y_{i,j} - \overline{Y}_{i\cdot}\right)^2,$$

plus the weighted sum of squared deviations of the group means from the predicted responses (known as the *lack-of-fit* sum of squares),

$$SS_\ell = \sum_{i} n_i \left(\overline{Y}_{i\cdot} - (\widehat{\alpha} + \widehat{\beta} x_i)\right)^2.$$

Pure error and lack-of-fit mean squares

The *pure error* mean square, MS_p, is defined as follows:

$$MS_p = \frac{1}{N-I} SS_p = \frac{1}{N-I} \sum_{i,j} \left(Y_{i,j} - \overline{Y}_{i\cdot}\right)^2.$$

MS_p is equal to the pooled estimate of the common variance. Theorem 6.1 can be used to demonstrate that $(N-I)MS_p/\sigma^2$ is a chi-square random variable with $(N-I)$ degrees of freedom.

The *lack-of-fit* mean square, MS_ℓ, is defined as follows:

$$MS_\ell = \frac{1}{I-2} SS_\ell = \frac{1}{I-2} \sum_{i} n_i \left(\overline{Y}_{i\cdot} - (\widehat{\alpha} + \widehat{\beta} x_i)\right)^2.$$

Properties of expectation can be used to demonstrate that

$$E\left(MS_\ell\right) = \sigma^2 + \frac{1}{I-2} \sum_{i} n_i \left(\mu_i - (\alpha + \beta x_i)\right)^2.$$

If the null hypothesis that the means follow a simple linear model is true, then the expected value of MS_ℓ is σ^2; otherwise, values of MS_ℓ will tend to be larger than σ^2. The following theorem relates the pure error and lack-of-fit mean squares.

Theorem 14.4 (Distribution Theorem). *Under the general assumptions of this section and if the null hypothesis is true, then the ratio $F = MS_\ell/MS_p$ has an f ratio distribution with $(I-2)$ and $(N-I)$ degrees of freedom.*

Table 14.1. *Goodness-of-fit analysis of the spruce tree data.*

Source	df	SS	MS	F	p value
Lack-of-Fit	4	132.289	33.072	0.120	0.975
Pure Error	47	12964.3	275.835		
Error	51	13096.5			

Goodness-of-fit test: Observed significance level

Large values of $F = MS_\ell/MS_p$ support the alternative hypothesis that the simple linear model does not hold. For an observed ratio, f_{obs}, the p value is $P(F \geq f_{\text{obs}})$.

For example, assume the spruce trees data (page 210) satisfy the general assumptions of this section. Table 14.1 shows the results of the goodness-of-fit test. There were 6 observed predictor values. The observed ratio of the lack-of-fit mean square to the pure error mean square is 0.120. The observed significance level, based on the f ratio distribution with 4 and 47 degrees of freedom, is 0.975. The simple linear model fits the data quite well.

14.3 Multiple linear regression

A *linear model* is a model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \epsilon,$$

where each X_i is independent of ϵ, and the distribution of ϵ has mean 0 and standard deviation σ. Y is called the *response* variable, each X_i is a *predictor* variable, and ϵ represents the measurement error.

The response variable Y can be written as a linear function of the $(p-1)$ predictor variables plus an error term. The linear prediction function has p parameters.

In multiple *linear regression*, the error distribution is assumed to be normal, and analyses are done conditional on the observed values of the predictor variables. Observations are called *cases*.

14.3.1 Least squares estimation

Assume that

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{p-1} x_{p-1,i} + \epsilon_i \text{ for } i = 1, 2, \ldots, N$$

are independent random variables with means

$$E(Y_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{p-1} x_{p-1,i} \text{ for all } i,$$

that the collection of errors $\{\epsilon_i\}$ is a random sample from a normal distribution with mean 0 and standard deviation σ, and that all parameters (including σ) are unknown.

The *least squares* (LS) estimators of the coefficients in the linear prediction function are obtained by minimizing the following sum of squared deviations of observed from expected responses:

$$S(\beta_0, \beta_1, \ldots, \beta_{p-1}) = \sum_{k=1}^{N} \left(Y_k - \left(\beta_0 + \beta_1 x_{1,k} + \beta_2 x_{2,k} + \cdots + \beta_{p-1} x_{p-1,k}\right)\right)^2$$
$$= \sum_{k=1}^{N} \left(Y_k - \sum_{j=0}^{p-1} \beta_j x_{j,k}\right)^2, \quad \text{where } x_{0,k} = 1 \text{ for all } k.$$

The first step in the analysis is to compute the partial derivative with respect to β_i for each i. Partial derivatives have the following form:

$$\frac{\partial S}{\partial \beta_i} = -2\left[\sum_{k=1}^{N} Y_k x_{i,k} - \sum_{j=0}^{p-1} \beta_j \left(\sum_{k=1}^{N} x_{j,k} x_{i,k}\right)\right].$$

The next step is to solve the p-by-p system of equations

$$\frac{\partial S}{\partial \beta_i} = 0, \text{ for } i = 0, 1, \ldots, p-1,$$

or, equivalently,

$$\sum_{j=0}^{p-1} \left(\sum_{k=1}^{N} x_{j,k} x_{i,k}\right) \beta_j = \sum_{k=1}^{N} Y_k x_{i,k}, \text{ for } i = 0, 1, \ldots, p-1.$$

In matrix notation, the system becomes

$$(\mathbf{X}^T\mathbf{X})\,\underline{\beta} = \mathbf{X}^T\underline{Y},$$

where $\underline{\beta}$ is the p-by-1 vector of unknown parameters, $\underline{Y}$ is the N-by-1 vector of response variables, $\mathbf{X}$ is the N-by-p matrix whose (i, j) element is $x_{j,i}$, and $\mathbf{X}^T$ is the transpose of the $\mathbf{X}$ matrix. The $\mathbf{X}$ matrix is often called the *design matrix*. Finally, the p-by-1 vector of LS estimators is

$$\widehat{\underline{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underline{Y}.$$

Estimates exist as long as $(\mathbf{X}^T\mathbf{X})$ is invertible.

The rows of the design matrix correspond to the observations (or cases). The columns correspond to the predictors. The terms in the first column of the design matrix are identically equal to one.

For example, in the simple linear case, the matrix product

$$(\mathbf{X}^T\mathbf{X}) = \begin{bmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}$$

has inverse

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{NS_{xx}} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & N \end{bmatrix}, \quad \text{where } S_{xx} = \sum_i (x_i - \bar{x})^2,$$

and the LS estimators are

$$\begin{bmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{bmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underline{Y} = \begin{bmatrix} \sum_i \left(\frac{1}{N} - \frac{(x_i-\bar{x})\bar{x}}{S_{xx}}\right) Y_i \\ \sum_i \left(\frac{(x_i-\bar{x})}{S_{xx}}\right) Y_i \end{bmatrix}.$$

The estimators here correspond exactly to those given on page 206.

Model in matrix form

The model can be written as

$$\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\epsilon},$$

where $\underline{Y}$ and $\underline{\epsilon}$ are N-by-1 vectors of responses and errors, respectively, $\underline{\beta}$ is the p-by-1 coefficient vector, and $\mathbf{X}$ is the N-by-p design matrix.

Theorem 14.5 (Parameter Estimation). *Given the assumptions and definitions above, the vector of LS estimators of $\underline{\beta}$ given on page 215 is a vector of ML estimators, and the vector*

$$\widehat{\underline{\epsilon}} = \underline{Y} - \mathbf{X}\widehat{\underline{\beta}} = (\mathbf{I} - \mathbf{H})\,\underline{Y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and $\mathbf{I}$ is the N-by-N identity matrix, is a vector of ML estimators of the error terms. Each estimator is a normal random variable, and each is unbiased. Further, the statistic

$$S^2 = \frac{1}{N-p}\sum_{i=1}^{N}\left(Y_i - \widehat{Y}_i\right)^2,$$

where $\widehat{Y}_i$ is the i^{th} estimated mean, is an unbiased estimator of σ^2.

The i^{th} estimated mean (or *predicted response*) is the random variable

$$\widehat{Y}_i = \widehat{\beta_0} + \widehat{\beta_1}x_{i,1} + \cdots + \widehat{\beta_{p-1}}x_{i,p-1} \text{ for } i = 1, 2, \ldots, N.$$

Further, the matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is often called the *hat matrix* since it is the matrix that transforms the response vector to the predicted response vector

$$\widehat{\underline{Y}} = \mathbf{X}\widehat{\underline{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underline{Y} = \mathbf{H}\underline{Y}$$

(the vector of Y_i's is transformed to the vector of Y_i *hats*).

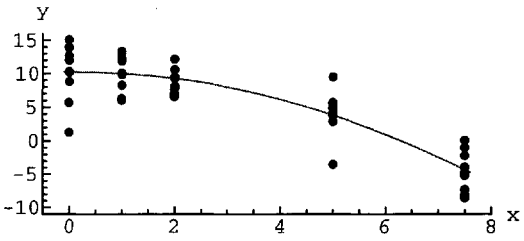


Figure 14.4. *Change in weight in grams (vertical axis) versus dosage level in* 100 *mg/kg/day (horizontal axis) for data from the toxicology study. The gray curve is the linear least squares fit,* $y = 10.2475 + 0.053421x - 0.2658x^2$.

Variability of LS estimators

If $\underline{V}$ is an m-by-1 vector of random variables and $\underline{W}$ is an n-by-1 vector of random variables, then $\Sigma(\underline{V}, \underline{W})$ is the m-by-n matrix whose (i, j) term is $Cov(V_i, W_j)$. The matrix $\Sigma(\underline{V}, \underline{W})$ is called a *covariance matrix.*

Theorem 14.6 (Covariance Matrices). *Under the assumptions of this section, the following hold:*

1. *The covariance matrix of the coefficient estimators is*

$$\sum\left(\underline{\hat{\beta}}, \underline{\hat{\beta}}\right) = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}.$$

2. *The covariance matrix of the error estimators is*

$$\sum\left(\underline{\hat{\epsilon}}, \underline{\hat{\epsilon}}\right) = \sigma^2(\mathbf{I} - \mathbf{H}).$$

3. *The covariance matrix of error estimators and predicted responses is*

$$\sum\left(\underline{\hat{\epsilon}}, \underline{\hat{Y}}\right) = \mathbf{0},$$

where $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ *is the hat matrix,* $\mathbf{I}$ *is the N-by-N identity matrix, and* $\mathbf{0}$ *is the N-by-N matrix of zeros.*

The diagonal elements of $\sum\left(\underline{\hat{\beta}}, \underline{\hat{\beta}}\right)$ and $\sum\left(\underline{\hat{\epsilon}}, \underline{\hat{\epsilon}}\right)$ are the variances of the coefficient and error estimators, respectively. The last statement in the theorem says that error estimators and predicted responses are uncorrelated.

Example: Toxicology study

To illustrate some of the computations, consider the data pictured in Figure 14.4, collected as part of a study to assess the adverse effects of a proposed drug for the treatment of tuberculosis [40].

Ten female rats were given the drug for a period of 14 days at each of five dosage levels (in 100 milligrams per kilogram per day). The vertical axis in the plot

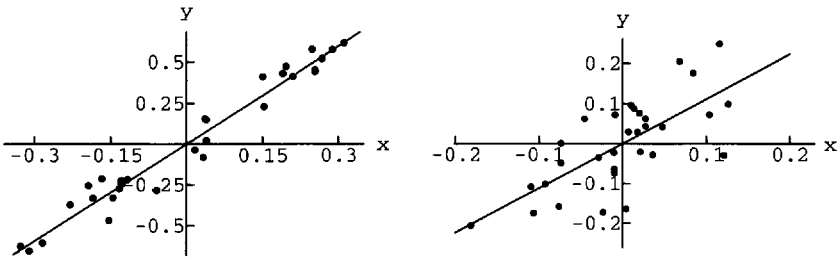


Figure 14.5. *Partial regression plots for the data from the timber yield study. The left plot pictures residuals of log-volume (vertical axis) versus log-diameter (horizontal axis) with the effect of log-height removed. The right plot pictures residuals of log-volume (vertical axis) versus log-height (horizontal axis) with the effect of log-diameter removed. The gray lines are $y = 1.983x$ in the left plot and $y = 1.117x$ in the right plot.*

shows the weight change in grams (WC), defined as the weight at the end of the period minus the weight at the beginning of the period; the horizontal axis shows the dose in 100 mg/kg/day. A linear model of the form

$$\text{WC} = \beta_0 + \beta_1 \text{ dose} + \beta_2 \text{ dose}^2 + \epsilon$$

was fit to the 50 (dose,WC) cases. (The model is linear in the unknown parameters and quadratic in the dosage level.) The LS prediction equation is shown in the plot.

Example: Timber yield study

As part of a study to find an estimate for the volume of a tree (and therefore its yield) given its diameter and height, data were collected on the volume (in cubic feet), diameter at 54 inches above the ground (in inches), and height (in feet) of 31 black cherry trees in the Allegheny National Forest [50, p. 159]. Since a multiplicative relationship is expected among these variables, a linear model of the form

$$\text{log-volume} = \beta_0 + \beta_1 \text{ log-diameter} + \beta_2 \text{ log-height} + \epsilon$$

was fit to the 31 (log-diameter,log-height,log-volume) cases, using the natural logarithm function to compute log values.

The LS prediction equation is

$$\text{log-volume} = -6.632 + 1.983 \text{ log-diameter} + 1.117 \text{ log-height}.$$

Figure 14.5 shows *partial regression* plots of the timber yield data.

(i) In the left plot, the log-volume and log-diameter variables are adjusted to remove the effects of log-height. Specifically, the residuals from the simple linear regression of log-volume on log-height (vertical axis) are plotted against the residuals from the simple linear regression of log-diameter on log-height (horizontal axis). The relationship between the adjusted variables can be described using the linear equation $y = 1.983x$.

(ii) In the right plot, the log-volume and log-height variables are adjusted to remove the effects of log-diameter. The relationship between the adjusted variables can be described using the linear equation $y = 1.117x$.

The slopes of the lines in the partial regression plots correspond to the LS estimates in the prediction equation above. The plots suggest that a linear relationship between the response variable and each of the predictors is reasonable.

14.3.2 Analysis of variance

The linear regression model can be reparametrized as follows:

$$Y_i = \mu + \sum_{j=1}^{p-1} \beta_j (x_{j,i} - \overline{x}_{j\cdot}) + \epsilon_i \text{ for } i = 1, 2, \ldots, N,$$

where μ is the overall mean

$$\mu = \frac{1}{N} \sum_{i=1}^{N} E(Y_i) = \beta_0 + \sum_{j=1}^{p-1} \beta_j \overline{x}_{j\cdot}$$

and $\overline{x}_{j\cdot}$ is the mean of the j^{th} predictor for all j. The difference $E(Y_i) - \mu$ is called the i^{th} *deviation* (or the i^{th} *regression effect*). The sum of the regression effects is zero.

This section develops an analysis of variance f test for the null hypothesis that the regression effects are identically zero (equivalently, a test of the null hypothesis that $\beta_i = 0$ for $i = 1, 2, \ldots, p-1$).

If the null hypothesis is accepted, then the $(p-1)$ predictor variables have no predictive ability; otherwise, they have some predictive ability.

Sources of variation; coefficient of determination

In the first step of the analysis, the sum of squared deviations of the response variables from the mean response (the *total* sum of squares),

$$SS_t = \sum_{i=1}^{N} (Y_i - \overline{Y})^2,$$

is written as the sum of squared deviations of the response variables from the predicted responses (the *error* sum of squares),

$$SS_e = \sum_{i=1}^{N} (Y_i - \widehat{Y}_i)^2,$$

plus the sum of squared deviations of the predicted responses from the mean response (the *model* sum of squares),

$$SS_m = \sum_{i=1}^{N} (\widehat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{N} \left(\sum_{j=1}^{p-1} \widehat{\beta}_j (x_{j,i} - \overline{x}_{j\cdot}) \right)^2.$$

The ratio of the model to the total sums of squares, $R^2 = SS_m/SS_t$, is called the *coefficient of determination*. R^2 is the proportion of the total variation in the response variable that is explained by the model.

> In the simple linear case, R^2 is the same as the square of the sample correlation.

Analysis of variance f test

The *error* mean square is the ratio $MS_e = SS_e/(N-p)$, and the *model* mean square is the ratio $MS_m = SS_m/(p-1)$. The following theorem relates these random variables.

Theorem 14.7 (Distribution Theorem). *Under the general assumptions of this section and if the null hypothesis is true, then the ratio $F = MS_m/MS_e$ has an f ratio distribution with $(p-1)$ and $(N-p)$ degrees of freedom.*

Large values of $F = MS_m/MS_e$ support the hypothesis that the proposed predictor variables have some predictive ability. For an observed ratio, f_{obs}, the p value is $P(F \geq f_{\text{obs}})$.

For the toxicology study example (page 217), $f_{\text{obs}} = 82.3$ and the p value, based on the f ratio distribution with 2 and 47 degrees of freedom, is virtually zero. The coefficient of determination is 0.778; the estimated linear model explains about 77.8% of the variation in weight change.

For the timber yield example (page 218), $f_{\text{obs}} = 613.2$ and the p value, based on the f ratio distribution with 2 and 28 degrees of freedom, is virtually zero. The coefficient of determination is 0.978; the estimated linear model explains about 97.8% of the variation in log-volume.

> It is possible for the f test to reject the null hypothesis and the value of R^2 to be close to zero. In this case, the potential predictors have some predictive ability, but additional (or different) predictor variables are needed to adequately model the response.

14.3.3 Confidence interval procedures

This section develops confidence interval procedures for the β parameters and for the mean response at a fixed combination of the predictor variables.

> Hypothesis tests can also be developed. Most computer programs automatically include both types of analyses.

Confidence intervals for β_i

Let v_i be the element in the (i, i) position of $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$, and let S^2 be the estimate of the common variance given in Theorem 14.5. Since the LS estimator $\widehat{\beta}_i$ is a normal random variable with mean β_i and variance $\sigma^2 v_i$, Theorem 6.2 can be used to demonstrate that

$$\widehat{\beta}_i \pm t_{N-p}(\gamma/2)\sqrt{S^2 v_i}$$

is a $100(1-\gamma)\%$ confidence interval for β_i, where $t_{N-p}(\gamma/2)$ is the $100(1-\gamma/2)\%$ point on the Student t distribution with $(N-p)$ degrees of freedom.

For example, if the data in the toxicology study (page 217) are the values of random variables satisfying the assumptions of this section, then a 95% confidence interval for β_2 is $[-0.43153, -0.10008]$. Since 0 is not in the confidence interval, the result suggests that the model with the dose^2 term is significantly better than a simple linear model relating dose to weight change.

Confidence intervals for mean response

The mean response $E(Y_o) = \sum_{i=0}^{p-1} \beta_i x_{i,0}$ at a new predictors-response case can be estimated using the statistic

$$\sum_{i=0}^{p-1} \widehat{\beta}_i x_{i,0} = \underline{x}_0^T \underline{\widehat{\beta}} = \underline{x}_0^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \underline{Y},$$

where $\underline{x}_0^T = (1, x_{1,0}, x_{2,0}, \ldots, x_{p-1,0})$. This estimator is a normal random variable with mean $E(Y_o)$ and variance

$$\sigma^2 \left(\mathbf{x_0}^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x_0}\right) = \sigma^2 \, v_o.$$

Thus, Theorem 6.2 can be used to demonstrate that

$$\sum_{i=0}^{p-1} \widehat{\beta}_i x_{i,0} \pm t_{N-p}(\gamma/2)\sqrt{S^2 \, v_0}$$

is a $100(1-\gamma)\%$ confidence interval for $E(Y_o)$, where S^2 is the estimate of the common variance given in Theorem 14.3 and $t_{N-p}(\gamma/2)$ is the $100(1-\gamma/2)\%$ point on the Student t distribution with $(N-p)$ degrees of freedom.

For example, an estimate of the mean log-volume of a tree with diameter 11.5 inches and height 80 inches is 3.106 log-cubic inches. If these data are the values of random variables satisfying the assumptions of this section, then a 95% confidence interval for the mean response at this combination of the predictors is [3.05944, 3.1525].

14.3.4 Regression diagnostics

Recall that the hat matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ is the matrix that transforms the vector of observed responses $\underline{Y}$ to the vector of predicted responses $\underline{\widehat{Y}}$. Each predicted response is a linear combination of the observed responses:

$$\widehat{Y}_i = \sum_{j=1}^{N} h_{i,j} Y_j \quad \text{for } i = 1, 2, \ldots, N,$$

where $h_{i,j}$ is the (i, j) element of $\mathbf{H}$. In particular, the diagonal element $h_{i,i}$ is the coefficient of Y_i in the formula for $\widehat{Y}_i$.

Leverage

The *leverage* of the i^{th} response is the value $h_i = h_{i,i}$. Leverages satisfy the following properties:

1. For each i, $0 \le h_i \le 1$.
2. $\sum_{i=1}^{N} h_i = p$, where p is the number of parameters.

Ideally, the leverages should be about p/N each (the average value).

Residuals and standardized residuals

Theorem 14.6 implies that the variance of the i^{th} estimated error (or *residual*) is $Var(\widehat{\epsilon}_i) = \sigma^2(1 - h_i)$, where h_i is the leverage. The i^{th} estimated *standardized residual* is defined as follows:

$$r_i = \widehat{\epsilon}_i / \sqrt{S^2(1 - h_i)} \text{ for } i = 1, 2, \ldots, N,$$

where S^2 is the estimate of the common variance given in Theorem 14.5.

Residuals and standardized residuals are used in diagnostic plots of model assumptions. See Section 14.2.2 for examples in the simple linear case.

Standardized influences

The *influence* of the i^{th} observation is the change in prediction if the i^{th} observation is deleted from the data set.

Specifically, the influence is the difference $\widehat{Y}_i - \widehat{Y}_i(i)$, where $\widehat{Y}_i$ is the predicted response using all N cases to compute parameter estimates, and $\widehat{Y}_i(i)$ is the prediction at a "new" predictor vector $\underline{x}_i$, where parameter estimates have been computed using the list of $N - 1$ cases with the i^{th} case removed.

For the model estimated using $N - 1$ cases only, linear algebra methods can be used to demonstrate that the predicted response is

$$\widehat{Y}_i(i) = \widehat{Y}_i - \widehat{\epsilon}_i \frac{h_i}{1 - h_i}$$

and the estimated common variance is

$$S^2(i) = \frac{1}{N - p - 1}\left((N - p)S^2 - \frac{\widehat{\epsilon}_i^2}{1 - h_i}\right).$$

The i^{th} standardized influence is the ratio of the influence to the standard deviation of the predicted response,

$$\frac{\widehat{Y}_i - \widehat{Y}_i(i)}{SD(\widehat{Y}_i)} = \frac{\widehat{\epsilon}_i h_i/(1 - h_i)}{\sqrt{\sigma^2 h_i}},$$

and the i^{th} estimated *standardized influence* is the value obtained by substituting $S^2(i)$ for σ^2:

$$\delta_i = \frac{\widehat{Y}_i - \widehat{Y}_i(i)}{\widehat{SD}(\widehat{Y}_i)} = \frac{\widehat{\epsilon}_i h_i/(1 - h_i)}{\sqrt{S(i)^2 h_i}} = \frac{\widehat{\epsilon}_i \sqrt{h_i}}{S(i)(1 - h_i)}.$$

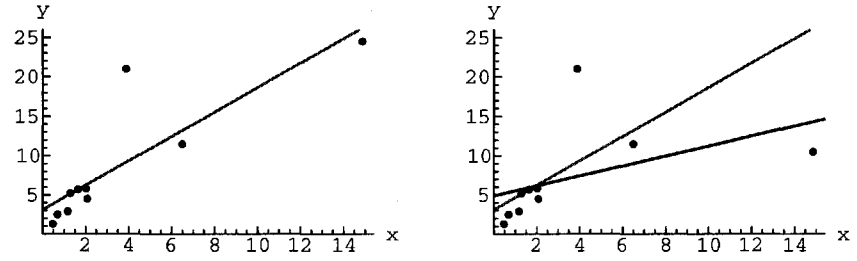


Figure 14.6. *Scatter plots of example pairs (left plot) and altered example pairs (right plot). The gray line in both plots has equation* $y = 3.11 + 1.55x$. *The black line in the right plot has equation* $y = 4.90 + 0.63x$.

Ideally, predicted responses should change very little if one case is removed from the list of N cases, and each δ_i should be close to zero. A general rule of thumb is that if $|\delta_i|$ is much greater than $2\sqrt{p/N}$, then the i^{th} case is highly influential.

Illustration

To illustrate the computations in the simple linear case, consider the following list of 10 (x, y) pairs:

x	0.47	0.69	1.17	1.28	1.64	2.02	2.08	3.88	6.50	14.86
y	1.30	2.50	2.90	5.20	5.70	5.80	4.50	21.00	11.50	24.50

The left plot in Figure 14.6 shows a scatter plot of the data pairs superimposed on the least squares fitted line, $y = 3.11 + 1.55x$. The following table gives the residuals, leverages, and standardized influences for each case:

i	1	2	3	4	5	6	7	8	9	10
$\widehat{\epsilon_i}$	−2.54	−1.68	−2.03	0.10	0.04	−0.45	−1.85	11.86	−1.72	−1.72
h_i	0.15	0.14	0.13	0.13	0.12	0.11	0.11	0.10	0.15	0.85
δ_i	−0.25	−0.16	−0.18	0.01	0.00	−0.04	−0.14	4.15	−0.17	−2.33

Based on the rule of thumb above, cases 8 and 10 are highly influential. Case 8 has a very large residual, and case 10 has a very large leverage value.

The right plot in Figure 14.6 illustrates the concept of leverage. If the observed response in case 10 is changed from 24.5 to 10.5, then the predicted response changes from 26.2 to 14.32. The entire line has moved to accommodate the change in case 10.

Different definitions of δ_i appear in the literature, although most books use the definition above. The rule of thumb is from [12], where the notation DFFITS_i is used for δ_i.

14.4 Bootstrap methods

Bootstrap resampling methods can be applied to analyzing the relationship between one or more predictors and a response. This section introduces two examples.

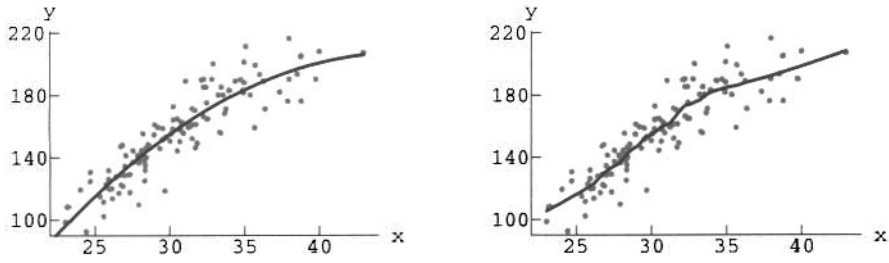


Figure 14.7. *Scatter plots of weight in pounds (vertical axis) versus waist circumference in inches (horizontal axis) for* 120 *physically active young adults. In the left plot, the curve* $y = -252.569 + 20.322x - 0.225x^2$ *is superimposed. In the right plot, the 25% lowess smooth is superimposed.*

Example: Unconditional analysis of linear models

If the observed cases are the values of a random sample from a joint distribution, then nonparametric bootstrap methods can be used to construct confidence intervals for parameters of interest without additional assumptions. (In particular, it is not necessary to condition on the observed values of the predictor variables.) Resampling is done from the list of N observed cases.

For example, the left plot in Figure 14.7 is a scatter plot of waist-weight measurements for 120 physically active young adults (derived from [53]) with a least squares fitted quadratic polynomial superimposed. An estimate of the mean weight for an individual with a 33-inch waist is 174.6 pounds. If the observed (x, y) pairs are the values of a random sample from a joint distribution satisfying a linear model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon,$$

then an approximate 95% confidence interval (based on 5000 resamples) for the mean weight when the waist size is 33 inches is [169.735, 176.944].

Example: Locally weighted regression

Locally weighted regression was introduced by W. Cleveland in the 1970's. Analysis is done conditional on the observed predictor values. In the single predictor case,

$$Y_i = g(x_i) + \epsilon_i \text{ for } i = 1, 2, \ldots, N$$

are assumed to be independent random variables, the function g is assumed to be a differentiable function of *unknown* form, and the collection $\{\epsilon_i\}$ is assumed to be a random sample from a distribution with mean 0 and standard deviation σ.

The goal is to estimate the conditional mean function, $y = g(x)$. Since $g(x)$ is differentiable, and a differentiable function is approximately linear on a small x-interval, the curve can be estimated as follows:

(i) For a given value of the predictor, say x_o, estimate the tangent line to $y = g(x)$ at $x = x_o$, and use the value predicted by the tangent line to estimate $g(x_o)$.

(ii) Repeat this process for all observed predictor values.

For a given x_o, the tangent line is estimated using a method known as weighted linear least squares. Specifically, the intercept and slope of the tangent line are obtained by minimizing the weighted sum of squared deviations

$$S(\alpha, \beta) = \sum_{i=1}^{N} w_i \left(Y_i - (\alpha + \beta x_i)\right)^2,$$

where the weights (w_i) are chosen so that pairs with x-coordinate near x_o have weight approximately 1; pairs with x-coordinate far from x_o have weight 0; and the weights decrease smoothly from 1 to 0 in a "window" centered at x_o.

The user chooses the proportion p of data pairs that will be in the "window" centered at x_o. When the process is repeated for each observed value of the predictor, the resulting estimated curve is called the $100p\%$ *lowess smooth*.

The right plot in Figure 14.7 shows the scatter plot of waist-weight measurements for the 120 physically active young adults with a 25% lowess smooth superimposed. The smoothed curve picks up the general pattern of the relationship between waist and weight measurements.

Lowess smooths allow researchers to approximate the relationship between predictor and response without specifying the function g. A bootstrap analysis can then be done, for example, to construct confidence intervals for the mean response at a fixed value of the predictor.

For the waist-weight pairs, a 25% smooth when $x = 33$ inches produced an estimated mean weight of 175.8 pounds. A bootstrap analysis (with 5000 random resamples) produced an approximate 95% confidence interval for mean weight when the waist size is 33 inches of [168.394, 182.650].

The lowess smooth algorithm implemented above uses tricube weights for smoothing and omits Cleveland's robustness step. For details about the algorithm, see [25, p. 121].

14.5 Laboratory problems

Laboratory problems for this chapter introduce *Mathematica* commands for linear regression analysis, permutation analysis of slope in the simple linear case, locally weighted regression, and diagnostic plots. Problems are designed to reinforce the ideas of this chapter.

14.5.1 Laboratory: Linear least squares analysis

In the main laboratory notebook (Problems 1 to 5), you will use simulation and graphics to study the components of linear least squares analyses; solve a problem on correlated and uncorrelated factors in polynomial regression; and apply linear least squares methods to three data sets from a study of sleep in mammals [3], [30].

14.5.2 Additional problem notebooks

Problems 6, 7, and 8 are applications of simple linear least squares (and other) methods. Problem 6 uses several data sets from an ecology study [32], [77]. Problem 7

uses data from an arsenic study [103]. Problem 8 uses data from a study on ozone exposure in children [113].

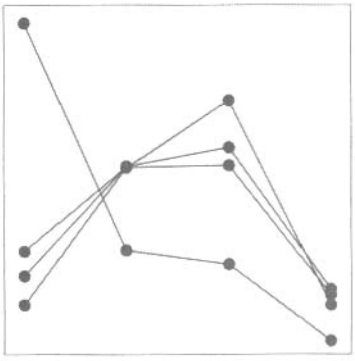
Problems 9, 10, and 11 are applications of multiple linear regression (and other) methods. In each case, the *adjusted* coefficient of determination is used to help choose an appropriate prediction model. Problem 9 uses data from a hydrocarbon-emissions study [90]. Problem 10 uses data from a study of factors affecting plasma beta-carotene levels in women [104]. Problem 11 uses data from a study designed to find an empirical formula for predicting body fat in men using easily measured quantities only [59].

Problem 12 applies the goodness-of-fit analysis in simple linear regression to several data sets from a physical anthropology study [50].

Problems 13 and 14 introduce the use of "dummy" variables in linear regression problems. In Problem 13, the methods are applied to a study of the relationship between age and height in two groups of children [5]. In Problem 14, the methods are applied to a study of the pricing of diamonds [26]. Problem 13 also introduces a permutation method for the same problem.

Note that the use of dummy variables in Problem 13 is an example of a *covariance analysis* and the use of dummy variables in Problem 14 is an example of the analysis of an *unbalanced* two-way layout.

Problems 15 and 16 are applications of locally weighted regression and bootstrap methods. Problem 15 uses data from a study of ozone levels in the greater Los Angeles area [28]. Problem 16 uses data from a cholesterol-reduction study [36].



Chapter 15
Contingency Table Analysis

This chapter introduces methods for analyzing data structured as I-by-J tables of frequencies. The row and column designations in the tables correspond to levels of two factors, and analyses focus on relationships between these factors.

Methods introduced in the first three sections generalize goodness-of-fit and permutation methods from Chapters 6, 11, and 13. Section 4 gives additional methods appropriate for tables with 2 rows and 2 columns. Section 5 outlines the laboratory problems. General references for this chapter are [1], [39], [69].

15.1 Independence analysis

This section considers methods appropriate for a single sample of size N, cross-classified as follows:

	$j=1$	$j=2$	$\cdots$	$j=J$	
$i=1$	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,J}$	$x_{1\cdot}$
$i=2$	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,J}$	$x_{2\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$i=I$	$x_{I,1}$	$x_{I,2}$	$\cdots$	$x_{I,J}$	$x_{I\cdot}$
	$x_{\cdot 1}$	$x_{\cdot 2}$	$\cdots$	$x_{\cdot J}$	N

In this table, $x_{i,j}$ is the number of observations at level i of factor 1 (the *row* factor) and level j of factor 2 (the *column* factor) for all i and j, and the numbers in the margins correspond to observed row and column totals.

15.1.1 Example: Pearson's chi-square test

The data are assumed to summarize N independent trials of a multinomial experiment with IJ outcomes and probabilities

$$p_{i,j} \text{ for } i = 1, 2, \ldots, J, j = 1, 2, \ldots, J,$$

where $p_{i,j}$ is the probability that an observation is at level i of factor 1 and level j of factor 2. Of interest is a test of the null hypothesis that the row and column factors are independent or, equivalently, a test of the null hypothesis

$$p_{i,j} = p_{i\cdot} \times p_{\cdot j} \text{ for } i = 1, 2, \ldots, I \text{ and } j = 1, 2, \ldots, J,$$

where $p_{i\cdot} = \sum_{j=1}^{J} p_{i,j}$ and $p_{\cdot j} = \sum_{i=1}^{I} p_{i,j}$ for each i, j, versus the general alternative that equality does not hold in at least one case.

Theorem 15.1 (Parameter Estimation). *Given the assumptions and definitions above, the following are ML estimators of model parameters:*

1. *Row probabilities:*

$$\widehat{p_{i\cdot}} = \frac{1}{N} X_{i\cdot} = \frac{1}{N} \sum_{j=1}^{J} X_{i,j} \text{ for each } i.$$

2. *Column probabilities:*

$$\widehat{p_{\cdot j}} = \frac{1}{N} X_{\cdot j} = \frac{1}{N} \sum_{i=1}^{I} X_{i,j} \text{ for each } j,$$

where $X_{i,j}$ is the number of observations at level i of the row factor and level j of the column factor, for each i and j.

Pearson's test

Pearson's goodness-of-fit test, introduced in Section 6.3, can be used to test the null hypothesis that the row and column factors are independent. Since $E(X_{i,j}) = Np_{i,j} = Np_{i\cdot}p_{\cdot j}$ under the null hypothesis, the form of the statistic is as follows:

$$\mathbf{X^2} = \sum_{i,j} \frac{(X_{i,j} - N\widehat{p_{i\cdot}}\widehat{p_{\cdot j}})^2}{N\widehat{p_{i\cdot}}\widehat{p_{\cdot j}}}.$$

Since $\sum_i p_{i\cdot} = 1$ and $\sum_j p_{\cdot j} = 1$, there are $(I-1) + (J-1)$ free parameters in the model. If N is large enough, the sampling distribution of $\mathbf{X^2}$ is approximately chi-square with $IJ - 1 - ((I-1) + (J-1)) = (I-1)(J-1)$ degrees of freedom.

Pearson's test can be shown to be an approximate likelihood ratio test.

Example: Alcohol-nicotine study

For example, as part of a study on factors affecting early childhood development, information was collected on 452 young mothers [78, p. 649]. The left part of Table 15.1 classifies the women by their alcohol intake before pregnancy (row factor) and their nicotine intake during pregnancy (column factor). Alcohol intake has four levels: no alcohol used, 0.01–0.10 ounces per day, 0.11–0.99 ounces per day, and 1

Table 15.1. *4-by-3 contingency table (left) and standardized residuals table (right) for the 452 women in the alcohol-nicotine study.*

	$j=1$	$j=2$	$j=3$
$i=1$	105	7	11
$i=2$	58	5	13
$i=3$	84	37	42
$i=4$	57	16	17

	$j=1$	$j=2$	$j=3$
$i=1$	2.45	−2.54	−2.44
$i=2$	0.96	−1.79	−0.26
$i=3$	−2.45	2.80	2.21
$i=4$	−0.45	0.85	0.12

or more ounces per day; nicotine intake has three levels: none, 1–15 milligrams per day, and 16 or more milligrams per day.

For these data, the observed value of Pearson's statistic is 42.25. The observed significance level, based on the chi-square distribution with 6 degrees of freedom, is virtually zero, suggesting a strong association between alcohol intake prior to pregnancy and nicotine intake during pregnancy.

The right part of Table 15.1 displays the standardized residuals for the test. The unusually high standardized residuals in the no alcohol/no nicotine group, and in the groups with 0.01–0.99 ounces per day of alcohol and 1 or more milligrams of nicotine per day, and unusually low standardized residuals in the remaining groups on the first and third rows, suggests that alcohol intake before pregnancy and nicotine intake during pregnancy are positively associated.

Note that the analysis above assumes that the 452 women chosen for the study are a simple random sample from the population of young mothers and that the population size is large enough to allow the table to be analyzed as a random observation from a multinomial model.

> A standard rule of thumb for using a multinomial approximation is that the number of individuals in the simple random sample is less than 5% of the total population size.

15.1.2 Example: Rank correlation test

In some studies, the levels of the row and column factors have a natural ordering. For example, in the alcohol-nicotine study above, the levels of the row factor correspond to increasing use of alcohol and the levels of the column factor correspond to increasing use of nicotine.

If the levels of the row and column factors are ordered, then the table can be analyzed using Spearman's rank correlation statistic, introduced in Section 11.3.2. For the first factor, $x_{1\cdot}$ observations are assumed to be tied at the lowest level, $x_{2\cdot}$ at the next level, etc. Similarly, for the second factor, $x_{\cdot 1}$ observations are assumed to be tied at the lowest level, $x_{\cdot 2}$ at the next level, etc.

The rank correlation test is appropriate under both population and randomization models (as discussed in Section 11.3). The null hypothesis is that any observed association between the factors is due to chance alone versus alternatives that the factors under study are positively or negatively associated.

For the alcohol-nicotine study example (page 228), the observed value of the rank correlation statistic is 0.218. The observed significance level for a two sided test of the null hypothesis of randomness is virtually zero. Since the observed rank correlation is positive, the results suggest a positive association between the factors.

15.2 Homogeneity analysis

This section considers methods appropriate for I row samples of sizes $r_1, r_2, \ldots, r_I$ and total sample size $N = \sum_i r_i$. The cross-classification is as follows:

	$j=1$	$j=2$	$\cdots$	$j=J$	
$i=1$	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,J}$	r_1
$i=2$	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,J}$	r_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$i=I$	$x_{I,1}$	$x_{I,2}$	$\cdots$	$x_{I,J}$	r_I
	$x_{\cdot 1}$	$x_{\cdot 2}$	$\cdots$	$x_{\cdot J}$	N

The I row samples correspond to the levels of the first factor, $x_{i,j}$ is the number of observations in sample i at level j of the second factor for all i and j, and the numbers in the bottom row correspond to observed column totals.

Equivalently, the table can be set up with J column samples of sizes $c_1, c_2, \ldots, c_J$ and total sample size $N = \sum_j c_j$. The numbers along the bottom row would correspond to the fixed column totals, and the numbers along the right column ($x_{i\cdot}$ for $i = 1, 2, \ldots, I$) would correspond to the observed row totals.

15.2.1 Example: Pearson's chi-square test

The data are assumed to summarize I independent random samples. For the i^{th} sample, the data summarize r_i independent trials of a multinomial experiment with J outcomes and probabilities

$$p_{i,j} \text{ for } j = 1, 2, \ldots, J,$$

where $p_{i,j}$ is the probability that an observation in sample i is at level j of the second factor. Of interest is a test of the null hypothesis that the I row models are equal or, equivalently, that

$$p_{1,j} = p_{2,j} = \cdots = p_{I,j} \text{ for } j = 1, 2, \ldots, J$$

versus the general alternative that, for at least one level of the second factor, some probabilities differ.

Let $(p_1, p_2, \ldots, p_j)$ be the model probabilities under the null hypothesis.

Theorem 15.2 (Parameter Estimation). *Given the definitions and assumptions above, the following are ML estimators of the model probabilities under the null*

Table 15.2. *2-by-4 contingency table (upper) and standardized residuals table (lower) for the 253 men in the chemotherapy study.*

	$j=1$	$j=2$	$j=3$	$j=4$
$i=1$	28	45	29	26
$i=2$	41	44	20	20

	$j=1$	$j=2$	$j=3$	$j=4$
$i=1$	−1.17	0.00	0.85	0.57
$i=2$	1.18	0.00	−0.86	−0.57

hypothesis:

$$\widehat{p_j} = \frac{1}{N} X_{\cdot j} = \frac{1}{N} \sum_{i=1}^{I} X_{i,j} \text{ for all } j,$$

where $X_{i,j}$ is the number of observations in sample i at level j of the second factor.

Pearson's test

An extension of Pearson's goodness-of-fit test can be used to test the null hypothesis that the row models are equal. Since $E(X_{i,j}) = r_i p_{i,j} = r_i p_j$ under the null hypothesis, the form of the statistic is as follows:

$$\mathbf{X^2} = \sum_{i,j} \frac{(X_{i,j} - r_i \widehat{p_j})^2}{r_i \widehat{p_j}}.$$

Since $\sum_j p_j = 1$, there are $(J-1)$ free parameters. Theorem 6.5 can be used to demonstrate that if N is large enough, the sampling distribution of $\mathbf{X^2}$ is approximately chi-square with $I(J-1) - (J-1) = (I-1)(J-1)$ degrees of freedom.

Pearson's test can be shown to be an approximate likelihood ratio test.

Example: Chemotherapy study

For example, as part of a study designed to compare treatments for small cell lung cancer, 253 male patients were randomized to one of two treatments [55]. The upper part of Table 15.2 classifies the men by the treatment received (row factor) and their response to treatment (column factor). A total of 128 men received the first treatment, where the same combination of drugs was administered at fixed times during the study period, and a total of 125 men received the second treatment, where three different combinations of drugs were alternated throughout the study period. The four levels of the response correspond to disease progression (the tumor increased in size), no change, partial regression (the tumor decreased in size), and complete remission (the tumor was not detectable).

For these data, the observed value of Pearson's statistic is 4.86. The observed significance level, based on the chi-square distribution with 3 degrees of freedom, is 0.182, suggesting that the response distributions for the two different treatment protocols are equal.

The lower part of Table 15.2 displays the standardized residuals for the test. There are no unusual standardized residuals.

Note that the analysis above assumes that the 253 men chosen for the study are a simple random sample from the population of men with small cell lung cancer and that the population size is large enough to allow the table to be analyzed as independent observations from two multinomial models.

> If the number of individuals in the simple random sample is less than 5% of the total population size, and if randomization is used to determine treatment assignments, then the table can be treated as independent observations from two multinomial models.

Comparison of Pearson statistics and tests

Given an I-by-J table of frequencies, the value of Pearson's statistic for the tests of independence, equality of I row models, and equality of J column models is the same. In each case, the estimate of $E(X_{i,j})$ reduces to the following:

$$(i^{\text{th}} \text{ row total}) \times (j^{\text{th}} \text{ column total})/N.$$

Further, if the null hypothesis is true and N is large, each statistic has an approximate chi-square distribution with $(I-1)(J-1)$ degrees of freedom.

For these reasons, most computer programs do not distinguish among these three types of tests for I-by-J contingency tables.

15.2.2 Example: Kruskal–Wallis test

In some studies, the levels of the column factor have a natural ordering. For example, in the chemotherapy study above, the levels are ordered by the change in tumor size during the study period.

If the levels of the column factor are ordered, then the table can be analyzed using the Kruskal–Wallis statistic, introduced in Section 13.1.2. In the test, $x_{\cdot 1}$ observations are assumed to be tied at the lowest level of the second factor, $x_{\cdot 2}$ at the next level, etc.

The Kruskal–Wallis test is appropriate under both population and randomization models. The null hypothesis is that observed differences in the I samples are due to chance alone versus the alternative that, for at least two samples, the values in one tend to be larger or smaller than those in the other.

For the chemotherapy study example (page 231), the observed value of the Kruskal–Wallis statistic is 4.26. The observed significance level, based on the chi-square distribution with 1 degree of freedom, is 0.039, suggesting that the response distributions differ. Since about 43% (55/128) of the men in the first treatment group had partial to full remission, compared to about 32% (40/125) of the men in the second treatment group, the first treatment protocol appears to be better.

By including the ordering of the response factor, the Kruskal–Wallis test was able to demonstrate a significant difference in the treatments at the 5% significance level, although the observed differences are not large.

15.3 Permutation chi-square tests

The chi-square approximation to the sampling distribution of Pearson's statistic is adequate when all estimated cell expectations are 5 or more. When some expectations are less than 5, permutation methods can be used to estimate a p value.

Independence analysis

When the data are a single sample of size N, the idea is to think of the observations as a list of N pairs,

$$x_{1,1} \text{ copies of } (1,1), \text{ followed by } x_{1,2} \text{ copies of } (1,2), \text{ and so forth.}$$

(The paired data are ordered by the levels of the row factor.) Let $\underline{v}$ be the list of first coordinates and $\underline{w}$ be the list of second coordinates.

The permutation distribution of $\mathbf{X^2}$ is obtained as follows: for each matching of a permutation of the $\underline{w}$ list to the ordered $\underline{v}$ list, an I-by-J table is constructed, and the value of $\mathbf{X^2}$ is computed.

The permutation chi-square test is appropriate under both population and randomization models. The null hypothesis is that any observed association between the factors is due to chance alone versus the general alternative that the association is not due to chance alone. Monte Carlo analysis is used to estimate p values in most situations.

Homogeneity analysis

When the data are I samples of sizes $r_1, r_2, \ldots, r_I$, the idea is to think of the observations in the i^{th} sample as a list of r_i values,

$$x_{i,1} \text{ copies of } 1, \text{ followed by } x_{i,2} \text{ copies of } 2, \text{ and so forth}$$

for each i. The permutation distribution of $\mathbf{X^2}$ is obtained as follows: for each partition of the I samples into distinguishable groups of sizes $r_1, r_2, \ldots, r_I$, an I-by-J table is formed, and the value of $\mathbf{X^2}$ is computed.

The permutation chi-square test is appropriate under both population and randomization models. The null hypothesis is that any observed differences in the I samples are due to chance alone versus the general alternative that observed differences are not due to chance alone. Monte Carlo analysis is used to estimate p values in most situations.

A similar analysis is possible if the data are a collection of J column samples.

Table 15.3. *2-by-2 contingency tables for the red dye study. The left table corresponds to animals who died before the end of the study period. The right table corresponds to animals who were sacrificed at the end of the study.*

	$j=1$	$j=2$
$i=1$	4	26
$i=2$	7	16

	$j=1$	$j=2$
$i=1$	0	14
$i=2$	7	14

Fixed row and column totals

In all three cases (the data are a single sample of size N, a collection of I row samples, or a collection of J column samples), each reordering of the data will produce an I-by-J table with the same row and column totals as the original table.

Most computer programs do not distinguish among the three types of permutation tests for I-by-J contingency tables.

Example: Red dye study

To illustrate the computations, consider the data in Table 15.3, collected as part of a study to determine if the food additive red dye 2 was a carcinogen [39, p. 53]. A total of 88 rats were randomly assigned to two different dosage groups: 44 were fed a low dosage of the food additive and 44 were fed a high dosage. The left part of the table corresponds to the 53 animals who died before the end of the study period and the right part to the remaining 35 animals who were sacrificed at the end of the study. In each part of the table, the levels of the row factor correspond to dosage level (low dosage, high dosage), and the levels of the column factor correspond to the presence or absence of tumors.

One way to determine if the information in the tables can be combined is to conduct a test of the homogeneity of row models in the following 2-by-4 table, where the rows correspond to the left and right parts of Table 15.3:

	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$
t_1	4	26	7	16
t_2	0	14	7	14

Since the estimated cell expectations in the first column are quite small, a permutation chi-square test was conducted. The observed value of Pearson's statistic is 4.23, and the permutation p value is 0.242, suggesting that the tables can be combined.

The left part of the table below classifies all 88 animals by low or high dosage group (row factor) and the presence or absence of tumors (column factor), and the right part shows the standardized residuals for a test of the equality of row models.

	$j=1$	$j=2$
$i=1$	4	40
$i=2$	14	30

	$j=1$	$j=2$
$i=1$	−1.67	0.85
$i=2$	1.67	−0.85

All estimated cell expectations in the combined table are greater than 5. The observed value of Pearson's statistic is 6.98. The observed significance level, based on the chi-square distribution with 1 degree of freedom, is 0.008, suggesting that red dye 2 is a carcinogen. Note that 9.1% (4/44) of the rats in the low dosage group developed tumors, compared to 35.0% (14/44) in the high dosage group.

15.4 Fourfold tables

This section considers additional methods for 2-by-2 contingency tables.

15.4.1 Odds ratio analysis

Assume that the levels of the first factor correspond to whether or not event A has occurred and the levels of the second factor correspond to whether or not event B has occurred, where A and B are events with $0 < P(A) < 1$ and $0 < P(B) < 1$.

Positive and negative association

A and B are said to be *positively associated* if one of the following equivalent conditions holds:

$$P(A \cap B) > P(A)P(B) \text{ or } P(B|A) > P(B|A^c) \text{ or } P(A|B) > P(A|B^c).$$

Similarly, A and B are said to be *negatively associated* if

$$P(A \cap B) < P(A)P(B) \text{ or } P(B|A) < P(B|A^c) \text{ or } P(A|B) < P(A|B^c).$$

Otherwise, A and B are independent.

Odds; odds ratio

If E is an event with $0 < P(E) < 1$, the *odds* of event E is defined to be the ratio of the probability of the event to the probability of its complement:

$$\text{Odds}(E) = \frac{P(E)}{P(E^c)}.$$

If A and B are events satisfying $0 < P(A) < 1$ and $0 < P(B) < 1$, the *odds ratio* (OR) is defined as the ratio of the odds of B given A to the odds of B given A^c (equivalently, the ratio of the odds of A given B to the odds of A given B^c):

$$\text{OR} = \frac{\text{Odds}(B|A)}{\text{Odds}(B|A^c)} = \frac{\text{Odds}(A|B)}{\text{Odds}(A|B^c)}.$$

Using the definition of conditional probability, each expression on the right above reduces to

$$\text{OR} = \frac{P(A \cap B)P(A^c \cap B^c)}{P(A \cap B^c)P(A^c \cap B)}.$$

The odds ratio can be used to measure the strength of the association between two events. Specifically, if the events are independent, then OR $= 1$; if the events are positively associated, then OR > 1; and if the events are negatively associated, then OR < 1.

For example, the following table shows values of the odds ratio for events with $P(A) = 0.30$, $P(B) = 0.40$, and $P(A \cap B) = 0.03, 0.06, \ldots, 0.21$:

$P(A \cap B)$	0.03	0.06	0.09	0.12	0.15	0.18	0.21
OR	0.10	0.26	0.54	1.00	1.80	3.27	6.26

Note that A and B are independent when $P(A \cap B) = P(A)P(B) = 0.12$.

Estimation

The odds ratio is an important measure of the association between events because it can be estimated in each of the following situations:

1. The data summarize a random sample of size N from a model with four outcomes. The four probabilities are
$$p_{1,1} = P(A \cap B), \quad p_{1,2} = P(A \cap B^c), \quad p_{2,1} = P(A^c \cap B), \quad p_{2,2} = P(A^c \cap B^c).$$
2. The data summarize independent random samples of sizes r_1 and r_2. The first row model has probabilities
$$p_{1,1} = P(B|A), \quad p_{1,2} = P(B^c|A),$$
and the second row model has probabilities
$$p_{2,1} = P(B|A^c), \quad p_{2,2} = P(B^c|A^c).$$
3. The data summarize independent random samples of sizes c_1 and c_2. The first column model has probabilities
$$p_{1,1} = P(A|B), \quad p_{2,1} = P(A^c|B),$$
and the second column model has probabilities
$$p_{1,2} = P(A|B^c), \quad p_{2,2} = P(A^c|B^c).$$

Let $X_{1,1}$, $X_{1,2}$, $X_{2,1}$, and $X_{2,2}$ be the number of observations in each cell of the 2-by-2 table.

Theorem 15.3 (Odds Ratio Estimation). *Under the assumptions of this section and if each $X_{i,j} > 0$, then the ML estimator of the odds ratio is*

$$\widehat{OR} = \frac{X_{1,1}X_{2,2}}{X_{1,2}X_{2,1}}.$$

Further, if each $X_{i,j}$ is large, then the distribution of the natural logarithm of the ML estimator, $\log(\widehat{OR})$, is approximately normal with mean $\log(OR)$ and standard deviation equal to the square root of the sum of the reciprocals

$$\sqrt{\frac{1}{X_{1,1}} + \frac{1}{X_{1,2}} + \frac{1}{X_{2,1}} + \frac{1}{X_{2,2}}}.$$

The odds ratio is often called the *cross-product ratio*. In each of the sampling situations above, $\text{OR} = (p_{1,1}p_{2,2})/(p_{1,2}p_{2,1})$. The ML estimate of OR follows the same pattern. That is, the estimate is the ratio of the product of the numbers on the main diagonal of the 2-by-2 contingency table to the product of the numbers on the off diagonal.

Approximate confidence intervals for odds ratio

Theorem 15.3 can be used to demonstrate that an approximate $100(1-\alpha)\%$ confidence interval for the odds ratio has the form

$$\exp\left(\log\left(\widehat{\text{OR}}\right) \pm z(\alpha/2)\sqrt{\frac{1}{X_{1,1}}+\frac{1}{X_{1,2}}+\frac{1}{X_{2,1}}+\frac{1}{X_{2,2}}}\right),$$

where log() is the natural logarithm function, exp() is the exponential function, and $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point on the standard normal distribution.

Example: Alcohol-nicotine study

Consider again the alcohol-nicotine study from page 228. Let A be the event that a mother did not use alcohol before becoming pregnant and B be the event that a mother did not smoke during pregnancy. The left part of the table below classifies the women using a 2-by-2 contingency table, and the right part shows the standardized residuals for a test of independence of events A and B.

	B	B^c
A	105	18
A^c	199	130

	B	B^c
A	2.45	−3.51
A^c	−1.50	2.15

For these data, the observed value of Pearson's statistic is 25.16. The observed significance level, based on the chi-square distribution with 1 degree of freedom, is virtually zero. The unusually large standardized residuals in the upper left and lower right corners suggest that the events are positively associated. Note that the estimate of $P(A)$ is 0.272 (123/452), of $P(B)$ is 0.673 (304/452), and of $P(A \cap B)$ is 0.232 (105/452).

The observed odds ratio is 3.90. An approximate 95% confidence interval for the odds ratio is [2.21, 6.58].

Example: Vitamin C study

As part of a study to determine the therapeutic value of vitamin C for treating the common cold, 279 skiers were randomized to one of two study groups: 139 received 1 gram of vitamin C per day during the study period, and 140 received "placebo" tablets (with no active ingredients) each day during the study period. Of interest is the relative frequencies of colds for the two groups [39, p. 8]. Let A be the event that the skier was in the vitamin C group and B be the event that the skier got a cold. The left part of the table below classifies the skiers using a 2-by-2 contingency table,

and the right part shows the standardized residuals for a test of the equality of row models.

	B	B^c
A	17	122
A^c	31	109

	B	B^c
A	−1.41	0.64
A^c	1.41	−0.64

For these data, the observed value of Pearson's statistic is 4.81. The observed significance level, based on the chi-square distribution with 1 degree of freedom, is 0.028. Although none of the standardized residuals is unusually large or small, the pattern of signs suggests that the events are negatively associated (equivalently, that vitamin C has some therapeutic value). Note that the estimate of $P(B|A)$ is 0.122 (17/139) and of $P(B|A^c)$ is 0.221 (31/140).

The observed odds ratio is 0.49. An approximate 95% confidence interval for the odds ratio is [0.26, 0.93].

15.4.2 Small sample analyses

This section introduces a permutation method, developed by R. A. Fisher in the 1930's, appropriate for tests of independence (or equality of row models or equality of column models) when sample sizes are small, and a small sample odds ratio confidence procedure. Analyses are done conditional on the row and column totals. Let r_1, r_2, c_1, c_2 be the fixed totals.

Fisher exact test

Following the general permutation strategy of Section 15.3, 2-by-2 tables can be constructed conditional on the row and column totals. The possible tables can be indexed using a single variable, X, as follows:

	B	B^c	
A	X	$r_1 - X$	r_1
A^c	$c_1 - X$	$N - r_1 - c_1 + X$	r_2
	c_1	c_2	N

Under all three sampling situations listed on page 236, the conditional distribution of X is hypergeometric with PDF

$$f(x) = P(X = x \mid r_1, r_2, c_1, c_2) = \frac{\binom{r_1}{x}\binom{r_2}{c_1 - x}}{\binom{N}{c_1}} = \frac{\binom{c_1}{x}\binom{c_2}{r_1 - x}}{\binom{N}{r_1}}$$

for all x in the range of the random variable (and 0 otherwise).

Fisher's test uses the PDF itself as test statistic. The p value for the test is the probability of observing a table as likely or less likely than the observed table,

$$P(f(X) \le f(x_{\text{obs}})),$$

where x_{obs} is the number of observations in the upper left corner of the table.

For example, the table below is based on information from a retrospective study of risk factors for cervical cancer [50, p. 247]. A total of 90 women ages 50 to 59 are represented: 14 women with cervical cancer (event B) and 76 women without the disease (event B^c). The rows of the table are related to levels of a potential risk factor, age at first pregnancy, where A corresponds to age at first pregnancy of 25 years or younger, and A^c corresponds to age at first pregnancy after age 25.

	B	B^c
A	13	46
A^c	1	30

A total of 15 tables have row totals 59, 31 and column totals 14, 76. Conditional on these totals, the probability of the observed table is 0.014 and the p value is $P(f(X) \leq 0.014) = 0.029$, suggesting a link between age at first pregnancy and cervical cancer. Note that the estimate of $P(A|B)$ is 0.93 (13/14) and $P(A|B^c)$ is 0.61 (46/76).

Odds ratio confidence procedure

Let X be the number in the upper left corner of the table and λ be the odds ratio. Assume that all row and column totals are positive.

Theorem 15.4 (Distribution Theorem). *Given the definitions and assumptions above, the conditional PDF of X has the form*

$$f(x) = P(X = x \mid r_1, r_2, c_1, c_2) = \frac{\binom{r_1}{x}\binom{r_2}{c_1 - x}\lambda^x}{D}$$

for all x in the range of the random variable (and 0 otherwise), where λ is the odds ratio and the denominator, D, is a sum of terms

$$D = \sum_y \binom{r_1}{y}\binom{r_2}{c_1 - y}\lambda^y$$

taken over all possible values of X. This conditional distribution is valid under all three sampling situations listed on page 236.

Confidence intervals are obtained using computer estimation and the conditional distribution above. The process is outlined in [1, p. 67] and has been implemented. For the cervical cancer example, a 95% confidence interval for the odds ratio is [1.14, 372.1], suggesting a positive association between A and B.

15.5 Laboratory problems

Laboratory problems for this chapter introduce *Mathematica* commands for analyzing contingency tables using both large sample and permutation methods. Problems are designed to reinforce the ideas of this chapter.

15.5.1 Laboratory: Contingency table analysis

In the main laboratory notebook (Problems 1 to 7), you will use simulation to study Pearson's chi-square, rank correlation, and Kruskal–Wallis tests for I-by-J tables; solve a problem involving odds ratios and probabilities in fourfold tables; and apply a variety of computational and graphical methods to four data sets: (1) data on the relationship between employment and marriage in men ages 25 to 44 [43], (2) data on the relationship between age at diagnosis and frequency of breast self-exam in women with breast cancer [96], (3) data from an experiment on treatments for nausea following surgery [68], and (4) data on smoking habits in male patients with lung cancer and with diseases other than lung cancer [1], [34].

15.5.2 Additional problem notebooks

Problems 8 through 13 are applications of large sample contingency table (and other) methods. Problem 8 uses data from a study on the relationship between disease and nutritional status in poor children [78]. Problem 9 uses data from a study on effects of smoking during pregnancy [21]. Problem 10 is a whimsical application of a variety of methods to the 1998 home run race between Mark McGwire of the St. Louis Cardinals and Sammy Sosa of the Chicago Cubs [98]. Problem 11 uses data from a study of factors influencing self-esteem in high school students [39]. Problem 12 uses data from a study of factors influencing hypertension in medical patients [69]. Problem 13 applies a variety of methods to study potential sex bias in graduate school admissions [14], [43]. The application in Problem 13 is an example of *Simpson's paradox*.

Problem 14 introduces the *risk ratio* and applies odds ratio and risk ratio methods to data from the Physicians' Health Study [84]. Problem 15 introduces *McNemar's test* for paired samples and applies the method to data from a study of the relationship between cellular telephone use and motor vehicle collisions [88].

Problem 16 considers small sample methods for fourfold tables and applies these methods to data on insulin dependence in diabetic patients [85].

Problem 17 introduces a method to construct the complete permutation distribution of Pearson's statistic in 2-by-J tables and applies a variety of methods to data from an ecology study [20]. Problem 18 demonstrates the general permutation strategy for analysis of I-by-J tables and applies a variety of methods to data from an ecology study [37].

Bibliography

[1] Agresti, A. (1990). Categorical data analysis. New York: John Wiley & Sons, Inc.

[2] Allison, D., Heshka, S., Sepulveda, D., & Heymsfield, S. (1993). Counting calories: Caveat emptor. *Journal of the American Medical Association*, 279, 1454–1456.

[3] Allison, T., & Cicchetti, D. V. (1976). Sleep in mammals: Ecological and constitutional correlates. *Science*, 194, 732–734.

[4] Anderson, J. W., Spencer, D. B., Hamilton, C. C., Smith, S. F., Tietyen, J., Bryant, C. A., & Oeltgen, P. (1990). Oat-bran cereal lowers serum total and LDL cholesterol in hypercholesterolemic men. *American Journal of Clinical Nutrition*, 52, 495–499.

[5] Anderson, S., Auquier, A., Hauck, W. W., Oakes, D., Vandaele, W., & Weisberg, H. I. (1980). Statistical methods for comparative studies. New York: John Wiley & Sons, Inc.

[6] Andrews, D. F., & Herzberg, A. M. (1985). Data: A collection of problems from many fields for the student and research worker. New York: Springer-Verlag.

[7] Bailer, A. J., & Oris, J. T. (1994). Assessing toxicity of pollutants in aquatic systems. In N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, & J. Greenhouse (Eds.), Case studies in biometry (pp. 25–40). New York: John Wiley & Sons, Inc.

[8] Bajorski, P., & Petkau, J. (1999). Non-parametric comparisons of changes on ordinal responses. *Journal of the American Statistical Association*, 94, 970–978.

[9] Baker, S. G. (1990). A simple EM algorithm for capture-recapture data with categorical covariates. *Biometrics*, 46, 1193–1200.

[10] Barlow, R. E., Toland, R. H., & Freeman, T. (1984). A Bayesian analysis of stress-rupture life of Kevlar/epoxy spherical pressure vessels. In T. D. Dwivedi (Ed.), Proceedings of the Canadian conference in applied statistics. New York: Marcel-Dekker.

[11] Barnicot, N. A., & Brothwell, D. R. (1959). The evaluation of metrical data in the comparison of ancient and modern bones. In G. E. W. Wolstenholme, & C. M. O'Connor (Eds.), Medical biology and Etruscan origins. Boston: Little, Brown, and Company.

[12] Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York: John Wiley & Sons, Inc.

[13] Berger, P. (1973). On the distribution of hand patterns in bridge: Man-dealt versus computer-dealt. *Canadian Journal of Statistics*, 1, 261–266.

[14] Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187, 398–404.

[15] Birnbaum, Z. W., & Saunders, S. C. (1958). A statistical model for the life-length of materials. *Journal of the American Statistical Association*, 53, 151–160.

[16] Bjerkdal, T. (1960). Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli. *American Journal of Hygiene*, 72, 130–148.

[17] Box, G. E. P., Hunter, W. G., & Hunter, S. J. (1978). Statistics for experimenters. New York: John Wiley & Sons, Inc.

[18] Brousseau, D., & Baglivo, J. (2000). Modeling seasonal proliferation of the parasite *Perkinsus marinus (Dermo)* in field populations of the oyster *Crassostrea virginica*. *Journal of Shellfish Research*, 19, 133–138.

[19] Brousseau, D., Filipowicz, A., & Baglivo, J. (2001). Laboratory investigations of the effects of predator sex and size on prey selection by the Asian shore crab, *Hemigrapsus sanguineus*. *Journal of Experimental Marine Biology and Ecology*, 262, 199–210.

[20] Brousseau, D., Baglivo, J., Filipowicz, A., Sego, L., & Alt, C. (2002). An experimental field study of site fidelity and mobility in the Asian shore crab, *Hemigrapsus sanguineus*. *Northeastern Naturalist*, 9, 381–390.

[21] Brown, P. J., Stone, J., & Ord-Smith, C. (1983). Toxaemic signs during pregnancy. *Applied Statistics*, 32, 69–72.

[22] Burgess, E. W., & Cottrell, L. S. (1955). The prediction of adjustment in marriage. In P. F. Lazarsfeld, & M. Rosenberg (Eds.), The language of social research (pp. 267–276). Glencoe, IL: The Free Press.

[23] Campbell, J. A., & Pelletier, O. (1962). Determinations of niacin (niacinamide) in cereal products. *Journal of the Association of Official Analytical Chemists*, 45, 449–453.

[24] Carver, W. A. (1927) A genetic study of certain chlorophyll deficiencies in maize, *Genetics*, 12, 126–134.

[25] Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). Graphical methods for data analysis. Belmont, CA: Wadsworth International Group.

[26] Chu, S. (2001). Pricing the c's of diamond stones. *Journal of Statistics Education.* Available at http://www.amstat.org/publications/jse/v9n2/datasets.chu.html.

[27] Clarke, R. D. (1946). An application of the Poisson distribution. *Journal of the Institute of Actuaries*, 72, 481.

[28] Cook, R. D., & Weisburg, S. (1994). An introduction to regression graphics. New York: John Wiley & Sons, Inc.

[29] Cressie, N. A. C. (1982). Playing safe with misweighted means. *Journal of the American Statistical Association*, 77, 754–759.

[30] The Data and Story Library. (1996). Available at FTP: lib.stat.cmu.edu/DASL/

[31] Davison, A. C., & Hinkley, D. V. (1997). Bootstrap methods and their application. Cambridge, England: Cambridge University Press.

[32] Denby, L., & Tukey, P. A. (1990). Kodiak island king crab data analysis exposition. In 1990 Proceedings of the Section on Statistical Graphics (pp. 94–98). Alexandria, VA: American Statistical Association.

[33] Diaconis, P. (1988). Group representations in probability and statistics. Hayward, CA: Institute of Mathematical Statistics.

[34] Doll, R., & Hill, A. B. (1952). A study of the aetiology of carcinoma of the lung. *British Medical Journal*, 2, 1271–1286.

[35] Ebel, J. (personal communication).

[36] Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. New York: Chapman & Hall, Inc.

[37] Erickson, W. P., Nick, T. G., & Ward, D. H. (1998). Investigating flight response of Pacific brant to helicopters at Izembek Lagoon, Alaska by using logistic regression. In R. Peck, L. D. Haugh, & A. Goodman (Eds.), Statistical case studies: A collaboration between academe and industry (pp. 155–170). Alexandria, VA: American Statistical Association.

[38] Fienberg, S. E. (1971). Randomization and social affairs: The 1970 draft lottery. *Science*, 171, 255–261.

[39] Fienberg, S. E. (1980). The analysis of cross-classified categorical data, second edition. Cambridge, MA: The MIT Press.

[40] Fine, J. B., & Bosch, R. J. (2000). Risk assessment with a robust probit model, with application to toxicology. *Journal of the American Statistical Association*, 95, 375–382.

[41] Fisher, R. A. (1935). The design of experiments. Edinburgh: Oliver & Boyd.

[42] Ford, S. (personal communication).

[43] Freedman, D., Pisani, R., Purves, R., & Adhikari, A. (1991). Statistics, second edition. New York: W. W. Norton & Company.

[44] Garthwaite, P. H. (1996). Confidence intervals for randomization tests. *Biometrics*, 52, 1387–1393.

[45] Gastwirth, J. (1987). The statistical precision of medical screening procedures. *Statistical Science*, 3, 213–222.

[46] Geissler, A. (1889). Beitrage zur Frage des Geschlechtsverhaltnisses der Gerboren. Zeitschrift des Koniglichen Sachsischen Statistichen Bureaus, 35, 1–24.

[47] Goldberg, J. D., & Wittes, J. T. (1978). The estimation of false negatives in medical screening. *Biometrics*, 34, 77–86.

[48] Habermam, S. J. (1978). Analysis of qualitative data: Vol. 1. Introductory topics. New York: Academic Press.

[49] Hamilton, L. C. (1992). Regression with graphics: A second course in applied statisics. Pacific Grove, CA: Brooks/Cole Publishing Company.

[50] Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., & Ostrowski, E. (1994). A handbook of small data sets. New York: Chapman & Hall.

[51] Hastings, K. J. (1997). Probability and statistics. New York: Addison-Wesley.

[52] Heckman, M. (1960). Flame photometric determination of calcium in animal feeds. *Journal of the Association of Official Analytical Chemists*, 43, 337–340.

[53] Heinz, G., Johnson, R. W., & Kerk, C. J. (2003). Exploring relationships in body measurements. *Journal of Statistics Education.* Available at http://www.amstat.org/publications/jse/v11n2/datasets.heinz.html.

[54] Hoeting, J. A., & Olsen, A. R. (1998). Are the fish safe to eat? Assessing mercury levels in fish in Maine lakes. In R. Peck, L. D. Haugh, & A. Goodman (Eds.), Statistical case studies: A collaboration between academe and industry (pp. 1–13). Alexandria, VA: American Statistical Association.

[55] Holtbrugge, W., & Schumacher, M. (1991). A comparison of regression models for the analysis of ordered categorical data. *Applied Statistics*, 40, 249–259.

[56] Hot dogs: There's not much good about them except the way they taste (June 1986). *Consumer Reports*, 51, 364–367.

[57] Jarrett, R. G. (1979). A note on the intervals between coal-mining disasters. *Biometrika*, 66, 191–193.

[58] Johnson, B. A. (1990). Red king crab catch per unit effort and spatial distribution. In 1990 Proceedings of the Section on Statistical Graphics (pp. 165–172). Alexandria, VA: American Statistical Association.

[59] Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education,* 4(1). Available by e-mail: archive@jse.stat.ncsu.edu, message: send jse/v4n2/datasets.johnson.

[60] Kafka, A. L., & Miller, P. E. (1996). Seismicity in the area surrounding two mesozoic rift basins in the northeastern United States. *Seismological Research Letters*, 67, 69–86.

[61] Kalwani, M. U., & Yim, C. K. (1992). Consumer price and promotion expectations: An experimental study. *Journal of Marketing Research*, 29, 90–100.

[62] Keller, J. (1973). A theory of competitive running. *Physics Today*, 26, 42–47.

[63] Khatri, B. O., McQuillen, M. P., Harrington, G. J., Schmoll, D., & Hoffman, R. G. (1985). Chronic progressive multiple sclerosis: Double-blind controlled study of plasmapheresis in patients taking immunosuppressive drugs. *Neurology*, 35, 312–319.

[64] Koopmans, L. (1987) Introduction to contemporary statisical methods. New York: Duxbury Press.

[65] Larsen, R. J., & Marx, M. L. (1986). An introduction to mathematical statistics and its applications, second edition. Englewood Cliffs, NJ: Prentice-Hall.

[66] Le Cam, L., & Neyman, J. (Eds.). (1967). Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Volume V: Weather modification. Berkeley: University of California Press.

[67] Lee, C., & Matzo, G. (1998). An evaluation of process capability for a fuel injector process using Monte Carlo simulation. In R. Peck, L. D. Haugh, & A. Goodman (Eds.), Statistical case studies: A collaboration between academe and industry (pp. 247–266). Alexandria, VA: American Statistical Association.

[68] Lehmann, E. L. (1975). Nonparametrics: Statistical methods based on ranks. San Francisco: Holden-Day, Inc.

[69] Leonard, T. (2000). A course in categorical data analysis. Boca Raton, FL: Chapman & Hall/CRC.

[70] Leung, M., Schachtel, G., & Yu, H. (1994). Scan statistics and DNA sequence analysis: The search for an origin of replication in a virus. *Nonlinear World,* 1, 445–471.

[71] Maguire, B. A., Pearson, E. S., & Wynn, A. H. A. (1952). The time intervals between industrial accidents. *Biometrika*, 39, 168–180.

[72] Magurran, A. E. (1988). Ecological diversity and its measurement. Princeton, NJ: Princeton University Press.

[73] Manly, B. F. J. (1997). Randomization, bootstrap and Monte Carlo methods in biology, second edition. London: Chapman & Hall.

[74] Maritz, J. S. (1995). Distribution-free statistical methods, second edition. London: Chapman & Hall.

[75] Matui, I. (1932). Statistical study of the distribution of scattered villages in two regions of the Tonami plain, Toyama prefecture. *Japanese Journal of Geology and Geography*, 9, 251–266.

[76] Miller, A. J., Shaw, D. E., & Veitch, L. G. (1979). Analyzing the results of a cloud-seeding experiment in Tasmania. *Communications in Statistics, Theory and Methods*, A8, 1017–1047.

[77] Miller, M. D., Wilkinson, J. L., & Willemain, T. R. (1990). Aspects of the Alaskan king crab life cycle. In 1990 Proceedings of the Section on Statistical Graphics (pp. 114–117). Alexandria, VA: American Statistical Association.

[78] Moore, D. S., & McCabe, G. P. (1999). Introduction to the practice of statistics, third edition. New York: W. H. Freeman and Company.

[79] Nolan, D. (1995). Lab 8: Herpes, DNA, and statistics. Course notes for Statistics 102, University of California at Berkeley.

[80] Olkin, I., Gleser, L. J., & Derman, C. (1994). Probability models and applications, second edition. New York: Macmillan College Publishing Company.

[81] Official sites of the International Olympic Committee. Available at www.olympic.org.

[82] Pagano, M., & Gauvreau, K. (1993). Principles of biostatistics. Belmont, CA: Duxbury Press.

[83] Pearson, K., & Lee, A. (1903). On the laws of inheritance in man. I. Inheritance of physical characters. *Biometrika*, 2, 357–462.

[84] Physicians' Health Study Group (1989). Final report on the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine*, 321, 129–135.

[85] Plackett, R. L. (1981). The analysis of categorical data, second edition. New York: Oxford University Press.

[86] Powell, G. L., & Russell, A. P. (1984). The diet of the eastern short-horned lizard in Alberta and its relationship to sexual size dimorphism. *Canadian Journal of Zoology*, 62, 428–440.

[87] Ramsey, F. L., McCracken, M., Crawford, J. A., Drut, M. S., & Ripple, W. J. (1994). Habitat association studies of the northern spotted owl, sage grouse, and flammulated owl. In N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, & J. Greenhouse (Eds.), Case studies in biometry (pp. 189–210). New York: John Wiley & Sons, Inc.

[88] Redelmeier, D. A., & Tibshirani, R. J. (1997). Association between cellular-telephone calls and motor vehicle collisions. *New England Journal of Medicine*, 336, 453–458.

[89] Regal, R. R., & Hook E. B. (1999). An exact test for all-way interaction in a 2^M contingency tale: Application to interval capture-recapture estimation of population size. *Biometrics*, 55, 1241–1246.

[90] Rice, J. A. (1995). Mathematical statistics and data analysis, second edition. Belmont, CA: Duxbury Press.

[91] Richardson, L. F. (1944) The distribution of wars in time. *Journal of the Royal Statistical Society*, 107, 242–250.

[92] Ricker, W. E. (1975). Computation and interpretation of biological statistics of fish populations. *Bulletin of the Fisheries Research Board of Canada*, 191, 83–86.

[93] Rutherford, E., & Geiger, M. (1910). The probability variations in the distribution of alpha-particles. *Philosophical Magazine*, Series 6, 20, 698–704.

[94] Schall, J. J., Bennett, A. F., & Putman, R. W. (1982). Lizards infected with malaria: Physiological and behavioral consequences. *Science*, 217, 1057–1059.

[95] Scott, D. W., Gotto, A. M., Cole, J. S., & Gorry, G. A. (1978). Plasma lipids as collateral risk factors in coronary artery disease: A study of 371 males with chest pain. *Journal of Chronic Diseases*, 31, 337–345.

[96] Senie, R. T., Rosen, P. P., Lesser, M. L., & Kinne, D. W. (1981). Breast self-examinations and medical examination relating to breast cancer stage. *American Journal of Public Health*, 71, 583–590.

[97] Shoemaker, A. L. (1996). What's normal? — Temperature, gender, and heart rate. *Journal of Statistics Education,* 4(2). Available by e-mail: archive@jse.stat.ncsu.edu, message: send jse/v4n2/datasets.shoemaker.

[98] Simonoff, J. S. (1998). Move over Roger Maris: Breaking baseball's most famous record. *Journal of Statistics Education,* 6(3). Available by e-mail: archive@jse.stat.ncsu.edu, message: send jse/v6n3/datasets.simonoff.

[99] Simpson, J., Olsen, A., & Eden, J. C. (1975). A Bayesian analysis of a multiplicative treatment effect in weather modification. *Technometrics*, 17, 161–166.

[100] Smolowitz, R. (personal communication).

[101] Solow, A. (1993). A simple test for change in community structure. *Journal of Animal Ecology*, 62, 191–193.

[102] Stuart, A. (1955) A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42, 412–416.

[103] Stukel, T. (1998a). Toenail samples as an indicator of drinking water arsenic exposure. Available at FTP: lib.stat.cmu.edu/datasets/.

[104] Stukel, T. (1998b). Determinants of plasma retinol and beta-carotene levels. Available at FTP: lib.stat.cmu.edu/datasets/.

[105] Syrjala, S. (1996). A statistical test for the difference between the spatial distributions of two populations. *Ecology*, 77, 75–80.

[106] Terman, L. M. (1919). The intelligence of school children. Boston: Houghton Mifflin Company.

[107] Thomson, A., & Randall-McGiver, R. (1905). Ancient races of the Thebaid. Oxford: Oxford University Press.

[108] Tibshirani, R. J. (1997). Who is the fastest man in the world? *The American Statistician*, 51, 106–111.

[109] Uhlenhuth, E. H., Lipman, R. S., Galter, M. B., & Stern, M. (1974). Symptom intensity and life stress in the city. *Archives of General Psychiatry*, 31, 759–764.

[110] United States Department of Agriculture (1991). Agricultural statistics.

[111] Waller, L., Turnbull, G., Clark, L., & Nasca, P. (1994). Spatial pattern analyses to detect rare disease clusters. In N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, & J. Greenhouse (Eds.), Case studies in biometry (pp. 3–24). New York: John Wiley & Sons, Inc.

[112] Wolfram, S. (2003). The Mathematica book, fifth edition. Wolfram Media, Inc.

[113] Wypij, D., & Liu, L.-J. S. (1994). Prediction models for personal ozone exposure assessment. In N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, & J. Greenhouse (Eds.), Case studies in biometry (pp. 41–56). New York: John Wiley & Sons, Inc.

[114] Yu, C., Waller, L., & Zelterman, D. (1998). Discrete distributions for use in twin studies. *Biometrics*, 54, 547–557.

[115] Ziv, G., & Sulman, F. G. (1972). Binding of antibiotics to bovine and ovine serum. *Antimicrobial Agents and Chemotherapy*, 2, 206–213.

This page intentionally left blank

Index