Probability Methods for Cost Uncertainty Analysis

A Systems Engineering Perspective

Paul R. Garvey

Probability Methods for Cost Uncertainty Analysis

A Systems Engineering Perspective

Paul R. Garvey
The MITRE Corporation
Bedford, Massachusetts

MARCEL DEKKER, INC. NEW YORK • BASEL

Library of Congress Cataloging-in-Publication Data

Garvey, Paul R.
Probability methods for cost uncertainty analysis: a systems engineering perspective/
Paul R. Garvey.
p. cm.
Includes bibliographical references and index.
ISBN 0-8247-8966-0 (alk. Paper)
1. Systems engineering—Costs. 2. Probabilities.

TA168 .G35 1999
658.15'52—dc21

99-051460

This book is printed on acid-free paper.

Headquarters
Marcel Dekker, Inc.
270 Madison Avenue, New York, NY 10016
tel: 212-696-9000; fax: 212-685-4540

Eastern Hemisphere Distribution
Marcel Dekker AG
Hutgasse 4, Postfach 812, CH-4001 Basel, Switzerland
tel: 41-61-261-8482; fax: 41-61-261-8896

World Wide Web
http://www.dekker.com

The publisher offers discounts on this book when ordered in bulk quantities. For more information, write to Special Sales/Professional Marketing at the headquarters address above.

Current printing (last digit):
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

To my wife Maura and daughters Alana and Kirsten.
I also dedicate this book to the memory of my daughter Katrina
and my parents Eva and Ralph...

Ad Majorem Dei Gloriam

This Page Intentionally Left Blank

Preface

Cost is a driving consideration in decisions that determine how systems are developed, produced, and sustained. Critical to these decisions is understanding how uncertainty affects a system's cost. The process of identifying, measuring, and interpreting these effects is known as *cost uncertainty analysis*. Used early, cost uncertainty analysis can expose potentially crippling areas of risk in systems. This provides managers time to define and implement corrective strategies. Moreover, the analysis brings realism to technical and managerial decisions that define a system's overall engineering strategy. In *Juan De Mairena* (1943), Antonio Machado wrote *"All uncertainty is fruitful...so long as it is accompanied by the wish to understand."* In the same way are insights gleaned from cost uncertainty analysis fruitful — provided they, too, are accompanied by the wish to understand and the will to take action.

Since the 1950s a substantial body of scholarship on this subject has evolved. Published material appears in numerous industry and government technical reports, symposia proceedings, and professional journals. Despite this, there is a need in the systems engineering community to synthesize prior scholarship and relate it to advances in technique and problem sophistication. This book addresses that need. It is a reference for systems engineers, cost engineers, management scientists, and operations research analysts. It is also a text for students of these disciplines.

As a text, this book is appropriate for an upper-level undergraduate (or graduate-level) course on the application of probability methods to cost engineering and analysis problems. It is assumed readers have a solid foundation in differential and integral calculus. An introductory background in probability theory, as well as systems and cost engineering, is helpful; however, the important concepts are developed as needed. A rich set of theoretical and applied exercises accompanies each chapter.

Throughout the book, detailed discussions on issues associated with cost uncertainty analysis are given. This includes the treatment of correlation between the cost of various system elements, how to present the analysis to decision-makers, and the use of bivariate probability distributions to capture the joint interactions between cost and schedule. Analytical techniques from probability theory are stressed, along with the Monte Carlo simulation method. Numerous examples and case discussions are provided to illustrate

the practical application of theoretical concepts. The numerical precision shown in some of the book's examples and case discussions is intended only for pedagogical purposes. In practice, analysts and engineers must always choose the level of precision appropriate to the nature of the problem being addressed.

Chapter 1 presents a general discussion of uncertainty and the role of probability in cost engineering and analysis problems. A perspective on the rich history of cost uncertainty analysis is provided. Readers are introduced to the importance of presenting the cost of a future system as a probability distribution.

Chapter 2 is an introduction to probability theory. Topics include the fundamental axioms and properties of probability. These topics are essential to understanding the terminology, technical development, and application of cost uncertainty analysis methods.

Chapter 3 presents the theory of expectation, moments of random variables, and probability inequalities. Examples derived from systems engineering projects illustrate key concepts.

Chapter 4 discusses modeling cost uncertainty by the probability formalism. A family of continuous univariate probability distributions, used frequently in cost uncertainty analysis, is fully described. A context for applying each distribution is also presented.

Chapter 5 introduces joint probability distributions, functions of random variables, and the central limit theorem. The application of these concepts to cost uncertainty analysis problems is emphasized. In addition, distributions are developed for a general form of the software cost-schedule model. The chapter concludes with a discussion of the Mellin transform, a useful (but little applied) method for working with cost functions that are products, or quotients, of two or more random variables.

Chapter 6 presents specific techniques for quantifying uncertainty in the cost of a future system. The reader is shown how methods from the preceding chapters combine to produce a probability distribution of a system's total cost. This is done from a work breakdown structure perspective. Case studies derived from systems engineering projects provide the application context.

Chapter 7 extends the discussion in chapter 6 by presenting a family of joint probability distributions for cost and schedule. This family consists of the classical bivariate normal, the bivariate normal-lognormal, and the bivariate lognormal distributions; the latter two distributions are rarely

discussed in the traditional literature. Examples are given to show the use of these distributions in a cost analysis context.

The book concludes with a summary of recommended practices and modeling techniques. They come from the author's experience and many years of collaboration with colleagues in industry, government, and academe.

The author gratefully acknowledges a number of distinguished engineers, scientists, and professors who contributed to this book. Their encouragement, enthusiasm, and insights have been instrumental in bringing about this work.

- *Stephen A. Book* — Distinguished Engineer, The Aerospace Corporation, Los Angeles, California. A long-time professional colleague, Dr. Book peer reviewed the author's major technical papers, some of which became chapters in this book. In addition, he independently reviewed and commented on many of the book's chapters as they evolved over the writing period.
- *Philip H. Young* — Director of Research, Lori Associates, Los Angeles, California, and formerly of The Aerospace Corporation, conducted a detailed review of selected areas in this book. Also a long-time professional colleague, Mr. Young shared with the author his formulas for the moments of the trapezoidal distribution (presented in chapters 4 and 5), as well as a derivation of the correlation function of the bivariate normal-lognormal distribution. This derivation is provided as theorem B-1, in appendix B.
- *Nancy E. Rallis* — Associate Professor of Mathematics, Boston College, led the book's academic review. For two years, Professor Rallis studied the entire text from a theoretical and computational perspective. Her years of experience as a statistical consultant and cost analyst at the NASA Goddard Spaceflight Center, TRW Inc., and the Jet Propulsion Laboratory (California Institute of Technology) brought a wealth of insights that greatly enhanced this book. *Sarah E. Quebec*, a graduate mathematics student at Boston College, assisted Professor Rallis' review. I am grateful for her diligence in checking the many examples and case discussions.
- *Wendell P. Simpson III* (Major, USAF-Ret) and *Stephen A. Giuliano* (Lieutenant Colonel, USAF-Ret) — Assistant Professors, United States Air Force Institute of Technology. Professors Simpson and Giuliano developed and taught the school's first graduate course on cost risk analysis. The course used early drafts of the manuscript as required reading. Their comments on the manuscript, as well as those from their students, contributed significantly to the book's content and presentation style.

- Colleagues at The MITRE Corporation...

  Chien-Ching Cho — Principal Staff, Economic and Decision Analysis Center. A long-time professional colleague, I am grateful to Dr. Cho for many years of technical discussions on theoretical aspects of this subject. I particularly appreciate his independent review of case discussion 6-2 and his commentary on Monte Carlo simulation, presented in chapter 6.

  Barbara E. Wolfinger — While a Group Leader in the Economic and Decision Analysis Center, Ms. Wolfinger reviewed original drafts of chapter 1 and chapter 2. A creative practitioner of cost uncertainty analysis, her experiences and analytical insights were highly valued, particularly in the early stages of this work.

  Neal D. Hulkower — While a Department Head in the Economic and Decision Analysis Center, Dr. Hulkower reviewed a number of the author's technical papers when they were early drafts. A veteran cost analyst, his leadership on the necessity of presenting a system's future cost as a probability distribution fostered the award-winning research contained in this book.

  William P. Hutzler — While Director of the Economic and Decision Analysis Center, Dr. Hutzler provided the senior managerial review and leadership needed to bring the manuscript into the public domain. His enthusiasm and encouragement for this work will always be gratefully appreciated.

  Francis M. Dello Russo and *John A. Vitkevich*, Jr. — Mr. Dello Russo (Department Head, Economic and Decision Analysis Center) and Mr. Vitkevich (Lead Staff, Economic and Decision Analysis Center) reviewed the book's first case discussion (chapter 3). From an engineering economics perspective, they provided valuable commentary on issues associated with cost-volume-profit analyses.

  Hank A. Neimeier — Principal Staff, Economic and Decision Analysis Center. Mr. Neimeier provided a careful review of the Mellin transform method (chapter 5) and independently checked the associated examples. His expertise in mathematical modeling provided a valuable context for the application of this method to cost engineering and analysis problems.

  Albert R. Paradis — Lead Staff, Airspace Management and Navigation. Dr. Paradis reviewed an early version of the manuscript. His comments were highly valued. They helped fine-tune the explanation of a number of important and subtle concepts in probability theory.

Raymond L. Fales — A long-time professional colleague, Dr. Fales introduced the author to cost uncertainty analysis. He was among the early practitioners of analytical methods at MITRE and a mentor to many technical staff in the Economic and Decision Analysis Center.

Ralph C. Graves — A gifted and insightful systems engineer, Mr. Graves and the author worked jointly on numerous cost studies for the United States Air Force. During these studies, he introduced the author to Monte Carlo simulation (chapter 6) as a practical approach for quantifying cost uncertainty.

The author also appreciates the staff at Marcel Dekker, Inc. for their diligence, professionalism, and enthusiasm for this work. Many thanks to Graham Garratt (Executive Vice President), Maria Allegra (Acquisitions Editor and Manager), Joseph Stubenrauch (Production Editor), and Regina Efimchik (Marketing and Promotions).

Paul R. Garvey

This Page Intentionally Left Blank

Contents

This Page Intentionally Left Blank

Reserved Notation

A list of reserved notation used in this book is provided below.

FY	Fiscal year
$K	Dollars thousand
$M	Dollars million
SM	Staff-months
$x_{PE_{Cost}}$	Point estimate of *Cost*
$x_{iPE_{X_i}}$	Point estimate of X_i
ℓ_r	Labor rate in dollars per SM
Eff	Effort for an activity (SM)
Eff_{SysEng}	Systems engineering effort (SM)
$Eff_{SysTest}$	System test effort (SM)
Eff_{SW}	Software development effort (SM)
I	The number of delivered source instructions (DSI)
P_r	Software development productivity rate in DSI per staff-month
T_{SW}	Software development schedule (months)

This Page Intentionally Left Blank

Probability Methods for Cost Uncertainty Analysis

This Page Intentionally Left Blank

Chapter 1

Uncertainty and the Role of Probability in Cost Analysis

The only certainty is uncertainty.
Pliny the Elder (Gaius Plinius Secundus)
Natural History

The public...demands certainties...
But there *are* no certainties.
Henry Louis Mencken
Prejudices, First Series [1919], ch. 3

1.1 Introduction and Historical Perspective

This book presents methods for quantifying the cost impacts of uncertainty in the engineering of systems. The term "systems" is used in this book to mean physical systems. Physical systems manifest themselves in physical terms and occupy physical space [1]. Radar systems, air traffic control systems, automobiles, and communication systems are examples of physical systems.

Systems engineering is a process that produces physical systems. It encompasses the scientific and engineering efforts needed to develop, produce, and sustain systems. Systems engineering is a highly complex technical and management undertaking. Integrating custom equipment with commercial products, designing external system interfaces, achieving user requirements, and meeting aggressive schedules while keeping within cost are among the many challenges faced in managing a systems engineering project.

When the cost of a future system* is considered, decision-makers often ask: "*What is the chance its cost will exceed a particular amount*?" "*How much could cost overrun*?" "*What are the uncertainties and how do they drive cost*?" Cost uncertainty analysis provides decision-makers insight into these and related questions. In general, *cost uncertainty analysis* is a process of

---

* This includes existing systems planned for modernization, consolidation, or re-engineering.

quantifying the cost impacts of uncertainties associated with a system's technical definition and cost estimation methodologies.

Throughout a system's life-cycle, cost uncertainty analysis provides motivation and structure for the vigorous management of risk. When appropriately communicated to decision-makers, the insights produced by the analysis directs management's attention to critical program risk-drivers. This enables risk mitigation strategies to be defined and implemented in a timely and cost-effective manner.

Cost uncertainty analysis has its genesis in a field known as military systems analysis [2], founded in the 1950s at the RAND Corporation. Shortly after World War II, military systems analysis evolved as a way to aid defense planners with long-range decisions on force structure, force composition, and future theaters of operation. Cost became a critical consideration in military systems analysis models and decision criteria. However, cost estimates of future military systems, particularly in the early planning phases, were often significantly lower than the actual cost or an estimate developed at a later phase. In the book "*Cost Considerations in Systems Analysis*," G. H. Fisher [3] attributes this difference to the presence of uncertainty; specifically, cost estimation uncertainty and requirements uncertainty.

Cost estimation uncertainty can originate from inaccuracies in cost-schedule estimation models, from the misuse (or misinterpretation) of cost-schedule data, or from misapplied cost-schedule estimation methodologies. Economic uncertainties that influence the cost of technology, the labor force, or geo-political policies further contribute to cost estimation uncertainty [4].

Requirements uncertainty can originate from changes in the system's mission objectives, from changes in performance requirements necessary to meet mission objectives, or from changes in the business or political landscapes that affect the need for the system. Requirements uncertainty most

often results in changes to the system's specified hardware-software configuration, which is also known as the system's architecture.

Uncertainty is also present in elements that define a system's configuration (or architecture). This is referred to as *system definition uncertainty*. Examples include uncertainties in the amount of software to develop, the extent code from another system can be reused, the number workstations to procure, or the delivered weight of an end-item (e.g., a satellite) [4].

The early literature on cost uncertainty analysis concentrated on defining the sources, scope, and types of uncertainties that impacted the cost of future systems. Technical papers published in the period between 1955 and 1962 were not explicitly focused on establishing and applying formal methods to quantify cost uncertainty. However, by the mid-1960s a body of techniques began to emerge. An objective of this book is to discuss these techniques, present advances in methodology, and illustrate how these methods apply from a systems engineering perspective.

1.2 The Problem Space

In systems engineering three types of uncertainties must be considered. Described in the preceding section they are cost estimation uncertainty, requirements uncertainty, and system definition uncertainty. Figure 1-1 [4] illustrates how these uncertainties are related.

The *n*-system configurations shown are in response to requirements uncertainty. For a given system configuration, cost-schedule probability models (as described in this book) capture *only* system definition and cost estimation uncertainties. They provide probability-based assessments of a system's cost and schedule for that system configuration. When requirements uncertainty necessitates defining an entirely new configuration, a new cost-schedule probability model is likely to be needed. The new model must be

tuned to capture the system definition and cost estimation uncertainties specific to the new configuration.

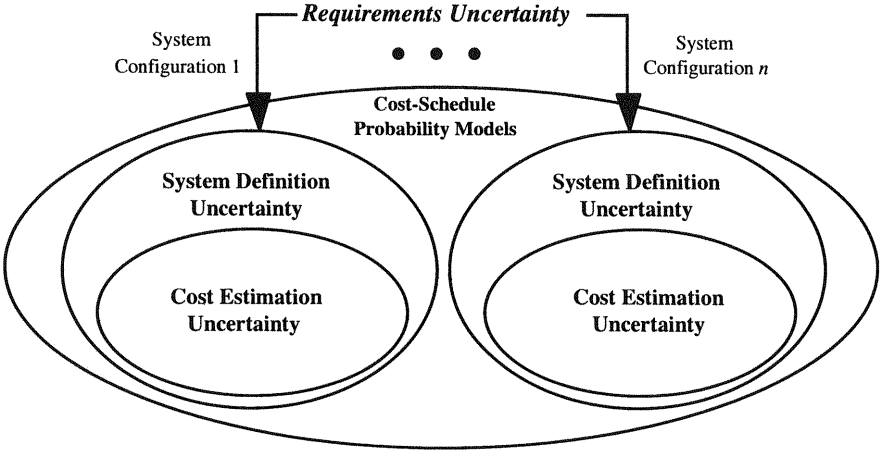


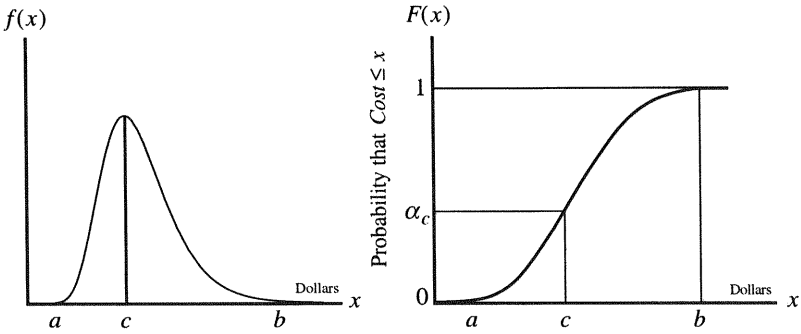
Figure 1-1. Types of Uncertainty Captured by Cost-Schedule Probability Models

1.3 Presenting Cost as a Probability Distribution

Cost is an uncertain quantity. It is highly sensitive to many conditions and assumptions that change frequently across a system's life-cycle. Examining the change in cost subject to varying certain conditions (while holding others constant) is known as *sensitivity analysis*. In a series of lectures to the United States Air Force (1962), Fisher [5] emphasized the importance of sensitivity analysis as a way to isolate cost drivers. He considered sensitivity analysis to be "*...a prime characteristic or objective in the cost analysis of advanced systems and/or force structure proposals.*" Although sensitivity analysis can isolate elements of a system that drive its cost, it is a deterministic procedure defined by a postulated set of "what-if" scenarios. Sensitivity analysis alone does not offer decision-makers insight into the question "*What is the chance*

of exceeding a particular cost in the range of possible costs?" A probability distribution is a way to address this question. Simply stated, a *probability distribution* is a mathematical rule associating a probability α to each possible outcome, or event of interest.

There are two ways to present a probability distribution. It can be shown as a probability density or as a cumulative probability distribution. Figure 1-2 presents one way to illustrate this concept from a cost perspective.



Cost
Probability Density
Figure 1-2a

Cost
Cumulative Probability Distribution
Figure 1-2b

Figure 1-2. Illustrative Probability Distributions

In figure 1-2, the range of possible values for *Cost* is given by the interval $a \le x \le b$. The probability *Cost* will not exceed a value $x = c$ is given by α_c. In figure 1-2a, this probability is the area under $f(x)$ between $x = a$ and $x = c$. In figure 1-2b, this probability is given by $F(c)$.

To develop a cost probability distribution, methods from probability theory were needed. Some of the earliest applications of probability theory to model cost uncertainty took place in the mid-1960s at the MITRE and RAND Corporations. In 1965, Steven Sobel [MITRE] published "*A Computerized Technique to Express Uncertainty in Advanced System Cost Estimates* [6]." It was among the earliest works on modeling cost uncertainty by the probability

formalism. Sobel pioneered using the method of moments technique to develop a probability distribution of a system's total cost.

Complementary to Sobel's analytical approach, in 1966 Paul F. Dienemann [RAND] published "*Estimating Cost Uncertainty Using Monte Carlo Techniques* [7]." The methodology applied Monte Carlo simulation, developed by operations analysts in World War II, to quantify the impacts of uncertainty on total system cost. With the advent of high-speed computers, Monte Carlo simulation grew in popularity and remains a primary approach for generating cost probability distributions. A discussion of Monte Carlo simulation is presented in chapter 6.

An overview of the cost uncertainty analysis process is shown in figure 1-3. The variables $X_1, X_2, X_3, \cdots, X_n$ are the costs of the n work breakdown structure (WBS)* cost elements that comprise the system. For instance, X_1 might represent the cost of the system's prime mission hardware and software; X_2 might represent the cost of the system's systems engineering and program management; X_3 might represent the cost of the system's test and evaluation. When specific values for these variables are uncertain, we can treat them as *random variables.* Probability distributions are developed for $X_1, X_2, X_3, \cdots, X_n$ which associate probabilities to their possible values. Such distributions are illustrated on the left-side of figure 1-3. The random variables $X_1, X_2, X_3, \cdots, X_n$ are summed to produce an overall probability distribution of the system's total cost, shown on the right-side of figure 1-3.

The "input" part of this process has many subjective aspects. Probability distributions for $X_1, X_2, X_3, \cdots, X_n$ are either specified directly or they are generated. Direct specification relies on expert judgment to characterize a distribution's shape. The probability density is the usual way to make this characterization.

---

* A full discussion of the work breakdown structure is presented in chapter 6.

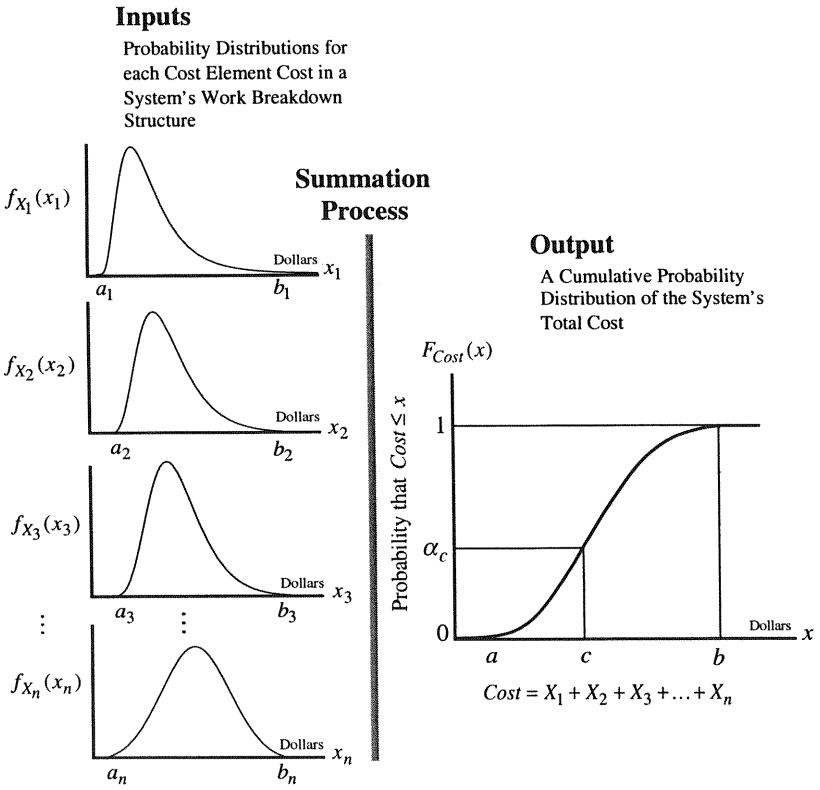


Figure 1-3. Cost Uncertainty Analysis Process

Generated distributions have shapes that are produced from a mathematical process. This is illustrated in the following discussion.

Suppose X_2 represents the cost of a system's systems engineering and program management (SEPM). Furthermore, suppose the cost of SEPM is derived as a function of three random variables* *Staff*, *PrgmSched*, and *LaborRate* as follows:

$$X_2 = Staff \cdot PrgmSched \cdot LaborRate \qquad (1\text{-}1)$$

---

* *Staff* (Persons), *PrgmSched* (Months), *LaborRate* ($/Person-Month)

Suppose the engineering team assessed ranges of possible (feasible) values for these variables and directly specified the shapes of their probability distributions, as shown in figure 1-4. Combining their distributions according to the rules of probability theory generates an overall distribution for X_2, which is the cost of SEPM. In this case, we say the probability distribution of X_2 has been generated by a mathematical process. Figure 1-4 illustrates this discussion.

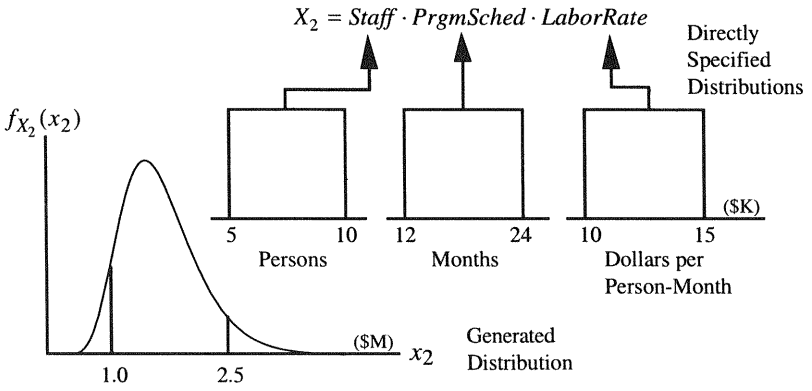


Figure 1-4. The Specification of Probability Distributions

Shown in figure 1-4, it is good practice to reserve the direct specification of distributions to their lowest level variables in a cost equation (e.g., equation 1-1). Often, expert judgment about the shapes and ranges of distributions are best at this level. Furthermore, this "specification" approach structures the overall analysis in a way that specific "cost-risk-driving" variables can be revealed. Identifying these variables, and quantifying how they affect a system's cost, are critical findings to communicate to decision-makers.

A term conventional to cost engineering and analysis is point estimate. The *point estimate* of a variable whose value is uncertain, is a single value for the variable in its range of possible values. From a mathematical perspective, the point estimate is simply one value among those that are feasible. In practice, a point estimate is established by an analyst (using appropriate cost analysis

methods) prior to an assessment of other possible values. It provides an "anchor" (i.e., a reference point) around which other possible values are assessed or generated. This is illustrated in figure 1-5.

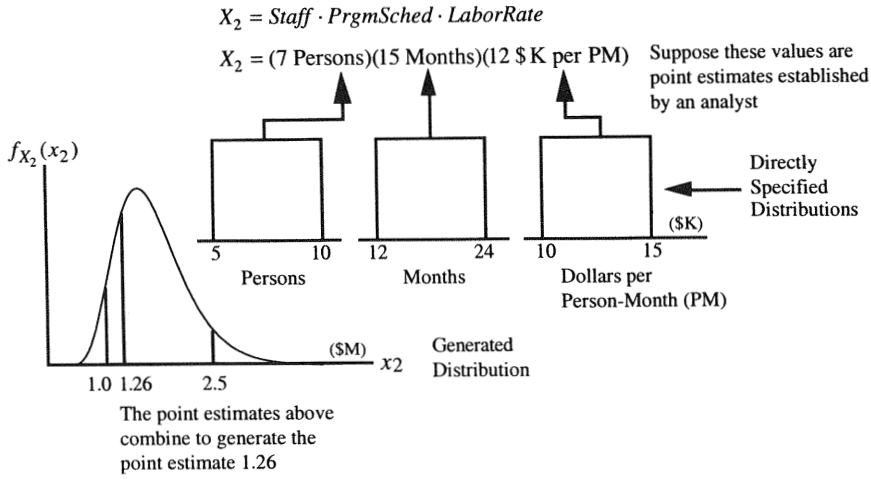


Figure 1-5. Point Estimates – An Illustration

In cost uncertainty analysis it is common to see more probability density to the right of a point estimate than to its left; this is seen in the generated distribution in figure 1-5. Although this is a common occurrence, the point estimate *can* fall anywhere along the variable's probability distribution; it is just one value among those that are feasible.

Suppose a system's total cost is given by

$$Cost = X_1 + X_2 + X_3 + \ldots + X_n$$

where the random variables $X_1, X_2, X_3, \cdots, X_n$ are the costs of the system's n WBS cost elements. Suppose point estimates are developed for each X_i ($i = 1, \ldots, n$). Their sum is the *point estimate of the cost of the system*. Let this sum be denoted by $x_{PE_{Cost}}$, where

$$x_{PE_{Cost}} = x_{1PE_{X_1}} + x_{2PE_{X_2}} + x_{3PE_{X_3}} + \ldots + x_{nPE_{X_n}} \quad (1\text{-}2)$$

and $x_{iPE_{X_i}}$ $(i = 1,\ldots,n)$ is the point estimate of X_i. Computing $x_{PE_{Cost}}$ according to equation 1-2, is known among practitioners as the "*roll-up*" procedure.*

In cost engineering and analysis, it is traditional to consider $x_{PE_{Cost}}$ a value for *Cost* that contains *no reserve dollars.* As a point estimate, $x_{PE_{Cost}}$ provides the "anchor" from which to choose a value for *Cost* that *contains reserve dollars.* Decision-makers tradeoff between $x_{PE_{Cost}}$ and the amount of reserve dollars to add to $x_{PE_{Cost}}$, such that the value of *Cost* determined by the expression [$x_{PE_{Cost}}$ + (reserve dollars)] has an acceptable probability of *not being exceeded.* Figure 1-6 illustrates this discussion.

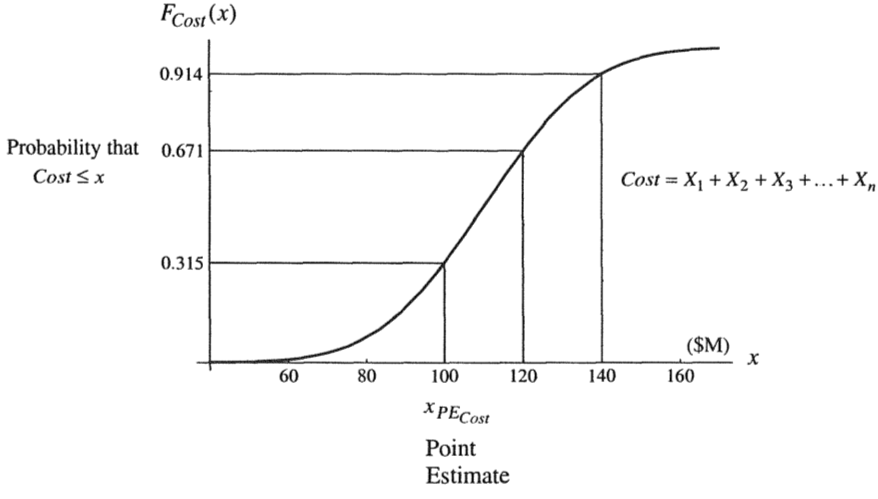


Figure 1-6. A Cumulative Probability Distribution of System Cost

* From a probability perspective there are important subtleties associated with the roll-up procedure. These subtleties are illustrated in case discussion 5-1 (chapter 5).

In figure 1-6, suppose the point estimate of a system's cost is 100 ($M); that is, $x_{PE_{Cost}} = 100$. This value of *Cost* has just over a 30 percent probability of not being exceeded. A reserve of 20 ($M) added to $x_{PE_{Cost}}$ is associated with a value of *Cost* that has a 67 percent probability of not being exceeded. A reserve of 40 ($M) added to $x_{PE_{Cost}}$ is associated with a value of *Cost* that has just over a 90 percent probability of not being exceeded.

It is possible for $x_{PE_{Cost}}$ to fall at a high confidence level on its associated distribution function. Such a circumstance may warrant the addition of no reserve dollars; it suggests there is a good chance for *Cost* to actually be lower than perhaps anticipated. However, it may also flag a situation where cost reserve was built, a priori, into the point estimate of each WBS cost element cost. These reserve dollars would be included in the "roll-up" of the individual point estimates. This result can make $x_{PE_{Cost}}$ hard to interpret, particularly if tradeoff studies are needed. In practice, it is recommended keeping $x_{PE_{Cost}}$ "clean" from reserve dollars. This provides analysts and decision-makers an anchor point that is "cost reserve-neutral" – one where the tradeoff between cost reserve and a desired level of confidence can be readily understood for various alternatives (or options) under consideration.

1.4 Benefits of Cost Uncertainty Analysis

Cost uncertainty analysis provides decision-makers many benefits and important insights. These include:

Establishing a Cost and Schedule Risk Baseline — Baseline probability distributions of a system's cost and schedule can be developed for a given system configuration, acquisition strategy, and cost-schedule estimation approach. This baseline provides decision-makers visibility into potentially high-payoff areas for risk reduction initiatives. Baseline distributions assist in determining a system's cost and schedule that simultaneously have a specified probability of not being exceeded (chapter 7). They can also provide decision-makers an assessment of the

likelihood of achieving a budgeted (or proposed) cost and schedule, or cost for a given feasible schedule [4].

Determining Cost Reserve — Cost uncertainty analysis provides a basis for determining cost reserve as a function of the uncertainties specific to a system. The analysis provides the direct link between the amount of cost reserve to recommend and the probability that a system's cost will not exceed a prescribed (or desired) magnitude. An analysis should be conducted to verify the recommended cost reserve covers fortuitous events (e.g., unplanned code growth, unplanned schedule delays) deemed possible by the system's engineering team [4].

Conducting Risk Reduction Tradeoff Analyses — Cost uncertainty analyses can be conducted to study the payoff of implementing risk reduction initiatives (e.g., rapid prototyping) on lessening a system's cost and schedule risks. Furthermore, families of probability distribution functions can be generated to compare the cost and cost risk impacts of alternative system requirements, schedule uncertainties, and competing system configurations or acquisition strategies [4].

The validity and meaningfulness of a cost uncertainty analysis relies on the engineering team's experience, judgment, and knowledge of the system's uncertainties. Formulating and documenting a supporting rationale, that summarizes the team's collective insights into these uncertainties, is *the critical part of the process.* Without a well-documented rationale, the credibility of the analysis can be easily questioned.

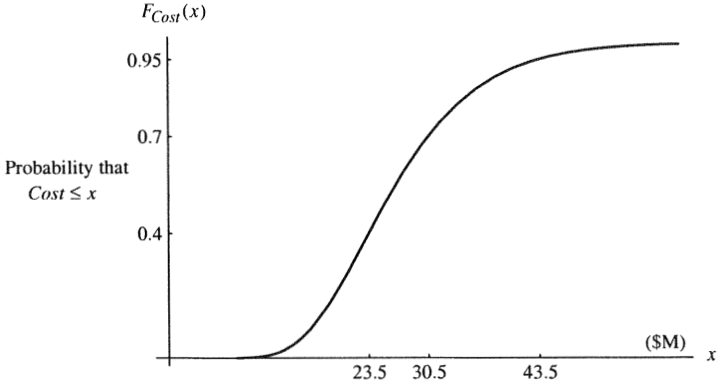
The details of the analysis methodology are important and should also be documented. The methodology *must be technically sound* and offer value-added problem structure, analyses, and insights otherwise not visible. Decisions that successfully eliminate uncertainty, or reduce it to acceptable levels, are ultimately driven by human judgment. This at best is aided by, not directed by, the methods presented in this book.

Exercises

1. State and define the three types of uncertainties that affect the cost of a systems engineering project. Give specific examples of each type.

2. Define, from a cost perspective, the term point estimate. How is the point estimate of a variable used to establish a range of other possible values? Explain what is meant by the "roll-up" procedure.

3. In the figure below, suppose the *point estimate* of a system's cost is 23.5 dollars million ($M). Assume the three values shown along the vertical axis are paired with the three values shown along the horizontal axis. How many reserve dollars are needed such that the value of *Cost* associated with that reserve has a 70 percent chance of *not being* exceeded? Similarly, what is the reserve needed such that the value of *Cost* has only a 5 percent chance of *being* exceeded?



Cumulative Probability Distribution for Exercise 3

References

1. Blanchard, B. S., and W. J. Fabrycky. 1990. *Systems Engineering and Analysis*, 2nd ed. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

2. Hitch, C. J. 1955. *An Appreciation of Systems Analysis*, P-699. Santa Monica, California: The RAND Corporation.

3. Fisher, G. H. 1971. *Cost Considerations in Systems Analysis*. New York: American Elsevier Publishing Company, Inc.

4. Garvey, P. R. 1996 (Spring). Modeling Cost and Schedule Uncertainties – A Work Breakdown Structure Perspective. *Military Operations Research*, V2, N1, pp. 37-43.

5. Fisher, G. H. 1962. *A Discussion of Uncertainty in Cost Analysis*, RM-3071-PR. Santa Monica, California: The RAND Corporation.

6. Sobel, S. 1965. *A Computerized Technique to Express Uncertainty in Advanced System Cost Estimates*, ESD-TR-65-79. Bedford, Massachusetts: The MITRE Corporation.

7. Dienemann, P. F. 1966. *Estimating Uncertainty Using Monte Carlo Techniques*, RM-4854-PR. Santa Monica, California: The RAND Corporation.

Chapter 2

Concepts of Probability Theory

If chance is the antithesis of law,
then we need to discover the laws of chance.
C. R. Rao
Indian Statistician

The theory of probabilities is
at bottom nothing but common
sense reduced to calculus.
Pierre Simon de Laplace
Oeuvres, vol. VII, Théorie Analytique des Probabilités

2.1 Introduction

Whether it's a storm's intensity, an arrival time, or the success of a financial decision, the words "probable" or "likely" have long been part of our language. Most people have a practical appreciation for the impact of chance on the occurrence of an event. In the last 300 years, the theory of probability has evolved to explain the nature of chance and how it may be studied.

Probability theory is the formal study of random events and random processes. Its origins trace to 17th century gambling problems. Games that involved playing cards, roulette wheels, and dice provided mathematicians a host of interesting problems. The solutions to many of these problems yielded the first principles of modern probability theory. Today, probability theory is of fundamental importance in science, engineering, and business.

2.2 Sample Spaces and Events

If a six-sided die* is tossed there are six possible outcomes for the number that appears on the upturned face. These outcomes can be listed as elements in the set $\{1,2,3,4,5,6\}$. The set of all possible outcomes of an experiment is called the *sample space*, which we will denote by Ω. The individual outcomes of Ω are called *sample points*, which we will denote by ω.

---

* Unless otherwise noted, dice are assumed in this book to be six-sided.

A sample space can be finite, countably infinite, or uncountable. A *finite sample space* is a set that consists of a finite number of outcomes. The sample space for the toss of a die is finite. A *countably infinite sample space* is a set whose outcomes can be arranged in a one-to-one correspondence with the set of positive integers. An *uncountable sample space* is one that is infinite but not countable. For instance, suppose the sample space for the duration t (in hours) of an electronic device is $\Omega = \{t : 0 \le t < 2500\}$; then Ω is an uncountable sample space; there are an infinite but not countable number of possible outcomes for t. Finite and countably infinite sample spaces are also known as *discrete sample spaces*. Uncountable sample spaces are known as *continuous sample spaces*.

An *event* is any subset of the sample space. An event is *simple* if it consists of exactly one outcome.* Simple events are also referred to as *elementary events* or *elementary outcomes*. An event is *compound* if it consists of more than one outcome. For instance, let B be the event an odd number appears and C be the event an even number appears in a single toss of a die. These are compound events, which may be expressed by the sets $B = \{1,3,5\}$ and $C = \{2,4,6\}$. Event B occurs if and only if one of the outcomes in B occurs; the same is true for event C.

Seen in this discussion, events can be represented by sets. New events can be constructed from given events according to the rules of set theory. The following presents a brief review of set theory concepts.

---

* As we shall see, probabilities associated with simple events are sensitive to the nature of the sample space. If Ω is a *discrete sample space*, the probability of an event is completely determined by the probabilities of the simple events in Ω; however, if Ω is a *continuous sample space*, the probability associated with each simple event in Ω is zero. This will be discussed further in chapter 3.

Union: For any two events A and B of a sample space Ω, the new event $A \cup B$ (which reads A *union* B) consists of all outcomes either in A *or* in B or in both A and B. The event $A \cup B$ occurs if *either* A *or* B *occurs*. To illustrate the union of two events, consider the following: if A is the event an odd number appears in the toss of a die and B is the event an even number appears, then the event $A \cup B$ is the set $\{1,2,3,4,5,6\}$, which is the sample space for this experiment.

Intersection: For any two events A and B of a sample space Ω, the new event $A \cap B$ (which reads A *intersection* B) consists of all outcomes that are *both* in A and in B. The event $A \cap B$ occurs *only if both* A *and* B *occur*. To illustrate the intersection of two events, consider the following: if A is the event a six appears in the toss of a die, B is the event an odd number appears, and C is the event an even number appears then the event $A \cap C$ is the simple event $\{6\}$; on the other hand, the event $A \cap B$ contains no outcomes. Such an event is called the *null event*. The null event is traditionally denoted by $\varnothing$. In general, if $A \cap B = \varnothing$, we say events A and B are *mutually exclusive* (*disjoint*). The intersection of two events A and B is sometimes written as AB, instead of $A \cap B$.

Complement: The *complement* of event A, denoted by A^c, consists of all outcomes in the sample space Ω that are *not* in A. The event A^c occurs if and only if A does not occur. The following illustrates the complement of an event. If C is the event an even number appears in the toss of a die, then C^c is the event an odd number appears.

Subset: Event A is said to be a *subset* of event B if all the outcomes in A are also contained in B. This is written as $A \subset B$.

Figure 2-1 illustrates these concepts in the form of Venn diagrams.

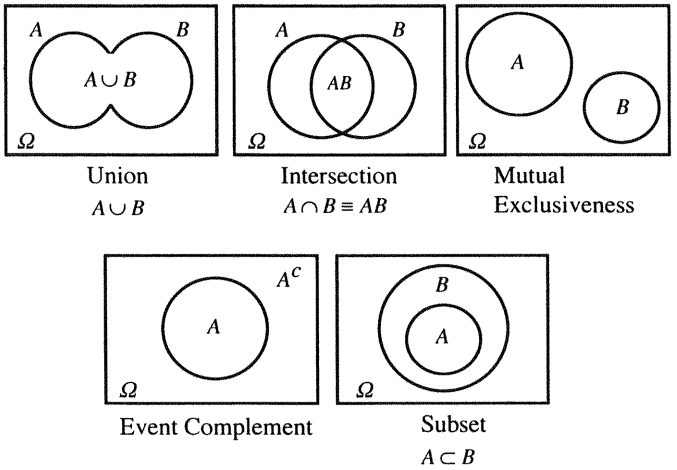


Figure 2-1. Venn Diagrams for Various Event Relationships

Operations involving the union and intersection of events follow the rules of set algebra. These rules are summarized below.

Identity Laws	$A \cup \varnothing = A$	$A \cap \varnothing = \varnothing$
	$A \cup \Omega = \Omega$	$A \cap \Omega = A$
De Morgan's Laws	$(A \cup B)^c = A^c \cap B^c$	$(A \cap B)^c = A^c \cup B^c$
Associative Laws	$A \cup B \cup C = (A \cup B) \cup C = A \cup (B \cup C)$	
	$A \cap B \cap C = (A \cap B) \cap C = A \cap (B \cap C)$	
Distributive Laws	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	
	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$	
Commutative Laws	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Idempotency Laws	$A \cup A = A$	$A \cap A = A$
Complementary Laws	$A \cup A^c = \Omega$	$A \cap A^c = \varnothing$

2.3 Interpretations and Axioms of Probability

In the preceding discussion, the sample space for the toss of a die was given by $\Omega = \{1,2,3,4,5,6\}$. If we *assume* the die is fair (which, unless otherwise

noted, is assumed throughout this book) then any outcome in the sample space is as likely to appear as any other. Given this assumption, it is reasonable to conclude the proportion of time each outcome is expected to occur is $\frac{1}{6}$. Thus, the probability of each simple event in the sample space is

$$P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6}$$

Similarly, suppose B is the event an odd number appears in a single toss of the die. This compound event is given by the set $B = \{1,3,5\}$. Since there are three ways event B can occur out of six possible, we conclude the probability of event B is

$$P(B) = \frac{3}{6} = \frac{1}{2}$$

The following presents a view of probability known as the equally likely interpretation.

Equally Likely Interpretation: In this view, if a sample space Ω consists of a finite number of outcomes n, which are all equally likely to occur, then the probability of each simple event is $\frac{1}{n}$. If an event A consists of m of these n outcomes, then the probability of event A is

$$P(A) = \frac{m}{n} \qquad (2\text{-}1)$$

In the above, it is assumed the sample space consists of a *finite* number of outcomes and all outcomes are equally likely to occur. What if the sample space is uncountable? What if the sample space is finite but the outcomes are *not* equally likely?* In these situations, probability might be measured in terms of how frequently a particular outcome occurs when the experiment is repeatedly performed under identical conditions. This leads to a view of probability known as the frequency interpretation.

---

* If a die is weighted in a particular way, then the outcomes of the toss are no longer considered fair, or equally likely.

Frequency Interpretation: In this view, the probability of an event is the limiting proportion of time the event occurs in a set of n repetitions of the experiment. In particular, we write this as

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

where $n(A)$ is the number of times in n repetitions of the experiment the event A occurs. In this sense $P(A)$ is the limiting frequency of event A. Probabilities measured by the frequency interpretation are referred to as *objective probabilities*. There are many circumstances where it is appropriate to work with objective probabilities. However, there are limitations with this interpretation of probability. It restricts events to those that can be subjected to repeated trials conducted under *identical conditions*. Furthermore, it is not clear how many trials of an experiment are needed to obtain an event's limiting frequency.

Axiomatic Definition: In 1933, the Russian mathematician Kolmogorov* presented a definition of probability in terms of three axioms. These axioms define probability in a way that encompasses the *equally likely and frequency interpretations* of probability. It is known as the axiomatic definition of probability. It is the view of probability adopted in this book. Under this definition it is assumed for each event A, in the sample space Ω, there exists a real number $P(A)$ that denotes the probability of A. In accordance with Kolmogorov's axioms, a probability is simply a numerical value (measure) that satisfies the following:

Axiom 1 $0 \leq P(A) \leq 1$ for any event A in Ω

Axiom 2 $P(\Omega) = 1$

---

* A. N. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung, Ergeb. Mat. und ihrer Grenzg.*, vol. 2, no. 3, 1933. Translated into English by N. Morrison, *Foundations of the Theory of Probability*, New York (Chelsea), 1956 [1].

Axiom 3 For any sequence of mutually exclusive events* $A_1, A_2, \ldots$ defined on Ω

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

For any *finite sequence* of mutually exclusive events $A_1, A_2, \ldots, A_n$ defined on Ω

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

The first axiom states the probability of any event is a nonnegative number in the interval zero to unity. In axiom 2, the sample space Ω is sometimes referred to as the *sure* or *certain event*; therefore, we have $P(\Omega)$ equal to unity. Axiom 3 states for any sequence of mutually exclusive events, the probability of at least one of these events occurring is the sum of the probabilities associated with each event A_i. In axiom 3, this sequence may also be finite. From these axioms come basic theorems of probability.

Theorem 2-1 The probability event A occurs is one minus the probability it will not occur; that is, $P(A) = 1 - P(A^c)$

Proof From the complementary law $\Omega = A \cup A^c$. From axiom 3 it follows that $P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$ since A and A^c are mutually exclusive events. From axiom 2, we know that $P(\Omega) = 1$; therefore, $1 = P(A) + P(A^c)$ and the result $P(A) = 1 - P(A^c)$ follows.

Theorem 2-2 The probability associated with the null event $\varnothing$ is zero

$$P(\varnothing) = 0$$

Proof From theorem 2-1 and axiom 2

$$P(\varnothing) = 1 - P(\varnothing^c) = 1 - P(\Omega) = 1 - 1 = 0$$

---

* That is, $A_i \cap A_j = \varnothing$ for $i \neq j$.

Theorem 2-3 If events A_1 and A_2 are mutually exclusive, then

$$P(A_1 \cap A_2) \equiv P(A_1 A_2) = 0$$

Proof Since A_1 and A_2 are mutually exclusive, $A_1 \cap A_2 = \varnothing$. This implies $P(A_1 \cap A_2) = P(\varnothing)$. From theorem 2-2, $P(\varnothing) = 0$; therefore, $P(A_1 \cap A_2) = 0$.

Theorem 2-4 For any two events A_1 and A_2

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

Proof The event $A_1 \cup A_2$, shown in figure 2-2, is written in terms of three mutually exclusive events, that is, $A_1 \cup A_2 = (A_1 A_2^c) \cup (A_1 A_2) \cup (A_1^c A_2)$. From axiom 3, $P(A_1 \cup A_2) = P(A_1 A_2^c) + P(A_1 A_2) + P(A_1^c A_2)$.

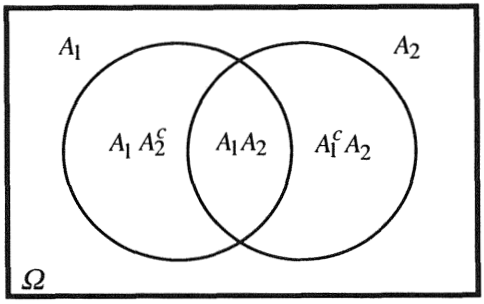


Figure 2-2. The Partition of A_1 Union A_2

From figure 2-2, A_1 can be written in terms of mutually exclusive events; that is, $A_1 = (A_1 A_2^c) \cup (A_1 A_2)$; similarly $A_2 = (A_1^c A_2) \cup (A_1 A_2)$. From axiom 3, it follows that $P(A_1) = P(A_1 A_2^c) + P(A_1 A_2)$ and $P(A_2) = P(A_1^c A_2) + P(A_1 A_2)$. Therefore, $P(A_1 \cup A_2)$ can be written as

$$P(A_1 \cup A_2) = P(A_1) - P(A_1 A_2) + P(A_1 A_2) + P(A_2) - P(A_1 A_2)$$

It follows that*

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 A_2) \blacklozenge$$

---

* The symbol ♦ is reserved in this book to signal, where it might not be clear, the completion of a proof, an example, or a case discussion.

If A_1 and A_2 were mutually exclusive events, theorem 2-4 simplifies to axiom 3, that is, $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ since $P(A_1A_2) \equiv P(A_1 \cap A_2) = P(\varnothing) = 0$.

Theorem 2-5 If event A_1 is a subset of event A_2 then

$$P(A_1) \leq P(A_2)$$

Proof Since A_1 is a subset of A_2, the event A_2 may be expressed as the union of two mutually exclusive events A_1 and $A_1^c A_2$. Refer to figure 2-3.

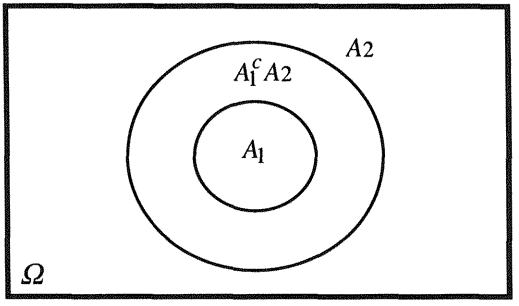


Figure 2-3. Event A_1 as a Subset of Event A_2

Since

$$A_2 = A_1 \cup A_1^c A_2$$

from axiom 3

$$P(A_2) = P(A_1) + P(A_1^c A_2)$$

Because $P(A_1^c A_2) \geq 0$ it follows that

$$P(A_1) \leq P(A_2) \blacklozenge$$

Example 2-1 The sample space Ω for an experiment that consists of tossing two dice is given by the 36 possible outcomes listed in table 2-1. The outcomes in table 2-1 are given by the pairs (d_1, d_2),* which we assume are equally likely. Let A, B, C, and D represent the following events:

---

* The outcomes from tossing two dice are recorded as (d_1, d_2), where d_1 and d_2 are the numbers appearing on the upturned faces of the first and second die, respectively.

A: The sum of the toss is odd

B: The sum of the toss is even

C: The sum of the toss is a number less than ten

D: The toss yielded the same number on each die's upturned face

Find $P(A)$, $P(B)$, $P(C)$, $P(A \cap B)$, $P(A \cup B)$, $P(B \cap C)$, and $P(B \cap C \cap D)$

Table 2-1. Sample Space for the Tossing of Two Dice

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	**(2,2)**	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	**(3,3)**	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	**(4,4)**	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	**(5,5)**	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	**(6,6)**

Solution The outcomes from the sample space in table 2-1 that make up event A are

{(1,2), (1,4), (1,6), (2,1), (2,3), (2,5), (3,2), (3,4), (3,6),
(4,1), (4,3), (4,5), (5,2), (5,4), (5,6), (6,1), (6,3), (6,5)}

The outcomes from the sample space in table 2-1 that make up event B are

{(1,1), (1,3), (1,5), (2,2), (2,4), (2,6), (3,1), (3,3), (3,5),
(4,2), (4,4), (4,6), (5,1), (5,3), (5,5), (6,2), (6,4), (6,6)}

The outcomes from the sample space in table 2-1 that make up event C are

{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3),
(2,4), (2,5), (2,6), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6),
(4,1), (4,2), (4,3), (4,4), (4,5), (5,1), (5,2), (5,3), (5,4),
(6,1), (6,2), (6,3)}

The outcomes from the sample space in table 2-1 that make up event D are

{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)}

Determination of $P(A)$, $P(B)$, and $P(C)$: From equation 2-1, we can compute

$$P(A)=\frac{18}{36}=\frac{1}{2} \qquad P(B)=\frac{18}{36}=\frac{1}{2} \qquad P(C)=\frac{30}{36}=\frac{15}{18}$$

Determination of $P(A\cap B)$: Observe event A and event B are mutually exclusive, that is, they share no elements in common. Therefore, from theorem 2-3

$$P(A\cap B)\equiv P(AB)=0$$

Determination of $P(A\cup B)$: From theorem 2-4

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

Since $P(A\cap B)=0$ and $P(A)=P(B)=\frac{1}{2}$ it follows that $P(A\cup B)=1$. Notice the event $A\cup B$ yields the sample space Ω for this experiment; by axiom 2 we know $P(\Omega)=1$.

Determination of $P(B\cap C)$: The event the sum of the toss is even and it is a number less than ten is given by $B\cap C$. This event contains the outcomes

$$\{(1,1),\ (1,3),\ (1,5),\ (2,2),\ (2,4),\ (2,6),\ (3,1),\ (3,3),\ (3,5),$$
$$(4,2),\ (4,4),\ (5,1),\ (5,3),\ (6,2)\}$$

from which $P(B\cap C)=14/36=7/18$.

Determination of $P(B\cap C\cap D)$: The event the sum of the toss is even and it is a number less than ten and the toss yielded the same number on each die's upturned face is given by $B\cap C\cap D$. This event contains the outcomes

$$\{(1,1),\ (2,2),\ (3,3),\ (4,4)\}$$

from which $P(B\cap C\cap D)=4/36=1/9$. Notice event $B\cap C\cap D$ is a subset of event $B\cap C$. From theorem 2-5 we expect $P(B\cap C\cap D)\le P(B\cap C)$.

Measure of Belief Interpretation: From the axiomatic view, probability need only be a number satisfying the three axioms stated by Kolomogorov. Given this, it is possible for probability to reflect a "measure of belief" in an event's occurrence. For instance, a software engineer might assign a probability of 0.70 to the event "*the radar software for the Advanced Air Traffic Control System (AATCS) will not exceed 100K delivered source instructions.*" We consider this event to be nonrepeatable. It is not practical, or possible, to build the AATCS *n*-times (and under identical conditions) to determine whether this probability is indeed 0.70. When an event such as this arises, its probability may be assigned. Probabilities assigned on the basis of personal judgment, or measure of belief, are known as *subjective probabilities.*

Subjective probabilities are the most common in systems engineering projects and cost analysis problems. Such probabilities are typically assigned by expert technical opinion. The software engineer's probability assessment of 0.70 is a subjective probability. Ideally, subjective probabilities should be based on available evidence and previous experience with similar events. Subjective probabilities risk becoming suspect if they are premised on limited insights or no prior experiences. Care is also needed in soliciting subjective probabilities. They must certainly be plausible; but even more, they must be *consistent* with Kolomogorov's axioms and the theorems of probability which stem from these axioms. Consider the following:

> The XYZ Corporation has offers on two contracts A and B. Suppose the proposal team made the following subjective probability assignments...the chance of winning contract A is 40 percent, the chance of winning contract B is 20 percent, the chance of winning contract A or contract B is 60 percent, and the chance of winning both contract A and contract B is 10 percent. It turns out this set of probability assignments is *not* consistent with the axioms and theorems of probability! Why is

this?* If the chance of winning contract B was changed to 30 percent, then this *particular set of probability assignments* would be consistent.

Kolmogorov's axioms, and the resulting theorems of probability, *do not suggest* how to assign probabilities to events; rather, they provide a way to verify the probability assignments (be they objective or subjective) are consistent.

Risk versus Uncertainty: There is an important distinction between the terms *risk* and *uncertainty*. Risk is the chance of loss or injury. In a situation that includes favorable and unfavorable events, risk is the *probability an unfavorable event occurs*. Uncertainty is the *indefiniteness about the outcome of a situation*. We analyze uncertainty *for the purpose of measuring risk*. In systems engineering the analysis might focus on measuring the risk of: failing to achieve performance objectives, overrunning the budgeted cost, or delivering the system too late to meet user needs. Conducting the analysis involves varying degrees of subjectivity. This includes defining the events of concern, as well as specifying their subjective probabilities. Given this, it is fair to ask whether it's meaningful to apply rigorous mathematical procedures to such analyses. In a speech before the 1955 Operations Research Society of America meeting, Charles Hitch addressed this question. He stated [2]:

> "Systems analyses provides a framework which permits the judgment of experts in many fields to be combined to yield results that transcend any individual judgment. The systems analyst [cost analyst] may have to be content with better rather than optimal solutions; or with devising and costing sensible methods of hedging; or merely with discovering critical sensitivities. We tend to be worse, in an absolute sense, in applying analysis or scientific method to broad context problems; but unaided intuition in such problems is also much worse in the absolute sense. Let's not deprive ourselves of any useful tools, however short of perfection they may fail."

---

* The answer can be seen from theorem 2-4; this is also exercise 6.

2.4 Conditional Probability

In many circumstances the probability of an event must be conditioned on knowing another event has taken place. Such a probability is known as a conditional probability. *Conditional probabilities* incorporate information about the occurrence of another event. The conditional probability of event A given an event B has occurred is denoted by $P(A|B)$. In example 2-1, it was shown if a pair of dice is tossed the probability the sum of the toss is even is 1/2; this probability is known as a *marginal* or *unconditional probability.* How would this unconditional probability change (i.e., be conditioned) if it was *known* the sum of the toss was a number less than ten? This is discussed in the following example.

Example 2-2 If a pair of dice is tossed and the sum of the toss is a number less than ten, compute the probability this sum is an even number.

Solution Returning to example 2-1, recall events B and C were given by

B: The sum of the toss is even

C: The sum of the toss is a number less than ten

The sample space Ω is given by the 36 outcomes in table 2-1. In this case, we want the subset of Ω containing *only* those outcomes whose toss yielded a sum less than 10. This subset is shown in table 2-2.

Table 2-2. Outcomes Associated With Event C

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	**(2,2)**	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	**(3,3)**	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	**(4,4)**	(4,5)	
(5,1)	(5,2)	(5,3)	(5,4)		
(6,1)	(6,2)	(6,3)			

Within table 2-2, 14 possible outcomes are associated with the event "*the sum of the toss is even, given the sum of the toss is a number less than ten.*"

$$\left\{\begin{matrix}(1,1),(1,3),(1,5),(2,2),(2,4),(2,6),(3,1),(3,3),(3,5)\\(4,2),(4,4),(5,1),(5,3),(6,2)\end{matrix}\right\}$$

Therefore, the probability of this event is $P(B|C)=\frac{14}{30}$ ♦

In example 2-2, observe $P(B|C)$ was obtained directly from a subset of the sample space Ω; furthermore, $P(B|C)=14/30<P(B)=1/2$ in example 2-2. If A and B are events in the same sample space Ω, then $P(A|B)$ is the probability of event A within the subset of the sample space defined by event B. Formally, the *conditional probability of event A given event B has occurred* is defined as

$$P(A|B)=\frac{P(A\cap B)}{P(B)}\equiv\frac{P(AB)}{P(B)} \qquad (2\text{-}2)$$

where $P(B)>0$. Likewise, the *conditional probability of event B given event A has occurred* is defined as

$$P(B|A)=\frac{P(B\cap A)}{P(A)}\equiv\frac{P(BA)}{P(A)} \qquad (2\text{-}3)$$

where $P(A) > 0$. In particular, relating equation 2-3 to example 2-2 (and referring to the computations in example 2-1) we have

$$P(B|C)=\frac{P(B\cap C)}{P(C)}=\frac{\frac{14}{36}}{\frac{30}{36}}=\frac{14}{30}$$

Example 2-3 A proposal team from XYZ Corporation has offers on two contracts A and B. The team made subjective probability assignments on the chances of winning these contracts. They assessed a 40 percent chance on the event winning contract A, a 50 percent chance on the event winning contract B, and a 30 percent chance on the event winning both contracts. Given this, what is the probability of

a) Winning at least one of these contracts?

b) Winning contract A and not winning contract B?

c) Winning contract A if the proposal team has won at least one contract?

Solution

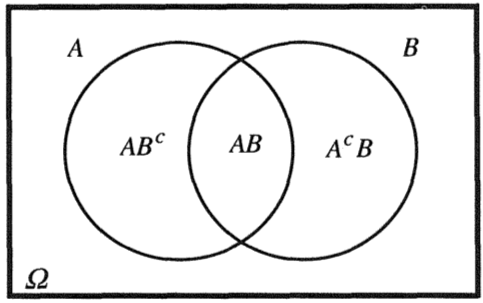
a) Winning at least one contract means winning either contract A *or* contract B or both contracts. This event is represented by the set $A \cup B$. From theorem 2-4

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

therefore $$P(A \cup B) = 0.40 + 0.50 - 0.30 = 0.60$$

b) The event winning contract A and not winning contract B is represented by the set $A \cap B^c$. From the Venn diagram below, observe that

$$P(A) = P((A \cap B^c) \cup (A \cap B))$$



Since the events $A \cap B^c$ and $A \cap B$ are disjoint, from theorem 2-4 we have

$$P(A) = P(A \cap B^c) + P(A \cap B)$$

This is equivalent to $P(A \cap B^c) = P(A) - P(A \cap B)$; therefore,

$$P(A \cap B^c) = P(A) - P(A \cap B) = 0.40 - 0.30 = 0.10$$

c) If the proposal team has won one of the contracts, the probability of winning contract A must be revised (or conditioned) on this information. This means we must compute $P(A|A \cup B)$. From equation 2-2

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

Since $P(A) = P(A \cap (A \cup B))$ we have

$$P(A|A \cup B) = \frac{P(A \cap (A \cup B))}{P(A \cup B)} = \frac{P(A)}{P(A \cup B)} = \frac{0.40}{0.60} = \frac{2}{3} \approx 0.67 \blacklozenge$$

A consequence of conditional probability is obtained if we multiply equations 2-2 and 2-3 by $P(B)$ and $P(A)$, respectively. This multiplication yields*

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A) \qquad (2\text{-}4)$$

Equation 2-4 is known as the *multiplication rule*. The multiplication rule provides a way to express the probability of the intersection of two events in terms of their conditional probabilities. An illustration of this rule is presented in example 2-4.

Example 2-4 A box contains memory chips of which 3 are defective and 97 are nondefective. Two chips are drawn at random, one after the other, without replacement. Determine the probability

a) Both chips drawn are defective.

b) The first chip is defective and the second chip is nondefective.

Solution

a) Let A and B denote the event the first and second chips drawn from the box are *defective*, respectively. From the multiplication rule, we have

$$\begin{aligned} P(A \cap B) &= P(AB) = P(A)P(B|A) \\ &= P(\text{1st chip defective})P(\text{2nd chip defective} \mid \text{1st chip defective}) \\ &= \frac{3}{100}\left(\frac{2}{99}\right) = \frac{6}{9900} \end{aligned}$$

---

* From the commutative law $P(A \cap B) = P(B \cap A)$, which is equivalent to $P(AB) = P(BA)$.

b) To determine the probability the first chip drawn is defective and the second chip is *nondefective*, let C denote the event the second chip drawn is nondefective. Thus,

$$P(A \cap C) = P(AC) = P(A)P(C|A)$$

$$= P(\text{1st chip defective})P(\text{2 nd chip nondefective} \mid \text{1st chip defective})$$

$$= \frac{3}{100}\left(\frac{97}{99}\right) = \frac{291}{9900} \blacklozenge$$

In this example the sampling was performed *without replacement*. Suppose the chips sampled were *replaced*; that is, the first chip selected was replaced before the second chip was selected. In that case, the probability of a defective chip being selected on the second drawing is independent of the outcome of the first chip drawn. Specifically,

$$P(\text{2nd chip defective}) = P(\text{1st chip defective}) = 3/100$$

So $P(A \cap B) = \frac{3}{100}\left(\frac{3}{100}\right) = \frac{9}{10000}$ and $P(A \cap C) = \frac{3}{100}\left(\frac{97}{100}\right) = \frac{291}{10000}$

Independent Events

Two events A and B are said to be *independent* if and only if

$$P(A \cap B) = P(A)P(B) \tag{2-5}$$

and *dependent* otherwise. The events $A_1, A_2, \ldots, A_n$ are (mutually) *independent* if and only if for every set of indices $i_1, i_2, \ldots, i_k$ between 1 and n, inclusive,

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_k}),\ (k = 2, \ldots, n)$$

For instance, events A_1, A_2, and A_3, are independent (or mutually independent) if the following equations are satisfied

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3) \tag{2-5a}$$

$$P(A_1 \cap A_2) = P(A_1)P(A_2) \quad \text{(2-5b)}$$

$$P(A_1 \cap A_3) = P(A_1)P(A_3) \quad \text{(2-5c)}$$

$$P(A_2 \cap A_3) = P(A_2)P(A_3) \quad \text{(2-5d)}$$

It is possible to have three events A_1, A_2, and A_3 for which equations 2-5b through 2-5d hold but equation 2-5a does not hold. Mutual independence implies pairwise independence, in the sense that equations 2-5b through 2-5d hold, but the converse is not true.

There is a close relationship between independent events and conditional probability. To see this, suppose events A and B are independent. This implies $P(AB) = P(A)P(B)$. From this, equations 2-2 and 2-3 become, respectively, $P(A|B) = P(A)$ and $P(B|A) = P(B)$. Thus, when two events are independent the occurrence of one event has no impact on the probability the other event occurs.

To illustrate the concept of independence, suppose a fair die is tossed. Let A be the event an odd number appears. Let B be the event one of these numbers $\{2,3,5,6\}$ appears, then $P(A) = 1/2$ and $P(B) = 2/3$. Since $A \cap B$ is the event represented by the set $\{3,5\}$, we can readily state $P(A \cap B) = 1/3$. Therefore, $P(A \cap B) = P(A)P(B)$ and we conclude events A and B are independent. Dependence can be illustrated by tossing two fair dice, as described in example 2-1. In that example, A was the event the sum of the toss is odd and B was the event the sum of the toss is even. In the solution to example 2-1, it was shown $P(A \cap B) = 0$ and $P(A)$ and $P(B)$ were each 1/2. Since $P(A \cap B) \neq P(A)P(B)$ we would conclude events A and B are dependent, in this case.

It is important not to confuse the meaning of independent events with mutually exclusive events. If events A and B are mutually exclusive, the event A *and* B is empty; that is, $A \cap B = \varnothing$. This implies $P(A \cap B) = P(\varnothing) = 0$. If

events A and B are *independent* with $P(A) \neq 0$ and $P(B) \neq 0$, then A and B cannot be mutually exclusive since $P(A \cap B) = P(A)P(B) \neq 0$.

Theorem 2-6 For any two independent events A_1 and A_2

$$P(A_1 \cup A_2) = 1 - P(A_1^c)P(A_2^c)$$

Proof From theorem 2-1 we can write

$$P(A_1 \cup A_2) = 1 - P((A_1 \cup A_2)^c)$$

From De Morgan's law (section 2.2) $(A_1 \cup A_2)^c = A_1^c \cap A_2^c$; therefore,

$$P(A_1 \cup A_2) = 1 - P(A_1^c \cap A_2^c) \equiv 1 - P(A_1^c A_2^c)$$

Since events A_1 and A_2 are independent, the above expression becomes

$$P(A_1 \cup A_2) = 1 - P(A_1^c)P(A_2^c) \blacklozenge \qquad (2\text{-}6)$$

To prove this theorem, we used a result that if A_1 and A_2 are independent then A_1^c and A_2^c are also independent. Showing this is left as an exercise for the reader. Extending theorem 2-6, it can be shown that if $A_1, A_2, \ldots, A_n$ are independent then

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n) &= 1 - P(A_1^c A_2^c A_3^c \ldots A_n^c) \\ &= 1 - P(A_1^c)P(A_2^c)P(A_3^c)\ldots P(A_n^c) \qquad (2\text{-}7) \end{aligned}$$

2.5 Bayes' Rule

Suppose we have a collection of events A_i representing possible conjectures about a topic. Furthermore, suppose we have some initial probabilities associated with the "truth" of these conjectures. Bayes' rule* provides a way to update (revise) initial probabilities when new information about these conjectures is evidenced.

Bayes' rule is a consequence of conditional probability. Suppose we partition a sample space Ω into a finite collection of three mutually exclusive

---

* Named in honor of Thomas Bayes (1702-1761), an English minister and mathematician.

events (see figure 2-4). Define these events as A_1, A_2, and A_3, where $A_1 \cup A_2 \cup A_3 = \Omega$. Let B denote an arbitrary event contained in Ω. From figure 2-4 we can write the event B as

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B) \tag{2-8}$$

Since the events $(A_1 \cap B), (A_2 \cap B), \text{and } (A_3 \cap B)$ are mutually exclusive, we can apply axiom 3 and write

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

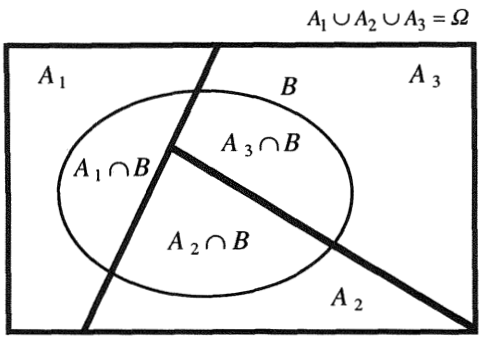


Figure 2-4. Partitioning Ω Into Three Mutually Exclusive Sets

From the multiplication rule given in equation 2-4, $P(B)$ can be expressed in terms of conditional probability as

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) \tag{2-9}$$

Equation 2-9 is known as the *total probability law*. Its generalization is

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i) \tag{2-10}$$

where $\Omega = \bigcup_{i=1}^{n} A_i$ and $A_i \cap A_j = \varnothing$ and $i \neq j$. The conditional probability for each event A_i given event B has occurred is

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{P(B)} \tag{2-11}$$

When the total probability law is applied to equation 2-11 we have

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)} \qquad (2\text{-}12)$$

Equation 2-12 is known as *Bayes' Rule.*

Example 2-5 The ChipyTech Corporation has three divisions D_1, D_2, and D_3 that each manufacture a specific type of microprocessor chip. From the total annual output of chips produced by the corporation, D_1 manufactures 35%, D_2 manufactures 20%, and D_3 manufactures 45%. Data collected from the quality control group indicate 1% of the chips from D_1 are defective, 2% of the chips from D_2 are defective, and 3% of the chips from D_3 are defective. Suppose a chip was randomly selected from the total annual output produced and it was found to be defective. What is the probability it was manufactured by D_1? By D_2? By D_3?

Solution Let A_i denote the *event* the selected chip was produced by division D_i ($i = 1,2,3$). Let B denote the event the selected chip is defective. To determine the probability the defective chip was manufactured by D_i we must compute the conditional probability $P(A_i|B)$ for $i = 1,2,3$. From the information provided, we have

$$P(A_1) = 0.35,\ P(A_2) = 0.20, \text{ and } P(A_3) = 0.45$$
$$P(B|A_1) = 0.01,\ P(B|A_2) = 0.02,\ P(B|A_3) = 0.03$$

The total probability law and Bayes' rule will be used to determine $P(A_i|B)$ for each i = 1, 2, and 3. Recall from equation 2-9 $P(B)$ can be written as

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3)$$
$$P(B) = 0.35(0.01) + 0.20(0.02) + 0.45(0.03) = 0.021$$

and from Bayes' rule we can write

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)} = \frac{P(A_i)P(B|A_i)}{P(B)}$$

from which

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(B)} = \frac{0.35(0.01)}{0.021} = 0.167$$

$$P(A_2|B) = \frac{P(A_2)P(B|A_2)}{P(B)} = \frac{0.20(0.02)}{0.021} = 0.190$$

$$P(A_3|B) = \frac{P(A_3)P(B|A_3)}{P(B)} = \frac{0.45(0.03)}{0.021} = 0.643$$

◆

Table 2-3 provides a comparison of $P(A_i)$ with $P(A_i|B)$ for each i = 1, 2, 3.

Table 2-3. Comparison of $P(A_i)$and $P(A_i|B)$

i	$P(A_i)$	$P(A_i\|B)$
1	0.35	0.167
2	0.20	0.190
3	0.45	0.643

The probabilities given by $P(A_i)$, are the probabilities the selected chip will have been produced by division D_i before it is randomly selected and before it is known whether or not the chip is defective. Therefore, $P(A_i)$ are the *prior*, or *a priori* (before the fact) probabilities. The probabilities given by $P(A_i|B)$ are the probabilities the selected chip was produced by division D_i after it is known the selected chip is defective. Therefore, $P(A_i|B)$ are the *posterior*, or *a posteriori* (after the fact) probabilities. Bayes' rule provides a means for the computation of posterior probabilities from the known prior probabilities $P(A_i)$ and the conditional probabilities $P(B|A_i)$ for a particular situation or experiment.

Bayes' rule established a philosophy in probability theory that became known as *Bayesian inference* and *Bayesian decision theory*. These areas play an important role in the application of probability theory to cost and systems engineering problems. In equation 2-10, we may think of A_i as representing possible states of nature to which an analyst or systems engineer assigns subjective probabilities. These subjective probabilities are the prior probabilities, which are often premised on personal judgments based on past experience. In general, Bayesian methods offer a powerful way to revise, or update, probability assessments as new (or refined) information becomes available.

Exercises

1. State the interpretation of probability implied by the following:
   a) The probability a tail appears on the toss of a fair coin is 1/2.
   b) After recording the outcomes of 50 tosses of a fair coin, the probability a tail appears is 0.54.
   c) It is with certainty the coin is fair!
   d) The probability is 60 percent that the stock market will close 500 points above yesterday's closing count.
   e) The design team believes there is less than a 5 percent chance the new microchip will require more than 12,000 gates.

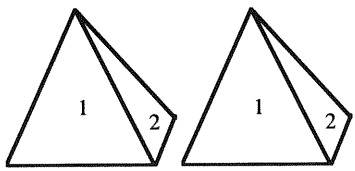
2. A sack contains 20 marbles exactly alike in size but different in color. Suppose the sack contains 5 blue marbles, 3 green marbles, 7 red marbles, 2 yellow marbles, and 3 black marbles. Picking a single marble from the sack and then replacing it, what is the probability of choosing the following:

a) Blue marble? b) Green marble? c) Red marble?
d) Yellow marble? e) Black marble? f) Non-blue marble
g) Red or non-red marble?

3. If a fair coin is tossed, what is the probability of not obtaining a head? What is the probability of the event: (a head or not a head)?

4. Show the probability of the event: (A or A complement) is *always* unity.

5. Suppose two tetrahedrons (4-sided polygons) are randomly tossed. Assuming the tetrahedrons are weighted fair, determine the set of all possible outcomes Ω. Assume each face is numbered 1, 2, 3, and 4.



Two Tetrahedron's for Exercise 5

Let the sets A, B, C, and D represent the following events

A: The sum of the toss is even

B: The sum of the toss is odd

C: The sum of the toss is a number less than 6

D: The toss yielded the same number on each upturned face

a) Find $P(A)$, $P(B)$, $P(C)$, $P(A \cap B)$, $P(A \cup B)$, $P(B \cup C)$, and $P(B \cap C \cap D)$.

b) Verify $P(A \cup B)^c = P(A^c \cap B^c)$ (De Morgan's Law).

6. The XYZ Corporation has offers on two contracts A and B. Suppose the proposal team made the following subjective probability assessments: the chance of winning contract A is 40 percent, the chance of winning

contract B is 20 percent, the chance of winning contract A or contract B is 60 percent, the chance of winning both contracts is 10 percent.

a) Explain why the above set of probability assignments is *inconsistent* with the axioms of probability.

b) What must $P(B)$ equal such that it and the set of other assigned probabilities specified above are consistent with the axioms of probability?

7. Suppose a coin is balanced such that tails appears 3 times more frequently than heads. Show the probability of obtaining a tail with such a coin is 3/4. What would you expect this probability to be if the coin was fair; that is, equally balanced?

8. Suppose the sample space of an experiment is given by $\Omega = A \cup B$. Compute $P(A \cap B)$ if $P(A) = 0.25$ and $P(B) = 0.80$.

9. If A and B are disjoint subsets of Ω show that

   a) $P(A^c \cup B^c) = 1$

   b) $P(A^c \cap B^c) = 1 - [P(A) + P(B)]$

10. Two missiles are launched. Suppose there is a 75 percent chance missile A hits the target and a 90 percent chance missile B hits the target. If the probability missile A hits the target is *independent* of the probability missile B hits the target, determine the probability missile A or missile B hits the target. Find the probability needed for missile A such that if the probability of missile B hitting the target remains at 90 percent, the probability missile A or missile B hits the target is 0.99.

11. Suppose A and B are independent events. Show that

a) The events A^c and B^c are independent.

b) The events A and B^c are independent.

c) The events A^c and B are independent.

12. Suppose A and B are independent events with $P(A) = 0.25$ and $P(B) = 0.55$. Determine the probability

a) At least one event occurs.

b) Event B occurs but event A does not occur.

13. Suppose A and B are independent events with $P(A) = r$ and the probability that "at least A or B occurs" is s. Show the only value for $P(B)$ is $(s-r)(1-r)^{-1}$.

14. In exercise 5, suppose event C has occurred. Enumerate the set of remaining possible outcomes. From this set compute $P(B)$. Compare this with $P(B|C)$ where $P(B|C)$ is determined from the definition of conditional probability.

15. At a local sweet shop, 10 percent of all customers buy ice cream, 2 percent buy fudge, and 1 percent buy both ice cream and fudge. If a customer selected at random bought fudge, what is the probability the customer bought an ice cream? If a customer selected at random bought ice cream, what is the probability the customer bought fudge?

16. For any two events A and B, show that $P(A|A \cap (A \cap B)) = 1$.

17. A production lot contains 1000 microchips of which 10 percent are defective. Two chips are successively drawn at random without replacement. Determine the probability

a) Both chips selected are nondefective.

b) Both chips are defective.

c) The first chip is defective and the second chip is nondefective.

d) The first chip is nondefective and the second chip is defective.

18. Suppose the sampling scheme in exercise 17 was with replacement, that is, the first chip is returned to the lot before the second chip is drawn. Show how the probabilities computed in exercise 17 are changed.

19. Spare power supply units for a communications terminal are provided to the government from three different suppliers A_1, A_2, and A_3. Thirty percent come from A_1, twenty percent come from A_2, and fifty percent come from A_3. Suppose these units occasionally fail to perform according to their specifications and the following has been observed: 2 percent of those supplied by A_1 fail, 5 percent of those supplied by A_2 fail, and 3 percent of those supplied by A_3 fail. What is the probability any one of these units provided to the government will perform *without* failure?

20. In a single day, ChipyTech Corporation's manufacturing facility produces 10,000 microchips. Suppose machines A, B, and C individually produce 3000, 2500, and 4500 chips daily. The quality control group has determined the output from machine A has yielded 35 defective chips, the output from machine B has yielded 26 defective chips, and the output from machine C has yielded 47 defective chips.

a) If a chip was selected at random from the daily output, what is the probability it is defective?

b) What is the probability a randomly selected chip was produced by machine A? By machine B? By machine C?

c) Suppose a chip *was* randomly selected from the day's production of

10,000 microchips and it was found to be defective. What is the probability it was produced by machine A? By machine B? By machine C?

References

1. Feller, W. 1968. *An Introduction to Probability Theory and Its Applications*, vol. 1, 3rd ed (revised). New York: John Wiley & Sons, Inc.

2. Hitch, C. J. 1955. *An Appreciation of Systems Analysis*, P-699. Santa Monica, California: The RAND Corporation.

Chapter 3

Distributions and the Theory of Expectation

There is only one thing about which I am certain, and this is that there is very little about which one can be certain.
W. Somerset Maugham
The Summing Up (1938)

We dance round in a ring and suppose
But the Secret sits in the middle and knows.
Robert Frost
The Secret Sits [1942]

3.1 Random Variables and Probability Distributions

Consider the experiment of tossing two fair dice described in example 2-1 (chapter 2). Suppose x represents the sum of the toss. Define X as a variable that takes on only values given by x. If the sum of the toss is 2 then $X = 2$; if the sum of the toss is 3 then $X = 3$; if the sum of the toss is 7 then $X = 7$. Numerical values of X are associated with *events* defined from the sample space Ω for this experiment, which was given in table 2-1 (chapter 2). In particular,

$X = 2$ is associated with only this simple event $\{(1,1)\}$*

$X = 3$ is associated with only these two simple events $\{(1,2)\},\{(2,1)\}$

$X = 7$ is associated with only these six simple events $\{(1,6)\},\{(2,5)\},\{(3,4)\},\{(4,3)\},\{(5,2)\},\{(6,1)\}$

In the above, we say X is a random variable. This is illustrated in figure 3-1. Formally, a *random variable* is a real-valued function defined over a sample space. The sample space is the *domain* of a random variable. Traditionally, random variables are denoted by capital letters such as X, W, and Z.

---

* The outcomes from tossing two dice are recorded as (d_1,d_2), where d_1 and d_2 are the numbers appearing on the upturned faces of the first and second die, respectively. Therefore, in this discussion, $x = d_1 + d_2$.

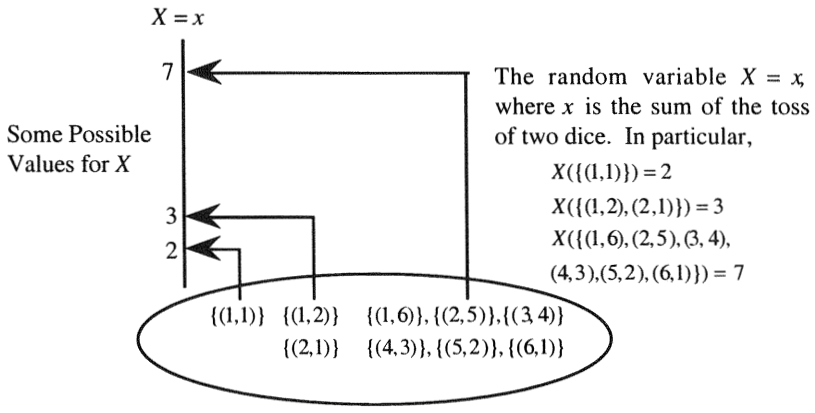


Figure 3-1. Some Possible Values of a Random Variable

The event $X = x$ is equivalent to

$$\{X = x\} \equiv \{\omega \in \Omega \,|\, X(\omega) = x\}$$

This represents a subset of Ω consisting of all sample points ω such that $X(\omega) = x$. In figure 3-1, the event $\{X = 3\}$ is equivalent to

$$\{X = 3\} \equiv \{(1,2),(2,1)\}$$

The probability of the event $\{X = x\}$ is equivalent to

$$P(\{X = x\}) \equiv P(\{\omega \in \Omega \,|\, X(\omega) = x\})$$

In figure 3-1, the probability of the event $\{X = 3\}$ is equivalent to

$$P(\{X = 3\}) \equiv P(\{(1,2),(2,1)\}) = 2/36$$

For convenience, the notation $P(\{X = x\}) \equiv P(X = x)$ is adopted in this book.

Random variables can be characterized as discrete or continuous. A random variable is *discrete* if its set of possible values is finite or countably infinite. A random variable is *continuous* if its set of possible values is uncountable.

Discrete Random Variables

Consider again the simple experiment of tossing a pair of fair dice. Let the random variable X represent the sum of the toss. The sample space Ω for this experiment consists of thirty-six outcomes given in table 2-1 (chapter 2). The random variable X is discrete since the *only* possible values are $x = 2,3,4,5,6,\ldots,12$. The function that describes probabilities associated with the event $\{X = x\}$, for all *feasible values* of x, is shown in figure 3-2. This function is known as the probability function of X.

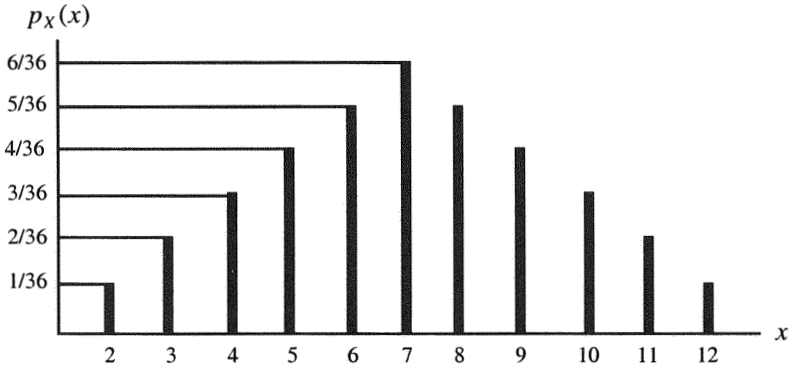


Figure 3-2. Probability Function for — the Sum of Two Dice Tossed

The *probability function* of a *discrete random variable* X is defined as

$$p_X(x) = P(X = x) \qquad (3\text{-}1)$$

The probability function is also referred to as the *probability mass function* or the *frequency function* of X. The probability function associates probabilities to events described by distinct (single) points of interest. Over all *feasible* (possible) values of x, probability functions satisfy, by the axioms of probability, the following conditions:

a) $p_X(x) \geq 0$ b) $\sum_x p_X(x) = 1$

If x is *not a feasible* value of X then

$$p_X(x) = P(X = x) = P(\varnothing) = 0$$

It is often of interest to determine probabilities associated with events of the form $\{X \leq x\}$. For instance, suppose we wanted the probability that the sum of the numbers resulting from the toss of two fair dice will not exceed seven. This is equivalent to computing $P(X \leq 7)$; in this instance, we have $P(X \leq 7) = P(\{X = 2\} \cup \{X = 3\} \cup \ldots \cup \{X = 7\})$. Thus, X can take a value not exceeding seven if and only if X takes on one of the values $2,3,\ldots,7$. Since the events $\{X = 2\},\{X = 3\},\ldots,\{X = 7\}$ are mutually exclusive, from axiom 3 (chapter 2) and figure 3-2 we have

$$P(X \leq 7) = P(X = 2) + P(X = 3) + \ldots + P(X = 7) = \frac{21}{36}$$

The function that produces probabilities for events of the form $\{X \leq x\}$ is known as the cumulative distribution function. Formally, if X is a discrete random variable then its *cumulative distribution function* (CDF) is defined by

$$F_X(x) = P(X \leq x) = \sum_{t \leq x} p_X(t) \quad (-\infty < x < \infty) \tag{3-2}$$

In terms of the above definition, we would write $P(X \leq 7)$ as

$$F_X(7) = P(X \leq 7) = \sum_{t \leq 7} p_X(t) = p_X(2) + p_X(3) + \ldots + p_X(7) = 21/36$$

where, from equation 3-1, $p_X(x) = P(X = x)$ for $x = 2,3,\ldots,7$.

The CDF for the random variable with probability function in figure 3-2 is pictured in figure 3-3. Notice the CDF is a "staircase" or "step" function. This is a characteristic of cumulative distribution functions for *discrete random variables*. The height of the "step" along the CDF is the probability the value associated with that step occurs. For instance, in figure 3-3, the probability that $X = 3$ is the height of the step (jump) between $X = 2$ and $X = 3$; that is, $P(X = 3) = \frac{3}{36} - \frac{1}{36} = \frac{2}{36}$.

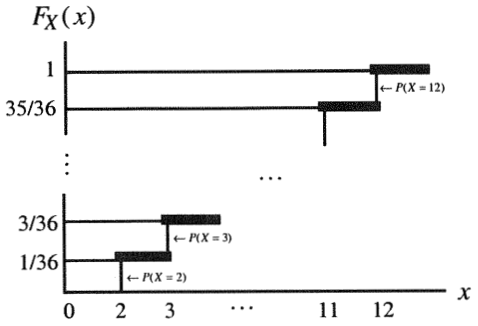


Figure 3-3. Cumulative Distribution Function for — the Sum of Two Dice Tossed

If X is a discrete random variable and a is any real number that is a feasible (or possible) value of X, then $P(X = a) = p_X(a)$ is equal to the height of the step (jump) of $F_X(x)$ at $x = a$.

The following presents theorems for determining probabilities from the CDF of a discrete random variable X. In the theorems below, a and b are real numbers with $a < b$.

Theorem 3-1 The probability of $\{X > a\}$ is $1 - F_X(a)$.

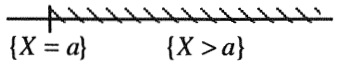
Proof Let A denote the event $\{X > a\}$; then $A^c = \{X \le a\}$; from theorem 2-1 and the definition given by equation 3-2, it immediately follows that

$$P(X > a) = 1 - P(X \le a) = 1 - F_X(a)$$

Theorem 3-2 The probability of $\{X \ge a\}$ is $1 - F_X(a) + P(X = a)$.

Proof We can write the event $\{X \ge a\}$ as the union of two mutually exclusive events $\{X = a\}$ and $\{X > a\}$; that is,

$$\{X \ge a\} = \{X = a\} \cup \{X > a\}$$



From theorems 2-4 and 3-1 we have

$$P(X \geq a) = P(\{X = a\} \cup \{X > a\}) = P(X = a) + P(X > a)$$
$$= P(X = a) + 1 - F_X(a) \equiv 1 - F_X(a) + P(X = a)$$

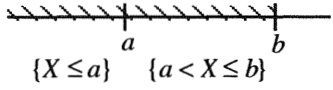
Theorem 3-3 The probability of $\{X < a\}$ is $F_X(a) - P(X = a)$.

Proof This is a direct consequence of theorems 3-2 and 3-1. The proof is left as an exercise for the reader.

Theorem 3-4 The probability of $\{a < X \leq b\}$ is $F_X(b) - F_X(a)$.

Proof We can write the event $\{X \leq b\}$ as the union of two mutually exclusive events $\{X \leq a\}$ and $\{a < X \leq b\}$; that is,

$$\{X \leq b\} = \{X \leq a\} \cup \{a < X \leq b\}$$



From theorem 2-4

$$P(X \leq b) = P(\{X \leq a\} \cup \{a < X \leq b\}) = P(X \leq a) + P(a < X \leq b)$$

Thus, $F_X(b) = F_X(a) + P(a < X \leq b)$

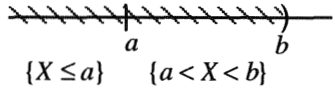
Therefore

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

Theorem 3-5 The probability of $\{a < X < b\}$ is $F_X(b) - F_X(a) - P(X = b)$.

Proof We can write the event $\{X < b\}$ as the union of two mutually exclusive events $\{X \leq a\}$ and $\{a < X < b\}$; that is,

$$\{X < b\} = \{X \leq a\} \cup \{a < X < b\}$$



From theorem 2-4

$$P(X < b) = P(\{X \leq a\} \cup \{a < X < b\}) = P(X \leq a) + P(a < X < b)$$

It follows that $P(X < b) - P(X \le a) = P(a < X < b)$

From theorem 3-3, $P(X < b) = F_X(b) - P(X = b)$; since $P(X \le a) = F_X(a)$ we have $P(a < X < b) = F_X(b) - F_X(a) - P(X = b)$, which was to be shown.

Theorem 3-6 The probability of $\{a \le X < b\}$ is

$$F_X(b) - F_X(a) + P(X = a) - P(X = b).$$

Proof We can write the event $\{a \le X < b\}$ as the union of two mutually exclusive events $\{X = a\}$ and $\{a < X < b\}$; that is,

$$\{a \le X < b\} = \{X = a\} \cup \{a < X < b\}$$

From theorem 2-4

$$P(a \le X < b) = P(\{X = a\} \cup \{a < X < b\}) = P(X = a) + P(a < X < b)$$

From theorem 3-5 $P(a < X < b) = F_X(b) - F_X(a) - P(X = b)$; therefore,

$$P(a \le X < b) = F_X(b) - F_X(a) + P(X = a) - P(X = b)$$

Theorem 3-7 The probability of $\{a \le X \le b\}$ is $F_X(b) - F_X(a) + P(X = a)$.

Proof We can write the event $\{a \le X \le b\}$ as the union of three mutually exclusive events $\{X = a\}$, $\{a < X < b\}$, and $\{X = b\}$. That is,

$$\{a \le X \le b\} = \{X = a\} \cup \{a < X < b\} \cup \{X = b\}$$

From axiom 3 (chapter 2) and theorem 3-5

$$\begin{aligned} P(a \le X \le b) &= P(\{X = a\} \cup \{a < X < b\} \cup \{X = b\}) \\ &= P(X = a) + P(a < X < b) + P(X = b) \\ &= P(X = a) + [F_X(b) - F_X(a) - P(X = b)] + P(X = b) \\ &= F_X(b) - F_X(a) + P(X = a) \blacklozenge \end{aligned}$$

The following presents the first of many case discussions in this book. The discussion addresses how a corporation might assess the chance of making a profit on a new electronics product.

*Case Discussion** *3-1* ChipyTech Corporation is a major producer and supplier of electronics products to industry world-wide. They are planning to bring a new product to the market. Management needs to know the product's potential for profit and loss during its first year on the market. In addition, they want to know the chance of *not making* a profit the first year. Suppose profit [1] is given by equation 3-3

$$Profit = (U_{Price} - U_{Cost})V \tag{3-3}$$

where U_{Price} is a discrete random variable that represents the product's unit price, U_{Cost} is a discrete random variable that represents the unit cost to manufacture the product, and V is a discrete random variable that represents the product's *sales* volume for year one, which is assumed to be nonzero. A profit exists when $U_{Price} > U_{Cost}$, a loss exists when $U_{Price} < U_{Cost}$, and *no profit* exists when $U_{Price} \leq U_{Cost}$. For purposes of this case discussion, we will assume U_{Price}, U_{Cost}, and V are *independent* random variables.

Suppose the corporation's sales, price, and cost histories for similar products have been analyzed. Further, suppose interviews were carefully conducted with subject matter experts from the engineering and marketing departments of ChipyTech. From the interviews and the historical data, possible values for the product's unit price, unit cost, and sales volume were established along with their respective probabilities of occurrence. Figure 3-4 presents these values for U_{Price}, U_{Cost}, and V.

---

* Adapted and expanded from an example in Park, W. R., and D. E. Jackson. 1984. *Cost Engineering Analysis – A Guide to Economic Evaluation of Engineering Projects*, 2nd ed. New York: John Wiley & Sons, Inc.

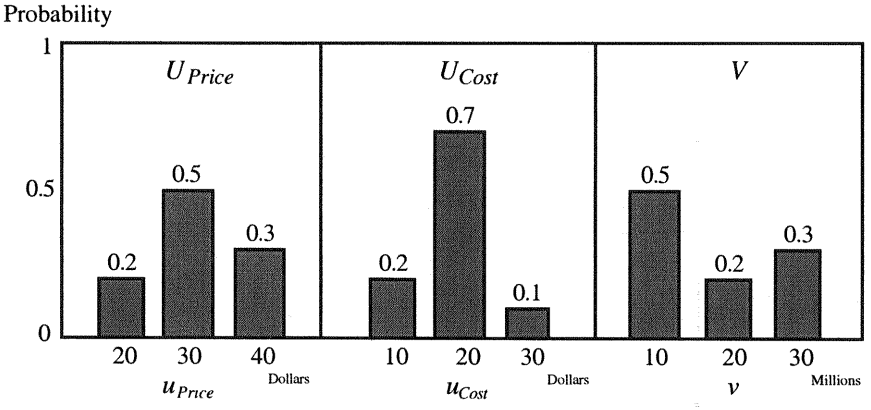


Figure 3-4. Possible Values for U_{Price}, U_{Cost}, and V

To find the dollar range on the product's profit or loss potential, we first list all possible combinations of U_{Price}, U_{Cost}, and V. This list is shown in table 3-1. Since U_{Price}, U_{Cost}, and V are given to be independent random variables,* the probability that any combination of U_{Price} and U_{Cost} and V will occur is

$$P(\{U_{Price} = u_{Price}\} \cap \{U_{Cost} = u_{Cost}\} \cap \{V = v\})$$
$$= P(U_{Price} = u_{Price})P(U_{Cost} = u_{Cost})P(V = v) \qquad (3\text{-}4)$$

where values for $P(U_{Price} = u_{Price})$, $P(U_{Cost} = u_{Cost})$, and $P(V = v)$ are given in figure 3-4. For example, the probability the new product will have a unit price of 20 dollars and a unit cost of 10 dollars and a sales volume of 10 million (the first year) is

$$P(\{U_{Price} = 20\} \cap \{U_{Cost} = 10\} \cap \{V = 10\})$$
$$= P(U_{Price} = 20)P(U_{Cost} = 10)P(V = 10) = 0.020 \qquad (3\text{-}5)$$

* When random variables are independent their associated events are independent. This is discussed further in chapter 5.

Table 3-1 summarizes the possible values for *Profit*. Table 3-1 also shows the probability *Profit* takes a value according to a specific combination of U_{Price}, U_{Cost}, and V. From table 3-1, observe there is a potential loss of as much as 300 ($M) and a potential gain of as much as 900 ($M). How probable are these extremes? What is the chance the corporation will *not make* a profit the first year? The following discussion addresses these questions.

From table 3-1 it can be seen there is less than a 1 percent chance (i.e., 0.6 percent) the new product will realize a loss of 300 ($M) during its first year on the market. Similarly, the maximum profit of 900 ($M) has just under a 2 percent chance (i.e., 1.8 percent) of occurring.

Table 3-1. Possible Profits and Their Probabilities

U_{Price} ($)	U_{Cost} ($)	V (Millions)	*Profit* ($M)	Probability
20	10	10	100	0.020
20	10	20	200	0.008
20	10	30	300	0.012
20	**20**	**10**	**0**	**0.070**
20	**20**	**20**	**0**	**0.028**
20	**20**	**30**	**0**	**0.042**
20	**30**	**10**	**-100**	**0.010**
20	**30**	**20**	**-200**	**0.004**
20	**30**	**30**	**-300**	**0.006**
30	10	10	200	0.050
30	10	20	400	0.020
30	10	30	600	0.030
30	20	10	100	0.175
30	20	20	200	0.070
30	20	30	300	0.105
30	**30**	**10**	**0**	**0.025**
30	**30**	**20**	**0**	**0.010**
30	**30**	**30**	**0**	**0.015**

Table 3-1. Possible Profits and Their Probabilities
(Concluded)

U_{Price} ($)	U_{Cost} ($)	V (Millions)	*Profit* ($M)	Probability
40	10	10	300	0.030
40	10	20	600	0.012
40	10	30	900	0.018
40	20	10	200	0.105
40	20	20	400	0.042
40	20	30	600	0.063
40	30	10	100	0.015
40	30	20	200	0.006
40	30	30	300	0.009
Total Probability				1

The corporation will *not make* a profit (i.e., $Profit \leq 0$) when $U_{Price} \leq U_{Cost}$. There are nine events in table 3-1 (shown by the bold-faced figures) that produce $Profit \leq 0$. Let these events be represented by $A_1, A_2, \ldots, A_9$, where

$$A_1 = \{\{U_{Price} = 20\} \cap \{U_{Cost} = 20\} \cap \{V = 10\}\}$$
$$A_2 = \{\{U_{Price} = 20\} \cap \{U_{Cost} = 20\} \cap \{V = 20\}\}$$
$$\vdots$$
$$A_9 = \{\{U_{Price} = 30\} \cap \{U_{Cost} = 30\} \cap \{V = 30\}\}$$

These events are mutually exclusive. Therefore, from axiom 3 (chapter 2) the probability that $Profit \leq 0$ is

$$P(Profit \leq 0) = P(\bigcup_{i=1}^{9} A_i) = \sum_{i=1}^{9} P(A_i) = 0.210 \qquad (3\text{-}6)$$

where each $P(A_i)$ is given in table 3-1.

Table 3-1 can also be used to develop the *probability function* for the random variable *Profit*. Since *Profit* is given (in this discussion) to be a discrete random variable, its probability function is

$$p_{Profit}(x) = P(Profit = x) \tag{3-7}$$

where feasible values of x are given in table 3-1. Figure 3-5 is the graph of $p_{Profit}(x)$. Among the many useful aspects of the probability function is identifying the value of x associated with the highest probability of occurrence. In figure 3-5, a profit of 200 ($M) has the highest probability of occurrence. A number of other computations can be determined from $p_{Profit}(x)$. For example, from figure 3-5 we have

$$P(Profit \le 0) = p_{Profit}(-300) + p_{Profit}(-200) + p_{Profit}(-100) + p_{Profit}(0) = 0.210$$

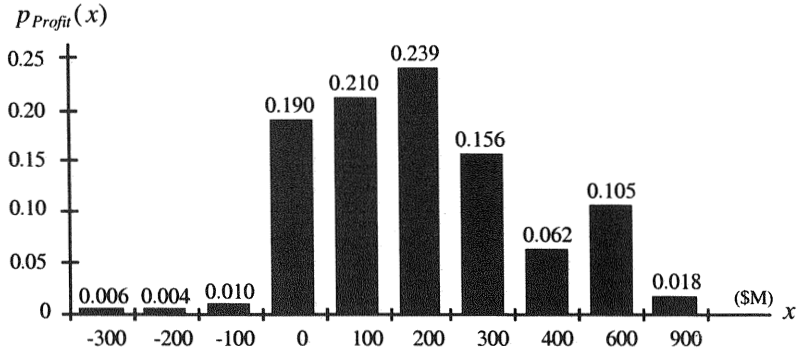


Figure 3-5. Probability Function for *Profit* — Case Discussion 3-1

Notice that $P(Profit \le 0)$ is really the value of the *cumulative distribution function* for *Profit* at $x = 0$. From equation 3-2, the CDF of *Profit* is

$$F_{Profit}(x) = P(Profit \le x) = \sum_{t \le x} p_{Profit}(t)$$

Equation 3-8 presents $F_{Profit}(x)$ for this case discussion.

$$F_{Profit}(x) = \begin{cases} 0 & \text{if } -300 < x \\ 0.006 & \text{if } -300 \le x < -200 \\ 0.010 & \text{if } -200 \le x < -100 \\ 0.020 & \text{if } -100 \le x < 0 \\ 0.210 & \text{if } 0 \le x < 100 \\ 0.420 & \text{if } 100 \le x < 200 \\ 0.659 & \text{if } 200 \le x < 300 \\ 0.815 & \text{if } 300 \le x < 400 \\ 0.877 & \text{if } 400 \le x < 600 \\ 0.982 & \text{if } 600 \le x < 900 \\ 1 & \text{if } 900 \ge x \end{cases} \tag{3-8}$$

The probability that ChipyTech will *not make* a profit can now be read directly from the CDF (equation 3-8), specifically

$$F_{Profit}(0) = P(Profit \le 0) = 0.210$$

A graph of $F_{Profit}(x)$ is presented with example 3-3 (figure 3-12). From equation 3-8, the probability *Profit* will fall within other intervals of interest can be determined. From theorems 3-2 through 3-7, with reference to figure 3-5 and equation 3-8, we have the following:

$$P(Profit \ge 200) = 1 - F_{Profit}(200) + P(Profit = 200) = 1 - 0.659 + 0.239 = 0.580$$

$$P(Profit < 200) = F_{Profit}(200) - P(Profit = 200) = 0.659 - 0.239 = 0.420$$

$$P(200 < Profit \le 600) = F_{Profit}(600) - F_{Profit}(200) = 0.982 - 0.659 = 0.323$$

$$P(200 < Profit < 600) = F_{Profit}(600) - F_{Profit}(200) - P(Profit = 600) = 0.982 - 0.659 - 0.105 = 0.218$$

$$P(200 \le Profit < 600) = F_{Profit}(600) - F_{Profit}(200) + P(Profit = 200) - P(Profit = 600) = 0.457$$

$$P(200 \le Profit \le 600) = F_{Profit}(600) - F_{Profit}(200) + P(Profit = 200) = 0.562$$

In summary, case discussion 3-1 illustrates how fundamental probability concepts such as the axioms, independence, the probability function, and the cumulative distribution function can provide decision-makers insights on profits and their associated probabilities.

Continuous Random Variables

Mentioned in the beginning of this chapter, a random variable is continuous if its set of possible values is uncountable. For instance, suppose T is a random variable representing the duration (in hours) of an electronic device. If the possible values of T are given by $\{t : 0 \le t \le 2500\}$, then T is a *continuous random variable.*

In general, we say X is a *continuous random variable* if there exists a *nonnegative function* $f_X(x)$, defined on the real line, such that for any interval A

$$P(X \in A) = \int_A f_X(x)\,dx$$

The function $f_X(x)$ is called the *probability density function* (PDF) of X. Unlike the probability function for a discrete random variable, the PDF *does not* directly produce a probability — $f_X(a)$ does not produce $p_X(a)$, defined by equation 3-1. In the above, the probability that X is contained in any subset of the real line is determined by integrating $f_X(x)$ over that subset. Since X *must* assume some value on the real line, it will always be true that

$$\int_{-\infty}^{\infty} f_X(x)\,dx \equiv P(X \in (-\infty, \infty)) = 1$$

In this case, the cumulative distribution function (CDF) of the random variable X is defined as

$$F_X(x) = P(X \le x) = P(X \in (-\infty, x]) = \int_{-\infty}^{x} f_X(t)\,dt \qquad (3\text{-}9)$$

A useful way to view equation 3-9 is shown by figure 3-6; if we assume $f_X(x)$ is a PDF, then from calculus we can interpret the probabilities of the events $\{X \le a\}$ and $\{a \le X \le b\}$ as the areas of the indicated regions in figure 3-6.

When X is a *continuous random variable*, the probability $X = a$ is zero; this is because

$$P(X = a) = P(a \le X \le a) = \int_a^a f_X(x)\,dx = 0 \qquad (3\text{-}10)$$

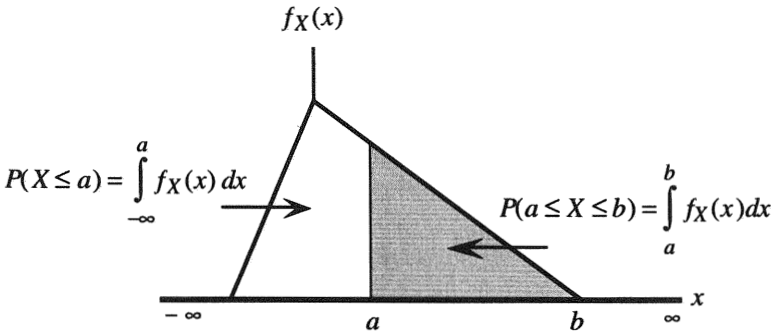


Figure 3-6. A Probability Density Function

From this it is seen the inclusion or exclusion of an interval's endpoints does not affect the probability X falls in the interval; thus, if a and b are any two real numbers

$$P(a < X \le b) = P(a < X < b)$$
$$= P(a \le X < b) = P(a \le X \le b) = F_X(b) - F_X(a) \qquad (3\text{-}11)$$

when X is a *continuous random variable*. Referring back to equation 3-9 note that $F_X(x)$ is determined from $f_X(x)$ by integration. From calculus, it follows that $f_X(x)$ is determined from $F_X(x)$ by differentiation; that is,

$$f_X(x) = \frac{d\left(F_X(x)\right)}{dx}$$

provided the derivative exists at all but a finite number of points.

Properties of $F_X(x)$ for Discrete or Continuous Random Variables

For *any* discrete or continuous random variable, the value of $F_X(x)$ at any x must be a number in the interval $0 \le F_X(x) \le 1$. The function $F_X(x)$ is always continuous from the right. It is nondecreasing as x increases; that is, if $x_1 < x_2$ then $F_X(x_1) \le F_X(x_2)$. Lastly,

$$\lim_{x \to -\infty} F_X(x) = 0 \text{ and } \lim_{x \to \infty} F_X(x) = 1$$

Example 3-1 Let I be a continuous random variable* that represents the size of a software application being developed for a data reduction task. Let I be expressed as the number of delivered source instructions (DSI). Suppose the state of the technical information about the application's functional and performance requirements is very sparse. Given this, suppose subject matter experts have assessed the "true" size of the application will fall somewhere in the interval $[1000, 5000]$. Furthermore, because of the sparseness of available information, suppose their size assessment is such that I could take any value in $[1000, 5000]$ with constant (uniform) probability density.

a) Compute the PDF and the CDF of I.

b) Determine a value x such that $P(I \leq x) = 0.80$.

Solution

a) Figure 3-7 presents a function with the property that its value is c (a constant) at any point in the interval $[1000, 5000]$. For this function to be a probability density, it is necessary to find c such that $\int_{-\infty}^{\infty} f_I(x)dx = 1$. It will then be true that all subintervals of $[1000, 5000]$ that are the same in length will occur with equal, or constant, probability (an exercise for the reader).

---

* In this example, and in many that follow, software size I is *treated* as a continuous random variable. In reality, the number of delivered source instructions for a software application is a positive integer —e.g., "*it takes 4,553 source instructions to pre-process the data stream passing into the radar's primary processor.*" If, for example, we treat software size as a discrete random variable, then each distinct value (assessed by subject matter experts as "possible") also requires an assessment of its probability of occurrence. Although this is a valid way to describe such a random variable, it is not clear how many distinct values (and their associated probabilities) are needed to adequately capture the overall distribution of possible values. In practice, a continuous distribution is often used to describe the range of possible values for a random variable such as software size. This enables subject matter experts to focus on the "shape" that best describes the distribution of probability, rather than assessing individual probabilities associated to each distinct possible value. If needed, the resulting continuous distribution could later be translated into a discrete form.

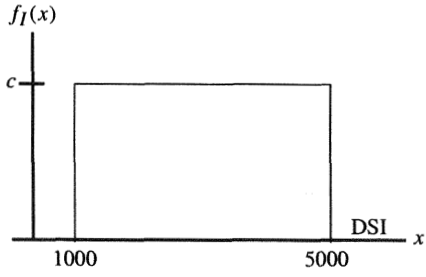


Figure 3-7. Probability Density Function for Example 3-1

From figure 3-7, $f_I(x)$ can be written as

$$f_I(x) = \begin{cases} c & \text{if } 1000 \le x \le 5000 \\ 0 & \text{otherwise} \end{cases} \tag{3-12}$$

For $f_I(x)$ to be a PDF, we need to find c such that

$$\int_{-\infty}^{\infty} f_I(x)dx = \int_{1000}^{5000} c\,dx = 4000c = 1$$

therefore $c = \frac{1}{4000}$. Thus, the PDF of the random variable I is

$$f_I(x) = \begin{cases} \frac{1}{4000} & \text{if } 1000 \le x \le 5000 \\ 0 & \text{otherwise} \end{cases} \tag{3-13}$$

To determine the CDF we must evaluate the integral

$$F_I(x) = P(I \le x) = \int_{-\infty}^{x} f_I(t)\,dt \quad \text{for} \quad -\infty < x < \infty$$

as x moves across the interval $-\infty < x < \infty$. From equation 3-9, and the PDF in equation 3-13, we can write the CDF as

$$F_I(x) = \begin{cases} 0 & \text{if } x < 1000 \\ \int_{1000}^{x} \frac{1}{4000}\,dt = (x-1000)/4000 & \text{if } 1000 \le x < 5000 \\ 1 & \text{if } x \ge 5000 \end{cases} \qquad (3\text{-}14)$$

Notice $F_I(x)$ is a straight line, as illustrated in figure 3-8.

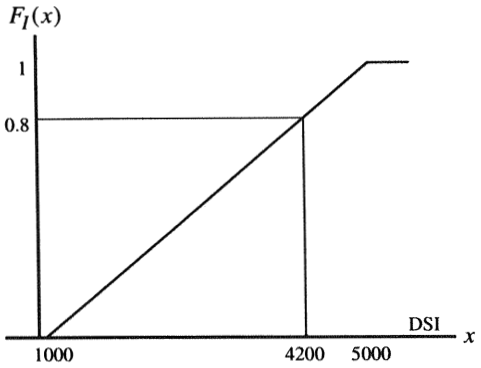


Figure 3-8. The Cumulative Distribution Function for Example 3-1

b) The value of x such that $P(I \le x) = 0.80$ is obtained from equation 3-14 by solving

$$\frac{x-1000}{4000} = 0.80$$

for x. The solution is $x = 4200$. Therefore, there is an 80 percent chance the "true" software size will be less than or equal to 4200 DSI.

Example 3-2 Suppose the probability density function for I in example 3-1 is now defined by the two regions shown in figure 3-9.

a) Find c such that $f_I(x)$ in figure 3-9 is a PDF.

b) Determine $F_I(x)$.

c) Compute $P(I \le 2000)$, $P(2000 < I < 4000)$, $P(2000 < I \le 5000)$.

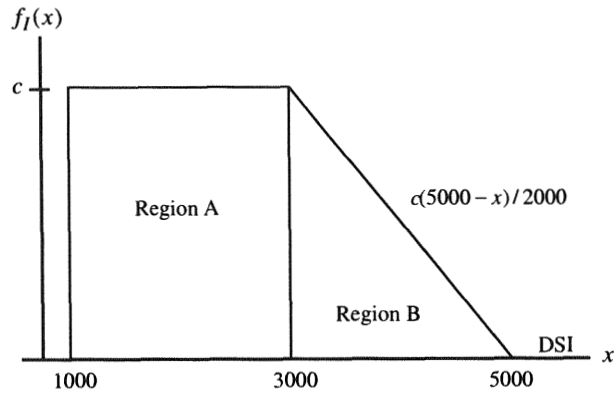


Figure 3-9. Probability Density Function for Example 3-2

Solution

a) From figure 3-9 it can be determined that

$$f_I(x) = \begin{cases} c & \text{if } 1000 \le x < 3000 \\ c(5000-x)/2000 & \text{if } 3000 \le x \le 5000 \end{cases} \tag{3-15}$$

For $f_I(x)$ to be a PDF there must exist a constant c such that

$$\int_{-\infty}^{\infty} f_I(x)\,dx = 1$$

This implies c is the solution to

$$\int_{1000}^{3000} c\,dx + \int_{3000}^{5000} c((5000-x)/2000)\,dx = 1$$

from which c = 1/3000. Thus, the probability density function is

$$f_I(x) = \begin{cases} 1/3000 & \text{if } 1000 \le x < 3000 \\ (5000-x)/6(10^6) & \text{if } 3000 \le x \le 5000 \end{cases} \tag{3-16}$$

b) To determine the cumulative distribution function $F_I(x)$, we must evaluate

$$F_I(x) = P(I \le x) = \int_{-\infty}^{x} f_I(t)\,dt \quad \text{for} \quad -\infty < x < \infty \tag{3-17}$$

as x moves across the interval $-\infty < x < \infty$. From the PDF given in equation 3-16, $F_I(x)$ is

$$F_I(x) = \begin{cases} 0 & \text{if } x < 1000 \\ \int_{1000}^{x} \frac{1}{3000}\,dt & \text{if } 1000 \le x < 3000 \\ \int_{1000}^{3000} \frac{1}{3000}\,dt + \int_{3000}^{x} (5000 - t)/6(10^6)\,dt & \text{if } 3000 \le x < 5000 \\ 1 & \text{if } x \ge 5000 \end{cases}$$

which is equal to

$$F_I(x) = \begin{cases} 0 & \text{if } x < 1000 \\ (x - 1000)/3000 & \text{if } 1000 \le x < 3000 \\ \frac{2}{3} - \frac{1}{12(10^6)}(x - 7000)(x - 3000) & \text{if } 3000 \le x < 5000 \\ 1 & \text{if } x \ge 5000 \end{cases} \tag{3-18}$$

c) Probabilities can be determined from equation 3-18. The probability I is less than or equal to 2000 DSI is

$$P(I \le 2000) = F_I(2000) = \frac{1}{3} = 0.333$$

The probability I will fall between 2000 and 4000 DSI is

$$P(2000 < I < 4000) = F_I(4000) - F_I(2000) = \frac{7}{12} = 0.583$$

The probability I will fall between 2000 and 5000 DSI is

$$P(2000 < I \le 5000) = F_I(5000) - F_I(2000) = \frac{2}{3} = 0.667 \blacklozenge$$

A graph of the CDF for this example is given in figure 3-10. When

examining such a CDF, it is often useful to determine the value of x associated with $F_X(x) = 0.50$. In figure 3-10, this value is 2500 (an exercise for the reader). A value of 2500 DSI for I has an equal probability of being larger or smaller.

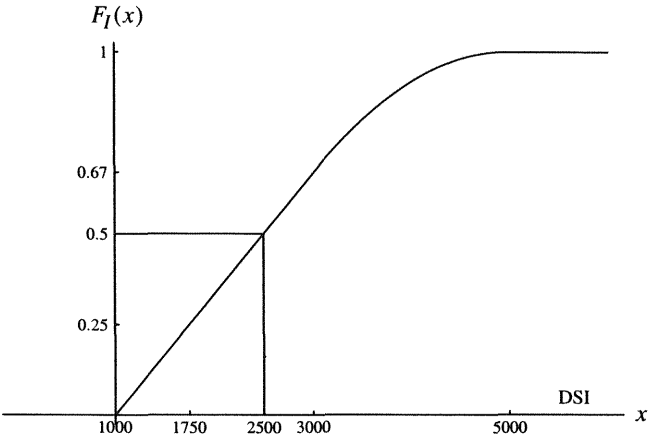


Figure 3-10. Cumulative Distribution Function for Example 3-2

This leads to the definition of an important measure about a distribution function known as the median. If X is a random variable with distribution function $F_X(x)$, a number x satisfying *both*

$$P(X \leq x) \geq 1/2 \quad \text{and} \quad P(X \geq x) \geq 1/2$$

is called the *median of X*. This will be denoted by $Med(X)$. Using theorem 3-2, the above inequalities combine to yield the expression [2]

$$\frac{1}{2} \leq F_X(x) \leq \frac{1}{2} + P(X = x) \tag{3-19}$$

If X is a continuous random variable, we know $P(X = x) = 0$ for all x; therefore, from expression 3-19, the median of X is the number x satisfying

$$F_X(x) = \frac{1}{2} \tag{3-19a}$$

When X is a continuous random variable its distribution function $F_X(x)$ is monotonically increasing, as seen in figure 3-10; therefore, there exists a unique value of x such that equation 3-19a is satisfied. When X is a *discrete* random variable, $Med(X)$ may not be unique. For instance, in figure 3-11 every point in the interval $3 \le x < 4$ is a median of X. From figures 3-10 and 3-11 we see that *every distribution function has at least one median.*

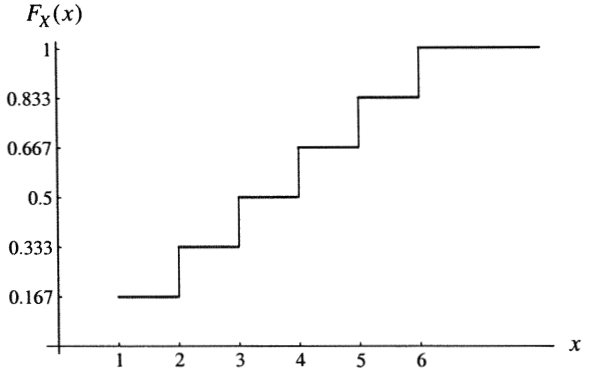


Figure 3-11. CDF of a Random Variable With Uncountably Many Medians

The median is one measure among a class of measures about a distribution known as *fractiles*. In general, the value x_α is called the α-fractile of X if $P(X \le x_\alpha) = \alpha$. The median is the 0.50-fractile of X; its value is given by $x_{0.50}$. In figure 3-10, we have $x_{0.50} = 2500$. Other α-fractiles of common interest are $x_{0.25}$ and $x_{0.75}$. Fractiles are one way to express *percentiles* of a distribution. In general, the α-fractile of X is the $\alpha(100)$th percentile of X. For instance, the median is the 50th percentile of X.

3.2 The Expectation of a Random Variable

When looking at the possible values of a random variable a useful value to determine is its expectation. The expectation of a random variable is also known as its mean. The *expectation* (or *mean*) of a *discrete* random variable X is defined as

$$E(X) \equiv \mu_X = \sum_x x p_X(x) \tag{3-20}$$

The expectation* of a random variable is the summation of all its possible values weighted by the probabilities associated with these values. The terms expectation and mean (usually denoted by the Greek symbol μ) are synonymous.

Example 3-3 Return to case discussion 3-1 and determine the following:

a) $P(Profit \geq E(Profit))$

b) $P(Profit = Med(Profit))$

Solution

a) First determine $E(Profit)$. From case discussion 3-1 the probability function for *Profit* is given in figure 3-5. Since *Profit* was defined by a discrete random variable, from equation 3-20 we have

$$\begin{aligned} E(Profit) &= \sum_{i=1}^{10} x_i \, p_{Profit}(x_i) \\ &= -300(0.006) + (-200)(0.004) + (-100)(0.010) \\ &\quad + 0(0.190) + 100(0.210) + 200(0.239) + 300(0.156) \\ &\quad + 400(0.062) + 600(0.105) + 900(0.018) \\ &= 216 \end{aligned}$$

Therefore, the expected profit is 216 ($M). From theorem 3-2, the probability *Profit* will be greater than or equal to its expected value is

$$P(Profit \geq E(Profit)) = 1 - F_{Profit}(E(Profit)) + P(Profit = E(Profit))$$

or $$P(Profit \geq 216) = 1 - F_{Profit}(216) + P(Profit = 216)$$

---

* The expectation $E(X)$ for a discrete random variable X exists if and only if the summation in equation 3-20 is absolutely convergent; that is, if and only if $\sum_x |x| p_X(x) < \infty$.

From equation 3-8 $F_{Profit}(216) = 0.659$; however $P(Profit = 216) = 0$ since the point $x = 216$ is not a feasible (possible) value of *Profit*; so

$$P(Profit \geq 216) = 1 - 0.659 + 0 = 0.341$$

b) First determine $Med(Profit)$. The median of *Profit* can be found by expression 3-19. Referring to equation 3-8 and figure 3-5, it can be seen that $x = 200$ satisfies both

$$P(Profit \leq x) \geq \frac{1}{2} \quad \text{and} \quad P(Profit \geq x) \geq \frac{1}{2}$$

From equation 3-8

$$P(Profit \leq 200) = F_{Profit}(200) = 0.659 \geq 1/2$$

therefore, the first inequality $P(Profit \leq x) \geq 1/2$ is true when $x = 200$. It now remains to verify that $P(Profit \geq x) \geq 1/2$ when $x = 200$. From theorem 3-2

$$\begin{aligned} P(Profit \geq 200) &= 1 - F_{Profit}(200) + P(Profit = 200) \\ &= 1 - 0.659 + 0.239 = 0.580 \geq 1/2 \end{aligned}$$

therefore, the second inequality is also true. It is left as an exercise for the reader to show that $x = 200$ is the *only* median of *Profit*, in this case. To complete part b) we need to determine $P(Profit = Med(Profit))$. Since it was established that $Med(Profit) = 200$, it can be readily seen from figure 3-5

$$P(Profit = Med(Profit)) = P(Profit = 200) = p_{Profit}(200) = 0.239$$

This result could also be obtained from the cumulative distribution function of *Profit*. Recall $P(X = a) = p_X(a)$ is the height of the jump of $F_X(x)$ at $x = a$, where a is a feasible value of X. From equation 3-8, the height of the jump of $F_{Profit}(x)$ at $x = 200$ is $0.659 - 0.420 = 0.239$. Figure 3-12 illustrates this probability and presents the cumulative distribution function for *Profit*, as described in case discussion 3-1.♦

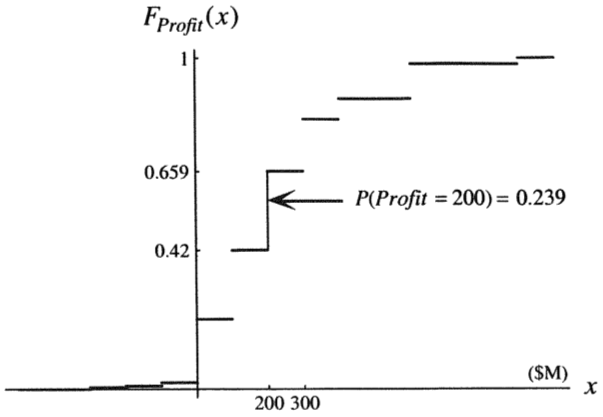


Figure 3-12. Cumulative Distribution Function for *Profit*— Defined in Case Discussion 3-1 and Example 3-3

Example 3-4 Suppose the probability function of the cost to develop an inspection system for radomes is given below.

a) What is the expected cost?

b) What is the 0.95-fractile of *Cost*?

Cost ($M)	40	65	80	95	105
Probability Function for *Cost*	0.30	0.20	0.25	0.20	0.05

Solution

a) From the information in the above table and equation 3-20

$$E(Cost) = 40(0.30) + 65(0.20) + 80(0.25) + 95(0.20) + 105(0.05) = 69.25$$

Therefore, the expected cost of the inspection system is 69.25 ($M).

b) We will use the cumulative distribution function to determine the 0.95-fractile of *Cost*. The table below expresses the probability function and the distribution function of *Cost*.

Cost ($M)	Probability Function	Cumulative Probability
40	0.30	0.30
65	0.20	0.50
80	0.25	0.75
95	0.20	0.95
105	0.05	1.00

From the above table, 95 ($M) is the 0.95-fractile of *Cost*; that is, $P(Cost \leq 95) = 0.95$ ♦

The above discussion focused on determining the expected value of a random variable for the discrete case. If X is a *continuous* random variable, the expectation* (or the mean) of X is defined as

$$E(X) \equiv \mu_X = \int_{-\infty}^{\infty} x f_X(x) dx \tag{3-21}$$

Example 3-5 Using equation 3-21, compute $E(I)$ in example 3-1.

Solution In example 3-1 the PDF of I was

$$f_I(x) = \begin{cases} \frac{1}{4000} & \text{if } 1000 \leq x \leq 5000 \\ 0 & \text{otherwise} \end{cases}$$

from equation 3-21

$$E(I) = \int_{-\infty}^{\infty} x f_I(x) dx = \int_{1000}^{5000} x \frac{1}{4000} dx = 3000 \text{ DSI}$$

Therefore, the expected (mean) size $E(I)$ of the software application described in example 3-1 is 3000 DSI. In figure 3-13, notice $E(I)$ falls exactly between the interval $[1000, 5000]$. In chapter 4, we will see when $f_X(x)$ is described by a *rectangular region*, within an interval $[a,b]$, then $E(X) = (a+b)/2$.

---

* The expectation $E(X)$ for a continuous random variable X exists if and only if the integral in equation 3-21 is absolutely convergent; that is, if and only if $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$.

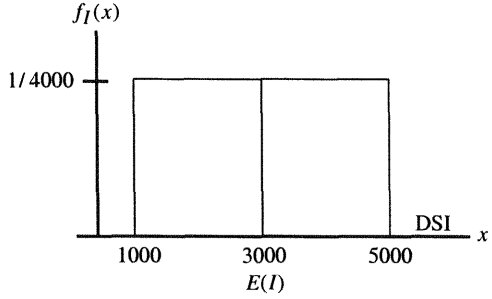


Figure 3-13. The Expectation of I for Example 3-5

Example 3-6 Compute $E(I)$ for the PDF in example 3-2.

Solution In example 3-2 the PDF of I was

$$f_I(x) = \begin{cases} 1/3000 & \text{if } 1000 \le x < 3000 \\ (5000 - x)/6(10^6) & \text{if } 3000 \le x \le 5000 \end{cases}$$

Using equation 3-21

$$E(I) = \int_{-\infty}^{\infty} x f_I(x)dx = \int_{1000}^{3000} x \tfrac{1}{3000} dx + \int_{3000}^{5000} x((5000 - x)/6(10^6))dx = 2555.56 \approx 2556 \text{ DSI}$$

Therefore, the expected (or mean) size $E(I)$ of the software application described in example 3-2 is approximately 2556 DSI. A graph illustrating the location of $E(I)$, in this example, is shown in figure 3-14.

Example 3-7 Let *Cost* denote the unit production cost of a transmitter synthesizer unit (TSU) for a communications terminal. Suppose there is uncertainty in the fabrication, assembly, inspection, and test hours per TSU. Because of this, suppose production engineering assessed that *Cost* is best described by the PDF in figure 3-15. Determine

a) $E(Cost)$ b) $P(Cost > E(Cost))$ c) $Med(Cost)$

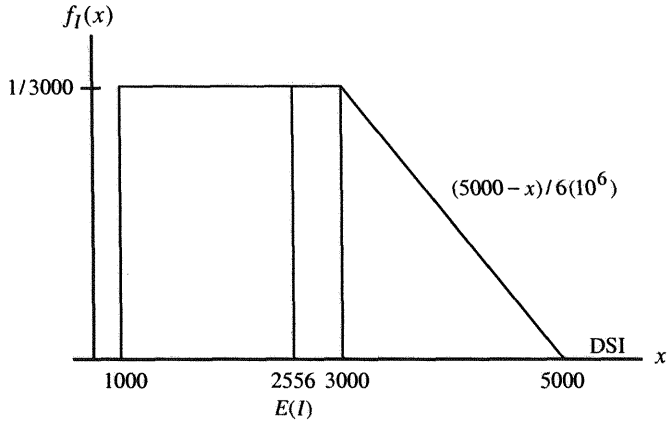


Figure 3-14. The Expectation of I for Example 3-6

Solution

a) To compute $E(Cost)$, it is necessary to determine the mathematical form of the PDF in figure 3-15. It is left to the reader to verify equation 3-22 is indeed a PDF.

$$f_{Cost}(x) = \begin{cases} (x-10000)/8(10^6) & 10000 \le x < 12000 \\ (18000-x)/24(10^6) & 12000 \le x \le 18000 \end{cases} \qquad (3\text{-}22)$$

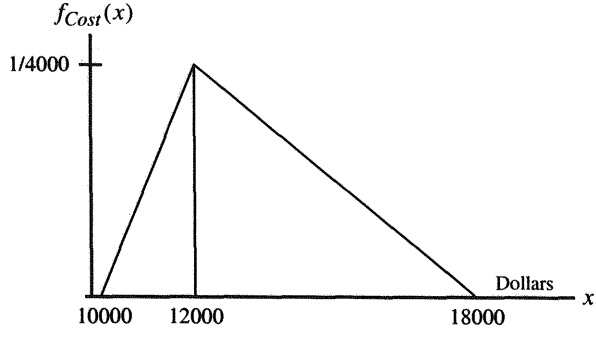


Figure 3-15. PDF for *Cost* in Example 3-7

From equation 3-21

$$E(Cost) \equiv \mu_{Cost} = \int_{-\infty}^{\infty} x f_{Cost}(x)\,dx$$

For this example

$$E(Cost) = \int_{10000}^{12000} x((x-10000)/8(10^6))dx + \int_{12000}^{18000} x((18000-x)/24(10^6))dx = 13333.3$$

Thus, the expected (mean) cost of the TSU is approximately 13,333 dollars.

b) To compute $P(Cost > E(Cost))$, recall from theorem 2-1 (chapter 2)

$$P(Cost > E(Cost)) = 1 - P(Cost \le E(Cost))$$
$$= 1 - F_{Cost}(E(Cost)) = 1 - F_{Cost}(13333.3)$$

From equation 3-9

$$F_{Cost}(13333.3) = \int_{-\infty}^{13333.3} f_{Cost}(t)dt$$

$$= \int_{10000}^{12000} ((t-10000)/8(10^6))dt + \int_{12000}^{13333.3} ((18000-t)/24(10^6))dt = 0.54629$$

therefore

$$P(Cost > E(Cost)) = 1 - 0.54629 = 0.45371$$

c) From equation 3-19a, the median of *Cost* is

$$Med(Cost) = P(Cost \le x) = 0.50$$

We need to find x such that

$$F_{Cost}(x) = P(Cost \le x) = \int_{-\infty}^{x} f_{Cost}(t)dt = 0.50$$

In figure 3-15, the area under the curve between $10000 \le x < 12000$ accounts for only 25 percent of the total area (which must equal unity) between $10000 \le x \le 18000$; that is,

$$P(Cost \le 12000) = \int_{10000}^{12000} ((t - 10000)/8(10^6))dt = 0.25$$

Therefore, the value of x that satisfies $P(Cost \le x) = 0.50$ must be to the right of $x = 12000$. To find this value we need to solve the equation below; specifically, we must find x such that

$$\int_{10000}^{12000} ((t - 10000)/8(10^6))dt + \int_{12000}^{x} ((18000 - t)/24(10^6))dt = 0.50$$

This expression simplifies to solving

$$\int_{12000}^{x} ((18000 - t)/24(10^6))dt = 0.25$$

for x. It turns out the only feasible value for x is 13101; showing this is left for the reader. Therefore, we say the median cost of the transmitter synthesizer unit is 13101; that is, $Med(Cost)$=13,101 dollars.*♦

Thus far, we have discussed the expectation (or mean) and the median of a random variable. Another value of interest is the mode. The mode of a random variable X, denoted by $Mode(X)$, is the value of X that occurs most frequently. It is often referred to as the most likely or most probable value of X. Formally, we say that a is the *mode of* X if

$$p_X(a) = \max_t p_X(t) \quad \text{when } X \text{ is a } \textit{discrete} \text{ random variable}$$

$$f_X(a) = \max_t f_X(t) \quad \text{when } X \text{ is a } \textit{continuous} \text{ random variable}$$

The mode of a random variable is not necessarily unique. The random variable described by the rectangular PDF in figure 3-7 does not have a *unique* mode. However, in example 3-7, x = 12000 is the *unique* mode of the

---

* Mentioned in the preface, the numerical precision shown in this example, and elsewhere in this book, is strictly for teaching purposes. Rounding results to a sensible level of precision is always applied in practice, particularly in the practice of cost analysis.

random variable *Cost*. The mean, median, and mode of a random variable are collectively known as *measures of central tendency*. Figure 3-16 illustrates these measures for the PDF in example 3-7.

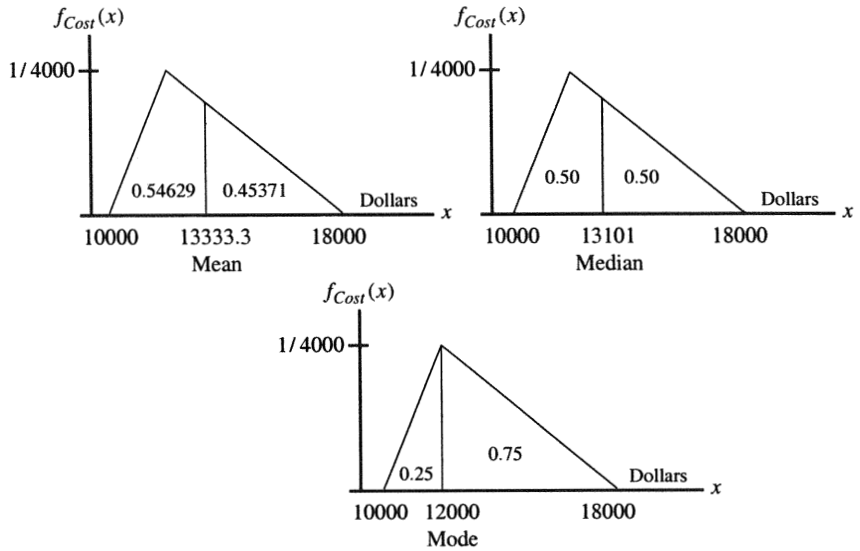


Figure 3-16. Central Tendency Measures for the PDF in Example 3-7

The term average is often used in the same context as the expected value (or mean) of a random variable. The following theorem explains this context.

Theorem 3-8 Let X be a random variable with mean $E(X)$. If an experiment is repeated n-times under identical conditions and X_i is the random variable X associated with the ith round of the experiment, then

$$P(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = E(X)) = 1$$

Theorem 3-8 is known as the *Strong Law of Large Numbers*. It states that for sufficiently large n, it is virtually certain the average of the observed values of $X_1, X_2, \ldots, X_n$ will be approximately the same as the expected value of X. For

example, it can be shown the expected value associated with tossing a fair six-sided die is 3.5. This does not mean we expect to obtain 3.5 on a toss; rather, the average value of many repeated tosses is expected to be approximately 3.5.

The Expected Value of a Function

The need to determine the expected value of a function arises frequently in practice. For instance, in cost analysis the effort Eff_{SW} (in staff-months) to develop software of size I might be given by*

$$Eff_{SW} = 2.8I^{1.2} \tag{3-23}$$

We might ask "*What is the expected software development effort*?" Assuming I is a continuous random variable, from equation 3-21 we could write the expected software development effort as

$$E(Eff_{SW}) = \int_{-\infty}^{\infty} u f_{Eff_{SW}}(u)du \tag{3-24}$$

To use equation 3-24 we need the PDF of Eff_{SW}. As we shall see in chapter 5, this can be difficult for certain kinds of functions. Is there another approach to computing $E(Eff_{SW})$? Note that Eff_{SW} is a function of I.

$$Eff_{SW} = 2.8I^{1.2} = g(I) \tag{3-25}$$

It follows that

$$E(Eff_{SW}) = E(g(I)) \tag{3-26}$$

The following proposition presents a general way to determine $E(Eff_{SW})$ from $E(g(I))$, where $E(g(I))$ is determined from the PDF of I.

---

* Boehm, B. W. 1981. *Software Engineering Economics*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc. In equation 3-23, I is in thousands of DSI.

Proposition 3-1 If X is a random variable and $g(x)$ is a real-valued function defined for all x that are feasible (possible) values of X, then

$$E(g(X)) = \begin{cases} \sum_x g(x)p_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x)f_X(x)dx & \text{if } X \text{ is continuous} \end{cases} \tag{3-27}$$

In the above, the summation and integral must be absolutely convergent. Applying proposition 3-1 to the discussion on Eff_{SW}, we have

$$E(Eff_{SW}) = E(g(I)) = \int_{-\infty}^{\infty} g(x)f_I(x)dx \tag{3-28}$$

Thus, the only information needed to determine $E(Eff_{SW})$ is the function $g(I)$ and $f_I(x)$, the PDF of I. For now, further discussion of this problem is deferred to chapter 5. In particular, case discussion 5-2 presents the determination of $E(Eff_{SW})$ in detail.

Theorem 3-9 If a and b are real numbers, then $E(aX + b) = aE(X) + b$

Proof Let $g(X) = aX + b$; if X is a discrete random variable, then from equation 3-27

$$\begin{aligned} E(aX+b) &= \sum_x (ax+b)p_X(x) = \sum_x axp_X(x) + \sum_x bp_X(x) \\ &= a\sum_x xp_X(x) + b\sum_x p_X(x) = aE(X) + b\cdot 1 = aE(X) + b \end{aligned}$$

If X is a continuous random variable, then from equation 3-27

$$\begin{aligned} E(aX+b) &= \int_{-\infty}^{\infty} (ax+b)f_X(x)dx = a\int_{-\infty}^{\infty} xf_X(x)dx + b\int_{-\infty}^{\infty} f_X(x)dx \\ &= aE(X) + b\cdot 1 = aE(X) + b \blacklozenge \end{aligned}$$

Directly from this proof it can be shown the expected value of a constant is the constant itself; that is, $E(b) = b$. From theorem 3-9, it can also be seen that $E(aX) = aE(X)$, where a is a real number. Showing these two results is an exercise for the reader.

Thus far, we have addressed the expectation (or mean) of a random variable. A quantity known as the variance measures its spread or dispersion (deviation) around the mean. The *variance* of a random variable X is

$$Var(X) \equiv \sigma_X^2 = E\left[(X - E(X))^2\right] \equiv E\left[(X - \mu_X)^2\right] \tag{3-29}$$

The positive square root of $Var(X)$ is known as the *standard deviation* of X, which is denoted by σ_X.

$$\sigma_X = \sqrt{Var(X)} \tag{3-29a}$$

Example 3-8 Let X represent the sum of the toss of a pair of fair dice.

a) Determine the expected sum.
b) Determine the variance of the sum.

Solution In this example, X is a discrete random variable.

a) From equation 3-20 and figure 3-2, the expected sum is

$$E(X) = \frac{1}{36}(2) + \frac{2}{36}(3) + \frac{3}{36}(4) + \frac{4}{36}(5) + \frac{5}{36}(6) + \frac{6}{36}(7)$$
$$+ \frac{5}{36}(8) + \frac{4}{36}(9) + \frac{3}{36}(10) + \frac{2}{36}(11) + \frac{1}{36}(12) = \frac{252}{36} = 7$$

b) From part a) we can write $Var(X) = E[(X-7)^2]$. If we let $g(X) = (X-7)^2$ then from equation 3-27

$$E[g(X)] = E[(X-7)^2] = \sum_x (x-7)^2 p_X(x) \quad x = 2,\ldots,12 \tag{3-30}$$

From figure 3-2, $p_X(2) = \frac{1}{36}, p_X(3) = \frac{2}{36}, \ldots, p_X(12) = \frac{1}{36}$. Working through the computation, equation 3-30 is equal to 5.833; therefore

$$Var(X) = E[g(X)] = E[(X-7)^2] = \sum_x (x-7)^2 p_X(x) = 5.833$$

The variance computed in example 3-8 could be interpreted as follows: the average value of the square of the deviations from the expected sum ($E(X) = 7$) of many repeated tosses of two dice is 5.833. In this case, what is the standard deviation of X?

From the definition of $Var(X)$ in equation 3-29, we can deduce the following theorems.

Theorem 3-10 $Var(X) = E(X^2) - (\mu_X)^2$

Proof The proof follows from the definition of $Var(X)$ and the properties of expectation, as presented in theorem 3-9.

$$\begin{aligned} Var(X) &= E[(X - E(X))^2] \\ &= E(X^2 - 2XE(X) + (E(X))^2) = E(X^2 - 2X\mu_X + (\mu_X)^2) \\ &= E(X^2) - E(2X\mu_X) + E(\mu_X)^2 \\ &= E(X^2) - 2\mu_X E(X) + (\mu_X)^2 \\ &= E(X^2) - 2(\mu_X)^2 + (\mu_X)^2 \\ &= E(X^2) - (\mu_X)^2 \end{aligned}$$

Theorem 3-10 is a convenient alternative for computing the variance of a random variable. It is left as an exercise for the reader to use this theorem to verify $Var(X) = 5.833$, where X is the random variable in example 3-8.

Theorem 3-11 If a and b are real numbers, then

$$Var(aX + b) = a^2 Var(X)$$

Proof The proof follows directly from the definition of $Var(X)$ and theorem 3-9, that is

$$\begin{aligned} Var(X) &= E[(aX + b - E(aX + b))^2] \\ &= E[(aX + b - aE(X) - b)^2] \end{aligned}$$

$$
\begin{aligned}
&= E[(aX - aE(X))^2] \\
&= E[(a(X - E(X)))^2] \\
&= E[a^2(X - E(X))^2] \\
&= a^2 E[(X - E(X))^2] \\
&= a^2 Var(X)
\end{aligned}
$$

This theorem demonstrates the variance of a random variable described by the *linear* function $aX + b$ is unaffected by the constant term b.

Example 3-9 For the communication terminal's transmitter synthesizer unit (TSU) described in example 3-7, compute

a) $Var(Cost)$ and σ_{Cost} using theorem 3-10.

b) Determine $P(|Cost - \mu_{Cost}| \leq \sigma_{Cost})$.

Solution

a) From example 3-7, the PDF for *Cost* was

$$
f_{Cost}(x) = \begin{cases} (x - 10000)/8(10^6) & 10000 \leq x < 12000 \\ (18000 - x)/24(10^6) & 12000 \leq x \leq 18000 \end{cases}
$$

From theorem 3-10 we have

$$Var(Cost) = E(Cost^2) - (\mu_{Cost})^2$$

From part a) in example 3-7, $\mu_{Cost} = E(Cost) = 13333.3$; therefore,

$$Var(Cost) = E(Cost^2) - (13333.3)^2$$

It remains to compute $E(Cost^2)$. From equation 3-27 we can write

$$
E(Cost^2) = \int_{10000}^{12000} x^2((x - 10000)/8(10^6))dx + \int_{12000}^{18000} x^2((18000 - x)/24(10^6))dx
$$

$$= 1.80667(10^8) \ (\$)^2$$

Therefore

$$Var(Cost) = \sigma_{Cost}^2 = 1.80667(10^8) - (13333.3)^2 = 2.88889(10^6) \ (\$)^2$$

from which

$$\sigma_{Cost} = \sqrt{Var(Cost)} = 1699.67 \approx 1700\ (\$)$$

The variance squares the units that define the random variable. Since $\$^2$ is not a useful way to look at *Cost*, the standard deviation σ_{Cost}, which is in dollar units, is usually a better way to interpret this deviation.

b) Probabilities associated with intervals* expressed in terms of the mean and standard deviation can be computed. For some positive real number k

$$P(|Cost - \mu_{Cost}| \le k\sigma_{Cost}) = P(\mu_{Cost} - k\sigma_{Cost} \le Cost \le \mu_{Cost} + k\sigma_{Cost})$$

From equation 3-11, we can express this probability in terms of F_{Cost} as

$$P(|Cost - \mu_{Cost}| \le k\sigma_{Cost}) = F_{Cost}(\mu_{Cost} + k\sigma_{Cost}) - F_{Cost}(\mu_{Cost} - k\sigma_{Cost})$$

For part b) we need $k = 1$; from part a) $\mu_{Cost} = 13333.3$, and $\sigma_{Cost} = 1700$

$$P(|Cost - \mu_{Cost}| \le \sigma_{Cost}) = P(11633.3 \le Cost \le 15033.3)$$
$$= F_{Cost}(15033.3) - F_{Cost}(11633.3)$$

where

$$F_{Cost}(15033.3) = \int_{-\infty}^{15033.3} f_{Cost}(t)dt$$

$$= \int_{10000}^{12000} ((t - 10000)/8(10^6))dt + \int_{12000}^{15033.3} ((18000 - t)/24(10^6))dt = 0.817$$

and

$$F_{Cost}(11633.3) = \int_{-\infty}^{11633.3} f_{Cost}(t)dt = \int_{10000}^{11633.3} ((t - 10000)/8(10^6))dt = 0.167$$

So

$$P(|Cost - \mu_{Cost}| \le \sigma_{Cost}) = 0.817 - 0.167 = 0.65 \blacklozenge$$

---

* Probability intervals are often given in the form $P(|X - a| \le b)$ or $P(|X - a| > b)$, where a and b are any two real numbers. In general, $P(|X - a| \le b) = P(-b \le X - a \le b) = P(a - b \le X \le a + b)$; furthermore, $P(|X - a| > b) = 1 - P(|X - a| \le b) = 1 - P(a - b \le X \le a + b)$.

The TSU cost falls within plus or minus one ($k = 1$) standard deviation (σ) around its expected (or mean) cost with probability 65 percent. The range of values for x associated with this probability is shown in the figure below. This range is sometimes referred to as the 1-sigma interval.

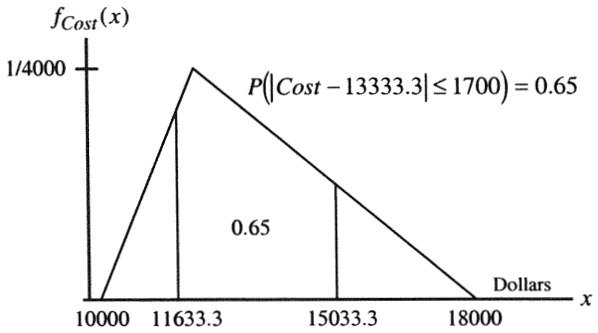


Figure 3-17. 1-Sigma Interval for the TSU Cost

A random variable can be standardized when its mean and variance are known. A *standardized* random variable has zero mean and unit variance. To see this, suppose X is a random variable with mean μ_X and variance σ_X^2. The *standard form* of X is the random variable $Y = (X - \mu_X)/\sigma_X$. From theorems 3-9 and 3-11, it can be shown Y has zero mean and unit variance; that is,

$$E(Y) = E\left(\frac{X-\mu_X}{\sigma_X}\right) = \frac{1}{\sigma_X}E(X-\mu_X) = \frac{1}{\sigma_X}[E(X)-\mu_X] = \frac{1}{\sigma_X}[\mu_X-\mu_X] = 0$$

$$Var(Y) = Var\left(\frac{X-\mu_X}{\sigma_X}\right) = \frac{1}{\sigma_X^2}Var(X-\mu_X) = \frac{1}{\sigma_X^2}Var(X) = \frac{\sigma_X^2}{\sigma_X^2} = 1$$

Referring to example 3-9, we have

$$E(Y) = E\left(\frac{(X - 13333.3)}{1700}\right) = 0$$

$$Var(Y) = Var\left(\frac{(X - 13333.3)}{1700}\right) = 1$$

3.3 Moments of Random Variables

Moments provide important information about the distribution function of a random variable. Such information includes the random variable's mean and variance, as well as the shape of its distribution function. Suppose X is a random variable and k is any positive integer. The expectation $E(X^k)$ is called the *kth moment of* X, which is given by equation 3-31. In general, we say the kth moment of X is

$$E(X^k) = \begin{cases} \sum_x x^k p_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^k f_X(x)\,dx & \text{if } X \text{ is continuous} \end{cases} \qquad (3\text{-}31)$$

In the above, the summation and integral must be absolutely convergent. The mean is the first moment of X. It can be considered that value of x which is the "balance point" or the "center of gravity" of the probability mass (or density) function. This is in contrast to the median. If the random variable is discrete, the median divides the entire mass of the distribution function into two equal parts; each part contains the mass 1/2. If the random variable is continuous, the median divides the entire area under the density function into equal parts. Each part contains an area equal to 1/2 (refer to figure 3-16).

The second moment of the random variable $(X - \mu_X)$ is $E[(X - \mu_X)^2]$. From equation 3-29 this is the variance of X, which provides a measure of the dispersion of X about its mean. What do higher moments of a random variable reveal about the shape of its distribution function?

Let Y be the standardized random variable of X; that is, $Y = (X - \mu_X)/\sigma_X$. The third and fourth moments of Y are known as the coefficients of *skewness* and *kurtosis*. These coefficients are given by equations 3-32 and 3-33, respectively.

$$\gamma_1 = E(Y^3) = E\left[\left(\frac{X - \mu_X}{\sigma_X}\right)^3\right] \tag{3-32}$$

$$\gamma_2 = E(Y^4) = E\left[\left(\frac{X - \mu_X}{\sigma_X}\right)^4\right] \tag{3-33}$$

Skewness, given by γ_1, is a measure of the symmetry of the distribution function of X about the mean of X. If this function has a long tail to the left, then γ_1 is usually negative and we say the distribution function is negatively skewed. If this function has a long tail to the right, then γ_1 is usually positive and we say the distribution function is positively skewed.

In cost analysis it is common to see distributions with $\gamma_1 > 0$. Such distributions have the property that the probability of exceeding the mode (often associated with the point estimate) is greater than the probability of falling below the mode. Experience suggests this is due to a variety of reasons. These include changing requirements, understating a project's true technical complexity, or planning the project against unrealistic cost and/or schedule objectives. Positively skewed distributions are often used to represent uncertainty in system definition variables, such as weight or software size. Point estimates for these variables, particularly in the early phases of a system's design, typically have a high probability of being exceeded.

If the distribution function of X is symmetric about the mean of X, then $\gamma_1 = 0$. The distribution function of X is symmetric about $x = a$ if

$$P(X \geq a + x) = P(X \leq a - x) \text{ for all } x \tag{3-34}$$

From theorem 3-2, equation 3-34 can be written as

$$F_X(a-x) = 1 - F_X(a+x) + P(X = a+x) \tag{3-35}$$

If equation 3-35 is true for all x, we say the distribution function $F_X(x)$ is symmetric with a as the *center of symmetry*. If the center of symmetry is the origin, then $a = 0$ and

$$F_X(-x) = 1 - F_X(x) + P(X = x) \tag{3-36}$$

If X is a continuous random variable, equation 3-36 simplifies to

$$F_X(-x) = 1 - F_X(x) \tag{3-37}$$

The distribution function of a continuous random variable X is symmetric with center a, if and only if

$$f_X(a-x) = f_X(a+x) \text{ for all } x \tag{3-38}$$

If $F_X(x)$ is a symmetric distribution, the center of symmetry is *always the median*. In certain symmetric distributions the mean and/or the mode may also equal the median. If the distribution function of a *continuous* random variable X is symmetric *and the mean of X exists*, then the median and mean of X are equal and they both locate the center of symmetry. The Cauchy distribution* is a symmetric distribution whose mean does not exist (i.e., it is not well-defined). It has a unique median and a unique mode, that equal each other. In the Cauchy distribution, both the median and the mode locate the center of symmetry. Figure 3-18 illustrates these and other cases of symmetric and skewed distributions.

---

* The Cauchy distribution is given by $f_X(x) = \{\pi b[1 + ((x-a)/b)^2]\}^{-1}$. The moments of X do not exist; however, X has a unique median and a unique mode, which both fall at $x = a$. In the Cauchy distribution the median and the mode are equal; they also locate the center of symmetry.

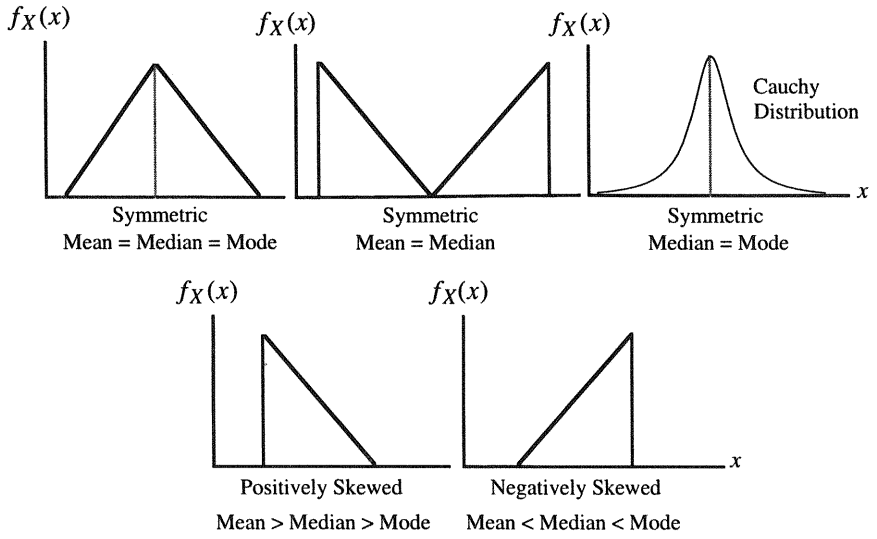


Figure 3-18. Illustrative Symmetric and Skewed Distributions

Kurtosis, given by γ_2 (equation 3-33), measures the peakedness of a random variable's distribution function around its mean. The kurtosis of a distribution function is usually compared with the value $\gamma_2 = 3$, which is the kurtosis of a standardized normal probability distribution (discussed in chapter 4). If $\gamma_2 > 3$, the distribution function of X has greater kurtosis (less peaked) than the normal probability distribution. If $\gamma_2 < 3$, the distribution function of X has less kurtosis (more peaked) than the normal probability distribution.

If we don't know exactly how a random variable is distributed, but we have knowledge about its mean, variance, skewness, and kurtosis, we can often guess its overall shape. In some instances, only the mean and variance of a random variable are needed to uniquely specify the form of its distribution.

3.4 Probability Inequalities Useful in Cost Analysis

Thus far, we have shown how probabilities can be computed from the distribution function of a random variable. However, circumstances frequently exist when the underlying distribution is unknown. This section presents inequalities that provide bounds on the probability of an event independent of the form of the underlying distribution function.

The *Markov Inequality*, due to A. A. Markov (1856-1922), can be used to compute an upper bound on the probability of an event when X is nonnegative and only its mean is known. The *Chebyshev Inequality*, derived by P. L. Chebyshev (1821-1894), bounds the probability that a random variable takes a value within k standard deviations around its mean. Chebyshev's inequality will be shown to be a consequence of Markov's inequality. Before discussing the details of these inequalities, we will first discuss the expected value of an *indicator function.*

The Indicator Function For a random variable X, the indicator function of the event $A = \{X \geq a\}$ is

$$I_A(X) = \begin{cases} 1 & \text{if event } \{X \geq a\} \text{ occurs} \\ 0 & \text{if event } \{X \geq a\} \text{ does not occur} \end{cases}$$

The expected value of $I_A(X)$ is the probability the event A occurs. This can be seen from the following argument. From equation 3-20, we can write

$$E(I_A(X)) = 1 \cdot P(X \geq a) + 0 \cdot [1 - P(X \geq a)] = P(A)$$

Markov's Inequality If X is a nonnegative random variable whose mean μ is positive, then $P(X \geq c\mu) \leq c^{-1}$ for any constant $c > 0$.

Proof

The random variable X is given to be nonnegative with positive mean μ. Since $c > 0$ it follows that $c\mu > 0$. Let

$$I_A(X) = \begin{cases} 1 & \text{if event } \{X \geq c\mu\} \text{ occurs} \\ 0 & \text{if event } \{X \geq c\mu\} \text{ does not occur} \end{cases}$$

where A is the event $\{X \geq c\mu\}$. From this it follows that

$$I_A(X) \leq \frac{X}{c\mu}$$

The expected value of $I_A(X)$ is

$$E(I_A(X)) \leq \frac{1}{c\mu} E(X)$$

Since $E(X) = \mu$ and $E(I_A(X)) = P(A)$ it follows immediately that

$$P(A) = P(X \geq c\mu) \leq 1/c \blacklozenge$$

Markov's inequality states the probability X takes a value greater than or equal to c times its mean cannot exceed $1/c$. For instance, if $c = 2$ then $P(X \geq 2\mu)$ can never exceed 1/2. If $c = 1$ then $P(X \geq \mu)$ is bounded by unity, which is consistent with the first axiom of probability (chapter 2). Markov's inequality is meaningless if c is less than one. Markov's inequality may also be written as

$$P(X \geq a) \leq \frac{1}{a} E(X)$$

where X is nonnegative and $a > 0$; this result follows immediately from the above proof (showing this is left as an exercise for the reader).

From a cost analysis perspective, Markov's inequality provides decision-makers an upper bound on the probability that *Cost* is greater than c times its mean. For instance, suppose the mean cost of a system is determined to be 100 million dollars ($M). Regardless of the underlying distribution function for *Cost*, Markov's inequality guarantees the probability that *Cost* takes a value greater than 200 ($M) can never exceed 1/2.

In general, the probability bound yielded by Markov's inequality is quite conservative. To illustrate this, suppose the random variable *Cost* is described

by the PDF in figure 3-19. This is a lognormal probability distribution* with mean 100 ($M) and standard deviation 25 ($M); it is slightly skewed to the right.

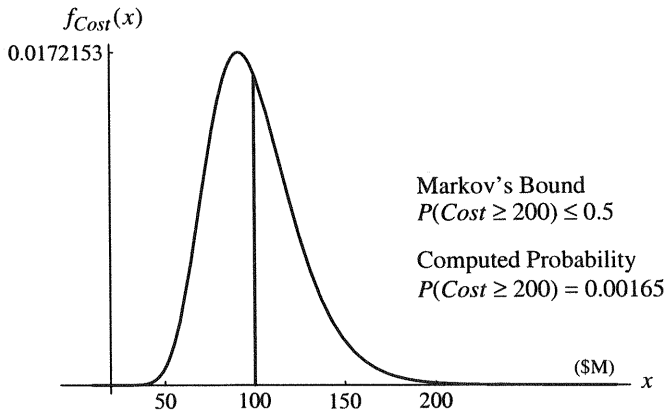


Figure 3-19. A LogNormal PDF for *Cost* with Mean 100 ($M)

The Markov bound is substantially larger than the computed probability of 0.00165 (shown in chapter 4, example 4-8). Such a wide disparity is not surprising since Markov's inequality relies only on the mean of a random variable. In systems engineering, decision-makers typically need more insight into the probability that *Cost* is likely to be exceeded than that provided by Markov's inequality. If values for the mean and variance of *Cost* are available, then Chebyshev's inequality provides probability bounds that improve on those obtained from Markov's inequality.

Chebyshev's Inequality If X is a random variable with finite mean μ and variance σ^2, then for $k \geq 1$

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2} \qquad (3\text{-}39)$$

* The lognormal distribution is often used in cost and economic analysis studies. It will be fully discussed in chapter 4, with additional applications in chapter 7.

Proof

Recall, in general, that

$$P(|X-a|\geq b)=1-P(|X-a|<b)=1-P(a-b<X<a+b) \qquad (3\text{-}40)$$

where a and b are real numbers. Suppose we let $a=\mu$ and $b=k\sigma$. Then

$$P(|X-a|\geq b)=P(|X-\mu|\geq k\sigma)$$

Now $(X-\mu)^2\geq k^2\sigma^2$ if and only if $|X-\mu|\geq k\sigma$; from Markov's inequality

$$P((X-\mu)^2\geq k^2\sigma^2)\leq\frac{1}{k^2\sigma^2}E((X-\mu)^2) \qquad (3\text{-}41)$$

Since $E((X-\mu)^2)=\sigma^2$ inequality 3-41 reduces to $P((X-\mu)^2\geq k^2\sigma^2)\leq\frac{1}{k^2}$, which is equivalent to

$$P(|X-\mu|\geq k\sigma)\leq\frac{1}{k^2} \qquad (3\text{-}42)$$

or

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

From equation 3-40

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

therefore

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

Chebyshev's inequality states that for any random variable X, the probability that X will assume a value within k standard deviations of its mean is at least $1\text{-}1/k^2$. From equation 3-39, the probability a random variable takes a value within 2 standard deviations of its mean will always be at least 0.75. If X is a continuous random variable, at least 95 percent of the *area under any probability density function* will always fall within 4.5 standard deviations of the mean.

Like Markov's inequality probabilities produced by Chebyshev's inequality are also conservative, but to a lesser extent. To illustrate this, consider once

again the random variable *Cost* with mean 100 ($M), standard deviation 25 ($M), and PDF given by figure 3-19. It can be *computed* that the interval

$$[\mu - 2\sigma, \mu + 2\sigma] = [50,150] \ (\$M)$$

accounts for nearly 96 percent (refer to chapter 4, example 4-8) of the total probability (area) under $f_{Cost}(x)$. This computed probability is in contrast to Chebyshev's inequality (equation 3-39), which indicates the interval [50,150] ($M) accounts for at least 75 percent of the total probability.

Various forms of Chebyshev's inequality are given below; in each form $a > 0$.

A. $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$

B. $P(|X - \mu| < a) \geq 1 - \frac{\sigma^2}{a^2}$

C. $P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$

D. $P(X - \mu \geq a) \leq P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$

Suppose $\mu = 100$, $\sigma = 25$, and $a = 100$. From form D of Chebyshev's inequality we have

$$P(Cost - 100 > 100) \leq \frac{(25)^2}{(100)^2}$$

$$\Rightarrow P(Cost > 200) \leq \frac{1}{16} = 0.0625$$

Thus, the probability *Cost* will exceed twice its mean will not be more than $\frac{1}{16}$ (or 0.0625). From the previous discussion, Markov's inequality revealed this bound could not be more than $\frac{1}{2}$. Although these results are consistent, form D of Chebyshev's inequality provides a significant refinement on the probability bound for this event. This is not surprising since additional information about the random variable *Cost*, specifically its variance, is taken into account. Because of this, Chebyshev's inequality will always provide a

tighter probability bound than that produced by Markov's inequality. Figure 3-20 summarizes this discussion and contrasts these probability bounds for the PDF given in figure 3-19.

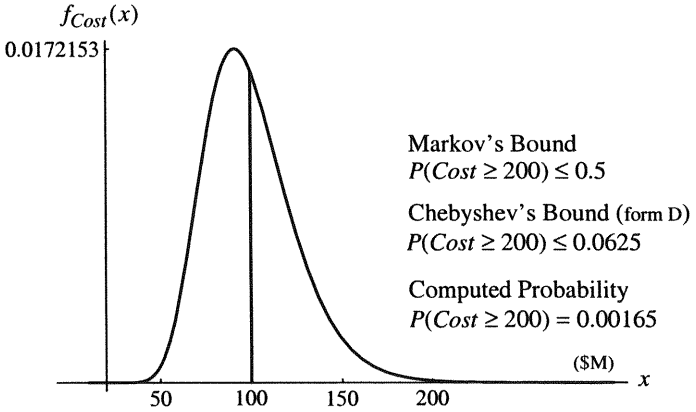


Figure 3-20. Some Probability Bounds on *Cost*

The probability inequalities presented here share the common characteristic that their bounds are valid for *any type* of distribution function. Although these bounds are conservative, they do offer decision-makers probabilities that are *independent of the underlying distribution.* When inequalities such as Chebyshev's are used in conjunction with an assumed or approximated distribution, decision-makers are provided alternative ways to view the probability associated with the same event.

3.5 A Cost Analysis Perspective

In cost uncertainty analysis two important statistical measures to determine are the expected (mean) cost and the standard deviation of cost. A classical way to view the relationship between a mean and a standard deviation is presented

in figure 3-21. Shown is a special distribution known as the *normal probability distribution* (chapter 4).

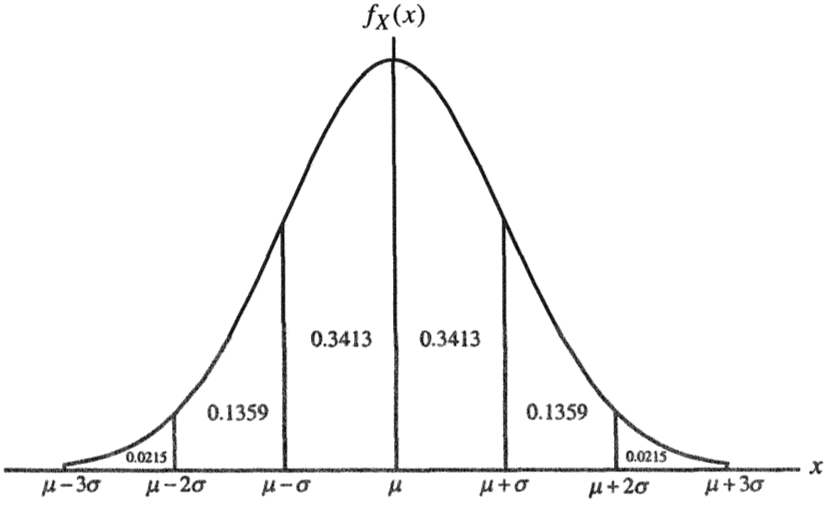


Figure 3-21. Areas Under the Normal Probability Distribution

The normal distribution is symmetric about its mean. It has the property that its mode and median equal its mean. In particular, the 1-sigma interval

$$[\mu - \sigma, \mu + \sigma]$$

will always account for slightly more than 68 percent of the total area under a *normal* probability density function. Similarly, the 2-sigma interval

$$[\mu - 2\sigma, \mu + 2\sigma]$$

will always account for slightly more than 95 percent of the total area under a *normal* probability density function. Although the mean is an important statistical measure that contributes many useful insights about the underlying distribution, it is just a single value among infinitely many that define the curve. Alone, the mean provides no direct view into the variability implicit to the distribution. For this reason, analysts and decision-makers must consider the mean and the standard deviation jointly. Figure 3-22 illustrates this point.

Comparing just the difference in the mean costs between system design alternatives *A* and *B*, it may appear to a decision-maker alternative *B* is the better choice.

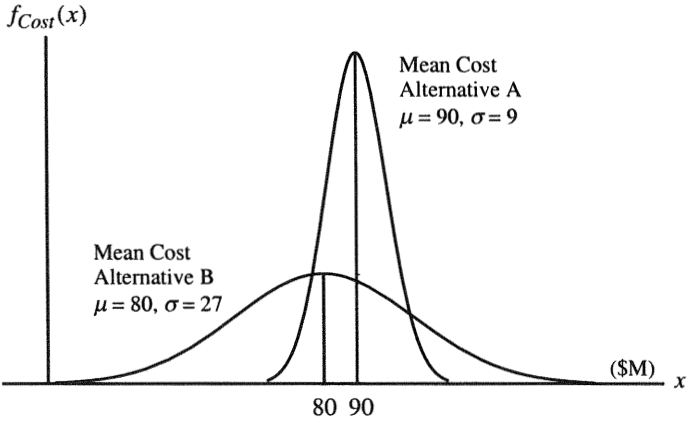


Figure 3-22. Comparing the Mean Costs of Alternatives

However, when the dispersion σ in cost is considered and the 1-sigma interval is determined for each alternative, the decision-maker may very well select alternative *A* instead. Specifically, the 1-sigma interval for alternative *A* (from figure 3-22) is

$$[\mu - \sigma, \mu + \sigma] = [81, 99] \text{ (\$M)}$$

The 1-sigma interval for alternative *B* is

$$[\mu - \sigma, \mu + \sigma] = [53, 107] \text{ (\$M)}$$

Thus, for the same level of confidence implied by the 1-sigma interval (68 percent) choosing alternative *B* implies accepting three times the variability in cost (54 ($M)) than that associated with alternative *A* (18 ($M)). Clearly, this result would not have been seen if comparing the mean costs was the sole criterion for selecting an alternative.

This discussion illustrates the usefulness of another statistic known as the *coefficient of dispersion.* Defined by equation 3-43, the coefficient of dispersion D is the ratio of the standard deviation to the mean.

$$D = \frac{\sigma}{\mu} \tag{3-43}$$

Consider again figure 3-22. The coefficient of dispersion for alternative A is 0.10. This implies the value of *Cost* at one standard deviation above the mean, will be 10 percent higher than the mean of *Cost*, which is 90 ($M) for alternative A. Similarly, the coefficient of dispersion for alternative B is 0.3375. This implies the value of *Cost* at one standard deviation above its mean will be nearly 34 percent higher than the mean of *Cost*, which is 80 ($M) for alternative B. Clearly, a significantly higher cost penalty exists at 1-sigma above the mean under alternative B than for alternative A. A decision-maker might consider this cost risk to be unacceptable. Although the cost mean for alternative A is 10 ($M) higher than the cost mean for alternative B, its significantly lower cost variance (i.e., less cost risk) may be the acceptable tradeoff.

Exercises

1. Let X denote the sum of the toss of two fair dice. Determine the following using the probability function in figure 3-2 and the appropriate theorems in section 3.1.

   a) $P(X < 7)$ b) $P(X > 7)$ c) $P(X \geq 7)$

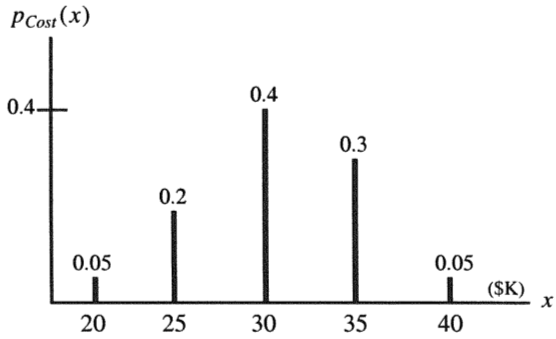
   d) $P(10 \leq X \leq 12)$ e) $P(10 \leq X < 12)$ f) $P(10 < X < 12)$

2. Suppose the probability function for the development and production cost of a microchip is given below. Determine the following:

a) The cumulative distribution function of *Cost*

b) $P(Cost \leq 35)$ c) $P(Cost > 25)$ d) $P(Cost \geq 25)$

e) $P(20 \leq Cost < 35)$ f) $P(20 < Cost < 35)$ g) $P(Cost < 35)$



Probability Function for Exercise 2

3. For any random variable X, show that $P(X < a) = F_X(a) - P(X = a)$.

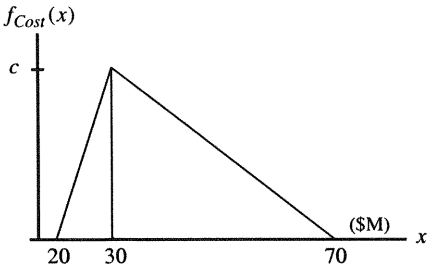
4. Refer to Case Discussion 3-1 and answer the following:

   a) Find $p_{Profit}(x)$ and $F_{Profit}(x)$ if $P(V = 5) = 0.1$, $P(V = 15) = 0.8$, and $P(V = 20) = 0.1$, where V is the sales volume (in millions).

   b) With what probability does $Profit = 0$?

5. Suppose the profit function to sell 10000 electronic widgets, with a unit price of \$10 per widget, is given by $Profit = (10)^4(10 - U_{Cost})$, where U_{Cost} is a discrete random variable that represents the unit cost (in dollars) of each widget. If U_{Cost} can take one of the values in the set $\{4,7,10\}$, where u_{Cost} represents one of these values, find the constant c such that $p_{Profit}(u_{Cost}) = c\,Profit$ is a *probability function.*

6. Suppose *Cost* is a continuous random variable whose possible values are given by the interval $20 \le x \le 70$, where x is in dollars million (\$M).
   a) Find c such that the function below is a probability density function.
   b) Compute $P(Cost \le 30)$, $P(30 < Cost < 70)$, $P(Cost = 30)$.



Function for Exercise 6

7. Show that $f_{Cost}(x)$ in exercise 6 is the derivative of the cumulative distribution function $F_{Cost}(x)$, where

$$F_{Cost}(x) = \begin{cases} 0 & \text{if } x < 20 \\ \frac{1}{500}(x-20)^2 & \text{if } 20 \le x < 30 \\ \frac{1}{5} + \frac{1}{50}\left[40 - \frac{(x-70)^2}{40}\right] & \text{if } 30 \le x < 70 \\ 1 & \text{if } x \ge 70 \end{cases}$$

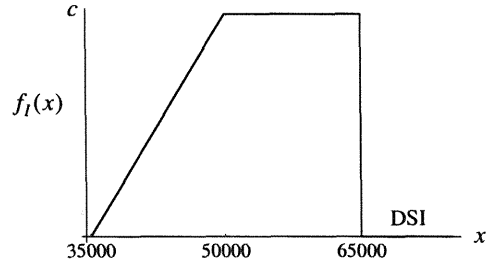
8. For the probability density function in example 3-1 (figure 3-7), show that all subintervals of [1000, 5000] that are the same in length will occur with equal probability.

9. a) Given the probability function in exercise 2, determine *Med*(*Cost*).
   b) From the *Profit* probability function in Case Discussion 3-1, show that $x = 200$ is the *only* value of x that satisfies the relationship

$$(1/2) \le F_{Profit}(x) \le (1/2) + P(Profit = x)$$

c) In example 3-2, show that $Med(I) = 2500$ DSI.

10. Suppose the uncertainty in the size I of a software application is expressed by the *probability density function* in the figure below.
    a) Determine the cumulative distribution function $F_I(x)$.
    b) Compute $P(I \leq 50000)$, $P(40000 \leq I \leq 60000)$, $P(50000 \leq I \leq 65000)$.



Function for Exercise 10

11. In exercise 10, show that $Med(I) = 53,750$ DSI.

12. Find the expected number of workstations purchased per month and the standard deviation if the probability function for the monthly demand is given in the following table.

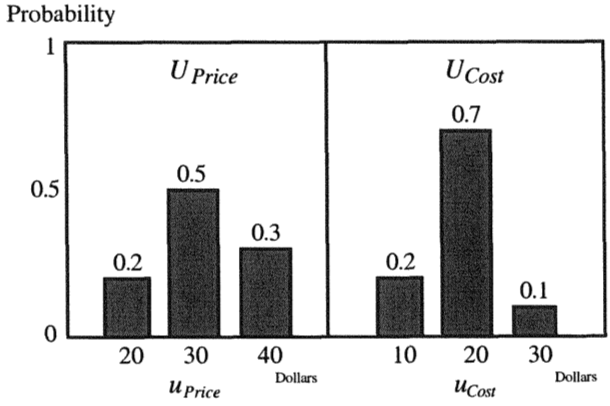
Workstations Purchased per Month	14	9	36	6	4
Probability	0.23	0.15	0.42	0.10	0.10

Probability Function for Exercise 12

13. From Case Discussion 3-1, the profit on a new electronics product manufactured and sold by ChipyTech Corporation was given by

$$Profit = (U_{Price} - U_{Cost})V$$

Suppose the product's sales volume V for its first year on the market is set at 30 million. Suppose the probability functions of U_{Price} and U_{Cost} are given in the figure below. Assume U_{Price} and U_{Cost} are independent.



Probability Functions for Exercise 13

Compute:

a) $p_{Profit}(x)$ and $F_{Profit}(x)$ b) $E(Profit)$

c) $Var(Profit)$ d) $P(Profit = E(Profit))$

e) $P(Profit < E(Profit))$ f) The probability of making *no profit.*

14. A random variable X takes the value 1 with probability p and the value 0 with probability $1-p$. Show that $E(X) = p$ and $Var(X) = p(1-p)$.

15. From exercise 10 compute the following:
a) $E(I)$ b) σ_I c) $P(|I - E(I)| > \sigma_I)$

16. Let Y be a random variable with probability function given in the table below. Compute
a) $E(3Y+1)$ b) $Var(3Y+1)$

y	1	2	3	4	5
$P(Y=y)$	1/4	1/8	1/4	1/4	1/8

Probability Function for Exercise 16

17. Suppose $E(X)=4$ and $Var(X)=E(X)/2$. Find the expectation and variance of the random variable $(1-2X)/2$.

18. a) If X has mean μ_X show that $E(X-\mu_X)$ is always zero.
    b) If a and b are constants, show that $E(b)=b$ and $E(aX)=aE(X)$.

19. a) Let X represent the value of the toss of a fair six-sided die. Show that $E(X)=3.5$. Determine $Var(X)$.
    b) If X is a random variable representing the sum of the toss of a pair of fair six-sided dice, use theorem 3-10 to verify that $Var(X)=5.833$.

20. Find a general formula for the kth moment of a continuous random variable X with density function $f_X(x)=(b-a)^{-1}$, where $a\le x\le b$.

21. Suppose X is a continuous random variable with $f_X(x)=1$ in the interval $0\le x\le 1$. Show that the coefficient of skewness for $f_X(x)$ is zero.

22. If the probability density function of X is given by

$$f_X(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2}\left[(x-\mu)^2/\sigma^2\right]}$$

    show that X is symmetric with center equal to μ.

23. If a is a constant, show that Markov's inequality can also be written in the form $P(X\ge a)\le a^{-1}E(X)$.

24. Let N_W be a random variable representing the number of widgets produced in a month. Suppose the expected number of widgets produced by a manufacturer during a month is 2000.

a) Find an upper bound on the probability this month's production will exceed 3200 widgets.

b) Suppose the standard deviation of a month's production is known to be 35 widgets. Find a and b such that the number of widgets produced this month falls in the interval $a < N_W < b$ with probability at least 0.75.

25. Suppose *Cost* is a random variable with $E(Cost) = 3$ and $Var(Cost) = 1$. Use Chebyshev's inequality to compute a lower bound on

a) $P(2 < Cost < 4)$

b) $P(|Cost - 3| < 5)$

References

1. Park, W. R., and D. E. Jackson. 1984. *Cost Engineering Analysis – A Guide to Economic Evaluation of Engineering Projects*, 2nd ed. New York: John Wiley & Sons, Inc.

2. Rohatgi, V. K. 1976. *An Introduction to Probability Theory and Mathematical Statistics*. New York: John Wiley & Sons, Inc.

Chapter 4

Special Distributions for Cost Uncertainty Analysis

All business proceeds on beliefs,
or judgments of probabilities,
and not on certainties.
Charles William Eliot, 1834-1926
President of Harvard University
The New Dictionary of Thoughts, 1957

Obviously, a man's judgment cannot
be better than the information
on which he has based it.
Arthur Hays Sulzberger
Address to the New York State Publishers
Association [August 30, 1948]

In probability theory there is a class of distribution functions known as special distributions. Special distributions are those that occur frequently in the theory and application of probability. A well-known special distribution is the Bernoulli distribution, a discrete distribution whose probability function is given by equation 4-1. The Bernoulli distribution can be used to study a random variable X representing the outcome of an experiment that succeeds, $\{X = 1\}$, with probability p or fails, $\{X = 0\}$, with probability $(1 - p)$.

$$p_X(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases} \tag{4-1}$$

Another well-known special distribution is the normal distribution, a continuous distribution discussed later in this chapter. Special distributions have been well-studied over the years and are fully described in a two volume text by Johnson and Kotz [1]. To avoid an extended exposition on the entire class of special distributions, this chapter focuses on a subset of these distributions which frequently arise in cost uncertainty analysis.

4.1 The Trapezoidal Distribution

The trapezoidal distribution is illustrated in figure 4-1. It is rarely presented in traditional, or classical, texts on probability theory. Despite this, the trapezoidal distribution is highly useful and flexible for many situations in

cost uncertainty analysis. Seen in figure 4-1, it can model a random variable whose probability density function increases in the interval $a \leq x < m_1$, remains constant across the interval $m_1 \leq x < m_2$, then decreases to zero in the interval $m_2 \leq x \leq b$.

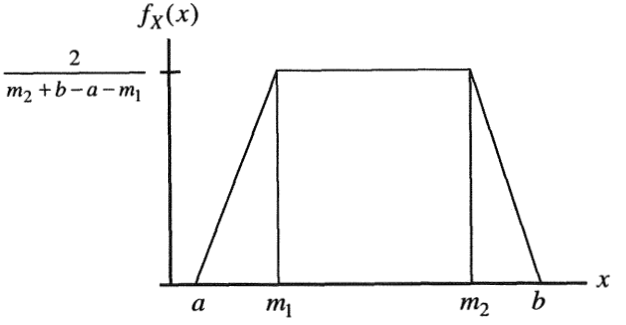


Figure 4-1. Trapezoidal Probability Density Function

Mathematically, a trapezoidal distribution can arise from the sum of two independent continuous random variables whose probability density functions are constants over closed intervals of the real line.* In cost uncertainty analysis, the trapezoidal distribution is primarily used to *directly* specify a range of possible values for a random variable. For instance, suppose an experienced software engineer was asked to assess the number of DSI needed to build a particular software application. The engineer may have solid technical reasons why this number would not exceed $x = b$ DSI or be less than $x = a$ DSI. However, the engineer may strongly believe it is more likely the number of DSI will fall in an interval of constant density between m_1 and m_2. Such a situation can be represented by a trapezoidal distribution.

A random variable X is said to have a *trapezoidal distribution* if its probability density function is given by equation 4-2 [2]

---

* Independent random variables are discussed in chapter 5. Refer to table 5-9 (chapter 5) for a further discussion on the sum of two continuous random variables with constant density.

$$f_X(x) = \begin{cases} \dfrac{2}{(m_2 + b - a - m_1)} \dfrac{1}{m_1 - a}(x - a) & \text{if } a \le x < m_1 \\ \dfrac{2}{(m_2 + b - a - m_1)} & \text{if } m_1 \le x < m_2 \\ \dfrac{2}{(m_2 + b - a - m_1)} \dfrac{1}{b - m_2}(b - x) & \text{if } m_2 \le x \le b \end{cases} \tag{4-2}$$

where $-\infty < a < m_1 < m_2 < b < \infty$. A trapezoidal probability density function is illustrated in figure 4-1. The numbers a and b represent the minimum and maximum possible values of X, respectively. Note that $f_X(x) = 0$ if $x < a$ or $x > b$. The mode of X is not unique. It is any value of x in the interval $m_1 \le x \le m_2$. For the remainder of this book, a random variable X with PDF given by equation 4-2 will be implied by the expression

$$X \sim Trap(a, m_1, m_2, b)^*$$

The cumulative distribution function of X is given by equation 4-3 [2].

$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \dfrac{1}{(m_2 + b - a - m_1)} \dfrac{1}{m_1 - a}(x - a)^2 & \text{if } a \le x < m_1 \\ \dfrac{1}{(m_2 + b - a - m_1)}(2x - a - m_1) & \text{if } m_1 \le x < m_2 \\ 1 - \dfrac{1}{(m_2 + b - a - m_1)} \dfrac{1}{b - m_2}(b - x)^2 & \text{if } m_2 \le x < b \\ 1 & \text{if } x \ge b \end{cases} \tag{4-3}$$

A graph of $F_X(x)$ is shown in figure 4-2.

---

* The symbol "~" means "is distributed as." In this case, we say X is distributed as a trapezoidal random variable with parameters a, m_1, m_2, and b. We might also say X is a trapezoidal random variable with PDF given by equation 4-2.

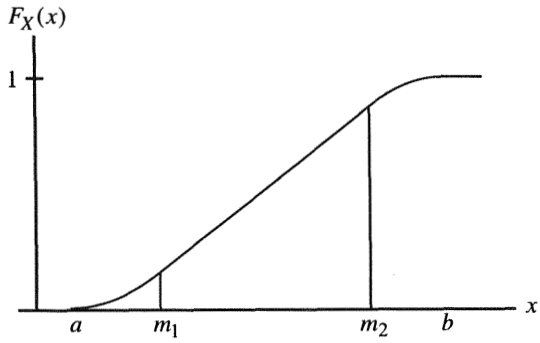


Figure 4-2. The Trapezoidal Cumulative Distribution Function

The CDF is linear in the interval $m_1 \le x < m_2$, where the density function is constant, and quadratic in the intervals $a \le x < m_1$ and $m_2 \le x < b$.

Theorem 4-1 [2] If X is a trapezoidal random variable then

$$E(X) = \frac{((m_2 + b)^2 - m_2 b) - ((a + m_1)^2 - a m_1)}{3(m_2 + b - a - m_1)}$$

$$Var(X) = \frac{(m_2^2 + b^2)(m_2 + b) - (a^2 + m_1^2)(a + m_1)}{6(m_2 + b - a - m_1)} - [E(X)]^2$$

Example 4-1 Let X represent the uncertainty in the number of delivered source instructions (DSI) of a new software application. Suppose this uncertainty is expressed as the trapezoidal density function in figure 4-3. Determine the following:

a) $E(X)$

b) $Med(X)$

c) $P(X \le E(X) + \sigma_X)$

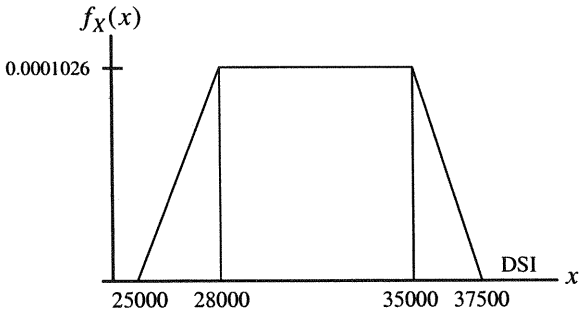


Figure 4-3. Trapezoidal Probability Density Function for Example 4-1

Solution

a) It is given that $X \sim Trap(25000, 28000, 35000, 37500)$; therefore, we have $a = 25000$, $m_1 = 28000$, $m_2 = 35000$, $b = 37500$. Substituting these values into the expectation formula in theorem 4-1 yields

$$E(X) = \frac{((m_2 + b)^2 - m_2 b) - ((a + m_1)^2 - a m_1)}{3(m_2 + b - a - m_1)} = 31363.24786 \approx 31363 \text{ DSI}$$

Since we need σ_X in part c) of this example, we will compute $Var(X)$ at this point; from theorem 4-1 we have

$$\sigma_X = \sqrt{Var(X)} = \sqrt{\frac{(m_2^2 + b^2)(m_2 + b) - (a^2 + m_1^2)(a + m_1)}{6(m_2 + b - a - m_1)} - [31363.24786]^2}$$

$$= 2925.26 \approx 2925 \text{ DSI}$$

b) To compute $Med(X)$, the median size of the software application, we need to find x such that $F_X(x) = 1/2$. It can be shown (left for the reader) that

$$P(25000 \le X \le 28000) = \frac{2}{13} < \frac{1}{2}$$

$$P(25000 \le X \le 35000) = \frac{2}{13} + \frac{28}{39} = \frac{34}{39} > \frac{1}{2}$$

Thus, the median of X will fall in the region of constant probability density; this is equivalent to finding x along the CDF of X such that

$$\frac{1}{(35000+37500-25000-28000)}(2x-25000-28000)=\frac{1}{2}$$

Solving the above yields $x=31375$; therefore $Med(X)=31375$ DSI.

c) To determine $P(X \le E(X)+\sigma_X)$ we have from part a) the result

$$E(X)+\sigma_X = 31363+2925 = 34288 \text{ DSI}$$

The value $x=34288$ falls in the linear region of $F_X(x)$; from equation 4-3

$$P(X \le E(X)+\sigma_X) = P(X \le 34288) = F_X(34288) = 0.798$$

Thus, there is nearly an 80 percent probability the amount of code to build the new software application will not exceed 34,288 DSI.

4.1.1 The Uniform Distribution

The uniform distribution can be considered a special case of the trapezoidal distribution.* In figure 4-1, as $(m_1 - a)$ and $(b - m_2)$ approach zero (in the limit), the trapezoidal distribution approaches a distribution with uniform (or constant) probability density, shown in figure 4-4.

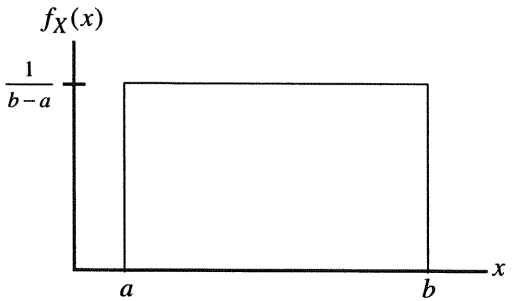


Figure 4-4. The Uniform Probability Density Function

* It is also a special case of the beta distribution, which is discussed later in this chapter.

A random variable X is said to have a *uniform distribution* (or *rectangular distribution*) if its probability density function is constant and given by

$$f_X(x) = \frac{1}{b-a} \quad \text{if } a \le x \le b \tag{4-4}$$

where $-\infty < a < b < \infty$. The numbers a and b are the minimum and maximum possible values of X, respectively. Note that $f_X(x) = 0$ if $x < a$ or $x > b$.

A random variable described by a uniform probability density function has the following interesting property. If the unit interval $0 \le x \le 1$ is the range of values for X, then $f_X(x) = 1$ and the probability X falls in any subinterval $a' \le x \le b'$ of $0 \le x \le 1$ is simply the length of that subinterval; specifically,

$$P(a' \le X \le b') = \int_{a'}^{b'} 1 \, dx = b' - a'$$

For the remainder of this book, a random variable X with PDF given by equation 4-4 will be implied by the expression

$$X \sim Unif(a,b)$$

The cumulative distribution function of X is given by equation 4-5.

$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \dfrac{1}{b-a}(x-a) & \text{if } a \le x < b \\ 1 & \text{if } x \ge b \end{cases} \tag{4-5}$$

A graph of $F_X(x)$ is shown in figure 4-5. Because the density function of X is constant in the interval $a \le x \le b$, the cumulative distribution is strictly a *linear function* of x in the interval $a \le x \le b$.

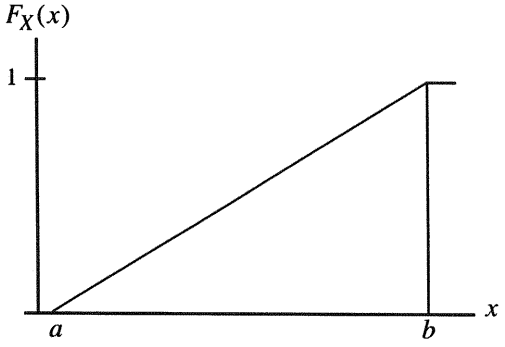


Figure 4-5. The Uniform Cumulative Distribution Function

The uniform distribution has no skew and no unique mode. From a cost analysis perspective, such random variables might be the number of DSI required for a new software application (refer to chapter 3, example 3-1), the weight of a new electronic device, or an unknown contractor's software productivity rate. In practice, the uniform distribution is used when a random variable is best described *only* by its extreme possible values. In cost analysis, this occurs most often in the very early stages of a system's design.

Theorem 4-2 If X is a uniform random variable then

$$E(X) = \frac{1}{2}(a+b)$$

$$Var(X) = \frac{1}{12}(b-a)^2$$

Example 4-2 If X has a uniform distribution, show that $Med(X) = E(X)$.

Solution Since $X \sim Unif(a,b)$ we know from the above discussion that

$$F_X(x) = \frac{1}{b-a}(x-a) \text{ if } a \le x < b$$

Since X is a continuous random variable we know X has a unique median. The median of X will be the value x such that

$$F_X(x) = \frac{1}{b-a}(x-a) = \frac{1}{2}$$

Solving the expression for x yields $x = (a+b)/2$, which is $Med(X)$. From theorem 4-2 we see that $Med(X) = (a+b)/2 = E(X)$, when $X \sim Unif(a,b)$.

4.1.2 The Triangular Distribution

The triangular distribution can also be considered a special case of the trapezoidal distribution. In the trapezoidal distribution if $m_1 = m_2 = m$ then the trapezoidal distribution becomes a triangular distribution, such as the one shown in figure 4-6.

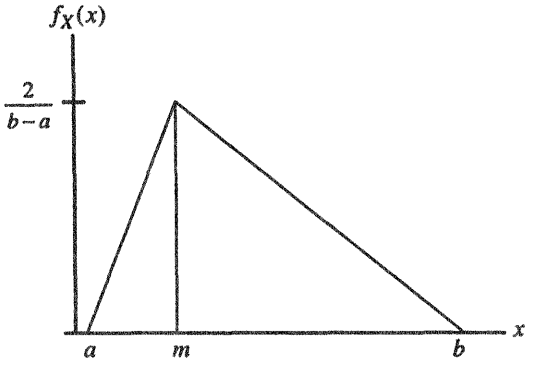


Figure 4-6. Triangular Probability Density Function

A random variable X is said to have a *triangular distribution* if its probability density function is given by

$$f_X(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(m-a)} & \text{if } a \le x < m \\ \dfrac{2(b-x)}{(b-a)(b-m)} & \text{if } m \le x \le b \end{cases} \qquad (4\text{-}6)$$

where $-\infty < a < m < b < \infty$. The numbers a, m, and b represent the minimum, the mode (most likely), and the maximum possible values of X, respectively.

Note that $f_X(x) = 0$ if $x < a$ or $x > b$. In cost analysis, the mode m is often regarded as the point estimate.*

For the remainder of this book, a random variable X with PDF given by equation 4-6 will be implied by the expression

$$X \sim Trng(a, m, b)$$

The cumulative distribution function of X is given by equation 4-7.

$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \dfrac{(x-a)^2}{(b-a)(m-a)} & \text{if } a \le x < m \\ 1 - \dfrac{(b-x)^2}{(b-a)(b-m)} & \text{if } m \le x < b \\ 1 & \text{if } x \ge b \end{cases} \tag{4-7}$$

A graph is shown in figure 4-7.

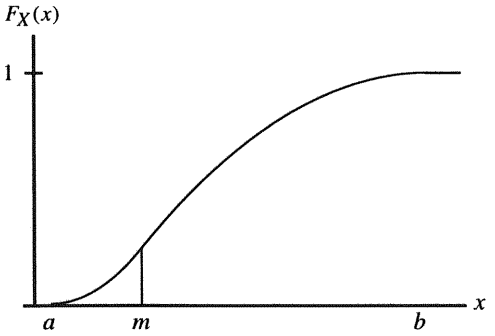


Figure 4-7. The Triangular Cumulative Distribution Function

The CDF is a quadratic function of x in the intervals $a \le x < m$ and $m \le x < b$.

---

* Associating the point estimate (defined in chapter 1) to the mode of a distribution is traditional in cost analysis; however, there are no strict reasons for doing so. An analyst might judge the point estimate is best represented by the median, or by the mean, of a distribution.

The location of m relative to a and b determines how much probability there is on either side of m. This is illustrated by the three triangular distributions in figure 4-8.

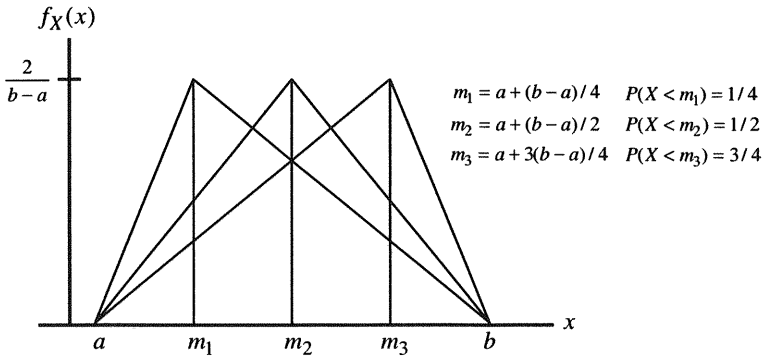


Figure 4-8. A Family of Triangular Probability Density Functions*

Seen in figure 4-8 the closer the mode is to the variable's maximum possible value b, the less likely it is the variable will exceed its mode. The closer the mode is to the variable's minimum possible value a, the more likely it is the variable will exceed its mode. For this reason the triangular distribution is often favored as a subjective probability distribution. Only three values (a, m, and b) are needed to specify the distribution. From these values, subject matter experts focus the distribution in a way that appropriately reflects the overall subjective distribution of probability for the variable under consideration.

Theorem 4-3 If X is a triangular random variable then

$$E(X) = (a+m+b)/3$$

$$Var(X) = \frac{1}{18}\left\{(m-a)(m-b)+(b-a)^2\right\}$$

---

* From Evans, M., N. Hastings (ed), and B. Peacock. 1993. *Statistical Distributions*, 2nd ed. New York: John Wiley & Sons, Inc.

Example 4-3 In example 3-7, the uncertainty in the unit production cost of a transmitter synthesizer unit (TSU), for a communications terminal, was given by the probability density function in figure 3-15. Use theorem 4-3 to show that $E(Cost) = 13333.3\ \$$ and $Var(Cost) = 2.89(10^6)\ \$^2$.

Solution Referring to example 3-7, we see the PDF for *Cost* can be written in the form given by equation 4-6 with $a = 10000$, $m = 12000$, and $b = 18000$. Substituting these values into the expected value and variance formulas given in theorem 4-3 yields

$$E(Cost) = (a + m + b)/3 = (10 + 12 + 18)10^3/3 = 13333.3\ \$$$

$$Var(Cost) = \frac{1}{18}\{(12-10)(12-18) + (18-10)^2\}(10^6) = 2.89(10^6)\ \$^2$$

4.2 The Beta Distribution

The beta distribution, like the distributions discussed in section 4.1, can be used to describe a random variable whose range of possible values is bounded by an interval of the real line. A random variable X is said to have a *beta distribution* if its probability density function is given by

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\dfrac{x-a}{b-a}\right)^{\alpha-1}\left(\dfrac{b-x}{b-a}\right)^{\beta-1} & a < x < b \\ 0 & \text{otherwise} \end{cases} \qquad (4\text{-}8)$$

where α and β ($\alpha > 0$ and $\beta > 0$) determine the shape of the density function and $\Gamma(\alpha)$ is the gamma function of the argument α.*

Beta distributions are in *standard form* when they are defined over the unit interval. A random variable Y is said to have a *standard beta distribution* if its probability density function is given by

---

* In general, $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$. If α is a positive integer, then $\Gamma(\alpha) = (\alpha-1)!$.

$$f_Y(y) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(y)^{\alpha-1}(1-y)^{\beta-1} & 0<y<1 \\ 0 \quad \text{otherwise} \end{cases} \tag{4-9}$$

For the remainder of this book, the random variables X and Y with density functions given by equations 4-8 and 4-9 will be implied by the expressions $X \sim Beta(\alpha,\beta,a,b)$ and $Y \sim Beta(\alpha,\beta)$, respectively. The transformation* of $X \sim Beta(\alpha,\beta,a,b)$ to its standard form $Y \sim Beta(\alpha,\beta)$ is done by letting $y=(x-a)/(b-a)$. Graphs of the standard beta probability density function for various α and β are illustrated in figure 4-9 and figure 4-10.

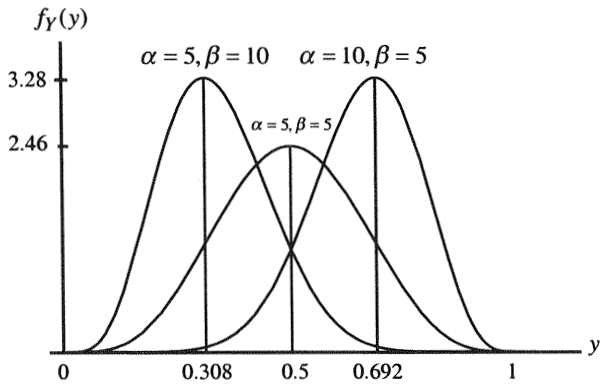


Figure 4-9. A Family of Standard Beta Probability Density Functions

Figure 4-9 illustrates several possible shapes associated with the standard beta density function. When $\alpha=\beta$ it is symmetric about $y=0.5$, which is the median of Y. When $\alpha=\beta$ the median, mean, and mode of Y are equal. If $\alpha>1$ and $\beta>1$ the mode of Y is unique and occurs at

$$y=\frac{1-\alpha}{2-\alpha-\beta} \tag{4-10}$$

---

* Transformations of random variables are formally discussed in chapter 5.

Figure 4-10 illustrates some other shapes associated with the standard beta density. For instance, the beta density is **U** shaped if $\alpha < 1$ and $\beta < 1$. If $\alpha = 1$ and $\beta = 1$ the beta density becomes the *Unif*(0,1) (uniform) density function. A *Beta*(1,2) density is a right-skewed triangular PDF, while a *Beta*(2,1) is a left-skewed triangular PDF.

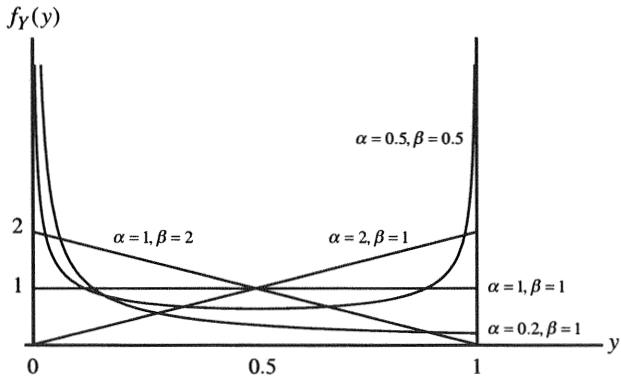


Figure 4-10. More Standard Beta Probability Density Functions

Seen in figure 4-9 and figure 4-10 the beta density can take a wide variety of shapes. This characteristic makes the beta density among the most diverse of the special distributions for describing (or modeling) a random variable whose range of possible values is bounded by an interval of the real line.

In general, from the transformation $y = (x-a)/(b-a)$ it can be shown the cumulative distribution function of X can be found from the cumulative distribution function of Y according to

$$F_X(x) = F_Y\left(\frac{x-a}{b-a}\right) = F_Y(y)$$

However, a closed form expression for the cumulative distribution function of Y (given by equation 4-11) does not exist.

$$F_Y(y) = \int_0^y f_Y(t)dt \quad \text{if } 0 < y < 1 \tag{4-11}$$

Values for $F_Y(y)$ are determined through a numerical integration procedure. A number of software applications, such as *Mathematica*® [3], are available for numerically computing the integral given by equation 4-11. A family of graphs for $F_Y(y)$ is presented in figure 4-11. These cumulative distribution functions are the integrals of the three beta densities given in figure 4-9.

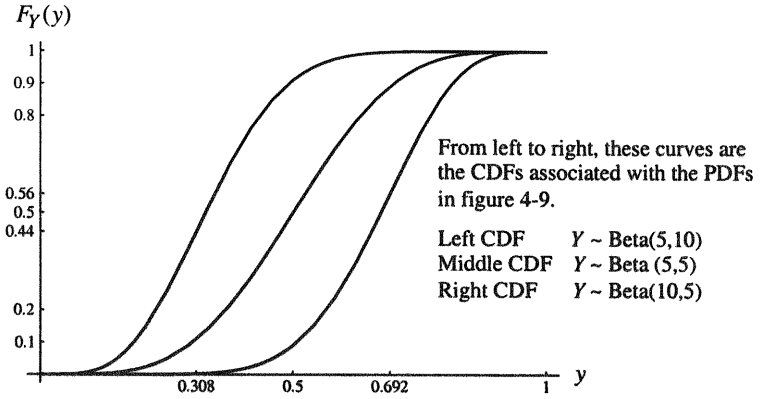


Figure 4-11. A Family of Standard Beta Cumulative Distribution Functions

Theorem 4-4 If $Y \sim Beta(\alpha, \beta)$ and $X \sim Beta(\alpha, \beta, a, b)$ then

$$E(Y) = \frac{\alpha}{\alpha + \beta} \tag{4-12}$$

$$E(X) = a + (b - a)E(Y) \tag{4-12a}$$

$$Var(Y) = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2} \tag{4-13}$$

$$Var(X) = (b - a)^2 Var(Y) \blacklozenge \tag{4-13a}$$

If the mean and variance of Y are known, it can be shown from theorem 4-4, the shape parameters of the beta distribution are uniquely determined by

$$\alpha = E(Y)\left[\frac{E(Y)(1-E(Y))}{Var(Y)} - 1\right] \tag{4-14}$$

$$\beta = \alpha\left(\frac{1-E(Y)}{E(Y)}\right) \tag{4-15}$$

Lastly, if $Y \sim Beta(\alpha,\beta)$ then $1-Y \sim Beta(\beta,\alpha)$. Discuss how this property is seen in figure 4-9 and in figure 4-11.

Example 4-4 Suppose the activity time X (in minutes) to complete the assembly of a microcircuit is beta distributed in the interval $4 < x < 9$, with shape parameters $\alpha = 5$ and $\beta = 10$. Determine $P(X \le Mode(X))$.

Solution From equation 4-10

$$Mode(Y) = \frac{1-\alpha}{2-\alpha-\beta} = \frac{1-5}{2-15} = \frac{4}{13} \approx 0.308$$

where Y is the standard beta density of X. This is in terms of the unit interval, that is, if $Y \sim Beta(5,10)$ then $Mode(Y) = 0.308$. From the transformation $y = (x-a)/(b-a)$, the value $y = 0.308$ in the unit interval is equivalent to the value $x = 5.54$ in the interval $4 < x < 9$ (where $a = 4$ and $b = 9$); therefore, $Mode(X) = 5.54$. To determine $P(X \le Mode(X))$ we have

$$P(X \le Mode(X)) = P\left(\frac{X-a}{b-a} \le \frac{Mode(X)-a}{b-a}\right) = P(Y \le Mode(Y))$$

Since $Y \sim Beta(5,10)$ we have

$$f_Y(y) = \frac{\Gamma(15)}{\Gamma(5)\Gamma(10)}(y)^4(1-y)^9 \quad 0 < y < 1$$

From numerical integration it can be shown

$$P(Y \le Mode(Y)) = F_Y(0.308) = \int_0^{0.308} \frac{\Gamma(15)}{\Gamma(5)\Gamma(10)} (y)^4 (1-y)^9 \, dy \approx 0.44$$

Since $P(X \le Mode(X)) = P(Y \le Mode(Y))$ we conclude

$$P(X \le Mode(X)) = 0.44$$

Therefore, with probability 0.44 the assembly time of the microcircuit will be less than or equal to 5.54 minutes. Discuss why this probability is also seen in figure 4-11.

4.3 The Normal Distribution

The distributions presented in section 4.1 and section 4.2 can be thought of as finite distributions. Random variables described by *finite distributions* have values that are restricted to a bounded interval of the real line. The trapezoidal, uniform, triangular, and beta distributions are examples of finite distributions. In contrast to these, a random variable described by a normal distribution is unbounded. Its values fall in the open interval given by the entire real line. The normal distribution is the first of two *infinite distributions* we will discuss in this chapter.

The trapezoidal, uniform, triangular, and beta PDFs are frequently used in cost analysis to *directly specify* the uncertainty in the value of a variable. Typically, such variables are inputs for deriving cost.* These variables might include the number of new DSI for a software function, the weight of a future hardware item (e.g., a satellite), or the time required to assemble a new electronic device. The normal distribution *could* be used in the same way; however, from a cost analysis perspective, the normal most often characterizes the *underlying distribution function of a derived cost.* In this sense, the normal distribution can reflect the shape of an "output" distribution —

---

* This is illustrated in the discussion associated with figure 1-4 (chapter 1).

particularly one generated from a summation of "input" distributions, like those discussed in sections 4.1 and 4.2. For instance, suppose the random variable *Cost* is derived from the sum of the cost of each work breakdown structure cost element X_i $(i = 1,\ldots,n)$ in a system. Specifically, if

$$Cost = X_1 + X_2 + X_3 + \cdots + X_n \tag{4-16}$$

then under certain conditions (discussed in chapters 5 and 6) the normal distribution will characterize the underlying distribution function of *Cost*.

A random variable X is said to be *normally distributed* if its probability density function is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left[(x-\mu)^2/\sigma^2\right]} \tag{4-17}$$

where $-\infty < x < \infty$ and $\sigma > 0$. Equation 4-17 is also known as the *Gaussian* distribution, named after the German mathematician Karl Friedrich Gauss (1777-1855). For the remainder of this book, a random variable X with PDF given by equation 4-17 will be implied by the expression $X \sim N(\mu, \sigma^2)$. The normal PDF is *uniquely defined* by two parameters μ and σ^2. Theorem 4-6 will show these parameters are the mean and variance of X, respectively. A graph of the normal PDF is presented in figure 4-12.

The normal distribution is symmetric about its mean μ. It has the property that its mode and median equal its mean. The numbers in figure 4-12 are the areas under the curve within the indicated intervals. In particular, we have

$$P(\mu - \sigma \le X \le \mu + \sigma) = \int_{\mu-\sigma}^{\mu+\sigma} f_X(x)dx = 0.6826 \tag{4-18}$$

where $f_X(x)$ is given by equation 4-17. Similarly,

$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0.9544 \tag{4-19}$$

$$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0.9973 \tag{4-20}$$

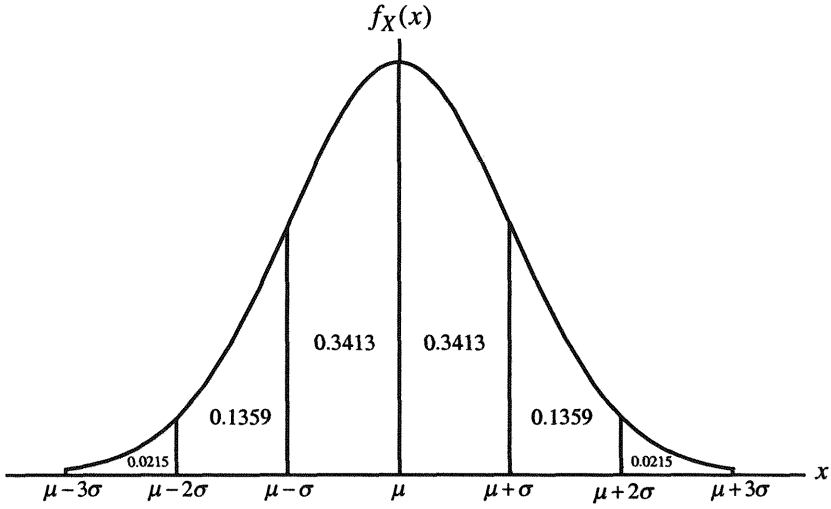


Figure 4-12. The Normal Probability Density

Thus, when X is *normally distributed* the probability X falls within $\pm 1\sigma$ from its mean is always 0.6826; the probability X falls within $\pm 2\sigma$ from its mean is always 0.9544; the probability X falls within $\pm 3\sigma$ from its mean is always 0.9973.

The peak of the normal PDF is governed only by the variance of X. Furthermore, *Mode*(X) occurs at $x = \mu$. The probability density function evaluated at $x = \mu$ is equal to $0.399/\sigma$. Decreasing σ increases the maximum height of the normal PDF and the concentration of probability around the mean μ. This is illustrated in figure 4-13.

If $X \sim N(\mu, \sigma^2)$ and $Z = (X - \mu)/\sigma$, the standard form of X, it can be shown (theorem 4-5) that Z has a normal distribution with mean 0 and variance 1. The density function of Z is known as the *standard normal density*, which is given by equation 4-21.

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad -\infty < z < \infty \tag{4-21}$$

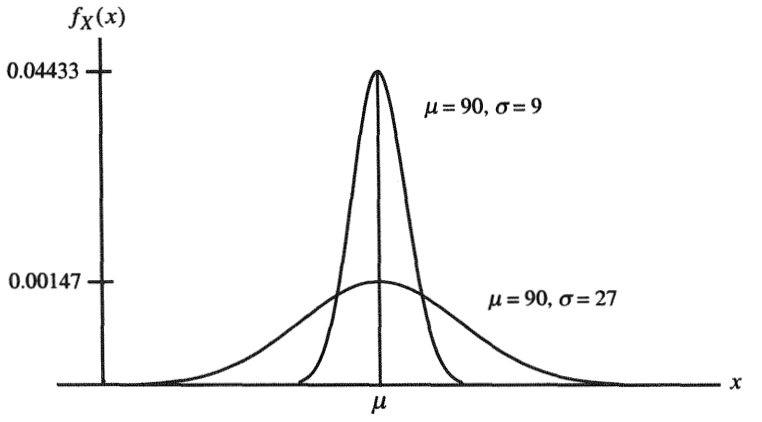


Figure 4-13. A Comparison of the Heights of Two Normal PDFs

For the remainder of this book, a random variable Z with PDF given by equation 4-21 will be implied by the expression $Z \sim N(0,1)$. A graph of $f_Z(z)$ is shown in figure 4-14. The peak of the *standard normal density* occurs at $z = 0$, which is *Mode*(Z). Since $Var(Z) = 1$ the standard normal probability density function evaluated at *Mode*(Z) is equal to 0.399.

Closed form expressions for the cumulative distribution functions $F_X(x)$ and $F_Z(z)$ do not exist. However, from the transformation $z = (x - \mu)/\sigma$ it can be shown that

$$F_X(x) = F_Z((x - \mu)/\sigma) = F_Z(z) \tag{4-22}$$

where

$$F_Z(z) = P(Z \leq z) = \int_{-\infty}^{z} f_Z(y)dy$$

and $f_Z(y)$ is given by equation 4-21. Thus, values for $F_X(x)$ can be obtained from values for $F_Z(z)$ by a numerical integration of $f_Z(y)$. The results of such an integration are summarized in table A-1 (presented in appendix A). A graph of $F_Z(z)$ is also shown in figure 4-14.

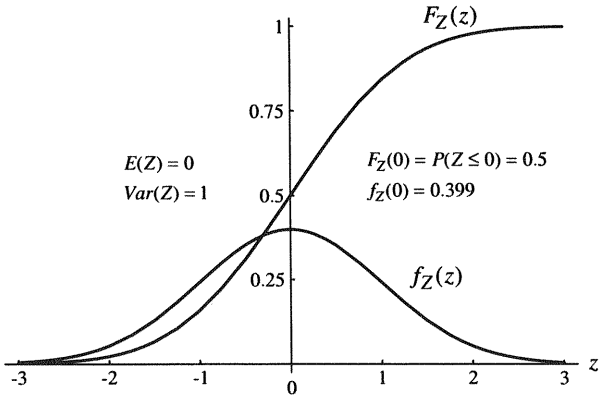


Figure 4-14. The Standard Normal PDF and CDF

Because the standard normal is symmetric about $z = 0$, we have $P(Z \le -k) = P(Z > k)$. In terms of the cumulative distribution function of Z this is equivalent to $F_Z(-k) = 1 - F_Z(k)$. In particular, if $X \sim N(\mu, \sigma^2)$ then the probability X is within $\pm k\sigma$ of the mean of X is

$$P(\mu - k\sigma \le X \le \mu + k\sigma) = P(-k \le Z \le k)$$
$$= F_Z(k) - F_Z(-k) = F_Z(k) - [1 - F_Z(k)] = 2F_Z(k) - 1 \qquad (4\text{-}23)$$

Example 4-5 Using table A-1, show that $P(\mu - \sigma \le X \le \mu + \sigma) = 0.6826$.

Solution From equation 4-23 we see that $k = 1$, in this case; so,

$$P(\mu - \sigma \le X \le \mu + \sigma) = P(-1 \le Z \le 1) = 2F_Z(1) - 1$$

From table A-1 $F_Z(1) = 0.8413$; therefore,

$$P(\mu - \sigma \le X \le \mu + \sigma) = P(-1 \le Z \le 1) = 2(0.8413) - 1 = 0.6826 \blacklozenge$$

If $X \sim N(\mu, \sigma^2)$ then probability statements about X can be written in terms of its standard form Z. From equation 4-22, we have the general relationship

$$P(a \le X \le b) = F_Z\left(\frac{b-\mu}{\sigma}\right) - F_Z\left(\frac{a-\mu}{\sigma}\right) \qquad (4\text{-}24)$$

Example 4-6 In figure 1-6 (chapter 1), the distribution function of a system's cost was normal with mean 110.42 ($M) and standard deviation 21.65 ($M). Given this, determine $P(100 \le Cost \le 140)$.

Solution We are given $Cost \sim N(110.42,(21.65)^2)$. In terms of equation 4-24

$$P(100 \le X \le 140) = F_Z\left(\frac{140-110.42}{21.65}\right) - F_Z\left(\frac{100-110.42}{21.65}\right)$$

$$= F_Z(1.37) - F_Z(-0.48)$$

Since $F_Z(-k) = 1 - F_Z(k)$ we have $F_Z(-0.48) = 1 - F_Z(0.48)$; therefore,

$$P(100 \le X \le 140) = F_Z(1.37) - [1 - F_Z(0.48)]$$

From table A-1 $F_Z(1.37) = 0.91465$ and $F_Z(0.48) = 0.68439$; so,

$$P(100 \le X \le 140) = 0.599 \approx 0.60$$

Thus, there is nearly a 60 percent chance the system's cost will fall between 100 and 140 million dollars.

Example 4-7 Suppose the uncertainty in a system's cost is described by the normal PDF shown in figure 4-15. Suppose there is a 5 percent chance the system's cost will not exceed 30.34 ($M) and an 85 percent chance its cost will not exceed 70.55 ($M). From this information determine the mean and standard deviation of the system's cost.

Solution We are given $P(Cost \le 30.34) = 0.05$ and $P(Cost \le 70.55) = 0.85$. Expressing the random variable *Cost* in standard form we have

$$P\left(Z \le \frac{30.34-\mu}{\sigma}\right) = 0.05 \quad \text{and} \quad P\left(Z \le \frac{70.55-\mu}{\sigma}\right) = 0.85$$

where μ and σ are the mean and standard deviation of *Cost*, respectively. We will first work with the probability

$$P\left(Z \le \frac{30.34-\mu}{\sigma}\right) = 0.05$$

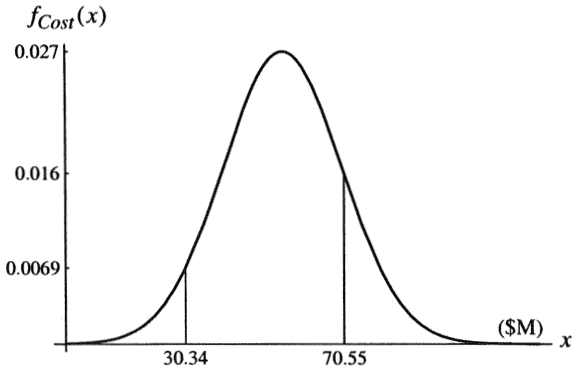


Figure 4-15. PDF for Example 4-7

From appendix A (table A-1) $P(Z \leq 1.645) = 0.95$; it follows that

$$1 - P(Z \leq 1.645) = 0.05$$

This is equivalent to $P(Z > 1.645) = 0.05$. Since the standard normal distribution is symmetric about $z = 0$, $P(Z > 1.645) = P(Z \leq -1.645)$; therefore, we have

$$\frac{30.34 - \mu}{\sigma} = -1.645 \tag{4-25}$$

Similar reasoning applies to the other probability. From appendix A (table A-1)

$$P\left(Z \leq \frac{70.55 - \mu}{\sigma}\right) = 0.85$$

is true when

$$\frac{70.55 - \mu}{\sigma} = 1.04 \tag{4-26}$$

Solving equations 4-25 and 4-26 simultaneously for μ and σ yields

$$\mu \approx 55 \text{ ($M)}$$

$$\sigma \approx 15 \text{ ($M)}$$

Theorem 4-5 If $X \sim N(\mu,\sigma^2)$, then $Z \sim N(0,1)$ where $Z = (X-\mu)/\sigma$.

Proof Since $X \sim N(\mu,\sigma^2)$ we have

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left[(t-\mu)^2/\sigma^2\right]} dt$$

By the definition of a cumulative distribution function we also have

$$F_Z(z) = P(Z \le z) = P\left(\frac{X-\mu}{\sigma} \le z\right) = P(X \le z\sigma + \mu)$$

$$= \int_{-\infty}^{z\sigma+\mu} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left[(x-\mu)^2/\sigma^2\right]} dx \qquad (4\text{-}27)$$

If we let $y = (x-\mu)/\sigma$ then $\sigma dy = dx$; substituting this change of variable into equation 4-27 yields

$$F_Z(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}y^2} \sigma dy = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy \qquad (4\text{-}28)$$

Equation 4-28 is the cumulative distribution function of the standard normal density; thus,

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Therefore, $Z \sim N(0,1)$. We will next show this result implies $E(Z) = 0$ and $Var(Z) = 1$, as well as a more general case.

Theorem 4-6 If $X \sim N(\mu,\sigma^2)$ then $E(X) = \mu$ and $Var(X) = \sigma^2$.

Proof Since $X \sim N(\mu,\sigma^2)$ we have

$$E(X)=\int_{-\infty}^{\infty} x\cdot\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{1}{2}\left[(x-\mu)^2/\sigma^2\right]}dx$$

By the change of variable $z=(x-\mu)/\sigma$ we have

$$E(X)=\int_{-\infty}^{\infty}(z\sigma+\mu)\cdot\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{1}{2}z^2}\,\sigma dz$$

which simplifies to

$$E(X)=\sigma\int_{-\infty}^{\infty} z\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz+\mu\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz \qquad (4\text{-}29)$$

The first integral in equation 4-29 is $E(Z)$. This integral is equal to zero since the integral exists and its integrand is an odd function; that is,

$$E(Z)=\int_{-\infty}^{\infty} z\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz=0$$

The second integral in equation 4-29 is unity since it is the integral of the standard normal density function. Therefore, equation 4-29 simplifies to

$$E(X)=\sigma E(Z)+\mu\cdot 1=\sigma\cdot 0+\mu=\mu$$

To show that $Var(X)=\sigma^2$, recall that $Var(X)=E(X^2)-(E(X))^2$. We know that

$$E(X^2)=\int_{-\infty}^{\infty} x^2\cdot\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{1}{2}\left[(x-\mu)^2/\sigma^2\right]}dx$$

From the family of integrals of exponential functions, presented in appendix A, note that

$$E(X^2)=\int_{-\infty}^{\infty} x^2\cdot\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{1}{2}\left[(x-\mu)^2/\sigma^2\right]}dx=\mu^2+\sigma^2$$

therefore, $Var(X)=\mu^2+\sigma^2-(\mu)^2=\sigma^2$.

4.4 The LogNormal Distribution

The lognormal probability distribution is the last of the infinite distributions we will discuss in this book. It has broad applicability in engineering, economics, and cost analysis. In engineering, the failure rates of mechanical or electrical components often follow a lognormal distribution. In economics, the random variation between the production cost of goods to capital and labor costs is frequently modeled after the lognormal distribution; the classical example is the Cobb-Douglas production function, given by equation 4-30.

$$Q = aW_1^{a_1} W_2^{a_2} \tag{4-30}$$

In the above, the production cost of goods Q is a function of capital cost W_1 and labor cost W_2; the terms a, a_1, and a_2 are real numbers. Under certain conditions Q can be shown to have a lognormal probability distribution. In cost analysis, Young [4] observed that the lognormal can approximate the probability distribution of a system's total cost — particularly when the cost distribution is positively skewed. Empirical studies by Garvey and Taub [5, 6] identify circumstances where the lognormal can approximate the combined (joint) distribution of a program's total cost and schedule.*

The lognormal distribution has a close relationship with the normal distribution. If X is a *nonnegative random variable* where the natural logarithm of X, denoted by $\ln X$, follows the normal distribution, then X is said to have a lognormal distribution. This is illustrated in figure 4-16. On the left-side of figure 4-16 the random variable X has a lognormal PDF, with $E(X) = 100$ and $Var(X) = 625$. On the right-side is the representation of X in logarithmic space. In logarithmic space X has a normal PDF, with

* This is fully discussed in chapter 7.

$E(\ln X) = 4.57486$ and $Var(\ln X) = 0.0606246$. How these latter two values were determined is discussed in theorem 4-8.

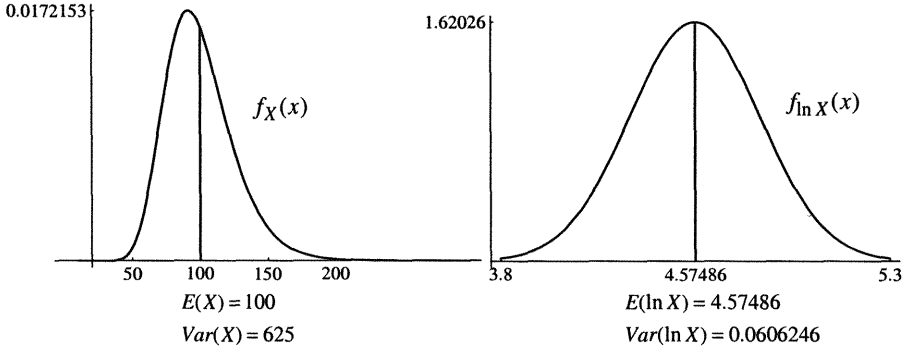


Figure 4-16. Probability Density Functions of X and $\ln X$

$X \sim LogN(100, 625)$ and $\ln X \sim N(4.57486, 0.0606246)$

Under certain conditions (discussed in chapter 5), the normal distribution can arise from a summation of many random variables (as illustrated by equation 4-16); the lognormal distribution can arise from a multiplicative combination of many random variables, as illustrated by equation 4-30.

A random variable X is said to be *lognormally distributed* if its probability density function is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma_Y}\frac{1}{x}e^{-\frac{1}{2}\left[(\ln x - \mu_Y)^2/\sigma_Y^2\right]} \qquad (4\text{-}31)$$

where $0 < x < \infty$, $\sigma_Y > 0$, $\mu_Y = E(\ln X)$, and $\sigma_Y^2 = Var(\ln X)$. For the remainder of this book, a random variable X with PDF given by equation 4-31 will be implied by the expression $X \sim LogN(\mu_Y, \sigma_Y^2)$. The parameters μ_Y and σ_Y^2 are the mean and variance of the normally distributed random variable $Y = \ln X$, which is the logarithmic representation of X (refer to figure 4-16). Graphs of a family of lognormal PDFs are presented in figure 4-17. Notice

the lognormal PDF is positively skewed and values for x are always nonnegative.

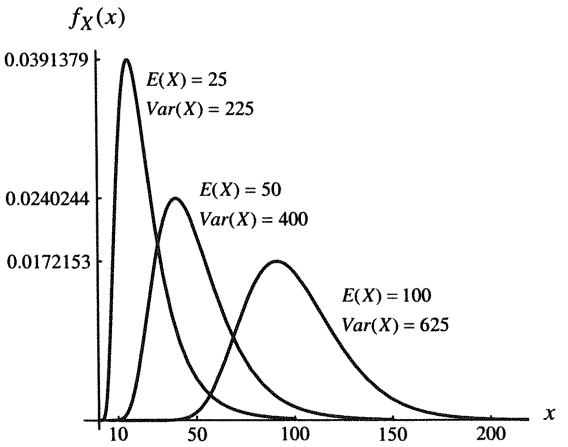


Figure 4-17. A Family of LogNormal Probability Density Functions

Theorem 4-7 If X is a lognormal random variable then $E(X) = \mu_X = e^{\mu_Y + \frac{1}{2}\sigma_Y^2}$ and $Var(X) = \sigma_X^2 = e^{2\mu_Y + \sigma_Y^2}(e^{\sigma_Y^2} - 1)$.

Proof Since X has a lognormal distribution, the PDF of X is given by equation 4-31; therefore

$$E(X) = \int_0^\infty x f_X(x)dx = \int_0^\infty x \cdot \frac{1}{\sqrt{2\pi}\,\sigma_Y} \frac{1}{x} e^{-\frac{1}{2}\left[(\ln x - \mu_Y)^2 / \sigma_Y^2\right]} dx \qquad (4\text{-}32)$$

Equation 4-32 simplifies to

$$E(X) = \int_0^\infty \frac{1}{\sqrt{2\pi}\,\sigma_Y} e^{-\frac{1}{2}\left[(\ln x - \mu_Y)^2 / \sigma_Y^2\right]} dx \qquad (4\text{-}33)$$

Suppose we set $y = \ln x - \mu_Y$; then $-\infty < y < \infty$, $x = e^y e^{\mu_Y}$, and $dx = e^y e^{\mu_Y} dy$. Substituting this into equation 4-33 we have

$$E(X) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_Y} e^{-\frac{1}{2}\left[y^2/\sigma_Y^2\right]} e^{y} e^{\mu_Y}\, dy \tag{4-34}$$

$$E(X) = e^{\mu_Y} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_Y} e^{-\frac{1}{2}\left[(y^2 - 2\sigma_Y^2 y)/\sigma_Y^2\right]} dy$$

$$E(X) = e^{\mu_Y} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_Y} e^{-\frac{1}{2\sigma_Y^2}\left[(y-\sigma_Y^2)^2 - \sigma_Y^4\right]} dy$$

$$E(X) = e^{\mu_Y} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_Y} e^{-\frac{1}{2\sigma_Y^2}\left[(y-\sigma_Y^2)^2\right]} e^{\frac{1}{2}\sigma_Y^2} dy$$

$$E(X) = e^{\mu_Y + \frac{1}{2}\sigma_Y^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_Y} e^{-\frac{1}{2\sigma_Y^2}\left[(y-\sigma_Y^2)^2\right]} dy = e^{\mu_Y + \frac{1}{2}\sigma_Y^2} \tag{4-35}$$

The integral in equation 4-35 is unity since it is the PDF of a $N(\sigma^2, \sigma^2)$ random variable. The above result can be generalized to the r-th moment of X; it is left to the reader to show that

$$E(X^r) = e^{r\mu_Y + \frac{1}{2}\sigma_Y^2 r^2} \tag{4-36}$$

To show that $Var(X) = e^{2\mu_Y + \sigma_Y^2}(e^{\sigma_Y^2} - 1)$ recall that

$$Var(X) = E(X^2) - \left(E(X)\right)^2 \tag{4-37}$$

Substituting equations 4-35 and 4-36 (with $r = 2$) into equation 4-37 it is easily shown that $Var(X) = e^{2\mu_Y + \sigma_Y^2}(e^{\sigma_Y^2} - 1)$.♦

This theorem can be illustrated by referring to figure 4-16. There, we have $\mu_Y = E(\ln X) = 4.57486$ and $\sigma_Y^2 = Var(\ln X) = 0.0606246$. From theorem 4-7

$$E(X) = e^{\mu_Y + \frac{1}{2}\sigma_Y^2} = e^{4.57486 + \frac{1}{2}(0.0606246)} = 100$$

$$Var(X) = e^{2\mu_Y + \sigma_Y^2}(e^{\sigma_Y^2} - 1) = e^{2(4.57486) + 0.0606246}(e^{0.0606246} - 1) = 625$$

Thus, when X is a lognormal random variable its mean and variance are defined in terms of the normally distributed random variable $Y = \ln X$. The same is true about the mode and median of X; in particular, if X is a lognormal random variable then

$$Mode(X) = e^{\mu_Y - \sigma_Y^2} \tag{4-38}$$

$$Median(X) = e^{\mu_Y} \tag{4-39}$$

In figure 4-16,

$$Mode(X) = e^{4.57486 - 0.0606246} = 91.307$$

$$Median(X) = e^{4.57486} = 97.014$$

The lognormal PDF peaks at the value

$$f_X(Mode(X)) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_Y} (e^{\frac{1}{2}\sigma_Y^2 - \mu_Y}) \tag{4-40}$$

Showing this is left as an exercise for the reader.

In cost analysis applications of the lognormal distribution we typically do not have values for $E(\ln X)$ and $Var(\ln X)$ (where X might represent the cost of a system or a particular work breakdown structure cost element). How do we specify the distribution function of a lognormal random variable X, when only $E(X)$ and $Var(X)$ are known? The next theorem addresses this question. Theorem 4-8 presents translation formulas for determining $E(\ln X)$ and $Var(\ln X)$ when only $E(X)$ and $Var(X)$ are known.

Theorem 4-8 If X is a lognormal random variable with mean $E(X) = \mu_X$ and $Var(X) = \sigma_X^2$ then

$$\mu_Y = E(\ln X) = \frac{1}{2} \ln \left[\frac{(\mu_X)^4}{(\mu_X)^2 + \sigma_X^2} \right] \tag{4-41}$$

and

$$\sigma_Y^2 = Var(\ln X) = \ln \left[\frac{(\mu_X)^2 + \sigma_X^2}{(\mu_X)^2} \right] \tag{4-42}$$

Proof From theorem 4-7 we have

$$\mu_X = e^{\mu_Y + \frac{1}{2}\sigma_Y^2}$$

$$\ln \mu_X = \mu_Y + \tfrac{1}{2}\sigma_Y^2 \tag{4-43}$$

$$2\ln \mu_X = 2\mu_Y + \sigma_Y^2 \tag{4-44}$$

We will first establish equation 4-42 in theorem 4-8 and then use that result to establish equation 4-41. From theorem 4-7

$$Var(X) = \sigma_X^2 = e^{2\mu_Y + \sigma_Y^2}(e^{\sigma_Y^2} - 1)$$

This is equivalent to $\ln(e^{\sigma_Y^2} - 1) = \ln \sigma_X^2 - (2\mu_Y + \sigma_Y^2)$

Using equation 4-44 $\ln(e^{\sigma_Y^2} - 1) = \ln \sigma_X^2 - 2\ln \mu_X$

$$\ln(e^{\sigma_Y^2} - 1) = \ln\left(\frac{\sigma_X^2}{\mu_X^2}\right)$$

$$e^{\sigma_Y^2} = \frac{\sigma_X^2}{\mu_X^2} + 1$$

Therefore $\sigma_Y^2 = Var(\ln X) = \ln\left[\dfrac{(\mu_X)^2 + \sigma_X^2}{(\mu_X)^2}\right]$

To establish equation 4-41, we can write μ_Y (in equation 4-43) as

$$\mu_Y = \ln \mu_X - \tfrac{1}{2}\sigma_Y^2$$

From equation 4-42 we have

$$\mu_Y = \ln \mu_X - \tfrac{1}{2}\ln\left[\frac{(\mu_X)^2 + \sigma_X^2}{(\mu_X)^2}\right]$$

$$\mu_Y = \frac{1}{2}\left(2\ln \mu_X - \ln\left[\frac{(\mu_X)^2 + \sigma_X^2}{(\mu_X)^2}\right]\right)$$

$$\mu_Y = \frac{1}{2}\left(\ln(\mu_X)^2 - \ln\left[\frac{(\mu_X)^2 + \sigma_X^2}{(\mu_X)^2}\right]\right)$$

Therefore $$\mu_Y = E(\ln X) = \frac{1}{2}\ln\left[\frac{(\mu_X)^4}{(\mu_X)^2 + \sigma_X^2}\right].$$

Using theorem 4-8 the parameters μ_Y and σ_Y^2, which uniquely specify the lognormal probability density function, can be determined from $E(X)$ and $Var(X)$. In figure 4-17, the left-most PDF has $E(X) = 25$ and $Var(X) = 225$; from theorem 4-8 this is equivalent to a lognormal PDF with parameters $\mu_Y = 3.06513$ and $\sigma_Y^2 = 0.307485$. The middle PDF (in figure 4-17) has $E(X) = 50$ and $Var(X) = 400$; from theorem 4-8 this is equivalent to a lognormal PDF with parameters $\mu_Y = 3.83781$ and $\sigma_Y^2 = 0.14842$. The right-most PDF (in figure 4-17) has $E(X) = 100$ and $Var(X) = 625$; from theorem 4-8 this is equivalent to a lognormal PDF with parameters $\mu_Y = 4.57486$ and $\sigma_Y^2 = 0.0606246$. Thus, the equations for the three PDFs in figure 4-17, from left to right, are as follows:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,(0.554513)}\frac{1}{x}e^{-\frac{1}{2}\left[(\ln x - 3.06513)^2/0.307485\right]}$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,(0.385253)}\frac{1}{x}e^{-\frac{1}{2}\left[(\ln x - 3.83781)^2/0.14842\right]}$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,(0.246221)}\frac{1}{x}e^{-\frac{1}{2}\left[(\ln x - 4.57486)^2/0.0606246\right]}$$

where the general form for $f_X(x)$ was given by equation 4-31.

The cumulative distribution function of a lognormal random variable is given by equation 4-45.

$$F_X(x) = P(X \le x) = \int_0^x \frac{1}{\sqrt{2\pi}\,\sigma_Y}\frac{1}{t}e^{-\frac{1}{2}\left[(\ln t - \mu_Y)^2/\sigma_Y^2\right]}dt \qquad (4\text{-}45)$$

Figure 4-18 presents a family of lognormal CDFs associated with the PDFs in figure 4-17.

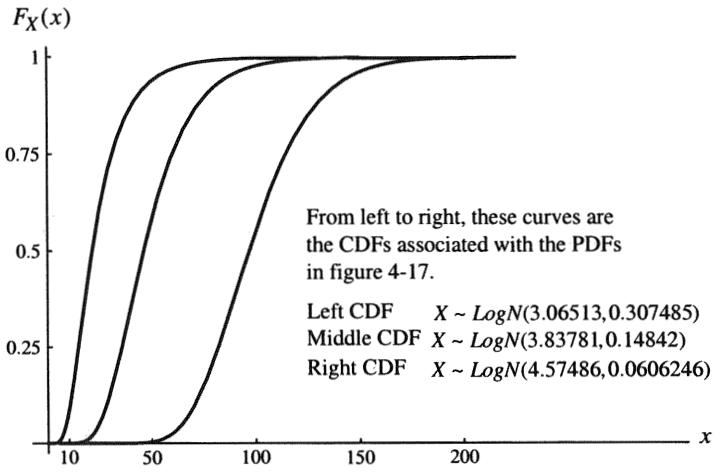


Figure 4-18. A Family of LogNormal CDFs

The cumulative distribution function given by equation 4-45 does not exist in closed form. It can be evaluated by a numerical integration procedure. An alternative to such a procedure involves using a table of values from the standard normal distribution. The following discusses this approach.

If $X \sim LogN(\mu_Y, \sigma_Y^2)$ then $Y = \ln X \sim N(\mu_Y, \sigma_Y^2)$; therefore,

$$P(X \le x) = P(\ln X \le \ln x) = P\left(\frac{\ln X - \mu_Y}{\sigma_Y} \le \frac{\ln x - \mu_Y}{\sigma_Y}\right) \qquad (4\text{-}46)$$

Since $Y = \ln X \sim N(\mu_Y, \sigma_Y^2)$, from theorem 4-5 it follows that

$$\frac{\ln X - \mu_Y}{\sigma_Y} \sim N(0,1)$$

This implies the random variable $\dfrac{\ln X - \mu_Y}{\sigma_Y}$ is equivalent to the standard normal random variable Z. From this result equation 4-46 is equivalent to

$$P(X \le x) = P\left(Z \le \frac{\ln x - \mu_Y}{\sigma_Y}\right) \qquad (4\text{-}47)$$

If X has a lognormal distribution, then probabilities associated with various intervals around X can be determined from a table of values of Z, the standard normal distribution.

Example 4-8 Suppose the uncertainty in a system's cost is described by a lognormal PDF with $E(Cost) = 100$ (\$M) and $Var(Cost) = 625$ $(\$M)^2$; this is the right-most PDF in figure 4-17. Determine

a) $P(Cost > 2E(Cost))$

b) $P(50 \le Cost \le 150)$

Solution

a) To determine $P(Cost > 2E(Cost))$ recall that

$$P(Cost > 2E(Cost)) = 1 - P(Cost \le 2E(Cost))$$

It is given that $E(Cost) = 100$; therefore

$$P(Cost > 200) = 1 - P(Cost \le 200)$$

In this example, the random variable *Cost* is given to have a lognormal distribution with $E(Cost) = 100$ and $Var(Cost) = 625$. Thus, the random variable $Y = \ln Cost$ is normally distributed with parameters (determined from theorem 4-8)

$$\mu_Y = E(\ln Cost) = 4.57486$$

$$\sigma_Y^2 = Var(\ln Cost) = 0.0606246$$

From equation 4-47

$$P(Cost \le 200) = P\left(Z \le \frac{\ln 200 - 4.57486}{0.246221}\right) = P(Z \le 2.938)$$

From table A-1 (appendix A) $P(Z \le 2.938) = 0.998348$, after some interpolation. Therefore,

$$P(Cost > 200) = 1 - P(Z \le 2.938) = 0.00165$$

This result is consistent with the Markov bound discussion in chapter 3 (section 3.4), as illustrated in figure 3-19.

b) To determine $P(50 \le Cost \le 150)$ note that

$$P(50 \le Cost \le 150) = P(\ln 50 \le \ln(Cost) \le \ln 150)$$

$$= P\left(\frac{\ln 50 - \mu_Y}{\sigma_Y} \le \frac{\ln Cost - \mu_Y}{\sigma_Y} \le \frac{\ln 150 - \mu_Y}{\sigma_Y} \right)$$

$$= P\left(\frac{\ln 50 - \mu_Y}{\sigma_Y} \le Z \le \frac{\ln 150 - \mu_Y}{\sigma_Y} \right)$$

$$= P(-2.69 \le Z \le 1.77)$$

where

$$Z = \frac{\ln Cost - \mu_Y}{\sigma_Y}$$

$\mu_Y = E(\ln Cost) = 4.57486$ (from theorem 4-8)

$\sigma_Y^2 = Var(\ln Cost) = 0.0606246$ (from theorem 4-8)

From theorem 4-5 we know $Z \sim N(0,1)$, thus

$$P(50 \le Cost \le 150) = P(-2.69 \le Z \le 1.77)$$
$$= F_Z(1.77) - F_Z(-2.69)$$
$$= F_Z(1.77) - [1 - F_Z(2.69)]$$

where $F_Z(-2.69) = 1 - F_Z(2.69)$. From table A-1 (appendix A)

$$P(50 \le Cost \le 150) = 0.961636 - [1 - 0.9964] = 0.958 \approx 0.96$$

Thus, the system's cost will fall between 50 and 150 million dollars with probability 0.96. This result is also consistent with the discussion presented in chapter 3 (section 3.4), as illustrated in figure 3-20.

Example 4-9 In figure 1-5 (chapter 1) the random variable X_2 represented the cost of a system's systems engineering and program management. Furthermore, the point estimate of X_2, denoted by $x_{2PE_{X_2}}$, was equal to 1.26 ($M). If X_2 can be approximated by a lognormal distribution, with $E(X_2) = 1.6875$ ($M) and $Var(X_2) = 0.255677$ ($M)^2, determine

a) $P(X_2 \le x_{2PE_{X_2}})$

b) $P(X_2 \le E(X_2))$

Solution

a) Since the distribution function of X_2 is approximated by a lognormal, from equation 4-47 we can write

$$P(X_2 \le x_{2PE_{X_2}}) = P\left(Z \le \frac{\ln x_{2PE_{X_2}} - \mu_Y}{\sigma_Y}\right)$$

where $Z \sim N(0,1)$, $\mu_Y = E(\ln X_2)$, and $\sigma_Y^2 = Var(\ln X_2)$. Since $E(X_2) = 1.6875$ and $Var(X_2) = 0.255677$, from theorem 4-8 $\mu_Y = 0.480258$ and $\sigma_Y^2 = 0.0859804$. Thus,

$$P(X_2 \le 1.26) = P\left(Z \le \frac{\ln 1.26 - 0.480258}{0.293224}\right) = P(Z \le -0.85)$$

From table A-1 (appendix A)

$$P(Z \le -0.85) = P(Z \ge 0.85) = 1 - P(Z < 0.85) = 1 - 0.802 = 0.198$$

thus, $$P(X_2 \le 1.26) = P(Z \le -0.85) = 0.198$$

Therefore, there is nearly a 20 percent chance the cost of the system's systems engineering and program management will be less than or equal to 1.26 ($M).

b) We are given $E(X_2) = 1.6875$, therefore $P(X_2 \le E(X_2)) = P(X_2 \le 1.6875)$.

From equation 4-47 we can write

$$P(X_2 \le 1.6875) = P\left(Z \le \frac{\ln 1.6875 - 0.480258}{0.293224}\right) = P(Z \le 0.1466)$$

From table A-1 (appendix A) $P(Z \le 0.1466) = 0.558$; thus,

$$P(X_2 \le 1.6875) = P(Z \le 0.1466) = 0.558$$

Therefore, there is nearly a 56 percent chance the cost of the system's systems engineering and program management will be less than or equal to 1.6875 ($M). For interest, the PDF and CDF of X_2, for this example, are shown in the figure below.

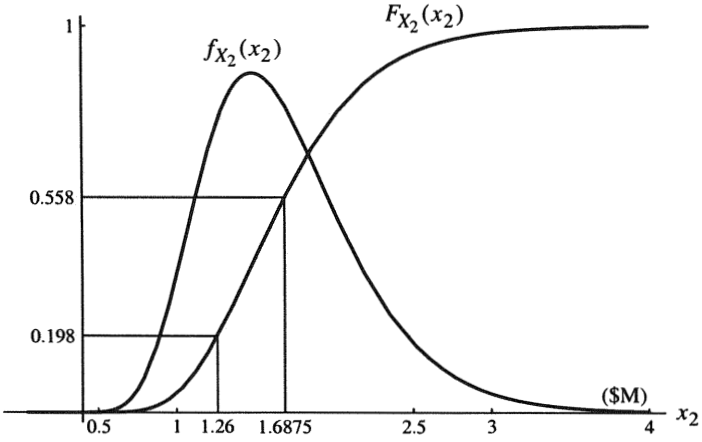


Figure 4-19. The PDF and CDF of X_2 in Example 4-9

This concludes the discussion of special probability distributions commonly used in cost uncertainty analysis. Chapters 5 through 7 will provide further illustrations of their application to modeling cost uncertainty from a system work breakdown structure perspective. In preparing for that discussion this chapter concludes with a presentation on how to specify some of these special distributions, when only partial information about them is available.

4.5 Specifying Continuous Probability Distributions

In systems engineering, probability distributions of variables whose values are uncertain must often be specified by expert technical opinion. This is particularly true in the absence of historical data. In such circumstances, expert opinion can be the only way to quantify a variable's uncertainty. Even when data exists its quality may be so suspect as to nullify its use altogether. This section discusses strategies for specifying probability distributions when expert subjective assessments are required. This is illustrated in the context of continuous probability distributions.* Before delving into the details of these strategies, we discuss further the concept of subjective probabilities and distribution functions (introduced in chapter 2).

Subjective Probabilities and Distribution Functions

In systems engineering, probabilities are often used to quantify uncertainties associated with a system's design parameters (e.g., weight), as well as uncertainties in cost and schedule. For reasons mentioned above, quantifying this uncertainty is often done in terms of subjective probabilities. Discussed in chapter 2, subjective probabilities are those assigned to events on the basis of personal judgment. They measure of a person's degree-of-belief that an event will occur. Subjective probabilities are most often associated with one-time, nonrepeatable, events — those whose probabilities cannot be objectively determined from a population of outcomes developed by repeated trials, observations, or experimentation. Subjective probabilities cannot be arbitrary; they *must* adhere to the axioms of probability [refer to chapter 2]. For instance, if an electronics engineer assigns a probability of 0.70 to the event

---

* In practice, a continuous distribution is often used to describe the range of possible values for a random variable. This enables subject matter experts to focus on the "shape" that best describes the distribution of probability, rather than assessing individual probabilities associated to each distinct possible value (needed for discrete distributions).

"*the number of gates for the new processor chip* ***will not exceed*** *12,000*," it must follow that the chip *will exceed* 12,000 gates with probability 0.30. Subjective probabilities are *conditional* on the state of the person's knowledge, which changes with time. To be credible, subjective probabilities should *only* be assigned to events by subject experts — persons with significant experience with events similar to the one under consideration. In addition, the rationale supporting the assigned probability *must be well documented.*

Instead of assigning a single subjective probability to an event, subject experts often find it easier to describe a function that depicts a subjective distribution of probabilities. Such a distribution is sometimes called a *subjective probability distribution.* Subjective probability distributions are governed by the properties of probability distributions associated with discrete or continuous random variables (refer to chapter 3). Because of their nature, subjective probability distributions can be thought of as "belief functions"— mathematical representations of a subject expert's best professional judgment in the distribution of probabilities for a particular event.

When formulating subjective probability distributions, subject experts often prefer specifying a range that contains most, but not all, possible values. That is, there is a small nonzero probability that values will occur outside the expert's specified range. One strategy for specifying a subjective probability distribution involves the direct assessment of the distribution's fractiles. Another strategy involves assigning a subjective probability to a subinterval of the range of the distribution function. The following illustrates these strategies. This is done in the context of the distributions presented in this chapter. We begin with the beta distribution.

Specifying a Beta Distribution

The beta distribution has long been the distribution of "choice" for subjective assessments. It can take a wide-variety of forms, as seen in figure 4-9 and figure 4-10. The following illustrates how the beta distribution can be specified from subjective assessments on the shape parameters α and β and *any* two fractiles.

Case 1 Specify a nonstandard beta distribution for the random variable X given the shape parameters α and β and any two fractiles x_i and x_j, where $(0 \le i < j \le 1)$. An illustration of this case is presented in figure 4-20.

Purpose(s) To determine the minimum and maximum possible values for X, where $X \sim Beta(\alpha, \beta, a, b)$. To compute $E(X)$ and $Var(X)$ from the specified distribution.

Required Information

Assessments of α and β and any two fractiles x_i and x_j.

Discussion

An assessment of the shape parameters α and β can be facilitated by having a subject expert look at a family of beta distributions, as shown in figure 4-9 and figure 4-10. From such a family, an α and β pair can be chosen that reasonably depicts the distribution of probability (e.g., skewed, symmetric) for the variable under consideration. With α and β and any two fractiles x_i and x_j, the minimum and maximum possible values of X are given by equations 4-48 and 4-49 (refer to exercise 25), respectively.

$$a = \frac{x_i y_j - x_j y_i}{y_j - y_i} \tag{4-48}$$

$$b = \frac{x_j(1-y_i) - x_i(1-y_j)}{y_j - y_i} \tag{4-49}$$

In the above, the terms x_i and x_j are the assessed values of X such that $P(X \le x_i) = i$ and $P(X \le x_j) = j$. The terms y_i and y_j are fractiles computed from the *standard beta distribution* associated with the given (as chosen by the subject expert) α and β.

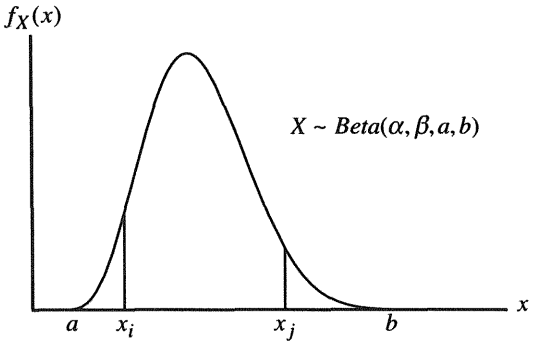


Figure 4-20. An Illustrative Beta Distribution — Case 1

Once a and b have been determined, theorem 4-4 can be used to compute $E(X)$ and $Var(X)$ associated with the specified distribution.

Example 4-10 Find the minimum and maximum possible values of X if $X \sim Beta(5,10,a,b)$, $x_{0.05} = 4.76359$, and $x_{0.95} = 6.70003$. Find $E(X)$ and $Var(X)$.

Solution

Since $X \sim Beta(5,10,a,b)$, the distribution function of X has shape parameters $\alpha = 5$ and $\beta = 10$. From equations 4-48 and 4-49 we can write

$$a = \frac{4.76359y_{0.95} - 6.70003y_{0.05}}{y_{0.95} - y_{0.05}} \tag{4-50}$$

$$b = \frac{6.70003(1 - y_{0.05}) - 4.76359(1 - y_{0.95})}{y_{0.95} - y_{0.05}} \qquad (4\text{-}51)$$

Since the random variable Y must have the standard beta distribution $Y \sim Beta(5,10)$, it can be determined* that $y_{0.05} = 0.152718$ and $y_{0.95} = 0.540005$. Substituting these values into equation 4-50 and 4-51 we have $a = 4$ and $b = 9$, which are the minimum and maximum possible values of X, respectively. The reader should notice this example is directly related to example 4-4 (section 4.2). Now that values for a and b are determined, the mean and variance of X can be determined directly from theorem 4-4. It is left to the reader to show that $E(X) = 5.67$ and $Var(X) = 0.347$, in this example.

Example 4-11 Suppose I represents the uncertainty in the number of delivered source instructions (DSI) for a new software application. Suppose a team of software engineers judged 100,000 DSI as a reasonable assessment of the 50th percentile of I and a size of 150,000 DSI as a reasonable assessment of the 95th percentile. Furthermore, suppose the distribution function in figure 4-21 was considered a good characterization of the uncertainty in the number of DSI. Given this,

a) Find the extreme possible values for I.

b) Compute the mode of I.

c) Compute $E(I)$ and σ_I.

Solution

a) In figure 4-21, I is given to be a beta distribution with shape parameters $\alpha = 2$ and $\beta = 3.5$. We are also given two probability assessments for I,

* Determined by the *Mathematica*® routine Quantile[BetaDistribution[5,10],k], where k is equal to 0.05 and 0.95.

specifically, $P(I \le 100{,}000) = 0.50$ and $P(I \le 150{,}000) = 0.95$; this is equivalent to the fractiles $x_{0.50} = 100{,}000$ and $x_{0.95} = 150{,}000$ (refer to figure 4-21). Since $\alpha = 2$ and $\beta = 3.5$, the *standard beta distribution* is $Y \sim Beta(2, 3.5)$. From this we can determine the fractiles $y_{0.50}$ and $y_{0.95}$. Using *Mathematica*®, $y_{0.50} = 0.346086$ and $y_{0.95} = 0.70189$ when $\alpha = 2$ and $\beta = 3.5$. Substituting $y_{0.50} = 0.346086$, $y_{0.95} = 0.70189$, $x_{0.50} = 100{,}000$, and $x_{0.95} = 150{,}000$ into equations 4-48 and 4-49 provides the minimum and maximum possible values for I. These values are denoted below by a and b.

$$a = \frac{(100000)0.70189 - (150000)0.346086}{0.70189 - 0.346086} = 51366$$

$$b = \frac{150000(1 - 0.346086) - 100000(1 - 0.70189)}{0.70189 - 0.346086} = 191892$$

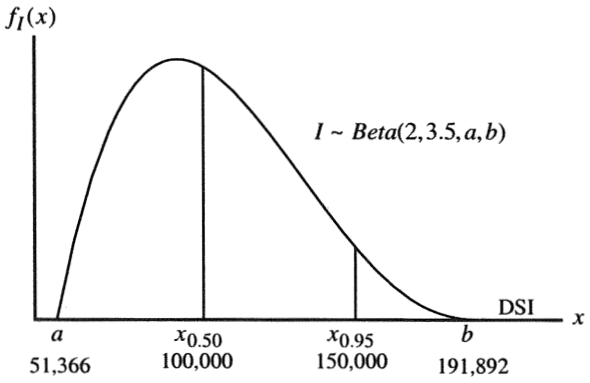


Figure 4-21. Beta Distribution for Example 4-11

b) Since $\alpha > 1$ and $\beta > 1$, from equation 4-10, the mode of $Y \sim Beta(2, 3.5)$ is

$$y = \frac{1 - \alpha}{2 - \alpha - \beta} = \frac{1 - 2}{2 - 2 - 3.5} = 0.2857$$

By the linear transformation $y=(x-a)/(b-a)$, where a and b were determined from part a) we have $Mode(I)=a+0.2857(b-a)=91{,}514$ DSI. Because the beta distribution in this example has a positive skew, the mode of I falls to the left of the 50th percentile of I.

c) From theorem 4-4 with $\alpha=2$, $\beta=3.5$, $a=51366$ DSI, and $b=191892$ DSI we have

$$E(I)=a+(b-a)E(Y)=a+(b-a)\frac{\alpha}{\alpha+\beta}$$

$$=51366+(191892-51366)\frac{2}{2+3.5}=102{,}466 \text{ DSI}$$

Once again, because the beta distribution in this example has a positive skew, the mean of I falls to the right of the 50th percentile of I. Lastly, from equations 4-13 it can be shown that $Var(Y)=0.0356$. From equation 4-13a this translates to $Var(I)=7.03(10)^8$ DSI^2; therefore,

$$\sigma_I=\sqrt{Var(I)}=26514 \text{ DSI}\blacklozenge$$

A nice feature of this approach is its flexibility to fully specify, for a given pair of shape parameters, a *nonstandard beta distribution* from *any two fractiles* of the distribution. This feature has strong practical utility. Subject experts often make "better" judgmental assessments of fractiles that fall near the middle of a distribution (e.g., the $x_{0.40}$ and $x_{0.60}$ fractiles) than out near its tails. Selecting shape parameters that "best" characterize the skewness (or symmetry) of the distribution has not been considered, in practice, too difficult. Shape parameters can be inferred by asking the expert to visually choose a distribution from a family of beta distributions plotted for various α and β. Representative plots of such a family are shown in figures 4-9 and 4-10. Visual representations of a variable's uncertainty by distribution functions can be an excellent way to communicate risk to decision-makers.

Specifying Uniform Distributions

The following presents strategies for specifying a uniform distribution, when a subject expert assigns a probability α to a subinterval of the distribution's range. In the cases below, assume the random variable X is *uniformly distributed* over the range $a \le x \le b$.

Case 2 Specify a uniform distribution for the random variable X given the subinterval $a \le x \le b'$ and α, where a is the minimum possible value of X, $b' < b$, and $\alpha = P(a \le X \le b')$. An illustration of this case is presented in figure 4-22.

Purpose(s) To determine the maximum possible value of X. To compute $E(X)$ and $Var(X)$ from the specified distribution.

Required Information

Assessments of α and the endpoints of the subinterval $a \le x \le b'$.

Discussion

In this case a subject expert defines the subinterval $a \le x \le b'$ of the range of possible values for X, given by $a \le x \le b$. In addition, an assessment is made on the probability X will fall in this subinterval.

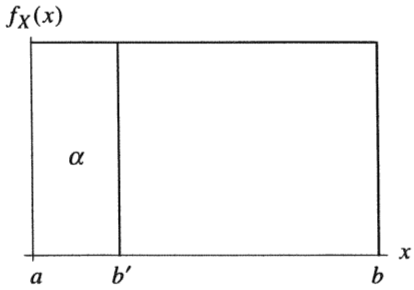


Figure 4-22. An Illustrative Uniform Distribution — Case 2

If $P(a \le X \le b') = \alpha < 1$ the maximum possible value of X is

$$b = a + \frac{1}{\alpha}(b' - a) \tag{4-52}$$

For example, if $\alpha = 0.25$, $a = 20$, and $b' = 30$ then, from equation 4-52, the maximum value of X must be $b = 60$. This is illustrated in figure 4-23.

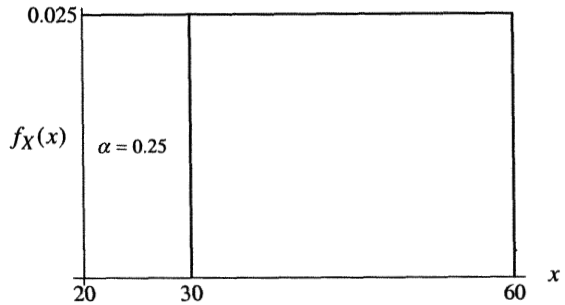


Figure 4-23. An Illustration of Case 2

For an application context, the random variable X might represent the uncertainty in the number of source instructions to develop for a new software application, or in the weight of a new electronic device, or in the number of labor hours to assemble a new widget.

Case 3 Specify a uniform distribution for the random variable X given the subinterval $a' \le x \le b'$ and α, where $a < a'$, $b' < b$, and $\alpha = P(a' \le X \le b')$. An illustration of this case is presented in figure 4-24.

Purpose(s) To determine the minimum and maximum possible values of X. To compute $E(X)$ and $Var(X)$ from the specified distribution.

Required Information

Assessments of α and the endpoints of the subinterval $a' \le x \le b'$. Furthermore, assume

$$a' - a = b - b'$$

for this case.

Discussion

In this case a subject expert defines the subinterval $a' \le x \le b'$ of the range of possible values for X, given by $a \le x \le b$. In addition, an assessment of the probability X will fall in the subinterval $a' \le x \le b'$ is made.

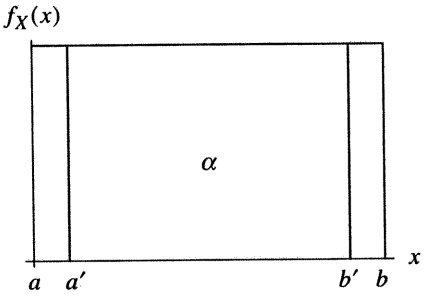


Figure 4-24. An Illustrative Uniform Distribution — Case 3

If $P(a' \le X \le b') = \alpha < 1$, the minimum and maximum possible values of X are

$$a = a' - \frac{1-\alpha}{2\alpha}(b' - a') \tag{4-53}$$

$$b = b' + \frac{1-\alpha}{2\alpha}(b' - a') \tag{4-54}$$

Notice that $a' - a = b - b'$. Furthermore, for this case we have

$$P(a \le X < a') = P(b' < X \le b) = \tfrac{1}{2}(1-\alpha)$$

For example, if $\alpha = 0.80$, $a' = 40$, and $b' = 60$ then, from equations 4-53 and 4-54, the minimum and maximum possible values of X are $a = 37.5$ and $b = 62.5$. This is illustrated in figure 4-25. An application context for this case is similar to the previous case.

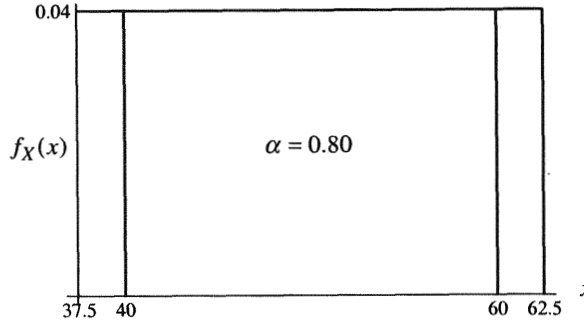


Figure 4-25. An Illustration of Case 3

It is possible, in this case, for a to become negative even when a' is positive. In applications where it is sensible that X *be nonnegative* (e.g., if X is the uncertainty in the weight of a new widget), such an occurrence signals a reassessment of a' and α is needed.

Specifying a Triangular Distribution*

The following illustrates one strategy for specifying a triangular distribution, when a subject expert assigns a probability α to a subinterval of the distribution's range. In the case below, assume the random variable X has a *triangular distribution* over the range $a \le x \le b$.

Case 4 Specify a triangular distribution for the random variable X given m, the subinterval $a' \le x \le b'$, and α where $a < a'$, $a' < m < b'$, $b' < b$, and $\alpha = P(a' \le X \le b')$. An illustration of this case is presented in figure 4-26.

Purpose(s) To determine the minimum and maximum possible values of X. To compute $E(X)$ and $Var(X)$ from the specified distribution.

---

* This case was developed by Dr. Chien-Ching Cho, The MITRE Corporation, Bedford, Massachusetts.

Required Information

Assessments of α and the endpoints of the subinterval $a' \le x \le b'$, where $a' < m < b'$. Furthermore, assume for this case

$$\frac{P(X \le a')}{P(X \ge b')} = \frac{P(X \le m)}{P(X \ge m)}$$

Discussion

In this case a subject expert defines the subinterval $a' \le x \le b'$ of the range of possible values for X, given by $a \le x \le b$. In addition, an assessment of the probability X will fall in the subinterval $a' \le x \le b'$ is made.

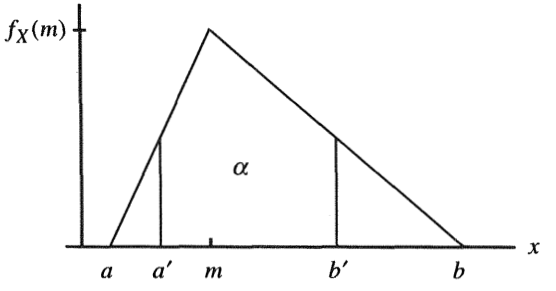


Figure 4-26. An Illustrative Triangular Distribution — Case 4

If $P(a' \le X \le b') = \alpha < 1$, the minimum and maximum possible values of X are

$$a = m - \frac{m - a'}{1 - \sqrt{1 - \alpha}} \tag{4-55}$$

$$b = m + \frac{b' - m}{1 - \sqrt{1 - \alpha}} \tag{4-56}$$

Equations 4-55 and 4-56 originate from the assumption (for this case) that

$$\frac{P(X \le a')}{P(X \ge b')} = \frac{P(X \le m)}{P(X \ge m)}$$

For example, if $\alpha = 0.75$, $a' = 25$, $m = 35$, and $b' = 60$ then, from equations

4-55 and 4-56, the minimum and maximum possible values of X are $a = 15$ and $b = 85$. This is illustrated in figure 4-27.

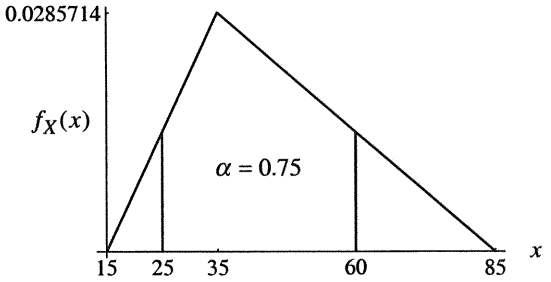


Figure 4-27. An Illustration of Case 4

An application context for this case is similar to the previous cases. It is also possible in this case for a to become negative, even when a' is positive. In applications where it is sensible that X *be nonnegative* (e.g., if X is the uncertainty in the weight of a new widget), such an occurrence signals a reassessment of a' and α is needed.

In Summary

Sir Josiah Stamp* once said...

> *"The government are very keen on amassing statistics. They collect them, raise them to the n-th power, take the cube root, and prepare wonderful diagrams. But one must never forget that every one of these figures comes in the first instance from the village watchman, who puts down what he damn pleases."*

Several techniques have been presented for quantifying uncertainty in terms of subjective probabilities and distributions. As discussed, the need to do so is unavoidable on systems engineering projects. An extensive body of social

---

* President of the Bank of England during the 1920s.

science research exists on techniques for eliciting subjective probabilities and distributions. The book *Uncertainty: A Guide to Dealing With Uncertainty in Quantitative Risk and Policy Analysis*, by Morgan and Henrion [7], provides an excellent summary of this research. Despite these studies, there remains a lack of consensus on the superiority of any particular elicitation technique.

Although the use of expert opinion is sometimes criticized, the basis of the criticism is often traceable to a) the subject expert was really the "village watchman" or b) the full scope of the problem being addressed by the expert was poorly described. To lessen the chance of a) or b) occurring, it is the prime responsibility of the project's cost and engineering team to collectively do the technical diligence needed to establish credible and defensible assessments.

For our purposes, it must be stressed that a key product from subjective assessment efforts must be a well documented set of assumptions, arguments, and supportive materials. Documentation enables similarly qualified persons (or teams) to conduct independent and objective reviews of the assessments. This alone is an important step towards objectivity and one that would surface the presence of "village watchmen." Credible analyses stem from credible and defensible assessments; credible and defensible assessments stem from credible expertise. Properly conducted and documented assessments, on areas of a project that drive cost, schedule, and technical uncertainties, are among the most important products cost uncertainty analysis drives to produce.

Exercises

1. Given the trapezoidal distribution in example 4-1, show that

   a) $P(25000 \le X \le 28000) = \frac{2}{13}$ b) $P(25000 \le X \le 35000) = \frac{34}{39}$

2. Suppose $X \sim Trap(a, m_1, m_2, b)$ with PDF given in figure 4-1.

   a) Show that $1 - P(X \le m_1) - P(X > m_2) = 2\dfrac{(m_2 - m_1)}{m_2 + b - a - m_1}$

   b) What region in figure 4-1 does the probability in exercise 2a) represent?

3. If $Cost \sim Unif(3,8)$, then answer the following:

   a) $P(Cost < 5)$

   b) $P(4 < Cost \le 7)$

   c) Find x such that $P(Cost \le x) = 0.80$.

4. If $X \sim Unif(a,b)$ show that

   a) $E(X) = \frac{1}{2}(a+b)$ b) $Var(X) = \frac{1}{12}(b-a)^2$

5. For the uniform distributions defined in case 2 and case 3, section 4.5, derive equations 4-52 (in case 2), 4-53 (in case 3), and 4-54 (in case 3).

6. If $X \sim Trng(a,m,b)$, then answer the following:

   a) Verify $f_X(x)$ given by equation 4-6 is a PDF.

   b) Show $F_X(x)$ changes concavity at $Mode(X)$.

   c) Prove that $E(X) = \frac{1}{3}(a+m+b)$.

7. Verify the probabilities in figure 4-8 by computing the areas under the appropriate regions of each triangle.

8. If $X \sim Trng(15,35,85)$, then answer the following:

   a) Compute $P(X \le 60)$.

   b) Compute $P(X \le 25)$.

   c) Show that $P(X \le 60) - P(X \le 25) = 0.75$ (as seen in figure 4-27).

9. If $X \sim Trng(0,1,1)$ compute

   a) $E(5X+1)$ b) $Var(3X-1)$

10. If $Y \sim Beta(\alpha,\beta)$, verify equations 4-14 and 4-15 if $E(Y)$ and $Var(Y)$ are known.

11. Suppose $Y \sim Beta(\alpha,\beta)$ and $f_Y(y) = 12y^2(1-y)$, where $0 < y < 1$

    a) Find α and β. b) Compute $E(Y)+\sigma_Y$.

    c) Determine $P(0.3 < Y \leq 0.7)$.

12. In example 4-4 (section 4.2)

    a) Determine whether the expected time (in minutes) to assemble the microcircuit is greater than or less than the most probable time.

    b) Compute the standard deviation (in minutes) of the assembly time.

13. If the cost of a system is *normally distributed* with mean 20 ($M) and standard deviation 4 ($M) determine

    a) $P(Cost \leq 17)$ b) $P(15 \leq Cost < 22)$ c) $P\left(|Cost-\mu| \geq \frac{1}{2}\right)$

14. Suppose the uncertainty in a system's cost is described by a *normal distribution.* Suppose there is a 5 percent chance the system's cost will not exceed 100 ($M) and an 85 percent chance its cost will not exceed 200 ($M). From this information determine the mean and standard deviation of the system's cost.

15. If $X \sim N(\mu,\sigma^2)$, then show the following is true

    a) $f_X(x)$ changes concavity at the points $x = \mu+\sigma$ and $x = \mu-\sigma$.

    b) $P(\mu-2\sigma \leq X \leq \mu+2\sigma) = 0.9544$

    c) $P(\mu-3\sigma \leq X \leq \mu+3\sigma) = 0.9973$

16. If X has a *lognormal distribution*, what does $P(\ln X \leq E(\ln X))$ always equal?

17. Compute the mean and variance of $\ln X$ for the three lognormal distributions in figure 4-17.

18. Suppose the uncertainty in a system's cost is described by a *lognormal* PDF with $E(Cost) = 25$ (\$M) and $Var(Cost) = 225$ $(\$M)^2$; this is the left-most PDF in figure 4-17. Determine
    a) $P(Cost > E(Cost))$ b) $P(Cost \leq 50)$

19. In figure 1-5 (chapter 1) the random variable X_2 represented the cost of a system's systems engineering and program management. The point estimate of X_2, denoted by $x_{2PE_{X_2}}$, was equal to 1.26 (\$M). If X_2 can be approximated by a *lognormal distribution*, with $E(X_2) = 1.6875$ (\$M) and $Var(X_2) = 0.255677$ $(\$M)^2$, determine
    a) $P(x_{2PE_{X_2}} \leq X_2 < E(X_2))$
    b) $P(1 \leq X_2 < 2.5)$ c) $P(X_2 \leq 2.5)$

20. If X is a lognormal random variable, show the maximum value *of its density function* is given by equation 4-40.

21. If X is a lognormal random variable, show that the r-th moment of X is given by $E(X^r) = e^{r\mu_Y + \frac{1}{2}\sigma_Y^2 r^2}$

22. Suppose I represents the uncertainty in the number of delivered source instructions (DSI) for a new software application. Suppose a team of software engineers judged 35,000 DSI as a reasonable assessment of the 50th percentile of I and a size of 60,000 DSI as a reasonable assessment

of the 95th percentile. Furthermore, suppose the distribution function in figure 4-21 was considered a good characterization of the uncertainty in the number of DSI. Given this,

a) Find the extreme possible values for I.

b) Compute the mode of I.

c) Compute $E(I)$ and σ_I.

23. Suppose W represents the uncertainty in the weight of a new unmanned spacecraft. Suppose a team of space systems engineers judged 1500 pounds as a reasonable assessment of the minimum possible weight. Furthermore, suppose this team also assessed the chance that W could fall between the minimum possible weight and 2000 pounds to be 80 percent. If the distribution function for W is *uniform*, determine the expected weight of the spacecraft.

24. Suppose I represents the uncertainty in the amount of new code for a software application. Suppose this uncertainty is characterized by the *triangular* PDF in figure 4-26. If the probability is 0.90 that the amount of code is between 20,000 DSI and 30,000 DSI, with 25,000 DSI as most probable, determine $E(I)$.

25. For the beta distribution defined in case 1, section 4.5, show that

$$a = \frac{x_i y_j - x_j y_i}{y_j - y_i} \text{ and } b = \frac{x_j(1-y_i) - x_i(1-y_j)}{y_j - y_i}$$

Hint: Solve for a and b from a simultaneous equation that involves the transformation $y = (x-a)/(b-a)$. Note that $P(Y \le y_i) = i = P(X \le x_i)$ and $P(Y \le y_j) = j = P(X \le x_j)$, in the context of case 1 (section 4.5).

References

1. Johnson, N. L., and S. Kotz. 1969. *Distributions in Statistics: Discrete Distributions*. 1970. *Continuous Univariate Distributions 1, Continuous Univariate Distributions 2*. 1972. *Continuous Multivariate Distributions*. New York: John Wiley & Sons, Inc.

2. Young, D. C., and P. H. Young. 1995. A Generalized Probability Distribution for Cost/Schedule Uncertainty in Risk Assessment. *Proceedings of the 1995 Western MultiConference on Business/MIS Simulation Education*, The Society for Computer Simulation. San Diego, California.

3. Wolfram, S. 1991. *Mathematica®: A System for Doing Mathematics by Computer*, 2nd ed. Reading, Massachusetts: Addison-Wesley Publishing Company, Inc.

4. Abramson, R. L., and P. H. Young. 1997 (Spring). FRISKEM–Formal Risk Evaluation Methodology. *The Journal of Cost Analysis*, pp. 29-38.

5. Garvey, P. R., and A. E. Taub. 1997 (Spring). A Joint Probability Model for Cost and Schedule Uncertainties. *The Journal of Cost Analysis*, pp. 3-27.

6. Garvey, P. R. 1996 (Spring). Modeling Cost and Schedule Uncertainties – A Work Breakdown Structure Perspective. *Military Operations Research*, V2, N1, pp. 37-43.

7. Morgan, M. G., and M. Henrion. 1990. *Uncertainty: A Guide to Dealing With Uncertainty in Quantitative Risk and Policy Analysis*. New York: Cambridge University Press.

Chapter 5

Functions of Random Variables and Their Application to Cost Uncertainty Analysis

When nothing is sure, everything is possible.
Margaret Drabble, 1939
English Novelist
The Middle Ground, 1980

Lest men suspect your tale untrue,
Keep probability in view.
John Gay, 1688-1732
English Poet, Dramatist
The Painter who pleased Nobody and Everybody

This chapter presents methods for studying the behavior of *functions of random variables*. Topics include joint probability distributions, linear combinations of random variables, the central limit theorem, and the development of distribution functions specific to a general class of software cost-schedule models.

5.1 Introduction

Functions of random variables occur frequently in cost engineering and analysis problems. For example, the first unit-cost UC of an unmanned spacecraft might be derived according to [1]

$$UC = 5.48(SC_{wt})^{0.94}(BOLP)^{0.30}$$

where SC_{wt} is the spacecraft's dry weight (pounds) and $BOLP$ is the beginning-of-life power (watts). If it's early in a new spacecraft's design the precise values for SC_{wt} and $BOLP$ might be unknown. The engineering team might better assess ranges of possible values for SC_{wt} and $BOLP$ instead of single point values. These ranges might be described by probability distributions, such as those presented in chapter 4. If the first unit-cost is a function of the random variables SC_{wt} and $BOLP$, a common question is "*What is the probability distribution of UC given probability distributions for* SC_{wt} *and* $BOLP$?" This

chapter presents methods to answer this and related questions. First, some mathematical preliminaries.

5.1.1 Joint and Conditional Distributions

When a function is defined by two or more random variables its probability distribution is called a *joint probability distribution*. Joint probability distributions generalize the concept of univariate distributions to functions of several random variables. Analogous to the univariate case, the *joint cumulative distribution function* of two random variables X and Y is

$$F_{X,Y}(x,y) = P(X \le x, Y \le y) \qquad -\infty < x, y < \infty \tag{5-1}$$

Discrete Random Variables

If X and Y are *discrete* random variables their *joint probability mass function* is defined as

$$p_{X,Y}(x,y) = P(X = x, Y = y) \tag{5-2}$$

Illustrated in figure 5-1, $p_{X,Y}(x,y)$ is the probability a possible pair of values (x_i, y_k) will occur.

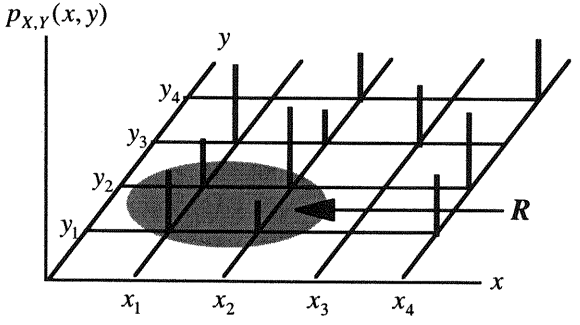


Figure 5-1. A Joint Probability Mass Function of X and Y

If R is any region in the xy-plane and X and Y are *discrete* random variables then

$$P((X,Y) \in R) = \sum_{(x,y)\in R} p_{X,Y}(x,y) \tag{5-3}$$

Equation 5-3 implies the probability of a *random point* falling in a region R is the sum of the heights of the vertical lines that correspond to the points contained in R. The heights of the lines are given by $p_{X,Y}(x,y)$. Joint probabilities are defined in terms of R and the joint probability mass function. For example, the probability X is less than Y is represented by the set of all points in the region where $x < y$. This can be written as

$$P((X,Y) \in \{(x,y): x < y\}) = \sum_{(x,y): x<y} p_{X,Y}(x,y) \tag{5-4}$$

If X and Y have a finite number of possible values, it is sometimes convenient to arrange the probabilities associated with these values in a *contingency table.* Table 5-1 illustrates a contingency table for two random variables that each have four values possible.

Table 5-1. A Contingency Table for X and Y

(X,Y)	y_1	y_2	y_3	y_4
x_1	$p_{X,Y}(x_1,y_1)$	$p_{X,Y}(x_1,y_2)$	$p_{X,Y}(x_1,y_3)$	$p_{X,Y}(x_1,y_4)$
x_2	$p_{X,Y}(x_2,y_1)$	$p_{X,Y}(x_2,y_2)$	$p_{X,Y}(x_2,y_3)$	$p_{X,Y}(x_2,y_4)$
x_3	$p_{X,Y}(x_3,y_1)$	$p_{X,Y}(x_3,y_2)$	$p_{X,Y}(x_3,y_3)$	$p_{X,Y}(x_3,y_4)$
x_4	$p_{X,Y}(x_4,y_1)$	$p_{X,Y}(x_4,y_2)$	$p_{X,Y}(x_4,y_3)$	$p_{X,Y}(x_4,y_4)$

The sum of all $p_{X,Y}(x_i,y_k)$ in a contingency table must equal unity. If X and Y are discrete random variables their *marginal probability mass functions* are given by

$$p_X(x) = P(X = x) = \sum_y p_{X,Y}(x,y) \tag{5-5}$$

$$p_Y(y) = P(Y = y) = \sum_x p_{X,Y}(x, y) \tag{5-6}$$

Equation 5-5 is the marginal probability mass function of X; equation 5-6 is the marginal probability mass function of Y.

Example 5-1 Suppose the effort (in staff-months) to modernize a management information system is given by $Eff_{SysEng} = XY$, where X is the number of systems engineering staff needed for Y months. Suppose a contingency table for X and Y is given below.

Table 5-2. Contingency Table for Example 5-1

	y_1 = 24 months	y_2 = 36 months	Total
x_1 = 15 staff	0.15	0.25	0.40
x_2 = 25 staff	0.20	0.40	0.60
Total	0.35	0.65	1.00

Compute

a) $P(X = 15, Y = 36)$ b) $P(X = 15)$

c) $P(Y = 36)$ d) $P(Eff_{SysEng} < 600)$

Solution

a) From equation 5-2

$$P(X = 15, Y = 36) = p_{X,Y}(15, 36) = 0.25$$

b) $P(X = 15)$ is a marginal probability; from equation 5-5

$$P(X = 15) = \sum_{k=1}^{2} p_{X,Y}(15, y_k) = 0.15 + 0.25 = 0.40$$

c) $P(Y = 36)$ is a marginal probability; from equation 5-6

$$P(Y = 36) = \sum_{t=1}^{2} p_{X,Y}(x_t, 36) = 0.25 + 0.40 = 0.65$$

d) From table 5-2 the region R where the event $\{Eff_{SysEng} < 600\}$ occurs contains only two points; specifically,

$$R = \{(x,y): xy < 600\} = \{(x_1,y_1),(x_1,y_2)\}$$

where $(x_1,y_1) = (15,24)$ and $(x_1,y_2) = (15,36)$. Referring to equation 5-3

$$P(Eff_{SysEng} < 600) = P(XY < 600)$$

$$= P((X,Y) \in \{(x,y): xy < 600\}) = \sum_{(x,y):\, xy<600} p_{X,Y}(x,y)$$

$$= p_{X,Y}(x_1,y_1) + p_{X,Y}(x_1,y_2) = 0.15 + 0.25 = 0.40 \blacklozenge$$

Continuous Random Variables

If X and Y are *continuous* random variables, the joint probability density function of X and Y, denoted by $f(x,y)$, satisfies for any set R in the two-dimensional plane

$$P((X,Y) \in R) = \iint_{(x,y)\in R} f(x,y)dxdy \qquad (5\text{-}7)$$

where $f(x,y) \geq 0$ and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$$

The probability associated with a univariate continuous random variable reflects an area under the variable's density function. The probability represented by the double integral in equation 5-7, is the *volume* over the region R between the xy-plane and the surface determined by $f(x,y)$. In particular,

$$P(a \leq X \leq b \text{ and } c \leq Y \leq d) = \int_a^b\int_c^d f(x,y)dydx \qquad (5\text{-}8)$$

With n continuous random variables, $X_1, X_2, X_3, \ldots, X_n$, we have

$$P(a_1 \le X_1 \le b_1 \cdots a_n \le X_n \le b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \cdots x_n) dx_n \cdots dx_1 \qquad (5\text{-}9)$$

The *marginal probability density functions* of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \text{ for } -\infty < x < \infty \qquad (5\text{-}10)$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx \text{ for } -\infty < y < \infty \qquad (5\text{-}11)$$

Example 5-2 Suppose the effort (in staff-months) to develop and implement a system's test plans and procedures is given by $Eff_{SysTest} = XY$, where X is the number of test engineering staff needed over Y months. Suppose X and Y are *continuous* random variables with joint PDF

$$f(x,y) = \begin{cases} \dfrac{1}{240} & 5 \le x \le 15, \quad 12 \le y \le 36 \\ 0 \quad \text{otherwise} \end{cases}$$

This joint PDF has marginal probability density functions in figure 5-2.

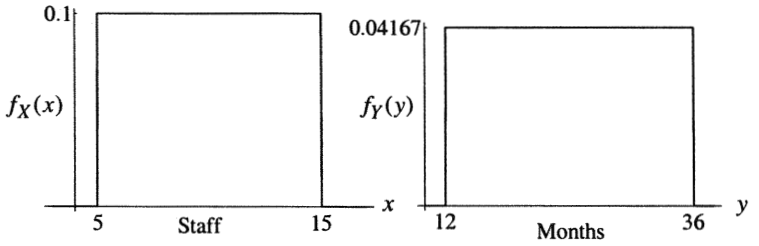


Figure 5-2. Marginal Distributions for X and Y

Determine

a) $P(Eff_{SysTest} \le 120)$ b) $P(Eff_{SysTest} \le 360)$

c) $P(Eff_{SysTest} \le 120)$ *given* the test engineering staff will not exceed 10 persons.

d) The probability $Eff_{SysTest} > 120$ staff-months and the test engineering staff and duration will not exceed 10 persons and 24 months, respectively.

Solution a) To determine the probability $Eff_{SysTest} \leq 120$, we first sketch the event space. This is shown in figure 5-3.

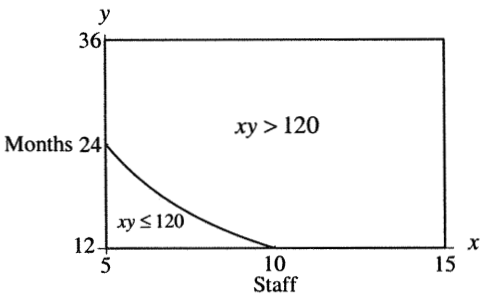


Figure 5-3. Event Space for $Eff_{SysTest} \leq 120$

From equation 5-8, we have

$$P(Eff_{SysTest} \leq 120) = \iint_{xy \leq 120} f(x,y)dxdy$$

$$= \int_{12}^{24} \int_{5}^{\frac{120}{y}} \frac{1}{240} dxdy = \int_{5}^{10} \int_{12}^{\frac{120}{x}} \frac{1}{240} dydx = 0.09657$$

b) To determine the probability $Eff_{SysTest} \leq 360$, we first sketch the event space. This is shown in figure 5-4.

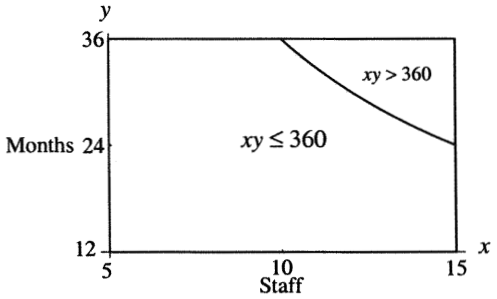


Figure 5-4. Event Space for $Eff_{SysTest} \leq 360$

From theorem 2-1

$$P(Eff_{SysTest} \le 360) = 1 - P(Eff_{SysTest} > 360)$$

It follows that

$$P(Eff_{SysTest} \le 360) = 1 - \iint\limits_{xy>360} f(x,y)dxdy$$

$$= 1 - \int_{24}^{36}\int_{\frac{360}{y}}^{15} \frac{1}{240}dxdy = 1 - \int_{10}^{15}\int_{\frac{360}{x}}^{36} \frac{1}{240}dydx = 0.858$$

c) The probability $Eff_{SysTest} \le 120$ staff-months *given* the test engineering staff-level will not exceed 10 persons is a conditional probability; specifically, the conditional probability is $P(Eff_{SysTest} \le 120 \mid X \le 10)$. From chapter 2, equation 2-2, we can write

$$P(Eff_{SysTest} \le 120 \mid X \le 10) = \frac{P(\{XY \le 120\} \cap \{X \le 10\})}{P(\{X \le 10\})}$$

In this case,

$$\frac{P(\{XY \le 120\} \cap \{X \le 10\})}{P(\{X \le 10\})} = \frac{\int_{5}^{10}\int_{12}^{\frac{120}{x}} \frac{1}{240}\,dydx}{\int_{5}^{10} \frac{1}{(15-5)}dx} = 2\int_{5}^{10}\int_{12}^{\frac{120}{x}} \frac{1}{240}\,dydx = 2(0.09657) = 0.193$$

The conditional probability, in this example, is twice its unconditional probability computed in part a). Why is this? The unconditional probability is associated with the joint distribution function

$$f(x,y) = (1/240),\ 5 \le x \le 15,\ 12 \le y \le 36$$

If it is given that $X \le 10$, the joint distribution function essentially becomes

$$f(x,y) = (1/120),\ 5 \le x \le 10,\ 12 \le y \le 36$$

With $f(x,y)=(1/120)$, and $5\le x\le 10$, $12\le y\le 36$, more probability exists in the region where $XY\le 120$ than in the same region with $f(x,y)=(1/240)$, and $5\le x\le 15$, $12\le y\le 36$.

d) To determine the probability $Eff_{SysTest}>120$ staff-months and the test engineering staff and duration will not exceed 10 persons and 24 months, define three events A, B, and C as

$$A=\left\{Eff_{SysTest}>120\right\}=\{XY>120\}$$

$$B=\{X\le 10\}$$

$$C=\{Y\le 24\}$$

Thus, the probability we want to determine is given by

$$P(A\cap B\cap C)=P\left(\{XY>120\}\cap\{X\le 10\}\cap\{Y\le 24\}\right)$$

$$=P\left(\left\{\frac{120}{Y}<X\right\}\cap\{X\le 10\}\cap\{Y\le 24\}\right)$$

$$=P\left(\left\{\frac{120}{Y}<X\le 10\right\}\cap\{Y\le 24\}\right)$$

From equation 5-8

$$P\left(\left\{\frac{120}{Y}<X\le 10\right\}\cap\{Y\le 24\}\right)=\int_{12}^{24}\int_{\frac{120}{y}}^{10}\frac{1}{240}\,dx\,dy=0.1534$$

The probability is just over 0.15 that the effort for system test will exceed 120 staff-months, and the test engineering staff-level and duration will not exceed 10 persons and 24 months. This probability is shown by the region R in figure 5-5.

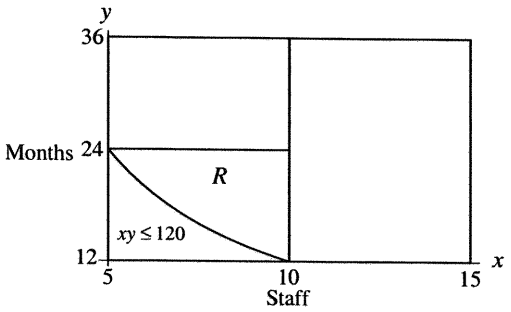


Figure 5-5. Region R Associated With Part C of Example 5-2

Example 5-3 Suppose the effort (staff-months) to develop a new software application is given by $Eff_{SW} = \frac{X}{Y}$, where X is the size of a software application (number of DSI) and Y is the development productivity rate (number of DSI per staff-month). Suppose X and Y are *continuous* random variables with joint PDF

$$f(x,y) = \begin{cases} \dfrac{1}{5(10^6)} & 50{,}000 \le x \le 100{,}000, \quad 100 \le y \le 200 \\ 0 \quad \text{otherwise} \end{cases}$$

This joint PDF has marginal probability density functions in figure 5-6.

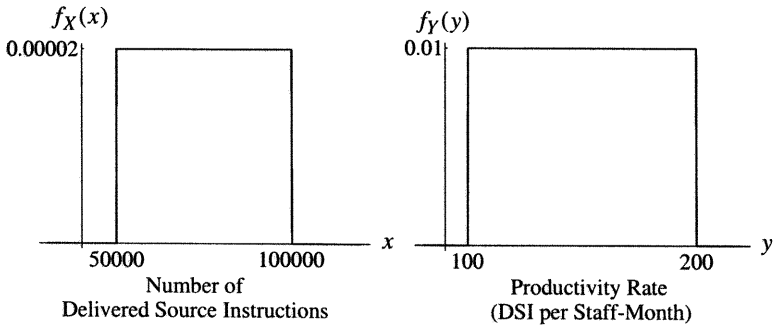


Figure 5-6. Marginal Distributions for X and Y

Determine the probability Eff_{SW} will not exceed 300 staff-months.

Solution To determine the probability Eff_{SW} will not exceed 300 staff-months, we first sketch the event space. This is shown in figure 5-7.

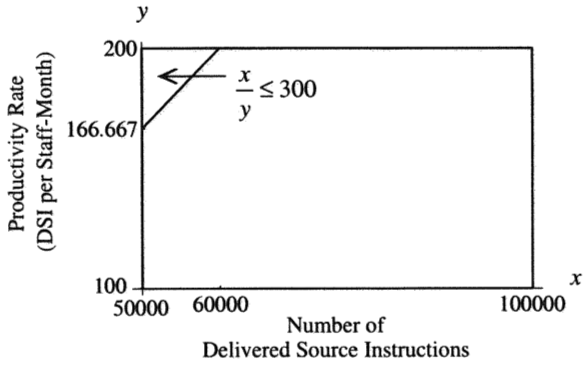


Figure 5-7. Event Space for $Eff_{SW} \le 300$

From equation 5-8, we have

$$P(Eff_{SW} \le 300) = \iint\limits_{\frac{x}{y} \le 300} f(x,y)dxdy$$

$$= \int_{166.667}^{200} \int_{50,000}^{300y} \frac{1}{5(10^6)} dxdy$$

$$= \int_{50,000}^{60,000} \int_{\frac{x}{300}}^{200} \frac{1}{5(10^6)} dydx = 0.0333 \blacklozenge$$

So far, we have introduced the concept of joint probability distributions for two random variables. Often, it is necessary to know the distribution of one random variable when the other takes a specific value. Such a distribution is known as a conditional probability distribution, which is discussed next in terms of discrete and continuous random variables.

Conditional Probability Mass Function

If two *discrete* random variables X and Y have joint probability mass function $p_{X,Y}(x,y)$, the *conditional probability mass function* of X given $Y = y$ is

$$p_{X|Y=y}(x) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \tag{5-12}$$

where $p_Y(y) > 0$. Similarly, the *conditional probability mass function* of Y given $X = x$ is

$$p_{Y|X=x}(y) = \frac{p_{X,Y}(x,y)}{p_X(x)} \tag{5-13}$$

where $p_X(x) > 0$. To illustrate this concept return to example 5-1; suppose we want the probability the number of systems engineering staff X will be 15 persons, *given* they are needed for 36 months. In this case we want $p_{X|Y=36}(15)$. From equation 5-12 and table 5-2 this is

$$p_{X|Y=36}(15) = \frac{p_{X,Y}(15,36)}{p_Y(36)} = \frac{0.25}{0.65} = \frac{5}{13} \approx 0.3846$$

This probability is conditioned on a fixed (or observed value) for Y. It has a value slightly less than the unconditioned probability $P(X = 15)$, which was shown in example 5-1 to be 0.40.

Conditional Probability Density Function

If two *continuous* random variables X and Y have joint density function $f(x,y)$, the *conditional probability density function* of X, given $Y = y$, is

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} \quad f_Y(y) > 0 \tag{5-14}$$

Similarly, the *conditional probability density function* of Y, given $X = x$, is

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} \quad f_X(x) > 0 \tag{5-15}$$

Example 5-4 In example 5-2, X and Y had joint probability density function

$$f(x,y) = \begin{cases} \dfrac{1}{240} & 5 \le x \le 15, \quad 12 \le y \le 36 \\ 0 & \text{otherwise} \end{cases}$$

Find the conditional probability density functions of X and Y.

Solution

From equation 5-14, the conditional probability density function of X is

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{\frac{1}{240}}{\frac{1}{24}} = \frac{1}{10} \quad 5 \le x \le 15$$

From equation 5-15, the conditional probability density function of Y is

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{\frac{1}{240}}{\frac{1}{10}} = \frac{1}{24} \quad 12 \le y \le 36 \blacklozenge$$

Conditional probability density functions provide a way to determine the conditional cumulative distribution function. Specifically,

$$F_{X|Y}(x = a|y) \equiv P(X \le a|Y = y) = \int_{-\infty}^{a} f_{X|Y}(x|y)\,dx \tag{5-16}$$

$$F_{Y|X}(y = b|x) \equiv P(Y \le b|X = x) = \int_{-\infty}^{b} f_{Y|X}(y|x)\,dy \tag{5-17}$$

5.1.2 Independent Random Variables

Two random variables X and Y are *independent* if for any two events $\{X \in A\}$ and $\{Y \in B\}$, where A and B are sets of real numbers, we have

$$P(\{X \in A\} \cap \{Y \in B\}) = P(\{X \in A\})P(\{Y \in B\}) \tag{5-18}$$

Equation 5-18 follows if and only if, for any x and y

$$P(\{X \le x\} \cap \{Y \le y\}) \equiv P(X \le x, Y \le y) = P(X \le x)P(Y \le y) \tag{5-19}$$

From equation 5-19, it follows that

$$F_{X,Y}(x,y) \equiv P(X \le x, Y \le y) = F_X(x)F_Y(y) \quad -\infty < x,y < \infty \tag{5-20}$$

If X and Y are independent *discrete* random variables, equation 5-18 becomes

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \tag{5-21}$$

It follows that

$$p_{X|Y=y}(x) = p_X(x) \tag{5-22}$$

$$p_{Y|X=x}(y) = p_Y(y) \tag{5-23}$$

Moreover, if equation 5-21 holds for two discrete random variables, then the random variables are independent. Similarly, X and Y are independent *continuous* random variables if and only if equation 5-24 holds for all feasible values of X and Y.

$$f(x,y) = f_X(x)f_Y(y) \tag{5-24}$$

It follows that

$$f_{X|Y}(x|y) = f_X(x) \tag{5-25}$$

$$f_{Y|X}(y|x) = f_Y(y) \tag{5-26}$$

From this, what do you conclude about the random variables X and Y in examples 5-2 and 5-3? A discussion of this is left as an exercise for the reader. *Dependent* random variables are those that are *not independent*.

5.1.3 Expectation and Correlation of Random Variables

In chapter 3, the expectation of a random variable was discussed. The expectation of two random variables is stated in the following proposition.

Proposition 5-1 If X and Y are random variables and $g(x,y)$ is a real-valued function defined for all x and y that are possible values of X and Y, then

$$E(g(X,Y)) = \begin{cases} \sum_x \sum_y g(x,y) \cdot p_{X,Y}(x,y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) \cdot f(x,y) dxdy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases} \quad (5\text{-}27)$$

In the above, the double summation and double integral must be absolutely convergent.

Example 5-5 Determine the expectation of $Eff_{SysTest}$ in example 5-2.

Solution We need to compute $E(Eff_{SysTest})$. From example 5-2, the joint distribution of X and Y is given as

$$f(x,y) = \begin{cases} \frac{1}{240} & 5 \le x \le 15, \quad 12 \le y \le 36 \\ 0 & \text{otherwise} \end{cases}$$

In example 5-2 $Eff_{SysTest}$ is a function of two random variables X and Y, that is, $Eff_{SysTest} = XY = g(X,Y)$. Therefore, in this case, $g(x,y) = xy$. Since X and Y are continuous random variables, from equation 5-27

$$E(Eff_{SysTest}) = E(XY) = E(g(X,Y)) = \int_5^{15} \int_{12}^{36} xy \cdot \frac{1}{240} dydx = 240 \text{ staff-months} \blacklozenge$$

It is often of interest to determine where the expected value of a random variable falls along the variable's cumulative distribution function. Mentioned in chapters 3 and 4, the expected value of a random variable *is not*, in general, equal to the median of the random variable. This is again illustrated with example 5-5. It is left as an exercise for the reader to show

$$P(Eff_{SysTest} \le E(Eff_{SysTest})) = P(Eff_{SysTest} \le 240) = 0.56$$

It is often necessary to know the degree to which two random variables associate or vary with each other. In cost analysis, questions such as "*How much is the*

variation in a new satellite's predicted weight attributable to the variation in its cost?" are common. *Covariance* is a measure of how two random variables vary together. Let X and Y be random variables with expected values (means) μ_X and μ_Y, respectively. The covariance of X and Y, denoted by $Cov(X,Y)$, is defined as

$$Cov(X,Y) \equiv \sigma_{XY} = E\{(X-\mu_X)(Y-\mu_Y)\} \tag{5-28}$$

Covariance can be positive, negative, or zero. If X and Y take values simultaneously larger than their respective means, the covariance will be positive. If X and Y take values simultaneously smaller than their respective means, the covariance will also be positive. If one random variable takes a value larger than its mean *and* the other takes a value smaller than its mean, the covariance will be negative. So, when two random variables simultaneously take values on the same sides as their respective means, the covariance will be positive. When two random variables simultaneously take values on opposite sides of their means, the covariance will be negative. The following theorems present useful properties of covariance. Theorem 5-1 presents a way to compute covariance that is easier than using the definition given by equation 5-28.

Theorem 5-1 If X and Y are random variables with means μ_X and μ_Y then

$$Cov(X,Y) = E(XY) - \mu_X\mu_Y$$

Theorem 5-2 If X and Y are random variables, then

a) $Cov(X,Y) = Cov(Y,X)$

b) $Cov(aX+b,cY+d) = acCov(X,Y)$ for any real numbers a, b, c, and d

Theorem 5-3 If X and Y are independent random variables then $Cov(X,Y) = 0$.

The proofs of these theorems are left as exercises for the reader.

Covariance as a measure of the degree two random variables covary can be hard to interpret. Suppose X_1 and Y_1 are random variables such that $X_2 = 2X_1$ and $Y_2 = 2Y_1$. From theorem 5-2 (part b), $Cov(X_2,Y_2) = 4Cov(X_1,Y_1)$. Although

X_1 and Y_1 and X_2 and Y_2 behave in precisely the same way with respect to each other, the random variables X_2 and Y_2 have a covariance four times greater than the covariance of X_1 and Y_1 [2]. A more convenient measure is one where the relationship between pairs of random variables could be interpreted along a common scale. The following discussion presents such a measure.

Suppose we have two standard random variables Z_X and Z_Y, where

$$Z_X = \frac{X - \mu_X}{\sigma_X} \text{ and } Z_Y = \frac{Y - \mu_Y}{\sigma_Y}$$

Using theorem 5-2, the covariance of Z_X and Z_Y reduces to

$$\begin{aligned} Cov(Z_X, Z_Y) &= Cov\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right) \\ &= \frac{1}{\sigma_X}\frac{1}{\sigma_Y} Cov(X - \mu_X, Y - \mu_Y) \\ &= \frac{1}{\sigma_X}\frac{1}{\sigma_Y} Cov(X, Y) \\ &= \rho_{X,Y} \end{aligned}$$

The term $\rho_{X,Y}$ is known as the Pearson correlation coefficient [2]. It is the traditional statistic to measure the degree to which two random variables correlate (or covary). Formally, the *Pearson correlation coefficient* between two random variables X and Y is

$$Corr(X, Y) \equiv \rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{5-29}$$

provided $\sigma_X > 0$ and $\sigma_Y > 0$. From theorem 5-1, equation 5-29 simplifies to

$$Corr(X, Y) \equiv \rho_{X,Y} = \frac{E(XY) - \mu_X \mu_Y}{\sigma_X \sigma_Y} \tag{5-30}$$

The correlation coefficient is dimensionless. Pearson's correlation coefficient measures the *strength of the linear relationship* between two random variables.

It is bounded by the interval $-1 \le \rho_{X,Y} \le 1$. If $Y = aX + b$, where a and b are real numbers and $a > 0$, then $\rho_{X,Y} = 1$; if $a < 0$ then $\rho_{X,Y} = -1$. When $\rho_{X,Y} = 0$, we say X and Y are *uncorrelated*. There is a complete absence of linearity between them. Figure 5-8 illustrates the types of correlation that can exist between random variables.

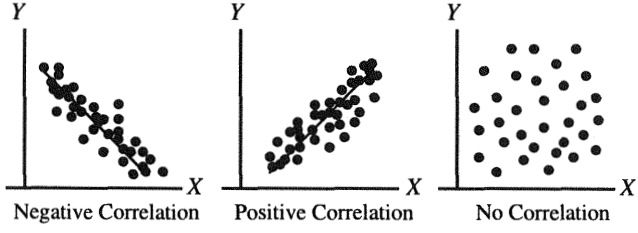


Figure 5-8. Correlation Between Random Variables X and Y

Example 5-6 If $Y = X^2$ and $X \sim Unif(-1,1)$, show that $\rho_{X,Y} = 0$.

Solution From equation 5-30

$$Corr(X,Y) \equiv Corr(X,X^2) \equiv \rho_{X,X^2} = \frac{E(XX^2) - \mu_X \mu_{X^2}}{\sigma_X \sigma_{X^2}}$$

Since $X \sim Unif(-1,1)$, we have $f_X(x) = \frac{1}{2}$ on $-1 \le x \le 1$ (chapter 4); therefore,

$$E(XX^2) = E(X^3) = \int_{-1}^{1} x^3 f_X(x)\,dx = \int_{-1}^{1} x^3 \tfrac{1}{2}\,dx = 0$$

$$\mu_X = E(X) = \int_{-1}^{1} x f_X(x)\,dx = \int_{-1}^{1} x \tfrac{1}{2}\,dx = 0$$

$$\mu_{X^2} = E(X^2) = \int_{-1}^{1} x^2 \tfrac{1}{2}\,dx = \tfrac{1}{3}$$

Therefore

$$Corr(X, X^2) \equiv \rho_{X,X^2} = \frac{0 - 0 \cdot \frac{1}{3}}{\sigma_X \sigma_{X^2}} = 0$$

In this example, we conclude there is a complete absence of linearity between X and Y.

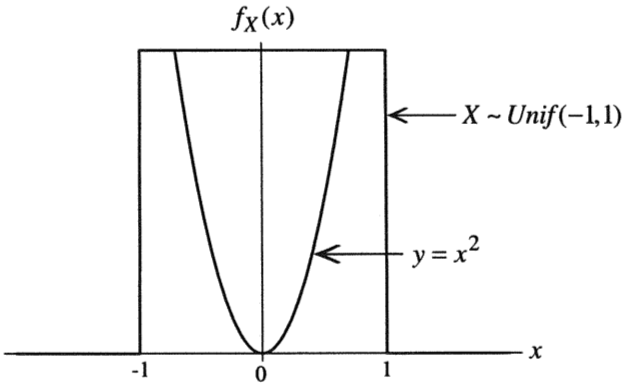


Figure 5-9. An Illustration of $Y = X^2$ and $X \sim Unif(-1,1)$

Theorem 5-4 If X and Y are independent random variables, then $\rho_{X,Y} = 0$.

Proof This follows from theorem 5-3 and equation 5-29. Since X and Y are independent random variables, from theorem 5-3 we have $Cov(X,Y) = 0$. From equation 5-29, if $Cov(X,Y) = 0$ it immediately follows that $\rho_{X,Y} = 0$.♦

The converse of theorem 5-4 is not true. If $\rho_{X,Y} = 0$ then X and Y are said to be *uncorrelated*. However, it *does not* follow that X and Y are independent. Again, if X is uniformly distributed in $-1 \le x \le 1$ and $Y = X^2$, then $\rho_{X,Y} = 0$; however, Y is dependent on X in this case. Theorem 5-4 gives rise to the following:

Theorem 5-5 If X and Y are independent random variables, then

$$E(XY) = E(X)E(Y) \tag{5-31}$$

Proof Since X and Y are independent random variables, from theorem 5-3 we have $Cov(X,Y)=0$. From theorem 5-1, this is equivalent to $E(XY)-\mu_X\mu_Y=0$; thus, $E(XY)=E(X)E(Y)$.

Theorem 5-6 If a_1 and a_2 are either both positive or both negative, and a_1, a_2, b_1, and b_2 are real numbers, then $Corr(a_1X+b_1,a_2Y+b_2)=Corr(X,Y)$.

Proof Let $Z=a_1X+b_1$ and $W=a_2Y+b_2$. We need to show

$$Corr(Z,W)=\frac{E(ZW)-\mu_Z\mu_W}{\sigma_Z\sigma_W}=Corr(X,Y) \tag{5-32}$$

From theorem 3-9 (chapter 3)

$$\begin{aligned}E(ZW)&=E((a_1X+b_1)(a_2Y+b_2))\\&=E(a_1a_2XY+a_1b_2X+a_2b_1Y+b_1b_2)\\&=a_1a_2E(XY)+a_1b_2E(X)+a_2b_1E(Y)+b_1b_2\end{aligned}$$

Also from theorem 3-9

$$\mu_Z\equiv E(Z)=a_1E(X)+b_1 \text{ and } \mu_W\equiv E(W)=a_2E(Y)+b_2$$

Further, from theorem 3-11

$$\sigma_Z^2=a_1^2\sigma_X^2 \text{ and } \sigma_W^2=a_2^2\sigma_Y^2$$

Combining the above

$$E(ZW)-\mu_Z\mu_W=a_1a_2E(XY)-a_1a_2E(X)E(Y)=a_1a_2(E(XY)-E(X)E(Y))$$

and

$$\sigma_Z=|a_1|\sigma_X,\ \sigma_W=|a_2|\sigma_Y$$

Substituting into equation 5-32 yields

$$Corr(Z,W)=\frac{a_1a_2(E(XY)-E(X)E(Y))}{a_1a_2\sigma_X\sigma_Y}=Corr(X,Y)\ \blacklozenge$$

This theorem states that the correlation between two random variables is unaffected by a linear change in either X or Y.

Example 5-7 Suppose X denotes the number of engineering staff required to test a new rocket propulsion system. Suppose X is uniformly distributed in the interval $5 \le x \le 15$. If the number of months Y required to design, conduct, and analyze the test is given by $Y = 2X + 3$, compute the expected test effort, measured in staff-months.

Solution We are given $X \sim Unif(5,15)$ and $Y = 2X + 3$. The test effort, in staff-months, is the product XY. To determine the *expected test effort*, we need to compute $E(XY)$. From equation 5-30, notice $E(XY)$ can be written as

$$E(XY) = \rho_{X,Y}\sigma_X\sigma_Y + \mu_X\mu_Y$$

Since Y is a linear function of X, we have $\rho_{X,Y} = 1$; thus,

$$E(XY) = \sigma_X\sigma_Y + \mu_X\mu_Y$$

Since $X \sim Unif(5,15)$, the mean and variance of X (theorem 4-2) is

$$\mu_X \equiv E(X) = \tfrac{1}{2}(5+15) = 10$$

$$\sigma_X^2 \equiv Var(X) = \tfrac{1}{12}(15-5)^2 = \frac{100}{12} \text{ and } \sigma_X = \frac{10}{\sqrt{12}}$$

Since $Y = 2X + 3$, the mean and variance of Y (theorems 3-9 and 3-11) is

$$\mu_Y \equiv E(Y) = E(2X+3) = 2E(X) + 3 = 2 \cdot 10 + 3 = 23$$

$$\sigma_Y^2 \equiv Var(Y) = Var(2X+3) = 2^2 Var(X) = 4\sigma_X^2 = 4 \cdot \frac{100}{12} = \frac{100}{3} \text{ and } \sigma_Y = \frac{10}{\sqrt{3}}$$

Substituting these values into $E(XY)$ we have

$$E(XY) = \sigma_X\sigma_Y + \mu_X\mu_Y = \frac{10}{\sqrt{12}} \cdot \frac{10}{\sqrt{3}} + 10 \cdot 23 = 246.7 \text{ staff-months}$$

Thus, the expected effort to test the new rocket's propulsion system is nearly 247 staff-months.

Example 5-8 Suppose the effort Eff_{SW} (in staff-months) to develop software is given by $Eff_{SW} = 2.8I^{1.2}$, where I is thousands of delivered source instructions (DSI) to be developed. If $I \sim Unif(20,60)$ determine $\rho_{Eff_{SW},I}$.

Solution

From equation 5-30

$$Corr(Eff_{SW}, I) \equiv \rho_{Eff_{SW},I} = \frac{E(Eff_{SW}I) - \mu_{Eff_{SW}}\mu_I}{\sigma_{Eff_{SW}}\sigma_I} \qquad (5\text{-}33)$$

Computation of $E(Eff_{SW}I)$

$$E(Eff_{SW}I) = E(2.8I^{1.2}I) = E(2.8I^{2.2})$$

From proposition 3-1,

$$E(2.8I^{2.2}) = \int_{20}^{60} 2.8x^{2.2} f_I(x)dx$$

Since $I \sim Unif(20,60)$, the probability density function of I is

$$f_I(x) = \frac{1}{40} \quad 20 \le x \le 60$$

Therefore,

$$E(Eff_{SW}I) = E(2.8I^{2.2}) = \int_{20}^{60} 2.8x^{2.2}\frac{1}{40}dx = 10397.385$$

Computation of $\mu_{Eff_{SW}}$

$$\mu_{Eff_{SW}} \equiv E(Eff_{SW}) = E(2.8I^{1.2}) = \int_{20}^{60} 2.8x^{1.2}\frac{1}{40}dx = 236.6106$$

Computation of $\sigma_{Eff_{SW}}$

$$\sigma_{Eff_{SW}} = \sqrt{Var(Eff_{SW})} = \sqrt{E((Eff_{SW})^2) - (\mu_{Eff_{SW}})^2} = 80.8256$$

It is left for the reader to show $E((Eff_{SW})^2) = 62517.36251$.

Computation of μ_I and σ_I

Since $I \sim Unif(20,60)$ it follows immediately from theorem 4-2 (chapter 4)

$$\mu_I \equiv E(I) = \tfrac{1}{2}(20+60) = 40$$

$$\sigma_I \equiv \sqrt{Var(I)} = \sqrt{\tfrac{1}{12}(60-20)^2} = 11.547$$

Computation of $\rho_{Eff_{SW},I}$

Substituting the above computations into equation 5-33 yields

$$Corr(Eff_{SW}, I) \equiv \rho_{Eff_{SW},I} = \frac{E(Eff_{SW} I) - \mu_{Eff_{SW}} \mu_I}{\sigma_{Eff_{SW}} \sigma_I} = \frac{932.961}{933.293} \approx 0.9996$$

Although the relationship between Eff_{SW} and I is nonlinear, a Pearson correlation coefficient of this magnitude suggests, in this case, the relationship is not distinguishably different from linear.

Rank Correlation

In 1904, statistician C. Spearman developed a correlation coefficient that uses the ranks of values observed for n-pairs of random variables. The coefficient is known as *Spearman's rank correlation coefficient.* Let

$$(U_1, V_1), (U_2, V_2), (U_3, V_3), \ldots, (U_n, V_n)$$

be n-pairs of random samples (a set of independent random variables from the same probability density function) from a continuous bivariate distribution. To determine the rank correlation between the pairs of random variables

$$(U_1, V_1), (U_2, V_2), (U_3, V_3), \ldots, (U_n, V_n)$$

the values of $U_1, U_2, U_3, \ldots, U_n$ and $V_1, V_2, V_3, \ldots, V_n$ are ranked among themselves. For instance, the values of $U_1, U_2, U_3, \ldots, U_n$ would be ranked in increasing order, with the smallest value receiving a rank of one. Likewise, the values of $V_1, V_2, V_3, \ldots, V_n$ would also be ranked in increasing order, with the

smallest value receiving a rank of one. The difference between these rankings is the basis behind Spearman's coefficient. Specifically, Spearman's rank correlation coefficient, denoted by r_s, is given by*

$$r_s = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} d_i^2 \tag{5-34}$$

where d_i is the difference in the ranks between U_i and V_i.

Where Pearson's correlation coefficient determines the degree of linearity between two random variables, Spearman's rank correlation coefficient measures their monotonicity. Like Pearson's correlation coefficient, Spearman's rank correlation coefficient is bounded by the interval $-1 \le r_s \le 1$. If r_s is close to 1, then larger values of U tend to be paired (or associated) with larger values of V. If r_s is close to -1, then larger values of U tend to be paired (or associated) with smaller values of V. An r_s near zero is expected when the ranks reflect a random arrangement.

In example 5-8, recall Eff_{SW} and I have a Pearson correlation of 0.9996 in the interval $20 \le I \le 60$. Mentioned in that example, this suggests the two random variables have a strong linear relationship in the interval indicated. Furthermore, since Eff_{SW} (given in example 5-8) is a strictly monotonically increasing function of I, the rank correlation between Eff_{SW} and I would be unity ($r_s = 1$). However, Pearson's correlation coefficient and Spearman's rank correlation coefficient can be very different. This is seen in figure 5-10. In figure 5-10 we have $Y = X^{100}$ and $X \sim Unif(0,1)$. Pearson's correlation coefficient between X and Y is 0.24 (showing this is left as an exercise for the reader), while their rank correlation is unity. Looking at figure 5-10, why (in this case) are these correlation coefficients so different?

---

* Keeping, E. S. 1962. *Introduction to Statistical Inference*. Princeton, New Jersey: D. Van Nostrand Company, Inc.

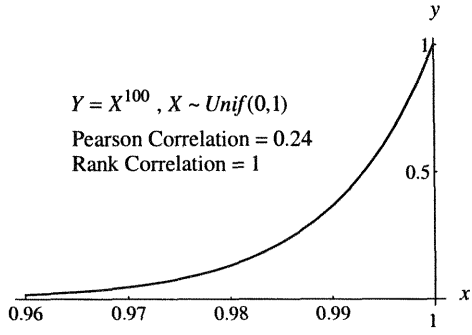


Figure 5-10. Illustrative Correlation Coefficients

Correlation is *not* causation. A strong positive correlation between two random variables does not necessarily imply large values for one *causes* large values for the other. Correlation close to unity *only* means two random variables are strongly associated and the hypothesis of a linear association (for Pearson's correlation coefficient) or a monotonic association (for Spearman's rank correlation coefficient) cannot be rejected.

5.2 Linear Combinations of Random Variables

It is often necessary to work with sums of random variables. Sums of random variables arise frequently in cost analysis. For instance, in figure 1-3 (chapter 1) a system's total cost can be expressed as

$$Cost = X_1 + X_2 + X_3 + \ldots + X_n \tag{5-35}$$

where $X_1, X_2, X_3, \cdots, X_n$ are random variables that represent the cost of the system's work breakdown structure cost elements. From this, we can often think of *Cost* as a linear combination of the random variables $X_1, X_2, X_3, \cdots, X_n$. In general, given a collection of n-random variables $X_1, X_2, X_3, \cdots, X_n$ and constants $a_1, a_2, a_3, \ldots, a_n$ the random variable

$$Y = a_1X_1 + a_2X_2 + a_3X_3 + \ldots + a_nX_n \tag{5-36}$$

is called a *linear combination* of $X_1, X_2, X_3, \cdots, X_n$.

Theorem 5-7 If $Y = a_1X_1 + a_2X_2 + a_3X_3 + \ldots + a_nX_n$ then

$$E(Y) = a_1E(X_1) + a_2E(X_2) + a_3E(X_3) + \ldots + a_nE(X_n) \qquad (5\text{-}37)$$

Theorem 5-7 is an extension of theorem 3-9. It states the expected value of a sum of random variables is the sum of the expected values of the individual random variables. Theorem 5-7 is valid whether or not the random variables $X_1, X_2, X_3, \cdots, X_n$ are independent.

Theorem 5-8 If $Y = a_1X_1 + a_2X_2 + a_3X_3 + \ldots + a_nX_n$ then

$$Var(Y) = \sum_{i=1}^{n} a_i^2 Var(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j \rho_{X_i,X_j}\sigma_{X_i}\sigma_{X_j} \qquad (5\text{-}38)$$

Theorem 5-8 is an extension of theorem 3-11. It states the variance of a sum of random variables is the sum of the variances of the individual random variables, plus the sum of the covariances between them. If the random variables $X_1, X_2, X_3, \cdots, X_n$ are *independent* then

$$Var(Y) = a_1^2 Var(X_1) + a_2^2 Var(X_2) + a_3^2 Var(X_3) + \ldots + a_n^2 Var(X_n) \quad (5\text{-}39)$$

Example 5-9 Suppose the total cost of a system is given by $Cost = X_1 + X_2 + X_3$. Let X_1 denote the cost of the system's prime mission product — PMP.* Let X_2 denote the cost of the system's systems engineering, program management, and system test. Suppose X_1 and X_2 are *dependent* random variables and $X_2 = \frac{1}{2}X_1$. Let X_3 denote the cost of the system's data, spare parts, and support equipment. Suppose X_1 and X_3 are *independent* random variables with distribution functions given in figure 5-11. Compute $E(Cost)$ and $Var(Cost)$.

---

* In systems cost analysis, PMP cost typically refers to the total cost of the system's hardware, software, and hardware-software integration. Chapter 6 provides a detailed discussion.

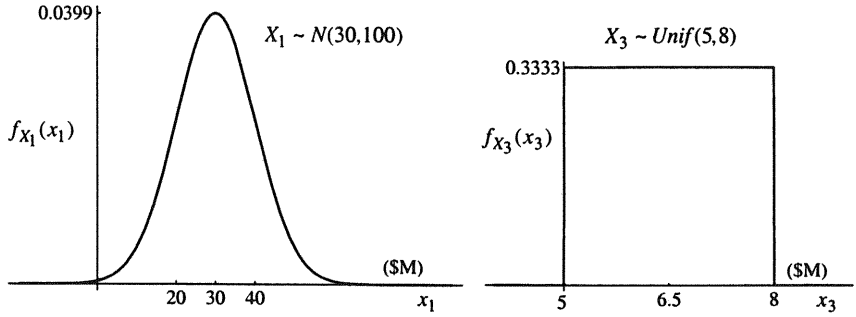


Figure 5-11. Density Functions for Example 5-9

Solution Since $X_1 \sim N(30,100)$, we have from theorem 4-6

$$E(X_1) = 30,\ Var(X_1) = 100,\ \sigma_{X_1} = \sqrt{Var(X_1)} = 10$$

From theorem 3-9 and theorem 3-11 we have

$$E(X_2) = E(\tfrac{1}{2}X_1) = \tfrac{1}{2}E(X_1) = 15$$

$$Var(X_2) = Var(\tfrac{1}{2}X_1) = \tfrac{1}{4}Var(X_1) = 25,\ \sigma_{X_2} = \sqrt{Var(X_2)} = 5$$

Since $X_3 \sim Unif(5,8)$, we have from theorem 4-2

$$E(X_3) = \tfrac{1}{2}(5+8) = 6.5,\ Var(X_3) = \tfrac{1}{12}(8-5)^2 = 0.75,\ \sigma_{X_3} = \sqrt{Var(X_3)} = \sqrt{0.75}$$

Computation of *E*(*Cost*)

From theorem 5-7 (for $i = 1,2,3$)

$$E(Cost) = E(X_1) + E(X_2) + E(X_3) = 30 + 15 + 6.5 = 51.5 \text{ (\$M)}$$

Computation of *Var*(*Cost*)

From theorem 5-8 (for $i = 1,2,3$)

$$Var(Cost) = Var(X_1) + Var(X_2) + Var(X_3)$$
$$+2\left[\rho_{X_1,X_2}\sigma_{X_1}\sigma_{X_2} + \rho_{X_1,X_3}\sigma_{X_1}\sigma_{X_3} + \rho_{X_2,X_3}\sigma_{X_2}\sigma_{X_3}\right]$$

Since X_1 and X_2 are *linearly related*, in this example, we know $\rho_{X_1,X_2} = 1$. Since X_1 and X_3 were given to be independent random variables, from theorem 5-4 we know $\rho_{X_1,X_3} = 0$. From this, and theorem 5-6, it follows that

$$\rho_{X_2,X_3} = \rho_{\frac{1}{2}X_1,X_3} = \rho_{X_1,X_3} = 0$$

Substituting these values into *Var*(*Cost*) we have

$$Var(Cost) = 100 + 25 + 0.75$$
$$+2\left[1(10)(5) + 0(10)(\sqrt{0.75}) + 0(5)(\sqrt{0.75})\right] = 225.75 \ (\$M)^2$$

The units of variance $(\$M)^2$ have little meaning; it is better to think of the range of dollars in terms of the standard deviation; that is,

$$\sigma_{Cost} = \sqrt{Var(Cost)} = 15.02 (\$M)$$

5.2.1 Cost Considerations on Correlation

In example 5-9, X_1 and X_2 were *dependent* random variables. Discussed above, the nature of their dependency was such that $\rho_{X_1,X_2} = 1$. Suppose X_1 and X_2 were *independent* random variables with $X_1 \sim N(30,100)$ and $X_2 \sim N(15,25)$. How would this impact *E*(*Cost*) and *Var*(*Cost*), as computed in example 5-9? The value of *E*(*Cost*) would remain the same. Why? However, if X_1 and X_2 are *independent* random variables then $\rho_{X_1,X_2} = 0$; the value of *Var*(*Cost*) reduces in magnitude; specifically,

$$Var(Cost) = Var(X_1) + Var(X_2) + Var(X_3) = 125.75 \ (\$M)^2$$

In example 5-9, the *dependency* between X_1 and X_2 results in a value for *Var*(*Cost*) nearly 80 percent greater than its value would be if X_1 and X_2 were *independent*. Seen in example 5-9, dependencies between random variables can significantly affect the variance of their sum. Since a system's total cost is essentially a sum of *n*-work breakdown structure cost element costs, that is,

$$Cost = X_1 + X_2 + X_3 + \ldots + X_n$$

it is *critically* important for cost analysts to capture dependencies among $X_1, X_2, X_3, \cdots, X_n$, particularly those with nonnegative correlations. Not doing so can significantly misstate the true variability (uncertainty) in a system's total cost. The following theorem illustrates how nonnegative correlation can affect the variance of a sum of n-random variables. Shown is how the variance increases dramatically with the number of random variables being summed and the extent that ρ approaches unity.

Theorem 5-9 [3] Let $Cost = X_1 + X_2 + X_3 + \ldots + X_n$ where $X_1, X_2, X_3, \cdots, X_n$ are random variables that represent a system's work breakdown structure (WBS) cost element costs. If each pair of $X_1, X_2, X_3, \ldots, X_n$ have common variance σ^2 and common nonnegative correlation ρ, then $Var(Cost) = \sigma^2[n + n(n-1)\rho]$.

Proof From theorem 5-8, we have

$$Var(Cost) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \rho_{X_i, X_j}\sigma_{X_i}\sigma_{X_j}$$

Each pair of $X_1, X_2, X_3, \ldots, X_n$ is given to have common variance σ^2 and common nonnegative correlation ρ; therefore

$$Var(Cost) = \sum_{i=1}^{n} \sigma^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \rho\sigma^2$$

$$= n\sigma^2 + n(n-1)\rho\sigma^2 = \sigma^2[n + n(n-1)\rho] \blacklozenge$$

Some interesting results follow from this theorem; in particular,

$$Var(Cost) = n\sigma^2 \text{ when } \rho = 0$$

$$Var(Cost) = \sigma^2[n + n(n-1)\rho] \text{ when } 0 < \rho < 1$$

$$Var(Cost) = n^2\sigma^2 \text{ when } \rho = 1$$

The following figure illustrates this theorem with $\sigma^2 = 1$.

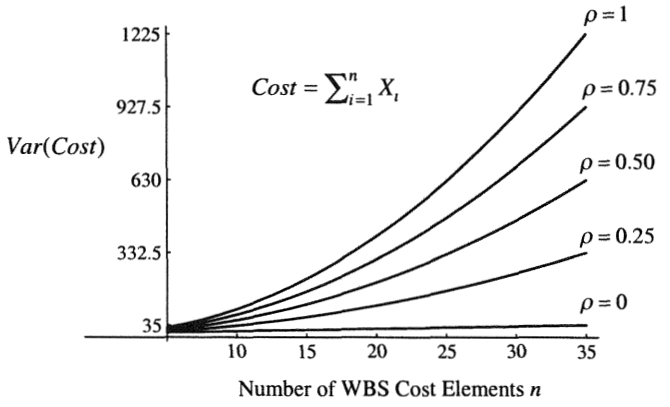


Figure 5-12. Theorem 5-9 with $\sigma^2 = 1$

5.3 The Central Limit Theorem and a Cost Perspective

This section describes one of the most important theorems in probability theory, the *central limit theorem.* The central limit theorem states that, under certain conditions, the distribution function of a sum of independent random variables approaches the normal distribution. From a cost analysis perspective this theorem has great practical importance. Mentioned previously, a system's total cost is a summation of work breakdown structure cost element costs $X_1, X_2, X_3, \cdots, X_n$. Because of this, the distribution function of a system's total cost will often be approximately normal. We will see many examples of this in the discussions and chapters that follow.

Theorem 5-10 *The Central Limit Theorem (CLT)*

Suppose $X_1, X_2, X_3, \cdots, X_n$ is a sequence of n *independent* random variables with $E(X_i) = \mu_i$ and $Var(X_i) = \sigma_i^2$ (each finite). If

$$Y = X_1 + X_2 + X_3 + \cdots + X_n$$

then, under certain conditions,* as $n \to \infty$ the random variable $Z = (Y - \mu)/\sigma$ approaches the standard normal, where

$$\mu = \sum_{i=1}^{n} \mu_i \text{ and } \sigma = \sqrt{\sum_{i=1}^{n} \sigma_i^2} \blacklozenge$$

Theorem 5-10 places no restriction on the types of distribution functions that characterize the random variables $X_1, X_2, X_3, \cdots, X_n$. However, for a given n, the "rate" that the distribution function of Y approaches the normal distribution is affected by the shapes of the distribution functions for $X_1, X_2, X_3, \cdots, X_n$. If these distributions are approximately "bell-shaped," then the distribution function of Y may approach the normal for small n. If they are asymmetric, then n may need to be large for Y to approach the normal distribution.

The central limit theorem is often cited to explain why the distribution function of a system's total cost is often approximately normal. This is illustrated in the following case discussion.

Case Discussion 5-1 The electronic components of a 20 watt solid state amplifier (SSA) for a satellite communication workstation are listed in table 5-3. Let the total component-level cost of the SSA be given by

$$Cost_{SSA} = X_1 + X_2 + X_3 + \ldots + X_{12} \tag{5-40}$$

Suppose $X_1, X_2, X_3, \ldots, X_{12}$ are independent random variables representing the costs of the SSA's components. Suppose the distribution function of each

---

* Informally, the individual random variables $X_1, X_2, X_3, \cdots, X_n$ that constitute Y should make only a small contribution to Y. In addition, none of the random variables $X_1, X_2, X_3, \cdots, X_n$ should dominate in standard deviation. For a further discussion of these conditions, as well as other forms of the central limit theorem, refer to Feller, W. 1968. *An Introduction to Probability Theory and Its Applications*, vol. 2, 3rd ed (revised). New York: John Wiley & Sons, Inc.

component is triangular, with parameters given in table 5-3. Furthermore, suppose the mode of X_i represents its point estimate, that is,

$$x_{iPE_{X_i}} = Mode(X_i) \quad i = 1,2,\ldots,12$$

From this, determine the mean and variance of $Cost_{SSA}$, as well as an approximation to its underlying distribution function.

Table 5-3. 20 Watt SSA Component Cost

Components		Cost ($K)			Mean ($K)	Variance ($K)^2
		Min	Mode	Max		
X_1	Transmitter Synthesizer	12.8	16.9	22.4	17.37	3.87
X_2	Receiver Synthesizer	12.8	16.9	22.4	17.37	3.87
X_3	Reference Generator	15.5	18.3	21.1	18.30	1.31
X_4	Receiver Loopback	7.4	9.2	11.1	9.23	0.57
X_5	BITE Control CCA	6.4	9.1	13.6	9.70	2.21
X_6	Power Supply	17.8	25.1	32.4	25.10	8.88
X_7	IMPATT Modules	36.4	66.5	100.5	67.80	171.41
X_8	Combiner Plate	15.2	18.7	22.7	18.87	2.35
X_9	SHF Upconverter	12.1	16.6	24.6	17.77	6.68
X_{10}	Chassis	21.1	29.6	44.8	31.83	24.03
X_{11}	Backplane	3.3	4.8	6.1	4.73	0.33
X_{12}	Wave Guide Components	4.8	6.7	8.7	6.73	0.63
Component Cost		165.6	238.4	330.4	244.8	226.13

Note: The sum of the modes is not the mode of the distribution function of $Cost_{SSA}$.

Since distribution function of each X_i is given to be triangular, theorem 4-3 can be applied to determine the mean and variance of each component's cost. For instance,

$$E(X_1) = \tfrac{1}{3}(a_1 + m_1 + b_1) = \tfrac{1}{3}(12.8 + 16.9 + 22.4) = 17.37 \ (\$K)$$

$$Var(X_1) = \tfrac{1}{18}\left[(m_1 - a_1)(m_1 - b_1) + (b_1 - a_1)^2\right] = 3.87 \ (\$K)^2$$

where a_1 is the minimum value of X_1, m_1 is the mode of X_1, and b_1 is the maximum value of X_1. Similar notation assumptions and calculations apply

to the other components in table 5-3. From theorem 5-7 and theorem 5-8 the mean and variance of the total component-level cost of the SSA is

$$E(Cost_{SSA}) = \mu_{Cost_{SSA}} = E\left(\sum_{i=1}^{12} X_i\right) = \sum_{i=1}^{12} E(X_i) = 244.8 \text{ (\$K)} \quad (5\text{-}41)$$

$$Var(Cost_{SSA}) = \sigma^2_{Cost_{SSA}} = Var\left(\sum_{i=1}^{12} X_i\right) = \sum_{i=1}^{12} Var(X_i) = 226.13 \text{ (\$K)}^2 \quad (5\text{-}42)$$

Since $X_1, X_2, X_3, \ldots, X_{12}$ are independent random variables (with finite means and variances), from the central limit theorem (theorem 5-10)

$$Z = \frac{Cost_{SSA} - E(Cost_{SSA})}{\sqrt{Var(Cost_{SSA})}} = \frac{Cost_{SSA} - 244.8}{\sqrt{226.13}} \text{ is approximately } N(0,1) \quad (5\text{-}43)$$

This is equivalent to saying

$$Cost_{SSA} \sim N(\mu_{Cost_{SSA}}, \sigma^2_{Cost_{SSA}}) \quad (5\text{-}44)$$

We will next assess the applicability of this theorem that suggests the distribution function for $Cost_{SSA}$ is approximately normal with parameters given by (5-44). Monte Carlo simulation is one way to make this assessment. In the context of case discussion 5-1, the Monte Carlo approach involves taking a random sample from each $X_1, X_2, X_3, \ldots, X_{12}$ and summing these sampled values according to equation 5-40. This produces one random sample for $Cost_{SSA}$. When this sampling process is repeated many thousands of times, an empirical frequency distribution of $Cost_{SSA}$ is produced. From the frequency distribution an empirical cumulative distribution function of $Cost_{SSA}$ can be established. In figure 5-13, the curve implied by the "points" is the empirical cumulative distribution function of $Cost_{SSA}$. The curve given by the *solid line* is an assumed normal distribution, with parameters given by (5-44). Observe how closely the "points" fall along the

solid line. On the basis of this evidence, it appears the central limit theorem is applicable in this case.

The analysis summarized in figure 5-13 provides empirical evidence *only* that the normal distribution is a reasonable form for the distribution function of $Cost_{SSA}$. It might next be asked "*Could the underlying distribution function for* $Cost_{SSA}$ *be normal*?" To answer this, a procedure known as the Kolmogorov-Smirnov (K-S) test can be used. The K-S test [4] applies *only* to continuous distribution functions. It is a formal statistical procedure for testing whether a sample of observations (such as samples generated by a Monte Carlo simulation) could come from a hypothesized theoretical distribution. The following illustrates the K-S test in the context of case discussion 5-1.

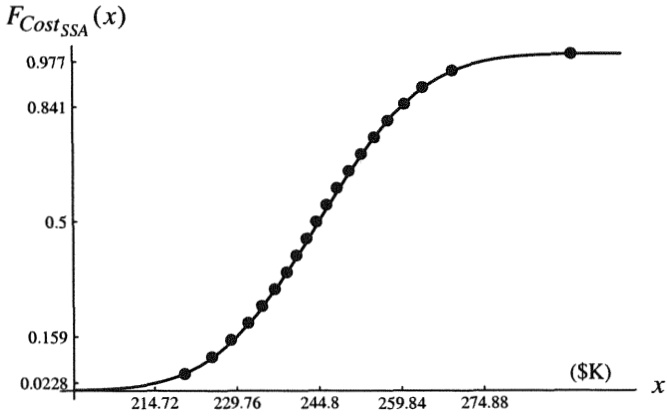


Figure 5-13. Cumulative Distribution Function of $Cost_{SSA}$

The Kolmogorov-Smirnov Test:

- Let $\hat{F}_{Cost_{SSA}}(x)$ represent the observed CDF of $Cost_{SSA}$ (equation 5-40) generated from a Monte Carlo sample of $n = 5000$ observations. This CDF is shown in table 5-4.

- Let $F_{Cost_{SSA}}(x)$ represent a hypothesized CDF. Suppose $F_{Cost_{SSA}}(x)$ is normal with mean 244.8(\$K) and variance 226.13(\$K)^2. Since the hypothesized distribution for $Cost_{SSA}$ is normal, values for $F_{Cost_{SSA}}(x)$ in table 5-4 reflect

$$P(Z \le \frac{x-244.8}{\sqrt{226.13}})$$

The mean and variance of the hypothesized CDF were *not* derived from the observations generated by the Monte Carlo samples.

- Compute the statistic $D = \underset{x}{Max} \left| F_{Cost_{SSA}}(x) - \hat{F}_{Cost_{SSA}}(x) \right|$. From table 5-4 it is seen that $D = 0.0129$.

- Suppose we wish to test the claim that the observed values summarized in table 5-4 come from the hypothesized distribution. Let α be the probability of rejecting the claim when it is actually true. Suppose we let $\alpha = 0.01$.

- Referring to table A-2 (appendix A), if

$$(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}) D > c_{1-\alpha}$$

reject the claim; otherwise, accept it. Since α was chosen to be 0.01 for this test, from table A-2 we have $c_{1-\alpha} = c_{0.99} = 1.628$. With $n = 5000$ and $D = 0.0129$ we have $(70.8322)(0.0129) = 0.9137 < c_{0.99} = 1.628$; thus, we accept the claim. ♦

In a strict sense, accepting the claim that the distribution function for $Cost_{SSA}$ is normal *only* means it is a plausible mathematical model of the underlying distribution. Acceptance does not mean the normal is the "best" or "unique" model form. Other hypothesized distributions might be accepted by the K-S test. It can be shown, in this case, the test also accepts the lognormal distribution as a plausible model of the underlying distribution of $Cost_{SSA}$. Showing this is left as an exercise for the reader.

In cost analysis the "precise" mathematical form of distribution functions, such as those for $Cost_{SSA}$, are rarely known. A credible analysis must provide decision-makers defensible analytical evidence that the form of a distribution

function is mathematically plausible. Looking into whether central limit theorem applies, plotting hypothesized versus simulated distribution functions (e.g., figure 5-13), and conducting statistical tests (i.e., the K-S test) are among the ways such evidence is established.

Table 5-4. Kolmogorov-Smirnov Test for Case Discussion 5-1
(Values in the Left-Most Column are in Dollars Thousand)

x	$\hat{F}_{Cost_{SSA}}(x)$	$F_{Cost_{SSA}}(x)$	$\left\|F_{Cost_{SSA}}(x)-\hat{F}_{Cost_{SSA}}(x)\right\|$
220.19	0.05	0.0509	0.0009
225.15	0.10	0.0957	0.0043
228.64	0.15	0.1413	0.0087
231.80	0.20	0.1937	0.0063
234.35	0.25	0.2436	0.0064
236.72	0.30	0.2955	0.0045
238.89	0.35	0.3472	0.0028
240.66	0.40	0.3915	0.0085
242.50	0.45	0.4392	0.0108
244.34	0.50	0.4878	0.0122
246.21	0.55	0.5374	0.0126
248.11	0.60	0.5871	$D=$ 0.0129
250.28	0.65	0.6422	0.0078
252.53	0.70	0.6964	0.0036
254.99	0.75	0.7510	0.0010
257.49	0.80	0.8006	0.0006
260.49	0.85	0.8516	0.0016
263.80	0.90	0.8968	0.0032
269.22	0.95	0.9478	0.0022

Further Considerations

Mentioned previously, the cost of a system can be expressed as

$$Cost = X_1 + X_2 + X_3 + \ldots + X_n \tag{5-45}$$

where $X_1, X_2, X_3, \ldots, X_n$ are random variables representing the costs of n work breakdown structure elements that constitute the system. From the preceding

case discussion, we saw a circumstance where $F_{Cost}(x)$ could be approximated by a normal distribution. This is sometimes viewed as a paradox. Since a system's cost historically exceeds the value anticipated, or planned, why is its distribution function not positively skewed? The normal distribution is symmetric about its mean; it has no skew.

There are many reasons why the cost of a system exceeds the value anticipated, or planned. A prime reason is a system's cost is often based *only* on its point estimate. From chapter 1 (equation 1-2) the point estimate of the cost of a system is given by

$$x_{PE_{Cost}} = x_1 PE_{X_1} + x_2 PE_{X_2} + x_3 PE_{X_3} + \ldots + x_n PE_{X_n}$$

where $x_{i\,PE_{X_i}}$ are the point estimates of each X_i $(i = 1,\ldots,n)$. Recall $x_{PE_{Cost}}$ is a value for *Cost* that traditionally contains *no reserve dollars* for uncertainties in a system's technical definition or cost estimation approaches. Because of this, $x_{PE_{Cost}}$ often falls below the 50th percentile of *Cost*; that is, $x_{PE_{Cost}}$ can have a high probability of being exceeded. This is illustrated by considering further case discussion 5-1. In this case discussion, $X_1, X_2, X_3, \ldots, X_{12}$ are *independent* random variables representing the costs of the SSA's twelve components. Suppose the point estimates of these components are the modes of X_i $(i = 1,\ldots,12)$, given in table 5-3. The point estimate of the cost of the SSA, denoted by $x_{PE_{Cost_{SSA}}}$ is

$$x_{PE_{Cost_{SSA}}} = Mode(X_1) + Mode(X_2) + Mode(X_3) + \ldots + Mode(X_{12}) \qquad (5\text{-}46)$$

From table 5-3, $x_{PE_{Cost_{SSA}}} = 238.4$ (\$K). Since the distribution function of $Cost_{SSA}$ is approximately normal, in this case, we have

$$P(Cost_{SSA} > x_{PE_{Cost_{SSA}}} = 238.4) = P(Z \leq \frac{238.4 - 244.8}{\sqrt{226.13}}) = 0.665$$

The normal probability density function of $Cost_{SSA}$ is shown in figure 5-14. Notice more probability exists to the right of $x_{PE_{Cost_{SSA}}}$ than to its left. If the cost of the SSA was anticipated, or planned, as the value given by $x_{PE_{Cost_{SSA}}}$, then there *is* a high probability (nearly 67 percent) it will be exceeded. This is true despite the distribution function of $Cost_{SSA}$ being approximately normal.

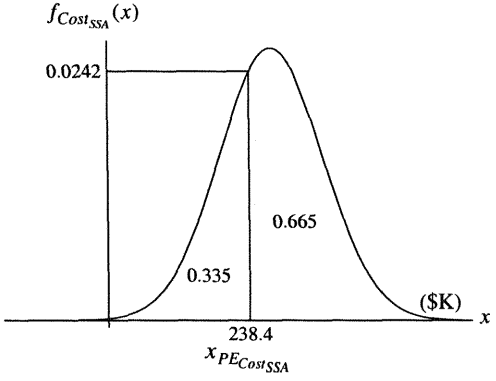


Figure 5-14. Probability Density Function for $Cost_{SSA}$

What drives this probability is the degree to which the distribution functions of each X_i $(i = 1,\ldots,12)$ are skewed. The greater the positive skew, the greater the probability that $x_{PE_{Cost_{SSA}}}$ (defined by equation 5-46) *will* be exceeded. The greater the negative skew, the greater the probability that $x_{PE_{Cost_{SSA}}}$ *will not* be exceeded. In either circumstance, the distribution function of the sum of these X_i's will, because of the central limit theorem, frequently approach a normal. This may seem nonintuitive; nonetheless, the sum of many random variables characterized by skewed distributions *can* result in a distribution function that has no skew at all.

Lastly, since $Cost_{SSA}$ is considered to have a normal distribution, in this case, the mode of $Cost_{SSA}$ is equal to its mean — 244.8 ($K). The sum of the modes of each X_i ($i = 1,\ldots,12$), seen in table 5-3, is 238.4 ($K). In general, the sum of the modes of n-random variables *will not equal* the mode of the distribution function of the sum of these variables; that is, if $Cost = X_1 + X_2 + X_3 + \ldots + X_n$ and

$$x_{PE_{Cost}} = Mode(X_1) + Mode(X_2) + Mode(X_3) + \ldots + Mode(X_n)$$

then $x_{PE_{Cost}} \neq Mode(Cost)$. If the distribution function of each X_i is normal, then $x_{PE_{Cost}} = Mode(Cost)$; in general, if the distribution function of each X_i is normal then $x_{PE_{Cost}} = Mode(Cost) = E(Cost) = Med(Cost)$.

5.4 Transformations of Random Variables

It is often necessary to determine the distribution function of a random variable that is a function (or transformation) of one or more random variables. For instance, the direct engineering hours to design a communication satellite may be a function of the satellite's weight W (pounds). Such a function might be given by equation 5-47.

$$Hours = 4 + 2\sqrt{W} \tag{5-47}$$

If W is a random variable then *Hours* is a function (or transformation) of the random variable W. In software cost analysis, the effort Eff_{SW} (staff-months) to develop software can be a function of the number of source instructions to develop. A general form of this function is

$$Eff_{SW} = c_1 I^{c_2} \tag{5-48}$$

where c_1 and c_2 are positive constants and I is the number of thousands of delivered source instructions (KSDI) to be developed.* If I is a random

* Section 5.4.2 presents a detailed discussion of the function given by equation 5-48.

variable then Eff_{SW} is a function of the random variable I. A question that might be asked is "*What is the 50th percentile of* Eff_{SW} *if the uncertainty in the number of source instructions to develop is characterized by a uniform distribution in the interval* $30\,\text{KDSI} \le x \le 80\,\text{KDSI}$?" To answer this question we need the distribution function of Eff_{SW} given the distribution function for I. In the preceding section we discussed a possible distribution function for the random variable *Cost*, where

$$Cost = X_1 + X_2 + X_3 + \ldots + X_n \tag{5-49}$$

and the X_i's ($i = 1,\ldots,n$) were random variables representing the costs of n work breakdown structure cost elements that constitute a system. In equation 5-49, *Cost* is a function of n random variables. From the central limit theorem, we saw the distribution function of *Cost* can, under certain conditions, be approximately normal. What if the central limit theorem does not apply? How is the distribution function determined for a random variable that is a function of other random variables? The following presents methods to address this question.

5.4.1 Functions of a Single Random Variable

This section presents how to determine the distribution function of a random variable that is a function of another random variable. This is presented in the context of continuous random variables.* Consider the following example.

Example 5-10 Suppose the direct engineering hours to design a new communication satellite is given by

$$Hours = 4 + 2\sqrt{W} \tag{5-50}$$

---

* Refer to case discussion 3-1 (chapter 3) for a view of this discussion from the perspective of discrete random variables.

where W is the satellite's weight, in pounds. Suppose the uncertainty in the satellite's weight is captured by a uniform distribution whose range of possible values is given by $1000 \le w \le 2000$. Suppose the satellite design team assessed 1500 pounds to be the point estimate for weight; that is, $w_{PE} = 1500$.*

a) Determine the cumulative distribution function of *Hours*.

b) Compute $P(Hours \le h_{PE})$, where $h_{PE} = 4 + 2\sqrt{w_{PE}}$.

c) Determine the probability density function of *Hours*.

Solution a) We are given $W \sim Unif(1000, 2000)$. From equation 4-4

$$f_W(w) = \frac{1}{1000} \quad 1000 \le w \le 2000$$

The cumulative distribution function of *Hours* is $F_{Hours}(h) = P(Hours \le h)$, where h denotes the possible values of *Hours*. Since $Hours = 4 + 2\sqrt{W}$, the interval $1000 \le w \le 2000$ is mapped onto the interval $67.2456 \le h \le 93.4427$. Thus, for h in the interval $67.2456 \le h \le 93.4427$

$$F_{Hours}(h) = P(Hours \le h) = P(4 + 2\sqrt{W} \le h) = P(W \le \left(\frac{h-4}{2}\right)^2)$$

$$= \int_{1000}^{[(h-4)/2]^2} f_W(w)dw = \frac{1}{1000}\left(\frac{h-4}{2}\right)^2 - 1$$

Thus, the CDF of *Hours*, presented in figure 5-15, is

$$F_{Hours}(h) = P(Hours \le h) = \begin{cases} 0 & h < 67.2456 \\ \frac{1}{1000}\left(\frac{h-4}{2}\right)^2 - 1 & 67.2456 \le h \le 93.4427 \\ 1 & h > 93.4427 \end{cases} \qquad (5\text{-}51)$$

---

* Instead of using w_{PEW} to denote the point estimate of the random variable W, we simplify the notation and let w_{PE} represent this value.

b) From equation 5-50 we have $h_{PE} = 4 + 2\sqrt{w_{PE}}$; thus, $h_{PE} = 81.46$ when $w_{PE} = 1500$. Therefore, $P(Hours \le h_{PE}) = P(Hours \le 81.46)$. From equation 5-51 this probability is

$$P(Hours \le h_{PE}) = P(Hours \le 81.46) = \frac{1}{1000}\left(\frac{81.46-4}{2}\right)^2 - 1 = 0.50$$

c) To compute the probability density function of *Hours*, we can differentiate $F_{Hours}(h)$ with respect to h. From chapter 3, recall that

$$f_{Hours}(h) = \frac{d}{dh}\left(F_{Hours}(h)\right)$$

It follows that $f_{Hours}(h) = \frac{1}{2000}(h-4) \quad 67.2456 \le h \le 93.4427$ ♦

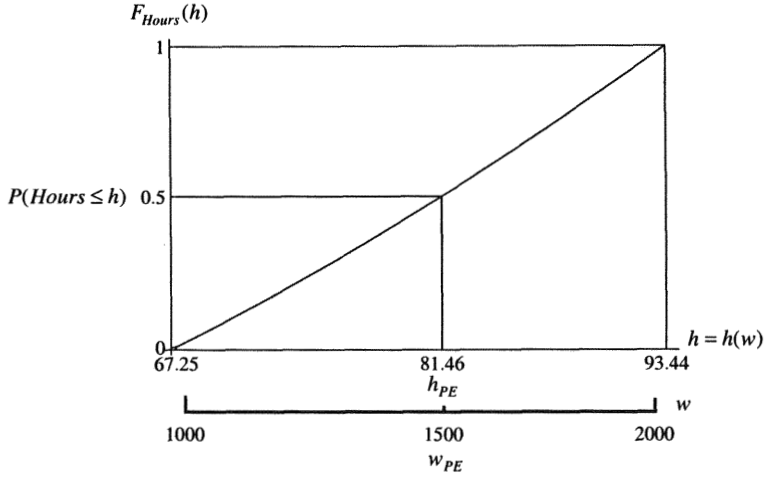


Figure 5-15. The Cumulative Distribution Function of *Hours*

In example 5-10, the procedures to develop $F_{Hours}(h)$ and $f_{Hours}(h)$ are generalized by the following theorem.

Theorem 5-11 Suppose X is a continuous random variable with probability density function $f_X(x) > 0$ for $a \le x \le b$. Consider the random variable $Y = g(X)$ where $y = g(x)$ is a strictly increasing or decreasing differentiable

function of x. Let the inverse of $y = g(x)$ be given by $x = v(y)$, then $Y = g(X)$ has probability density function

$$f_Y(y) = \begin{cases} f_X(v(y)) \cdot \left| \dfrac{d[v(y)]}{dy} \right| & g(a) \le y \le g(b) \quad \text{if } g(x) \text{ increasing} \\ f_X(v(y)) \cdot \left| \dfrac{d[v(y)]}{dy} \right| & g(b) \le y \le g(a) \quad \text{if } g(x) \text{ decreasing} \end{cases} \tag{5-52}$$

If $y = g(x)$ is strictly increasing

$$F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X \le v(y)) = F_X(v(y)) = F_X(x) \tag{5-53}$$

If $y = g(x)$ is strictly decreasing

$$F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X > v(y)) = 1 - F_X(v(y)) = 1 - F_X(x) \tag{5-54}$$

Discussion of Theorem 5-11

Applying theorem 5-11 to example 5-10 yields the following:

$$f_{Hours}(h) = f_W(v(h)) \cdot \left| \frac{d[v(h)]}{dh} \right| \quad g(1000) \le h \le g(2000) \tag{5-55}$$

where $h = g(w) = 4 + 2\sqrt{w}$ and $w = v(h) = \left(\dfrac{h-4}{2} \right)^2$. Since

$$f_W(w) = \frac{1}{1000} \quad 1000 \le w \le 2000$$

we have $f_W(v(h)) = f_W\left(\left(\dfrac{h-4}{2} \right)^2 \right) = \dfrac{1}{1000}$ and $\left| \dfrac{d[v(h)]}{dh} \right| = \dfrac{h-4}{2}$. Substituting into equation 5-55 yields

$$f_{Hours}(h) = \frac{1}{1000} \cdot \frac{h-4}{2} \quad 67.2456 \le h \le 93.4427 \tag{5-56}$$

which is the same as the PDF in part c) of example 5-10.

Theorem 5-11 also provides insight into the fractiles of a distribution function. In example 5-10, $h = g(w) = 4 + 2\sqrt{w}$ is a *strictly increasing* differentiable function of w. From theorem 5-11, this implies

$$F_{Hours}(h) = F_W(w)$$

Thus, the value of h associated with the α-fractile of W will also be the α-fractile of *Hours*. For example, in figure 5-15 observe that

$$F_W(1500) = 0.50 = F_{Hours}(81.46)$$

Here, the value of h associated with the 0.50-fractile of W is the 0.50-fractile of *Hours*. Specifically,

$$w_{0.50} = 1500 \text{ and } P(W \le w_{0.50}) = 0.50$$

$$h_{0.50} = 81.46 = 4 + 2\sqrt{w_{0.50}} \text{ and } P(Hours \le h_{0.50}) = 0.50$$

Similarly, it can be shown (left as an exercise for the reader) that

$$F_W(1750) = 0.75 = F_{Hours}(87.67)$$

The practical value of this aspect of theorem 5-11 is high, because cost-related equations (e.g., equation 5-50) are often simple increasing or decreasing differentiable functions of one variable. When $Y = g(X)$ and theorem 5-11 applies, the cumulative distribution function of Y is *not needed* to determine its fractiles. The α-fractiles of Y are, in fact, completely determined from the α-fractiles of X. In practice, not having to determine the cumulative distribution function of Y, either analytically or through Monte Carlo simulation, can save a great deal of mathematical effort. When possible, cost analysts should readily take advantage of this aspect of theorem 5-11.

Example 5-11 From the information in example 5-10 compute

a) $E(Hours)$ b) σ_{Hours}

Solution a) Two approaches are shown.

Approach 1

From equation 3-21, we can write

$$E(Hours) = \int_{67.2456}^{93.4427} h \cdot f_{Hours}(h)dh = \int_{67.2456}^{93.4427} h \cdot \frac{1}{2000}(h-4)dh = 81.09 \text{ hours}$$

Approach 2

Since $Hours = g(W) = 4 + 2\sqrt{W}$, it follows from proposition 3-1

$$E(Hours) = E(g(W)) = \int_{1000}^{2000} g(w) \cdot f_W(w)dw = \int_{1000}^{2000} (4 + 2\sqrt{w}) \cdot \frac{1}{1000} dw = 81.09 \text{ hours}$$

b) To determine σ_{Hours}, from theorem 3-10 we have

$$Var(Hours) = E(Hours^2) - [E(Hours)]^2$$

Since

$$E(Hours^2) = \int_{1000}^{2000} [g(w)]^2 \cdot f_W(w)dw$$

$$= \int_{1000}^{2000} [(4 + 2\sqrt{w})]^2 \cdot \frac{1}{1000} dw = 6632.75 \text{ (hours)}^2$$

we have

$$Var(Hours) = E(Hours^2) - [E(Hours)]^2 = 6632.75 - (81.09)^2 = 57.1619 \text{ (hours)}^2$$

therefore

$$\sigma_{Hours} = \sqrt{Var(Hours)} = 7.56 \text{ hours}$$

The reader should also verify that $E(Hours^2)$ can be computed by

$$E(Hours^2) = \int_{67.2456}^{93.4427} h^2 \cdot f_{Hours}(h)dh = \int_{67.2456}^{93.4427} h^2 \cdot \frac{1}{2000}(h - 4)dh$$

5.4.2 Applications to Software Cost-Schedule Models

This section presents a further discussion on functions of a single random variable as they apply to software cost-schedule models. These models are often used in cost analysis to determine the effort (staff-months), cost (dollars), and schedule (months) of a software development project. The general forms of these models are given below.

$$Eff_{SW} = c_1 I^{c_2} \tag{5-57}$$

$$Cost_{SW} = \ell_r Eff_{SW} \tag{5-58}$$

$$T_{SW} = k_1 (Eff_{SW})^{k_2} \tag{5-59}$$

In equation 5-57, Eff_{SW} is a random variable representing the software project's development effort (staff-months), c_1 and c_2 are positive constants, and I is a random variable representing the number of *thousands of delivered source instructions* (KDSI) to be developed.* In equation 5-58, $Cost_{SW}$ is a random variable representing the software project's development cost (dollars) and ℓ_r is a constant** representing a labor rate (dollars per staff-month). Notice $Cost_{SW}$ can also be expressed as a function of I, that is,

$$Cost_{SW} = \ell_r (c_1 I^{c_2}) \tag{5-60}$$

In equation 5-59, T_{SW} is a random variable representing the software project's development schedule (months) and k_1 and k_2 are positive constants. Notice T_{SW} can also be expressed as a function of I, that is,

$$T_{SW} = k_1 (c_1 I^{c_2})^{k_2} \tag{5-61}$$

Equations 5-57 through 5-61 represent one approach [5] for determining a software development project's effort, cost, and schedule; there are others. For instance, Eff_{SW} might be determined as the ratio of two random variables I and P_r as shown by equation 5-62. Here, P_r is the software project's development productivity rate (e.g., the number of DSI per staff-month).

$$Eff_{SW} = \frac{I}{P_r} \tag{5-62}$$

Equation 5-62 is an example of a function of two random variables. Working with such functions is discussed in section 5.4.3.

---

* Throughout this book, when I appears in the formula given by equation 5-57 it is assumed that I is always in KDSI. It is further assumed that I is always greater than zero.

** In this section, we treat ℓ_r as a constant to keep the discussion focused on functions of a single random variable; however, in practice, ℓ_r is often treated as a random variable.

Case Discussion 5-2 If the development effort Eff_{SW} for a software project is defined by $Eff_{SW} = c_1 I^{c_2}$, and $I \sim Unif(a,b)$, determine $F_{Eff_{SW}}(s)$, $f_{Eff_{SW}}(s)$, $E(Eff_{SW})$, and $Var(Eff_{SW})$.

Determination of $F_{Eff_{SW}}(s)$

We want the distribution function of Eff_{SW} given the distribution function for I is *uniform*, in the interval $a \le x \le b$. From equation 4-4 (chapter 4) we know

$$f_I(x) = \frac{1}{b-a} \quad a \le x \le b$$

where a and b represent the minimum and maximum possible values of I. By definition

$$F_{Eff_{SW}}(s) = P(Eff_{SW} \le s)$$

$$F_{Eff_{SW}}(s) = P(Eff_{SW} \le s) = P(c_1 I^{c_2} \le s)$$

$$= P(I \le \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}}) = \int_a^{\left(\frac{s}{c_1}\right)^{\frac{1}{c_2}}} f_I(x)dx = \frac{1}{b-a}\left[\left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} - a\right] \quad a \le \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} \le b$$

therefore,

$$F_{Eff_{SW}}(s) = P(Eff_{SW} \le s) = \frac{1}{b-a}\left[\left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} - a\right] \quad a \le \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} \le b \qquad (5\text{-}63)$$

Determination of $f_{Eff_{SW}}(s)$

Given $Eff_{SW} = g(I) = c_1 I^{c_2}$ we can write $s = g(x) = c_1 x^{c_2}$, which is a strictly increasing differentiable function of x. Let the inverse of x be given by

$$x = v(s) = \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}}$$

From theorem 5-11, we have

$$f_{Eff_{SW}}(s) = f_I(v(s)) \cdot \frac{d[v(s)]}{ds} \quad g(a) \le s \le g(b)$$

Therefore

$$f_{Eff_{SW}}(s) = \frac{1}{b-a} \cdot \frac{1}{c_1 c_2} \cdot \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}-1} \qquad c_1 a^{c_2} \le s \le c_1 b^{c_2} \qquad (5\text{-}64)$$

which is also the derivative of $F_{Eff_{SW}}(s)$ with respect to s. It is left to the reader to verify equation 5-64 is a density function.

Determination of $E(Eff_{SW})$

From proposition 3-1, the expected software development effort is

$$E(Eff_{SW}) = E(g(I)) = \int_a^b g(x) f_I(x)\, dx$$

$$= \int_a^b c_1 x^{c_2} \cdot \frac{1}{b-a}\, dx = c_1 \cdot \frac{1}{b-a} \int_a^b x^{c_2}\, dx$$

therefore $$E(Eff_{SW}) = \frac{c_1}{c_2+1} \cdot \frac{1}{b-a}\left[b^{c_2+1} - a^{c_2+1}\right] \qquad (5\text{-}65)$$

Alternatively, equation 5-65 could have been derived as follows:

$$E(Eff_{SW}) = \int_{c_1 a^{c_2}}^{c_1 b^{c_2}} s \cdot f_{Eff_{SW}}(s)\, ds = \int_{c_1 a^{c_2}}^{c_1 b^{c_2}} s \cdot \frac{1}{b-a} \cdot \frac{1}{c_1 c_2} \cdot \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}-1} ds$$

Determination of $Var(Eff_{SW})$

From theorem 3-10, we know

$$Var(Eff_{SW}) = E(Eff_{SW}^2) - \left[E(Eff_{SW})\right]^2$$

Now

$$E(Eff_{SW}^2) = E(g(I)^2) = \int_a^b g(x)^2 f_I(x)\, dx$$

$$= \int_a^b (c_1 x^{c_2})^2 \frac{1}{b-a}\, dx = \frac{1}{b-a} \frac{c_1^2}{2c_2+1}\left[b^{2c_2+1} - a^{2c_2+1}\right]$$

Therefore

$$Var(Eff_{SW}) = \frac{1}{b-a}\frac{c_1^2}{2c_2+1}\left[b^{2c_2+1} - a^{2c_2+1}\right] - \left[E(Eff_{SW})\right]^2 \qquad (5\text{-}66)$$

where

$$E(Eff_{SW}) = \frac{c_1}{c_2+1}\cdot\frac{1}{b-a}\left[b^{c_2+1} - a^{c_2+1}\right]$$

This concludes case discussion 5-2. The following illustrates how these results can be applied to a software development project.

Example 5-12 Suppose the effort (staff-months) to develop software for a new system is determined by $Eff_{SW} = 2.8I^{1.2}$. Suppose the uncertainty in I, the number of *thousands* of delivered source instructions (KDSI), is represented by the distribution $I \sim Unif(30,80)$. Determine

a) $P(Eff_{SW} \le 300)$

b) $P(Eff_{SW} \le E(Eff_{SW}))$

c) $\sigma_{Eff_{SW}}$

d) $P(Cost_{SW} \le 4{,}500{,}000)$ given $\ell_r = 15{,}000$ dollars per staff-month.

Solution

a) Given $Eff_{SW} = 2.8I^{1.2}$, we know from equation 5-57 that $c_1 = 2.8$, $c_2 = 1.2$. Since $I \sim Unif(30,80)$, from equation 5-63

$$P(Eff_{SW} \le 300) = \frac{1}{80-30}\left[\left(\frac{300}{2.8}\right)^{\frac{1}{1.2}} - 30\right] \quad 30 \le 49.16 \le 80$$

$$= 0.383$$

Figure 5-16 shows this region of probability for Eff_{SW}, as well as the PDF of Eff_{SW}. The PDF comes from equation 5-64 (in case discussion 5-2).

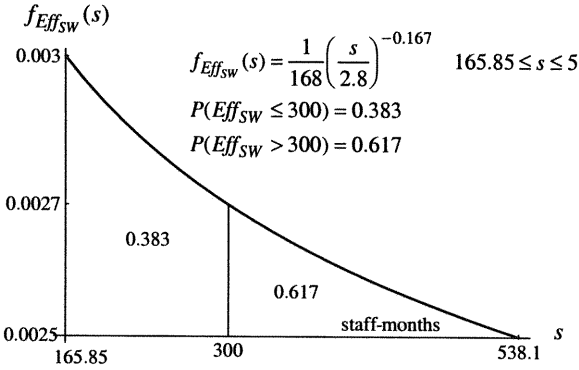


Figure 5-16. The Probability Density Function of Eff_{SW} in Example 5-12

b) From equation 5-65

$$E(Eff_{SW}) = \frac{2.8}{1.2+1} \cdot \frac{1}{50}\left[80^{1.2+1} - 30^{1.2+1}\right] = 346.12 \text{ staff-months}$$

From equation 5-63

$$P(Eff_{SW} \le E(Eff_{SW})) = P(Eff_{SW} \le 346.12) = F_{Eff_{SW}}(346.12) = 0.508$$

c) From equation 5-66

$$Var(Eff_{SW}) = \frac{1}{50}\frac{(2.8)^2}{2(1.2)+1}\left[80^{2(1.2)+1} - 30^{2(1.2)+1}\right] - [346.12]^2$$

$$= 11608.65 \text{ (staff-months)}^2$$

Therefore $\sigma_{Eff_{SW}} = \sqrt{Var(Eff_{SW})} = 107.7$ staff-months.

d) Given $\ell_r = 15{,}000$ dollars per staff-month, we have

$$P(Cost_{SW} \le 4{,}500{,}000) = P(\ell_r \cdot Eff_{SW} \le 4{,}500{,}000)$$

$$= P(Eff_{SW} \le \frac{4{,}500{,}000}{\ell_r}) = P(Eff_{SW} \le \frac{4{,}500{,}000}{15{,}000})$$

$$= P(Eff_{SW} \le 300) = 0.383$$

Example 5-13 Once again, suppose the effort (staff-months) to develop software for a new system is determined by $Eff_{SW} = 2.8I^{1.2}$, where $I \sim Unif(30,80)$. If the software development schedule (months) is given by $T_{SW} = 2.5(Eff_{SW})^{0.32}$, determine the schedule that has a 95 percent chance of *not* being exceeded.

Solution

Three solution approaches are presented.

Approach 1

This approach operates from the cumulative distribution function of I. From the information given in this example, we have

$$\begin{aligned} T_{SW} &= 2.5(Eff_{SW})^{0.32} \\ &= 2.5(2.8I^{1.2})^{0.32} = 3.48I^{0.384} \end{aligned} \tag{5-67}$$

Since $T_{SW} = g(I) = 3.48I^{0.384}$, and $I > 0$, we can write $t = g(x) = 3.48x^{0.384}$ where t and x are the values possible for T_{SW} and I, respectively. Since t is a strictly increasing differentiable function of x, in this example, from theorem 5-11

$$F_{T_{SW}}(t) = F_I(x) \tag{5-68}$$

The value of t associated with the 0.95-fractile of I will equal the 0.95-fractile of T_{SW}. From equation 4-5 (chapter 4), we know

$$F_I(x) = \frac{x-30}{80-30} = \frac{x-30}{50} \quad 30 \le x \le 80$$

The 0.95-fractile of I is $x_{0.95}$ such that $F(x_{0.95}) = P(I \le x_{0.95}) = 0.95$, that is, $x_{0.95}$ is the solution to

$$\frac{x_{0.95}-30}{50} = 0.95 \tag{5-69}$$

Solving equation 5-69 for $x_{0.95}$ yields $x_{0.95} = 77.5$ KDSI; thus

$$x_{0.95} = 77.5 \text{ and } P(I \le x_{0.95}) = 0.95$$

$$t_{0.95} = 18.5 = 3.48 x_{0.95}{}^{0.384} \text{ and } P(T_{SW} \le t_{0.95}) = 0.95$$

This is equivalent to

$$F_I(77.5) = F_{T_{SW}}(18.5) = 0.95$$

Therefore, 18.5 months is the software development schedule that has a 95 percent chance of not being exceeded.

Approach 2

This approach operates from the cumulative distribution function of Eff_{SW}. Since $T_{SW} = g(Eff_{SW}) = 2.5(Eff_{SW})^{0.32}$, we can write $t = g(s) = 2.5s^{0.32}$ where t and s are the values possible for T_{SW} and Eff_{SW}, respectively. Since t is a strictly increasing differentiable function of s, from theorem 5-11

$$F_{T_{SW}}(t) = F_{Eff_{SW}}(s)$$

Thus, the value of t associated with the 0.95-fractile of Eff_{SW} will equal the 0.95-fractile of T_{SW}. From case discussion 5-2, the general formula for $F_{Eff_{SW}}(s)$ is given by equation 5-63. It is left as an exercise for the reader to show, for this example $F_{Eff_{SW}}(518) = F_{T_{SW}}(18.5) = 0.95$.

Approach 3

This approach involves explicitly determining the functional form of $F_{T_{SW}}(t)$ and then solving the expression $F_{T_{SW}}(t_{0.95}) = 0.95$ for $t_{0.95}$. It is left as an exercise for the reader to show, for this example,

$$F_{T_{SW}}(t) = \frac{1}{50}\left[\left(\frac{t}{3.48}\right)^{\frac{1}{0.384}} - 30\right] \quad 12.8 \le t < 18.7^*$$

From the above expression it follows, after rounding, that $F_{T_{SW}}(18.5) = 0.95$.

---

* These endpoints are rounded from the interval $12.8467 \le t < 18.7226$.

Example 5-14 If $T_{SW} = 2.5(Eff_{SW})^{0.32}$ and $Eff_{SW} = 2.8I^{1.2}$ then write a general formula for $P(T_{SW} \le t)$, if $I \sim Trng(30,50,80)$.

Solution Notice T_{SW} can be written as $T_{SW} = 2.5(2.8I^{1.2})^{0.32} = 3.48I^{0.384}$. This implies $t = g(x) = 3.48x^{0.384}$, where t and x are possible values of T_{SW} and I, respectively. Notice t is a strictly increasing differentiable function of x; therefore, from theorem 5-11

$$F_{T_{SW}}(t) = P(T_{SW} \le t) = P(g(I) \le t) = P(I \le v(t)) = F_I(v(t)) = F_I(x)$$

where

$$x = v(t) = \left(\frac{t}{3.48}\right)^{\frac{1}{0.384}}$$

In the above, x is the inverse of $t = g(x) = 3.48x^{0.384}$. Since I is given to have a triangular distribution function, from equation 4-7 (chapter 4)

$$F_I(x) = \begin{cases} 0 & \text{if } x < a \\ \dfrac{(x-a)^2}{(b-a)(m-a)} & \text{if } a \le x < m \\ 1 - \dfrac{(b-x)^2}{(b-a)(b-m)} & \text{if } m \le x < b \\ 1 & \text{if } x \ge b \end{cases}$$

thus

$$P(T_{SW} \le t) = F_I(v(t)) = \begin{cases} 0 & \text{if } \left(\dfrac{t}{3.48}\right)^{\frac{1}{0.384}} < 30 \\ \dfrac{1}{50}\dfrac{1}{20}\left(\left(\dfrac{t}{3.48}\right)^{\frac{1}{0.384}} - 30\right)^2 & \text{if } 30 \le \left(\dfrac{t}{3.48}\right)^{\frac{1}{0.384}} < 50 \\ 1 - \dfrac{1}{50}\dfrac{1}{30}\left(80 - \left(\dfrac{t}{3.48}\right)^{\frac{1}{0.384}}\right)^2 & \text{if } 50 \le \left(\dfrac{t}{3.48}\right)^{\frac{1}{0.384}} < 80 \\ 1 & \text{if } \left(\dfrac{t}{3.48}\right)^{\frac{1}{0.384}} \ge 80 \end{cases}$$

Tables 5-5 and 5-6 present a summary of some general probability formulas for the software effort and schedule models described in this section.

Table 5-5. Software Effort Probability Formulas [6]

$Eff_{SW} = c_1 I^{c_2}$ and $c_1, c_2, I > 0$ (refer to equation 5-57)

Distribution of I	***Unif(a,b)***		
Cumulative Distribution Function	$F_{Eff_{SW}}(s) = P(Eff_{SW} \le s) = \begin{cases} 0 & \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} < a \\ \frac{1}{b-a}\left[\left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} - a\right] & a \le \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} < b \\ 1 & \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} \ge b \end{cases}$		(5-70)
Probability Density Function	$f_{Eff_{SW}}(s) = \frac{1}{b-a} \cdot \frac{1}{c_1 c_2} \cdot \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}-1} \quad a \le \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} \le b$		(5-71)
Mean	$E(Eff_{SW}) = \frac{c_1}{c_2+1} \cdot \frac{1}{b-a}\left[b^{c_2+1} - a^{c_2+1}\right]$	staff-months	(5-72)
Variance	$Var(Eff_{SW}) = \frac{1}{b-a}\frac{c_1^2}{2c_2+1}\left[b^{2c_2+1} - a^{2c_2+1}\right] - \left[E(Eff_{SW})\right]^2$	(staff-months)^2	(5-73)

Distribution of I $\quad Trng(a,m,b)$

Cumulative Distribution Function

$$F_{Eff_{SW}}(s) = P(Eff_{SW} \le s) = \begin{cases} 0 & \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} < a \\ \frac{1}{b-a}\frac{1}{m-a}\left[\left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} - a\right]^2 & a \le \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} < m \\ 1 - \frac{1}{b-a}\frac{1}{b-m}\left[b - \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}}\right]^2 & m \le \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} < b \\ 1 & \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} \ge b \end{cases} \tag{5-74}$$

Probability Density Function

$$f_{Eff_{SW}}(s) = \begin{cases} \frac{2}{b-a}\frac{1}{m-a}\frac{1}{c_2}\frac{1}{c_1}\left(\frac{s}{c_1}\right)^{\frac{1}{c_2}-1}\left[\left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} - a\right] & a \le \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} < m \\ \frac{2}{b-a}\frac{1}{b-m}\frac{1}{c_2}\frac{1}{c_1}\left(\frac{s}{c_1}\right)^{\frac{1}{c_2}-1}\left[b - \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}}\right] & m \le \left(\frac{s}{c_1}\right)^{\frac{1}{c_2}} \le b \end{cases} \tag{5-75}$$

Mean

$$E(Eff_{SW}) = c_1\frac{2}{b-a}\frac{1}{m-a}\left[\frac{m^{c_2+2}-a^{c_2+2}}{c_2+2} + \frac{a^{c_2+2}-am^{c_2+1}}{c_2+1}\right] + c_1\frac{2}{b-a}\frac{1}{m-b}\left[\frac{b^{c_2+2}-m^{c_2+2}}{c_2+2} + \frac{bm^{c_2+1}-b^{c_2+2}}{c_2+1}\right] \quad \text{staff-months} \tag{5-76}$$

Variance

$$Var(Eff_{SW}) = c_1^2\frac{2}{b-a}\frac{1}{m-a}\left[\frac{m^{2c_2+2}-a^{2c_2+2}}{2c_2+2} + \frac{a^{2c_2+2}-am^{2c_2+1}}{2c_2+1}\right] + c_1^2\frac{2}{b-a}\frac{1}{m-b}\left[\frac{b^{2c_2+2}-m^{2c_2+2}}{2c_2+2} + \frac{bm^{2c_2+1}-b^{2c_2+2}}{2c_2+1}\right] - \left[E(Eff_{SW})\right]^2 \quad \text{(staff-months)}^2 \tag{5-77}$$

Table 5-6. Software Schedule Probability Formulas

$T_{SW} = k_1 (Eff_{SW})^{k_2} \equiv T_{SW} = k_1 (c_1 I^{c_2})^{k_2}$ and $c_1, c_2, k_1, k_2, I > 0$ (refer to equations 5-59 and 5-61)

Distribution of I	$\boldsymbol{Unif(a,b)}$		
Cumulative Distribution Function	$F_{T_{SW}}(t) = P(T_{SW} \le t) = \begin{cases} 0 & \left(\frac{t}{k_1 c_1^{k_2}}\right)^{\frac{1}{c_2 k_2}} < a \\ \frac{1}{b-a}\left[\left(\frac{t}{k_1 c_1^{k_2}}\right)^{\frac{1}{c_2 k_2}} - a\right] & a \le \left(\frac{t}{k_1 c_1^{k_2}}\right)^{\frac{1}{c_2 k_2}} < b \\ 1 & \left(\frac{t}{k_1 c_1^{k_2}}\right)^{\frac{1}{c_2 k_2}} \ge b \end{cases}$		(5-78)
Probability Density Function	$f_{T_{SW}}(t) = \frac{1}{b-a} \cdot \frac{1}{k_1 k_2} \frac{1}{c_2 c_1^{k_2}} \cdot \left(\frac{t}{k_1 c_1^{k_2}}\right)^{\frac{1}{c_2 k_2} - 1} \quad a \le \left(\frac{t}{k_1 c_1^{k_2}}\right)^{\frac{1}{c_2 k_2}} \le b$		(5-79)
Mean	$E(T_{SW}) = \frac{k_1 c_1^{k_2}}{c_2 k_2 + 1} \frac{1}{b-a}\left[b^{c_2 k_2 + 1} - a^{c_2 k_2 + 1}\right]$	months	(5-80)
Variance	$Var(T_{SW}) = \frac{k_1^2 c_1^{2k_2}}{2c_2 k_2 + 1} \frac{1}{b-a}\left[b^{2c_2 k_2 + 1} - a^{2c_2 k_2 + 1}\right] - \left[E(T_{SW})\right]^2$	$(\text{months})^2$	(5-81)

Distribution of I $\quad$ $\boldsymbol{Trng(a,m,b)}$

Cumulative Distribution Function

$$F_{T_{SW}}(t)=P(T_{SW}\le t)=\begin{cases}0 & \left(\frac{t}{k_1c_1^{k_2}}\right)^{\frac{1}{c_2k_2}}<a\\ \frac{1}{b-a}\frac{1}{m-a}\left[\left(\frac{t}{k_1c_1^{k_2}}\right)^{\frac{1}{c_2k_2}}-a\right]^2 & a\le\left(\frac{t}{k_1c_1^{k_2}}\right)^{\frac{1}{c_2k_2}}<m\\ 1-\frac{1}{b-a}\frac{1}{b-m}\left[b-\left(\frac{t}{k_1c_1^{k_2}}\right)^{\frac{1}{c_2k_2}}\right]^2 & m\le\left(\frac{t}{k_1c_1^{k_2}}\right)^{\frac{1}{c_2k_2}}<b\\ 1 & \left(\frac{t}{k_1c_1^{k_2}}\right)^{\frac{1}{c_2k_2}}\ge b\end{cases} \tag{5-82}$$

Probability Density Function

$$f_{T_{SW}}(t)=\begin{cases}\frac{2}{b-a}\frac{1}{m-a}\frac{1}{c_2k_2}\frac{1}{k_1c_1^{k_2}}\left(\frac{t}{k_1c_1^{k_2}}\right)^{\frac{1}{c_2k_2}-1}\left[\left(\frac{t}{k_1c_1^{k_2}}\right)^{\frac{1}{c_2k_2}}-a\right] & a\le\left(\frac{t}{k_1c_1^{k_2}}\right)^{\frac{1}{c_2k_2}}<m\\ \frac{2}{b-a}\frac{1}{b-m}\frac{1}{c_2k_2}\frac{1}{k_1c_1^{k_2}}\left(\frac{t}{k_1c_1^{k_2}}\right)^{\frac{1}{c_2k_2}-1}\left[b-\left(\frac{t}{k_1c_1^{k_2}}\right)^{\frac{1}{c_2k_2}}\right] & m\le\left(\frac{t}{k_1c_1^{k_2}}\right)^{\frac{1}{c_2k_2}}\le b\end{cases} \tag{5-83}$$

Mean

$$E(T_{SW})=k_1c_1^{k_2}\frac{2}{b-a}\frac{1}{m-a}\left[\frac{m^{c_2k_2+2}-a^{c_2k_2+2}}{c_2k_2+2}+\frac{a^{c_2k_2+2}-am^{c_2k_2+1}}{c_2k_2+1}\right]$$
$$+k_1c_1^{k_2}\frac{2}{b-a}\frac{1}{m-b}\left[\frac{b^{c_2k_2+2}-m^{c_2k_2+2}}{c_2k_2+2}+\frac{bm^{c_2k_2+1}-b^{c_2k_2+2}}{c_2k_2+1}\right]\quad\text{months} \tag{5-84}$$

Variance

$$Var(T_{SW})=(k_1c_1^{k_2})^2\frac{2}{b-a}\frac{1}{m-a}\left[\frac{m^{2c_2k_2+2}-a^{2c_2k_2+2}}{2c_2k_2+2}+\frac{a^{2c_2k_2+2}-am^{2c_2k_2+1}}{2c_2k_2+1}\right]$$
$$+(k_1c_1^{k_2})^2\frac{2}{b-a}\frac{1}{m-b}\left[\frac{b^{2c_2k_2+2}-m^{2c_2k_2+2}}{2c_2k_2+2}+\frac{bm^{2c_2k_2+1}-b^{2c_2k_2+2}}{2c_2k_2+1}\right]-\left[E(T_{SW})\right]^2\quad(\text{months})^2 \tag{5-85}$$

Example 5-15 Suppose the effort and schedule of a software project are given by $Eff_{SW} = c_1 I^{c_2}$ and $T_{SW} = k_1 (Eff_{SW})^{k_2}$.

a) Develop the correlation formula between Eff_{SW} and T_{SW}, if $I \sim Unif(a,b)$.

b) Compute this correlation if $c_1 = 2.8$, $c_2 = 1.2$, $k_1 = 2.5$, $k_2 = 0.32$ and $I \sim Unif(30,80)$.

c) Discuss what the correlation implies about Eff_{SW} and T_{SW}.

Solution From equation 5-30, the correlation between Eff_{SW} and T_{SW} is

$$\rho_{Eff_{SW},T_{SW}} = \frac{E(Eff_{SW}T_{SW}) - E(Eff_{SW})E(T_{SW})}{\sigma_{Eff_{SW}}\sigma_{T_{SW}}} \tag{5-86}$$

The first term in the numerator can be written as

$$E(Eff_{SW}T_{SW}) = E(c_1 I^{c_2} \cdot k_1 (c_1 I^{c_2})^{k_2}) = k_1 c_1^{k_2+1} E(I^{c_2(k_2+1)})$$

Since

$$E(I^{c_2(k_2+1)}) = \int_a^b t^{c_2(k_2+1)} f_I(t)\,dt = \frac{1}{b-a}\frac{1}{c_2(k_2+1)+1}\left[b^{c_2(k_2+1)+1} - a^{c_2(k_2+1)+1}\right]$$

we have

$$E(Eff_{SW}T_{SW}) = k_1 c_1^{k_2+1}\frac{1}{b-a}\frac{1}{c_2(k_2+1)+1}\left[b^{c_2(k_2+1)+1} - a^{c_2(k_2+1)+1}\right]$$

From equation 5-65, we have

$$E(Eff_{SW}) = \frac{c_1}{c_2+1}\cdot\frac{1}{b-a}\left[b^{c_2+1} - a^{c_2+1}\right] \tag{5-87}$$

From equation 5-66, we have

$$\sigma_{Eff_{SW}} = \sqrt{\frac{1}{b-a}\frac{c_1^2}{2c_2+1}\left[b^{2c_2+1} - a^{2c_2+1}\right] - \left[E(Eff_{SW})\right]^2} \tag{5-88}$$

It is left as an exercise for the reader to show that

$$E(T_{SW}) = \frac{k_1 c_1^{k_2}}{c_2 k_2 + 1} \frac{1}{b-a}\left[b^{c_2 k_2 + 1} - a^{c_2 k_2 + 1}\right] \tag{5-89}$$

$$\sigma_{T_{SW}} = \sqrt{\frac{k_1^2 c_1^{2k_2}}{2c_2 k_2 + 1} \frac{1}{b-a}\left[b^{2c_2 k_2 + 1} - a^{2c_2 k_2 + 1}\right] - \left[E(T_{SW})\right]^2} \tag{5-90}$$

Thus, if $I \sim Unif(a,b)$, then the general formula for the correlation between Eff_{SW} and T_{SW} is

$$\rho_{Eff_{SW},T_{SW}} = \frac{k_1 c_1^{k_2+1} \frac{1}{b-a} \frac{1}{c_2(k_2+1)+1}\left[b^{c_2(k_2+1)+1} - a^{c_2(k_2+1)+1}\right] - E(Eff_{SW})E(T_{SW})}{\sqrt{\frac{c_1^2}{2c_2+1}\frac{1}{b-a}\left[b^{2c_2+1} - a^{2c_2+1}\right] - \left[E(Eff_{SW})\right]^2}\sqrt{\frac{k_1^2 c_1^{2k_2}}{2c_2 k_2 + 1}\frac{1}{b-a}\left[b^{2c_2 k_2 + 1} - a^{2c_2 k_2 + 1}\right] - \left[E(T_{SW})\right]^2}} \tag{5-91}$$

b) Substituting $c_1 = 2.8$, $c_2 = 1.2$, $k_1 = 2.5$, $k_2 = 0.32$, $a = 30$, and $b = 80$ into the above expressions yields

$$E(Eff_{SW}T_{SW}) = 5736.2323$$
$$E(Eff_{SW}) = 346.12,\ Var(Eff_{SW}) = 11610.31$$
$$E(T_{SW}) = 16.055,\ Var(T_{SW}) = 2.798$$

Therefore, from equation 5-91, the correlation between Eff_{SW} and T_{SW} is $\rho_{Eff_{SW},T_{SW}} = 0.995$.

c) Although the true relationship between Eff_{SW} and T_{SW} is nonlinear, a correlation coefficient this close to unity indicates the relationship is not statistically significantly different from linear in the region $165.85 \le s \le 538.10$. This is illustrated in figure 5-17.

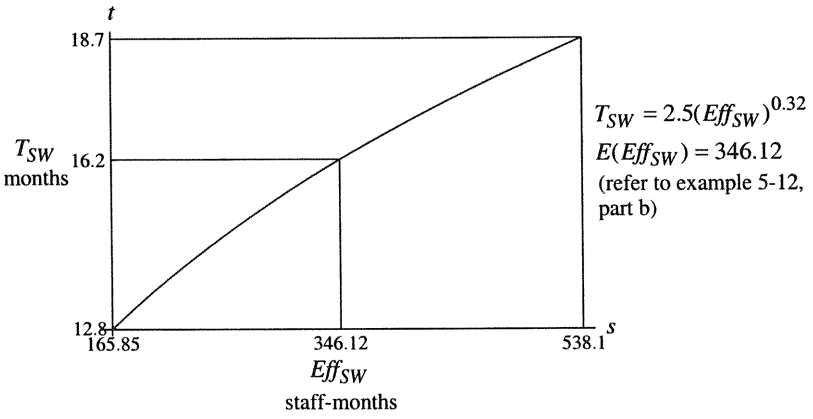


Figure 5-17. A Plot of T_{SW} Versus Eff_{SW} From Example 5-15

Example 5-16 Suppose a new radar system requires developing 14 software functions listed in table 5-7. Let the uncertainties in the amount of code to develop be represented by the random variables $I_1, I_2, I_3, \ldots, I_{14}$, where each I is in thousands of delivered source instructions (KDSI). Assume each I is characterized by a triangular distribution function. Suppose $I_1, I_2, I_3, \ldots, I_{14}$ are *independent random variables* and $I_{Total} = I_1 + I_2 + I_3 + \ldots + I_{14}$.

a) What is the mean and variance of I_{Total}?

b) What distribution function approximates the distribution of I_{Total}?

c) Determine the 0.50-fractile of $Eff_{SW} = 2.8(I_{Total})^{1.2}$.

Table 5-7. Radar Software Functions and Size Uncertainty Assessments

	Min (KDSI)	Mode (KDSI)	Max (KDSI)
Post Processor			
Radar Rpt Proc I_1	3.6	4.0	4.8
Radar Control Proc I_2	5.4	6.0	7.2
Seco Proc I_3	1.8	2.0	2.4
Auto Monitoring I_4	4.5	5.0	6.0
Network Interfacing I_5	1.8	2.0	2.4

Table 5-7. Radar Software Functions and Size Uncertainty Assessments (Concluded)

		Min (KDSI)	Mode (KDSI)	Max (KDSI)
System Control Processor				
Mode Control	I_6	10.8	12.0	14.4
Display Console	I_7	13.5	15.0	18.0
Missile Impact Prediction				
OS and Utilities	I_8	12.6	14.0	16.8
Operational Prgm	I_9	27.0	30.0	36.0
Satellite Test Pgm	I_{10}	12.6	14.0	16.8
Library	I_{11}	10.8	12.0	14.4
Data Reduction	I_{12}	29.7	33.0	39.6
Seco Support	I_{13}	14.4	16.0	19.2
Communications	I_{14}	6.3	7.0	8.4
Total	I_{Total}	154.8	172.0	206.4

Note: The sum of the modes is not the mode of the distribution function of I_{Total}.

Solution

a) We are given the distribution function for each I is triangular, that is,

$$I_1 \sim Trng(3.6, 4.0, 4.8), I_2 \sim Trng(5.4, 6.0, 7.2),$$

$$I_3 \sim Trng(1.8, 2.0, 2.4), \cdots, I_{14} \sim Trng(6.3, 7.0, 8.4)$$

From theorem 5-7 (equation 5-37)

$$E(I_{Total}) = E(I_1) + E(I_2) + E(I_3) + \ldots + E(I_{14}) \tag{5-92}$$

Since each I has a triangular distribution, from theorem 4-3

$$E(I_1) = \tfrac{1}{3}(3.6 + 4.0 + 4.8) = 4.13, \; E(I_2) = \tfrac{1}{3}(5.4 + 6.0 + 7.2) = 6.2$$

$$E(I_3) = \tfrac{1}{3}(1.8 + 2.0 + 2.4) = 2.067, \cdots, \; E(I_{14}) = \tfrac{1}{3}(6.3 + 7.0 + 8.4) = 7.23$$

Substituting these values into equation 5-92 yields

$$E(I_{Total}) = 4.13 + 6.2 + 2.067 + \ldots + 7.23 = 177.73 \text{ KDSI}$$

Since $I_1, I_2, I_3, \ldots, I_{14}$ are *independent*,* from theorem 5-8 (equation 5-39)

---

* From theorem 5-4, since $I_1, I_2, I_3, \ldots, I_{14}$ are independent random variables the correlation between each pair of $I_1, I_2, I_3, \ldots, I_{14}$ is zero.

$$Var(I_{Total}) = Var(I_1) + Var(I_2) + Var(I_3) + \ldots + Var(I_{14})$$

From theorem 4-3

$$Var(I_1) = \frac{1}{18}\left\{(4.0-3.6)(4.0-4.8) + (4.8-3.6)^2\right\} = 0.0622$$

Following a similar set of calculations for $I_2, I_3, \ldots, I_{14}$, it can be shown that

$$Var(I_{Total}) = 12.77 \text{ KDSI}^2$$

b) Since $I_1, I_2, I_3, \ldots, I_{14}$ are given to be independent random variables, the total size of the radar software I_{Total} is the sum of 14 independent random variables. By the central limit theorem (theorem 5-10), it is reasonable to assume the distribution function of I_{Total} will be approximately normal. From part a) this means $I_{Total} \sim N(E(I_{Total}), Var(I_{Total})) = N(177.73, 12.77)$.

c) In this example we are given $Eff_{SW} = 2.8(I_{Total})^{1.2}$. If x and s are the values possible for I_{Total} and Eff_{SW}, respectively, then $s = 2.8x^{1.2}$ is a strictly increasing differentiable function of x. From theorem 5-11, this implies

$$F_{I_{Total}}(x) = F_{Eff_{SW}}(s) \qquad (5\text{-}93)$$

From part b) we know that $F_{I_{Total}}(x) = 0.50$ when $x = 177.73$ KDSI; therefore, $x_{0.50} = 177.73$, which is the 0.50-fractile of I_{Total}. From equation 5-93

$$F_{I_{Total}}(177.73) = 0.50 = F_{Eff_{SW}}(s)$$

Since $s = 2.8x^{1.2}$, when $x = x_{0.50} = 177.73$ we have $s = 1402.4$; thus

$$F_{I_{Total}}(177.73) = 0.50 = F_{Eff_{SW}}(1402.4)$$

In summary, the 0.50-fractile of Eff_{SW} is 1402.4 staff-months. Note this is the same as saying $Med(Eff_{SW}) = 1402.4$ staff-months. It is left as an exercise for the reader to determine the 0.25 and 0.75 fractiles of Eff_{SW}.

5.4.3 Functions of Two Random Variables

Thus far, we have focused on deriving the probability distribution function for a function of a single random variable. Functions of two or more random variables commonly occur in cost uncertainty analysis. For instance, if the unit cost of an unmanned spacecraft is determined by

$$UC = 5.48(SC_{wt})^{0.94}(BOLP)^{0.30}$$

then UC is a function of two random variables — spacecraft weight SC_{wt} and beginning-of-life power $BOLP$. Likewise, if the software development effort for a project is determined by

$$Eff_{SW} = \frac{I}{P_r} \tag{5-94}$$

then Eff_{SW} is a function of two random variables — the amount of code to develop I (in DSI) and the development productivity P_r (in DSI per staff-month). The following theorem provides a set of general integral formulas for determining the density functions of sums, differences, products, and quotients of two random variables. We shall see that determining this density function, in closed form, can be computationally challenging. In many cases a closed form is not even possible. In such circumstances, computer-based methods (e.g., Monte Carlo simulation) are often used to approximate the density function.

Theorem 5-12 [7] Let X and Y be continuous random variables with joint density $f(x,y)$. If U is a function of X and Y with density function $g(u)$, then

$$U = X + Y \text{ has density } g(u) = \int_{-\infty}^{\infty} f(x, u-x)\,dx = \int_{-\infty}^{\infty} f(u-y, y)\,dy$$

$$U = X - Y \text{ has density } g(u) = \int_{-\infty}^{\infty} f(x, x-u)\,dx = \int_{-\infty}^{\infty} f(u+y, y)\,dy$$

$U = XY$ has density $\quad g(u) = \int_{-\infty}^{\infty} \frac{1}{|x|} f(x, \frac{u}{x}) dx = \int_{-\infty}^{\infty} \frac{1}{|y|} f(\frac{u}{y}, y) dy$

$U = X/Y$ has density $\quad g(u) = \int_{-\infty}^{\infty} |x| f(ux, x) dx = \int_{-\infty}^{\infty} |y| f(uy, y) dy$

The reader is directed to [7] for a proof of this theorem. Theorem 5-12 provides a number of interesting results. For instance, suppose U_1, U_2, and U_3 are *independent* random variables with $U_1 \sim Unif(0,1)$, $U_2 \sim Unif(0,1)$, and $U_3 \sim Unif(0,1)$. If $U = U_1 + U_2$ then the density function for U can be shown to be triangular [8]. Furthermore, if $U = U_1 + U_2 + U_3$ then the density function for U is "bell-shaped" — but not yet normally distributed (i.e., gaussian). Figure 5-18 [8] illustrates these results.

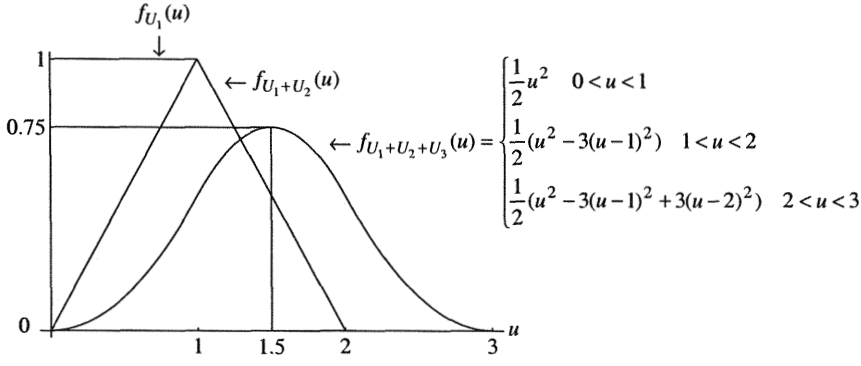


Figure 5-18. Sums of Independent *Unif* (0,1) Random Variables

Continuing the above, suppose the random variable U is defined by

$$U = U_1 + U_2 + U_3 + \ldots + U_n$$

where $U_1, U_2, U_3, \ldots, U_n$ are independent random variables and $U_i \sim Unif(0,1)$ for $i = 1, \ldots, n$. By the central limit theorem as n increases the distribution function of U will rapidly approach a normal distribution. This remarkable result is further discussed and illustrated in appendix A (section A.4).

The following presents an application of theorem 5-12. A probability density function for software development effort, defined by equation 5-94, is derived.

Example 5-17 In example 5-3, the effort Eff_{SW} to develop a new software application was given by

$$Eff_{SW} = \frac{X}{Y}$$

where $X = I$ is the amount of code to develop (in DSI) and $Y = P_r$ is the development productivity (in DSI per staff-month). Suppose X and Y are continuous random variables with joint PDF

$$f(x,y) = \begin{cases} \dfrac{1}{5(10^6)} & 50{,}000 \le x \le 100{,}000, \quad 100 \le y \le 200 \\ 0 \quad \text{otherwise} \end{cases}$$

a) Use theorem 5-12 to find the PDF of Eff_{SW}.

b) From part a), verify $P(Eff_{SW} \le 300) = 0.0333$ and $P(Eff_{SW} \le 610) \approx 0.75$.

c) From part a), determine $E(Eff_{SW})$.

Solution

a) Since Eff_{SW} is a ratio of two random variables, from theorem 5-12 Eff_{SW} has probability density function $g(u)$, where

$$g(u) = \int_{-\infty}^{\infty} |y| f(uy, y)\, dy$$

In the above, u represents feasible values of the random variable Eff_{SW} (in staff-months). To use the integral given by $g(u)$, it is necessary to define the regions of integration specific to this example. These regions are shown in figure 5-19.

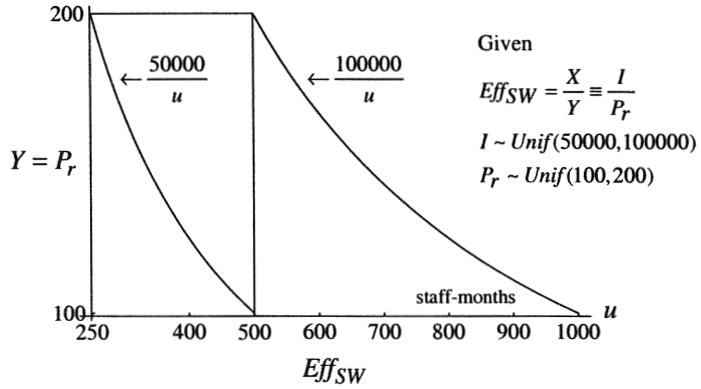


Figure 5-19. Regions of Integration for $g(u)$ in Example 5-17

From figure 5-19, we see that

$$g(u) = \begin{cases} \displaystyle\int_{\frac{50,000}{u}}^{200} \frac{1}{5(10^6)} y\,dy & 250 \le u \le 500 \\ \displaystyle\int_{100}^{\frac{100,000}{u}} \frac{1}{5(10^6)} y\,dy & 500 \le u \le 1000 \end{cases}$$

The probability density function of Eff_{SW} is, therefore, given by equation 5-95.

$$g(u) = \begin{cases} \frac{1}{2}\frac{1}{5(10^6)}\left\{(200)^2 - \left(\frac{50,000}{u}\right)^2\right\} & 250 \le u \le 500 \\ \frac{1}{2}\frac{1}{5(10^6)}\left\{\left(\frac{100,000}{u}\right)^2 - (100)^2\right\} & 500 \le u \le 1000 \end{cases} \tag{5-95}$$

A plot of this PDF is shown in figure 5-20.

b) Using equation 5-95, probabilities associated with various values of Eff_{SW}

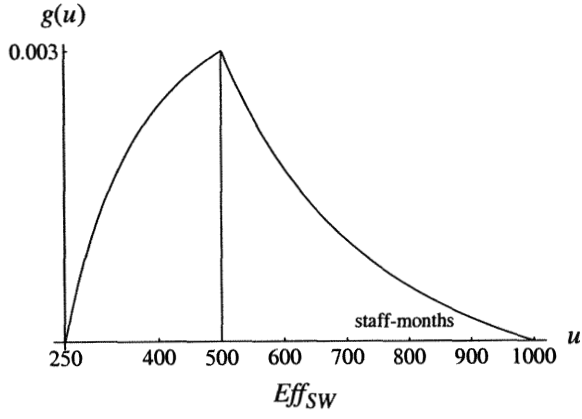


Figure 5-20. Probability Density Function for Eff_{SW}

can be computed. For instance, the probability that $Eff_{SW} \leq 300$ staff-months is

$$P(Eff_{SW} \leq 300) = \int_{250}^{300} \frac{1}{2} \frac{1}{5(10^6)} \left\{ (200)^2 - \left(\frac{50,000}{u} \right)^2 \right\} du = 0.0333$$

This result is consistent with example 5-3. The probability $Eff_{SW} \leq 610$ staff-months is

$$P(Eff_{SW} \leq 610) = \int_{250}^{500} \frac{1}{2} \frac{1}{5(10^6)} \left\{ (200)^2 - \left(\frac{50,000}{u} \right)^2 \right\} du$$

$$+ \int_{500}^{610} \frac{1}{2} \frac{1}{5(10^6)} \left\{ \left(\frac{100,000}{u} \right)^2 - (100)^2 \right\} du = 0.50 + 0.250656 \approx 0.75$$

A family of boundary curves for Eff_{SW} is presented in figure 5-21. Shown are values of Eff_{SW} for various combinations of the number of DSI to develop $X = I$ and the development productivity rate $Y = P_r$ (DSI per staff-month).

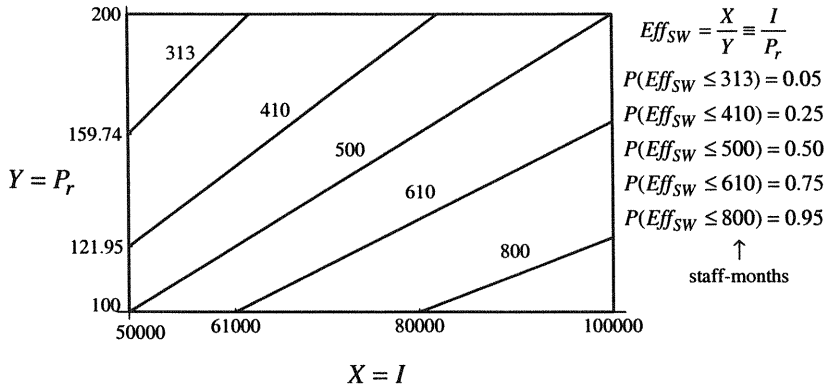


Figure 5-21. Boundary Curves for Eff_{SW} and Associated Probabilities

c) Lastly, from equation 5-95 the expected effort can be computed; specifically,

$$E(Eff_{SW}) = \int_{250}^{500} u \cdot \frac{1}{2}\frac{1}{5(10^6)}\left\{(200)^2 - \left(\frac{50,000}{u}\right)^2\right\} du$$

$$+ \int_{500}^{1000} u \cdot \frac{1}{2}\frac{1}{5(10^6)}\left\{\left(\frac{100,000}{u}\right)^2 - (100)^2\right\} du = 519.86 \text{ staff-months} \blacklozenge$$

Theorem 5-12 provides a way to determine the probability density function of sums, differences, products, and quotients of two random variables. The integrals in theorem 5-12 are classically known as *convolution integrals*. In many applied problems these integrals are hard to determine. In cost uncertainty analysis, conditions often prevail that enable analysts to approximate the form of a probability density function. If an approximation can be found (or theoretically claimed), then it is unnecessary to compute a convolution integral. For instance, we know (from the central limit theorem) the sum of a sufficiently large number of independent random variables will approach the normal distribution. Similarly, from the central limit theorem, we know the product of a sufficiently large number of independent random

variables will approach the lognormal distribution.

The last topic discussed in this chapter is the Mellin transform. The Mellin transform is a useful technique for computing the moments of products and quotients of many random variables. The application of the Mellin transform to cost functions comprised of two or more random variables is emphasized.

5.5 The Mellin Transform and its Application to Cost Functions

This section presents a little known technique for determining moments of products and quotients of random variables. Known as the Mellin transform [9, 10], it works on random variables that are continuous, independent, and nonnegative*. The Mellin transform is well suited to cost functions since *Cost* is essentially a nonnegative random variable. The following defines the Mellin transform. Examples are provided to illustrate its use from a cost perspective.

Definition If X is a nonnegative random variable, $0 < x < \infty$, the Mellin transform of its probability density function $f_X(x)$ is

$$M_X(s) = \int_0^\infty x^{s-1} f_X(x)\,dx \tag{5-96}$$

for all s for which the integral exists. From equation 5-96 it can be seen that

$$M_X(1) = \int_0^\infty f_X(x)\,dx = 1 \tag{5-97}$$

$$M_X(2) = \int_0^\infty x\, f_X(x)\,dx = E(X) \tag{5-98}$$

---

* An extension of the Mellin transform technique to random variables that are not everywhere positive is discussed in reference 9.

$$M_X(3) = \int_0^\infty x^2 f_X(x)dx = E(X^2) \tag{5-99}$$

From the above, it follows from equation 3-31 that

$$M_X(s) = E(X^{s-1}) \tag{5-100}$$

It also immediately follows that

$$Var(X) = M_X(3) - [M_X(2)]^2 \tag{5-101}$$

The Mellin transform is very useful when dealing with random variables raised to a power. For example, if for any real a we have $Y = X^a$ then

$$\begin{aligned} M_Y(s) &= E(Y^{s-1}) = E((X^a)^{s-1}) = E((X^{as-a})) \\ &= E((X^{(as-a+1)-1})) = M_X(as-a+1) \end{aligned} \tag{5-102}$$

As an illustration, consider the Mellin transform of $Eff_{SW} = 2.8I^{1.2}$. This yields

$$\begin{aligned} M_{Eff_{SW}}(s) &= E(Eff_{SW}{}^{s-1}) = E((2.8I^{1.2})^{s-1}) = E((2.8^{s-1}I^{1.2s-1.2})) \\ &= 2.8^{s-1}E((I^{(1.2s-1.2+1)-1})) = 2.8^{s-1}M_I(1.2s-1.2+1) \end{aligned} \tag{5-103}$$

therefore

$$M_{Eff_{SW}}(s) = 2.8^{s-1}M_I(1.2s-1.2+1) \tag{5-104}$$

Equation 5-104 provides a way to generate moments of the random variable Eff_{SW}. For instance, the expected effort $E(Eff_{SW})$ can be written in terms of equation 5-104 as follows:

$$E(Eff_{SW}) = M_{Eff_{SW}}(2) = 2.8M_I(2.2)$$

For example, if $I \sim Unif(30,80)$ then from equation 5-96

$$M_I(s) = \int_0^\infty t^{s-1} f_I(t)dt = \int_{30}^{80} t^{s-1}\frac{1}{50}dt = \frac{1}{50}\left[\frac{80^s - 30^s}{s}\right]$$

where $s \neq 0$. Therefore,

$$E(Eff_{SW}) = M_{Eff_{SW}}(2) = 2.8 M_I(2.2)$$

$$= (2.8)\frac{1}{50}\left[\frac{80^{2.2} - 30^{2.2}}{2.2}\right] = 346.12 \text{ staff-months} \qquad (5\text{-}105)$$

This value agrees with the value of $E(Eff_{SW})$ computed by equation 5-65 in example 5-12. Furthermore, notice equation 5-105 is a specific application of the general formula for $E(Eff_{SW})$ given by equation 5-65. The following presents an important convolution property of the Mellin transform.

Theorem 5-13 [10] Let X, Y, and W be independent random variables with probability density functions $f_X(x)$, $f_Y(y)$, and $f_W(w)$. If α, β_1, β_2, β_3 are constants and

$$Z = \alpha X^{\beta_1} Y^{\beta_2} W^{\beta_3}$$

then

$$M_Z(s) = \alpha^{s-1} M_X(\beta_1 s - \beta_1 + 1) M_Y(\beta_2 s - \beta_2 + 1) M_W(\beta_3 s - \beta_3 + 1) \blacklozenge$$

From theorem 5-13, if $Z = XY$ then

$$M_Z(s) = M_X(s) M_Y(s) \qquad (5\text{-}106)$$

Similarly, from theorem 5-13, if $Z = \frac{X}{Y}$ then

$$M_Z(s) = M_X(s) M_Y(2 - s) \qquad (5\text{-}107)$$

Table 5-8 provides Mellin transforms for three distribution functions defined in chapter 4. In table 5-8, it is assumed that $s \neq 0$. Exercise 19b, at the end of this chapter, examines how equation 5-108 (in table 5-8) is modified for the case when $s = 0$.

Table 5-8. Mellin Transforms for Selected Distribution Functions ($s \neq 0$)

Distribution of X	Mellin Transform of X	
$Unif(a,b)$	$M_X(s) = \frac{1}{s(b-a)}\left(b^s - a^s\right)$	(5-108)
$Trng(a,m,b)$	$M_X(s) = \frac{2}{s(s+1)(b-a)}\left[\frac{b(b^s - m^s)}{b-m} - \frac{a(m^s - a^s)}{m-a}\right],\ s \neq -1$	(5-109)
$Trap(a,m_1,m_2,b)$	$M_X(s) = L_1 L_3 \frac{\left[m_1^s(sm_1 - (s+1)a) + a^{s+1}\right]}{s(s+1)} + L_1 \frac{(m_2^s - m_1^s)}{s} + L_1 L_2 \frac{\left[m_2^s(sm_2 - (s+1)b) + b^{s+1}\right]}{s(s+1)},\ s \neq -1$	(5-110)

where $L_1 = \frac{2}{(m_2 + b - a - m_1)}$, $L_2 = \frac{1}{(b - m_2)}$, $L_3 = \frac{2}{(m_1 - a)}$

*Example 5-18** Let the unit cost UC of an unmanned spacecraft be given by

$$UC = 5.48(SC_{wt})^{0.94}(BOLP)^{0.30}$$

where UC is a function of SC_{wt} (the spacecraft's weight in pounds) and $BOLP$ (the spacecraft's beginning-of-life power in watts). Suppose point estimates for weight and power are 6500 pounds and 2000 watts; that is,

$$w_{PE_{SC_{wt}}} = 6500 \text{ and } j_{PE_{BOLP}} = 2000$$

where possible values for SC_{wt} and $BOLP$ are given by w and j, respectively. If the uncertainties around these point estimates are described by the probability density functions in figure 5-22, use the Mellin transform to compute the expected unit cost $E(UC)$.

---

* This example is an adaptation from Lurie, P. M., and M. S. Goldberg. 1993. *A Handbook of Cost Risk Analysis Methods*, P-2734. Alexandria, Virginia: The Institute for Defense Analyses.

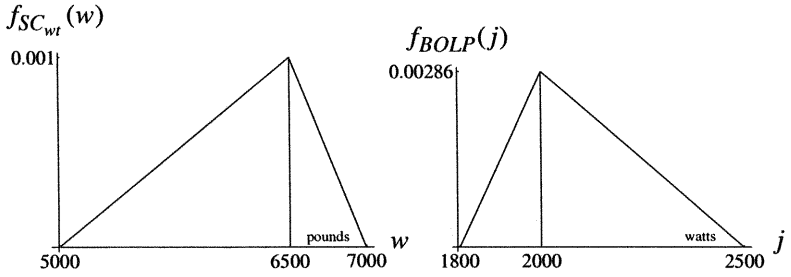


Figure 5-22. Probability Density Functions for SC_{wt} and $BOLP$

Solution To simplify notation let $X = SC_{wt}$, $Y = BOLP$, and $Z = UC$. We then need to compute $E(Z)$, where $Z = 5.48X^{0.94}Y^{0.30}$. From theorem 5-13, the Mellin transform of Z is

$$M_Z(s) = 5.48^{s-1} M_X(0.94s - 0.94 + 1) M_Y(0.30s - 0.30 + 1)$$

From equation 5-100

$$E(UC) = E(Z) = M_Z(2) = 5.48 M_X(1.94) M_Y(1.30)$$

Since the probability density functions for weight and power are triangular, from table 5-8 (equation 5-109)

$$M_X(1.94) = \frac{2}{1.94(2.94)(2000)}\left[\frac{7000(7000^{1.94} - 6500^{1.94})}{7000 - 6500} - \frac{5000(6500^{1.94} - 5000^{1.94})}{6500 - 5000}\right] = 3652.486$$

$$M_Y(1.30) = \frac{2}{1.30(2.30)(700)}\left[\frac{2500(2500^{1.30} - 2000^{1.30})}{2500 - 2000} - \frac{1800(2000^{1.30} - 1800^{1.30})}{2000 - 1800}\right] = 9.918$$

therefore

$$E(UC) = E(Z) = M_Z(2) = 198.5 \text{ (\$K)} \blacklozenge$$

Let's discuss this example further. If the point estimates for SC_{wt} and $BOLP$ were substituted into UC, then

$$UC_{PE} = 5.48(6500)^{0.94}(2000)^{0.30} = 205.7 \text{ (\$K)}$$

In this example, why is $E(UC) < UC_{PE}$? Seen in figure 5-22 the skew of SC_{wt} is negative. There is far more probability the spacecraft's weight will

fall to the left of 6500 pounds than to the right of 6500 pounds. Furthermore, the variance of SC_{wt} is significantly greater than the variance of *BOLP*; showing this is left for the reader. For these reasons, we have an expected cost that is less than the point estimate of the unit cost.

Example 5-19 A new software application is to be developed. Suppose the application consists of a mixture of new code I_{New} and reused code I_{Reused}. Let the effort associated with developing the application be a function of the equivalent size I_{Equiv}, where (from [11])

$$I_{Equiv} = I_{New} + I_{Reused}^{0.857} \tag{5-111}$$

Suppose values for I_{New} and I_{Reused} are uncertain. If I_{New} and I_{Reused} are independent random variables with probability density functions given in figure 5-23, use the Mellin transform to compute $E(I_{Equiv})$ and $\sigma_{I_{Equiv}}$.

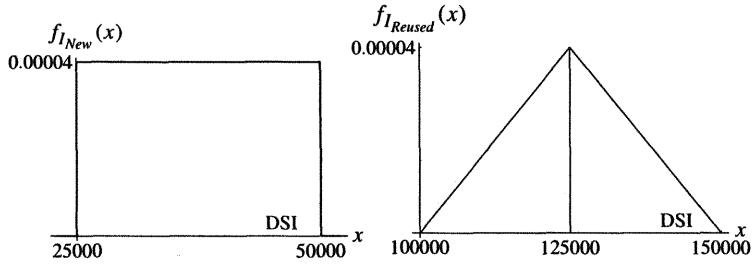


Figure 5-23. Probability Density Functions for I_{New} and I_{Reused}

Solution We are given $I_{Equiv} = I_{New} + I_{Reused}^{0.857}$. From theorems 5-7 and 5-8

$$E(I_{Equiv}) = E(I_{New}) + E(I_{Reused}^{0.857})$$

$$Var(I_{Equiv}) = Var(I_{New}) + Var(I_{Reused}^{0.857})$$

Computing $E(I_{Equiv})$

We have $E(I_{Equiv}) = E(I_{New}) + E(I_{Reused}^{0.857})$. From equation 5-98 $E(I_{New}) = M_{I_{New}}(2)$. Suppose we let $Z = I_{Reused}^{0.857}$, then from theorem 5-13

$$M_Z(s) = M_{I_{Reused}}(0.857(s-1)+1)$$

$$E(I_{Reused}^{0.857}) = M_Z(2) = M_{I_{Reused}}(0.857(2-1)+1) = M_{I_{Reused}}(1.857)$$

From this, we have

$$E(I_{Equiv}) = M_{I_{New}}(2) + M_{I_{Reused}}(1.857)$$

Since $I_{New} \sim Unif(25000, 50000)$, from equation 5-108 $M_{I_{New}}(2) = 37,500$; similarly, since $I_{Reused} \sim Trng(100000, 125000, 150000)$, from equation 5-109 $M_{I_{Reused}}(1.857) = 23,327.8$; therefore

$$E(I_{Equiv}) = 37,500 + 23,327.8 = 60,827.8 \text{ DSI} \approx 61 \text{ KDSI}$$

Computing $\sigma_{I_{Equiv}}$

To compute $\sigma_{I_{Equiv}}$, we begin by computing $Var(I_{Equiv})$. Since I_{New} and I_{Reused} are independent random variables

$$Var(I_{Equiv}) = Var(I_{New}) + Var(I_{Reused}^{0.857})$$

From equation 5-101

$$Var(I_{New}) = M_{I_{New}}(3) - (M_{I_{New}}(2))^2$$

We can write

$$Var(I_{Reused}^{0.857}) = E((I_{Reused}^{0.857})^2) - (E(I_{Reused}^{0.857}))^2 = E(I_{Reused}^{1.714}) - (M_{I_{Reused}}(1.857))^2$$

Suppose we let $W = I_{Reused}^{1.714}$, then from theorem 5-13

$$M_W(s) = M_{I_{Reused}}(1.714(s-1)+1)$$

$$E(I_{Reused}^{1.714}) = M_W(2) = M_{I_{Reused}}(1.714(2-1)+1) = M_{I_{Reused}}(2.714)$$

Therefore $Var(I_{Reused}^{0.857}) = M_{I_{Reused}}(2.714) - (M_{I_{Reused}}(1.857))^2$. From which

$$Var(I_{Equiv}) = M_{I_{New}}(3) - (M_{I_{New}}(2))^2 + M_{I_{Reused}}(2.714) - (M_{I_{Reused}}(1.857))^2$$

From equation 5-108 $M_{I_{New}}(3) = 1.45833(10)^9$ and $M_{I_{New}}(2) = 37,500$; from equation 5-109 $M_{I_{Reused}}(2.714) = 5.46856(10)^8$ and $M_{I_{Reused}}(1.857) = 23,327.8$. Substituting these values into $Var(I_{Equiv}) = M_{I_{New}}(3) - (M_{I_{New}}(2))^2 + M_{I_{Reused}}(2.714) - (M_{I_{Reused}}(1.857))^2$ produces

$$\sigma_{I_{Equiv}} = \sqrt{Var(I_{Equiv})} = 7,399.49 \text{ DSI} \approx 7.4 \text{ KDSI} \blacklozenge$$

Case Discussion 5-3 In example 5-2 the effort for system test was given by the $Eff_{SysTest} = XY$, where X is staff-level and Y is the number of months. Suppose X and Y are independent random variables with distribution functions shown in figure 5-24.*

a) Use a convolution integral in theorem 5-12 to develop a general formula for the probability density function of $Eff_{SysTest}$. Plot the density function.

b) Using the probability density function of $Eff_{SysTest}$ compute the mean of $Eff_{SysTest}$, $P(Eff_{SysTest} \le E(Eff_{SysTest}))$, and $P(Eff_{SysTest} \le 173)$.

c) Use the *Mellin transform* to compute the mean and variance of $Eff_{SysTest}$.

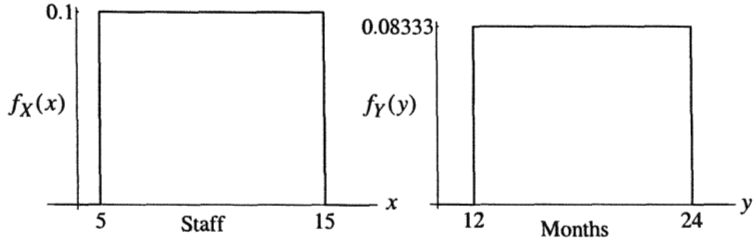


Figure 5-24. Marginal Distribution for X (Staff) and Y (Months)

Discussion

a) Since X and Y are independent, their joint distribution function is

$$f(x,y) = \frac{1}{10}\frac{1}{12} = \frac{1}{120} \quad 5 \le x \le 15,\ 12 \le y \le 24 \tag{5-112}$$

Let $Eff_{SysTest} = U = XY$. Let $g(u)$ represent the probability density function of $Eff_{SysTest}$. Since $Eff_{SysTest}$ is a product of two random variables, from theorem 5-12

---

* This is a slight variation from example 5-2, where the range of possible values for Y was given as 12-36 months. It is left to the reader to study how the problem solution presented in case discussion 5-3 changes, if Y varies from 12-36 months instead of 12-24 months.

$$g(u) = \int_{-\infty}^{\infty} \frac{1}{|y|} f(\tfrac{u}{y}, y)\, dy \qquad (5\text{-}113)$$

The regions of integration for $g(u)$ are shown in figure 5-25.

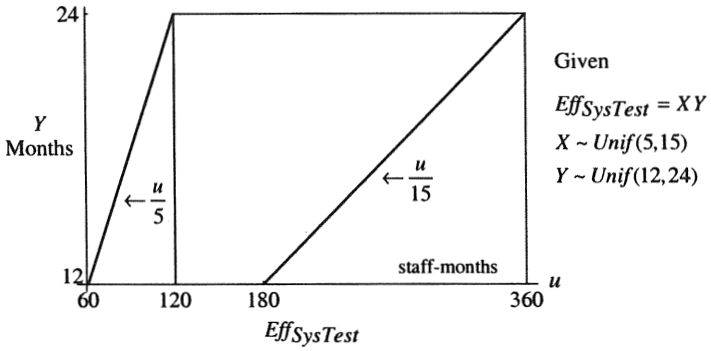


Figure 5-25. Region of Integration for $g(u)$

From figure 5-25, and equation 5-113, the probability density function of $Eff_{SysTest}$ is given by the three integrals over the following regions:

$$g(u) = \begin{cases} \int_{12}^{\frac{u}{5}} \frac{1}{y}\frac{1}{120}\, dy & 60 \le u \le 120 \\ \int_{12}^{24} \frac{1}{y}\frac{1}{120}\, dy & 120 \le u \le 180 \\ \int_{\frac{u}{15}}^{24} \frac{1}{y}\frac{1}{120}\, dy & 180 \le u \le 360 \end{cases} = \begin{cases} \frac{1}{120}\ln\left(\frac{u}{60}\right) & 60 \le u \le 120 \\ \frac{1}{120}\ln(2) & 120 \le u \le 180 \\ \frac{1}{120}\ln\left(\frac{360}{u}\right) & 180 \le u \le 360 \end{cases} \qquad (5\text{-}114)$$

Equation 5-114 is the probability density function of $Eff_{SysTest}$. It is left to the reader to check that $g(u)$ has unit area over the interval $60 \le u \le 360$. Figure 5-26 shows a plot of this density function.

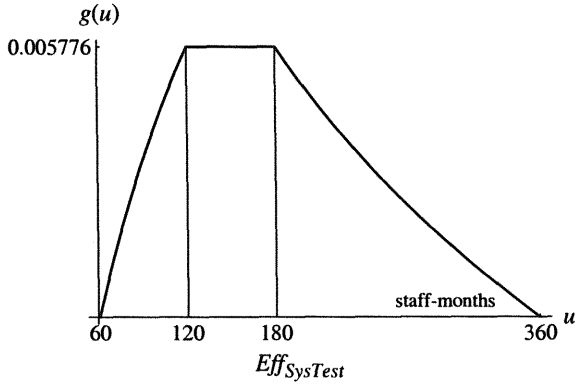


Figure 5-26. Probability Density Function of $Eff_{SysTest}$

b) From the density function we can compute the mean effort for system test, as well as various probabilities. These computations are given below.

$$E(Eff_{SysTest}) = \int_{60}^{360} u g(u)\,du = \int_{60}^{120} u\frac{1}{120}\ln\left(\frac{u}{60}\right)du + \int_{120}^{180} u\frac{1}{120}\ln(2)\,du + \int_{180}^{360} u\frac{1}{120}\ln\left(\frac{360}{u}\right)du = 180$$

Knowledge of the density function facilitates computing various probabilities of interest. From equation 5-114

$$P(Eff_{SysTest} \le E(Eff_{SysTest})) = P(Eff_{SysTest} \le 180)$$

$$= \int_{60}^{120} \frac{1}{120}\ln\left(\frac{u}{60}\right)du + \int_{120}^{180} \frac{1}{120}\ln(2)\,du = 0.19315 + 0.34657 = 0.54$$

Similarly,

$$P(Eff_{SysTest} \le 173) = \int_{60}^{120} \frac{1}{120}\ln\left(\frac{u}{60}\right)du + \int_{120}^{173} \frac{1}{120}\ln(2)\,du$$

$$= 0.19315 + 0.30613 \approx 0.50$$

In this case, the median test effort is approximately 173 staff-months.

Shown in figure 5-27 are curves of constant effort for various pairs of x (staff) and y (months). A probability associated with each effort is also shown.

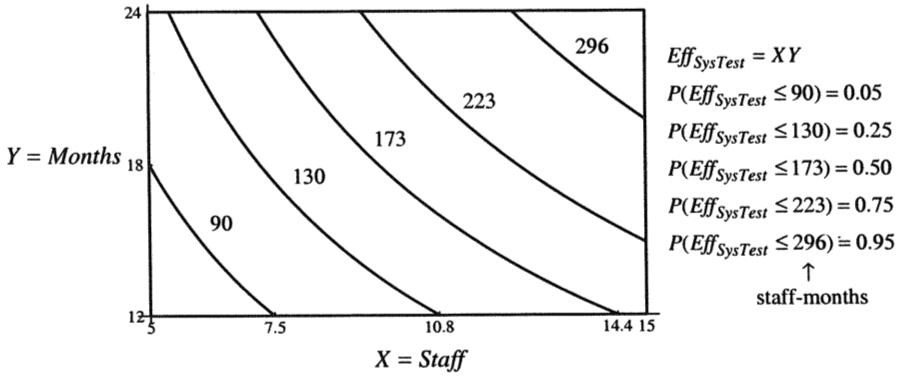


Figure 5-27. Boundary Curves for $Eff_{SysTest}$

Seen in the above discussion, developing a general formula for the probability density function of $Eff_{SysTest}$ involved some tricky mathematics. A slight alteration in the problem statement can further complicate the mathematics. If, for instance, the distribution function of X was triangular instead of uniform, it would be quite difficult to develop an analytical form of $g(u)$.

c) The following illustrates how the Mellin transform applies to this case discussion. The first two moments, which lead to the mean and variance of the test effort, are developed. It is given that

$$Eff_{SysTest} = U = XY \quad 0 < x < \infty \quad 0 < y < \infty$$

From theorem 5-13

$$M_{Eff_{SysTest}}(s) = M_U(s) = M_X(s)M_Y(s) \tag{5-115}$$

From equation 5-98

$$E\left(Eff_{SysTest}\right) = E(U) = M_U(2) = M_X(2)M_Y(2) \tag{5-116}$$

From equation 5-101

$$Var\left(Eff_{SysTest}\right) = Var(U) = M_U(3) - \left[M_U(2)\right]^2$$
$$= M_X(3)M_Y(3) - \left[M_X(2)M_Y(2)\right]^2 \qquad (5\text{-}117)$$

Since the distribution functions for X and Y are uniform with parameters shown in figure 5-24, from equation 5-108 (table 5-8) $M_X(2) = 10$, $M_Y(2) = 18$, $M_X(3) = 108.333$, and $M_Y(3) = 336$. Substituting these values into equations 5-116 and 5-117 yields

$$E\left(Eff_{SysTest}\right) = E(U) = 180 \text{ staff-months}$$
$$Var\left(Eff_{SysTest}\right) = Var(U) = 4000 \text{ (staff-months)}^2$$
$$\sigma_{Eff_{SysTest}} = \sqrt{Var\left(Eff_{SysTest}\right)} = 63.25 \text{ staff-months}$$

The Mellin transform is clearly a convenient way to compute the moments of $Eff_{SysTest}$. The density function of $Eff_{SysTest}$ is not needed.♦

Next, a final case discussion is presented. It will show how concepts throughout this chapter combine to produce useful results. Specifically, formulas for the mean and variance of a ratio of two uniformly distributed random variables and two beta distributed random variables are developed. Seen in previous examples, ratios of random variables can arise frequently in cost uncertainty analysis.

Case Discussion 5-4* Suppose I and P_r are independent random variables. Develop general formulas for $E(Eff_{SW})$ and $Var(Eff_{SW})$ if $Eff_{SW} = \dfrac{I}{P_r}$ and

a) $I \sim Unif(a_1, b_1)$ and $P_r \sim Unif(a_2, b_2)$

b) $I \sim Beta(\alpha_1, \beta_1, a_1, b_1)$ and $P_r \sim Beta(\alpha_2, \beta_2, a_2, b_2)$

---

* Assume only positive values for a_2 and b_2 are permitted.

Discussion Since I and P_r are independent, from theorem 5-5

$$E(Eff_{SW}) = E\left(\frac{I}{P_r}\right) = E(I)E\left(\frac{1}{P_r}\right) = \mu_I E\left(\frac{1}{P_r}\right) \tag{5-118}$$

By definition

$$Var(Eff_{SW}) = E(Eff_{SW}{}^2) - [E(Eff_{SW})]^2$$

$$= E\left(I^2 \frac{1}{P_r^2}\right) - \mu_I^2\left[E\left(\frac{1}{P_r}\right)\right]^2 = E(I^2)E\left(\frac{1}{P_r^2}\right) - \mu_I^2\left[E\left(\frac{1}{P_r}\right)\right]^2$$

By definition $Var(I) = E(I^2) - [E(I)]^2$. This is equivalent to

$$E(I^2) = \sigma_I^2 + \mu_I^2$$

Substituting into $Var(Eff_{SW})$ yields

$$Var(Eff_{SW}) = (\sigma_I^2 + \mu_I^2)E\left(\frac{1}{P_r^2}\right) - \mu_I^2\left[E\left(\frac{1}{P_r}\right)\right]^2 \tag{5-119}$$

a) We are interested in using these equations to develop general formulas for the mean and variance of Eff_{SW}, when I and P_r are uniformly distributed random variables. It has just been shown that

$$E(Eff_{SW}) = E\left(\frac{I}{P_r}\right) = E(I)E\left(\frac{1}{P_r}\right) = \mu_I E\left(\frac{1}{P_r}\right)$$

Since $I \sim Unif(a_1, b_1)$ we know $\mu_I = \frac{1}{2}(a_1 + b_1)$; therefore

$$E(Eff_{SW}) = \tfrac{1}{2}(a_1 + b_1)E\left(\frac{1}{P_r}\right)$$

To produce a general formula for $E(Eff_{SW})$, it remains to determine

$$E\left(\frac{1}{P_r}\right) \equiv E((P_r)^{-1})$$

Determining $E((P_r)^{-1})$ will be accomplished from the probability density function of $(P_r)^{-1}$. Let $Z=(P_r)^{-1}$; therefore,

$$Z=g(P_r)\Rightarrow z=g(y)=\frac{1}{y}\Rightarrow y=v(z)=\frac{1}{z}$$

Since $g(y)$ is a strictly decreasing differentiable function of y, from theorem 5-11

$$f_Z(z)=f_{P_r}(v(z))\cdot\left|\frac{d[v(z)]}{dz}\right| \quad g(b_2)\le z\le g(a_2)$$

$$f_Z(z)=\frac{1}{b_2-a_2}\frac{1}{z^2} \qquad \frac{1}{b_2}\le z\le\frac{1}{a_2}$$

A picture of this density function is shown in figure 5-28.

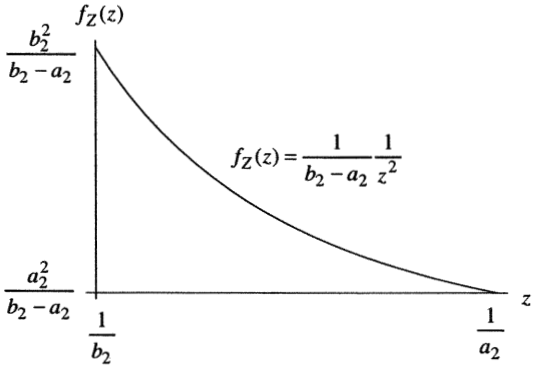


Figure 5-28. Probability Density Function of $Z=\frac{1}{P_r}$

From the probability density function we know that

$$E\left(\frac{1}{P_r}\right)=E(Z)=\int_{\frac{1}{b_2}}^{\frac{1}{a_2}} zf_Z(z)dz=\int_{\frac{1}{b_2}}^{\frac{1}{a_2}} z\cdot\frac{1}{b_2-a_2}\frac{1}{z^2}dz=\frac{1}{b_2-a_2}\ln\left(\frac{b_2}{a_2}\right)$$

therefore

$$E(Eff_{SW}) = \frac{1}{2}\frac{a_1 + b_1}{b_2 - a_2}\ln\left(\frac{b_2}{a_2}\right) \tag{5-120}$$

Next, we will develop a formula for the variance of Eff_{SW}. By definition

$$Var(Eff_{SW}) = E(Eff_{SW}^2) - [E(Eff_{SW})]^2$$

From equation 5-120

$$Var(Eff_{SW}) = E(Eff_{SW}^2) - \left[\frac{1}{2}\frac{a_1 + b_1}{b_2 - a_2}\ln\left(\frac{b_2}{a_2}\right)\right]^2$$

It remains, then, to determine $E(Eff_{SW}^2)$; this will be done by the Mellin transform technique. Let

$$Q = Eff_{SW}^2 = \frac{I^2}{P_r^2} \Rightarrow E(Q) = E(Eff_{SW}^2) = E\left(\frac{I^2}{P_r^2}\right)$$

From theorem 5-13

$$M_Q(s) = M_I(2s-1)M_{P_r}(3-2s) \tag{5-121}$$

$$E(Q) = M_Q(2) = M_I(3)M_{P_r}(-1) \tag{5-122}$$

Since I and P_r are uniformly distributed random variables, from table 5-8

$$M_I(3) = \frac{1}{3(b_1 - a_1)}(b_1^3 - a_1^3) \text{ and } M_{P_r}(-1) = \frac{1}{(-1)(b_2 - a_2)}(b_2^{-1} - a_2^{-1})$$

Following some algebraic manipulation we have

$$E(Q) = \frac{1}{3}\frac{1}{b_2 a_2}\left(b_1^2 + b_1 a_1 + a_1^2\right)$$

Therefore

$$Var(Eff_{SW}) = \frac{1}{3}\frac{1}{b_2 a_2}\left(b_1^2 + b_1 a_1 + a_1^2\right) - \left[\frac{1}{2}\frac{a_1 + b_1}{b_2 - a_2}\ln\left(\frac{b_2}{a_2}\right)\right]^2 \tag{5-123}$$

Equations 5-120 and 5-123 are general formulas for the mean and variance of Eff_{SW}, if I and P_r are independent uniformly distributed random variables. Suppose we apply these formulas to example 5-17; this implies $a_1 = 50{,}000$,

$b_1 = 100{,}000$, $a_2 = 100$, $b_2 = 200$. Substituting these values into equations 5-120 and 5-123 yields:

$$E(Eff_{SW}) = 519.86 \text{ staff-months}$$
$$Var(Eff_{SW}) = 21411.8 \text{ (staff-months)}^2$$
$$\sigma_{Eff_{SW}} = \sqrt{Var(Eff_{SW})} = 146.328 \text{ staff-months}$$

b) In this part, formulas are developed for the mean and variance of Eff_{SW} if I and P_r are *each beta distributed.* From chapter 4 (equation 4-8) a random variable X is beta distributed with shape parameters α and β ($\alpha > 0$ and $\beta > 0$) if its probability density function is

$$f_X(x \mid \alpha, \beta) = \begin{cases} \dfrac{1}{b-a}\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\dfrac{x-a}{b-a}\right)^{\alpha-1}\left(\dfrac{b-x}{b-a}\right)^{\beta-1} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

Continuing with this case discussion, let

$$Z = Eff_{SW} = \frac{I}{P_r} \Rightarrow M_Z(s) = M_I(s) M_{P_r}(2-s)$$

therefore, $$E(Z) = E(Eff_{SW}) = M_Z(2) = M_I(2) M_{P_r}(0) \qquad (5\text{-}124)$$

The Mellin transform of X is, in this case,

$$M_X(s) = \int_a^b x^{s-1} f_X(x \mid \alpha, \beta)\, dx$$
$$= \frac{1}{b-a}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_a^b x^{s-1}\left(\frac{x-a}{b-a}\right)^{\alpha-1}\left(\frac{b-x}{b-a}\right)^{\beta-1} dx$$
$$= \frac{1}{(b-a)^{\alpha+\beta-1}}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_a^b x^{s-1}(x-a)^{\alpha-1}(b-x)^{\beta-1}\, dx \qquad (5\text{-}125)$$

We are given $I \sim Beta(\alpha_1, \beta_1, a_1, b_1)$. From theorem 4-4, we know that

$$M_I(2) = E(I) = a_1 + (b_1 - a_1)\frac{\alpha_1}{\alpha_1 + \beta_1} \qquad (5\text{-}126)$$

Given $P_r \sim Beta(\alpha_2, \beta_2, a_2, b_2)$, from equation 5-125

$$M_{P_r}(0) = \xi = \frac{1}{(b_2 - a_2)^{\alpha_2+\beta_2-1}} \frac{\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)} \int_{a_2}^{b_2} y^{-1}(y-a_2)^{\alpha_2-1}(b_2-y)^{\beta_2-1}\,dy \tag{5-127}$$

Substituting equations 5-126 and 5-127 into equation 5-124, we have

$$E(Z) = E(Eff_{SW}) = \xi\left(a_1 + (b_1 - a_1)\frac{\alpha_1}{\alpha_1+\beta_1}\right) \tag{5-128}$$

As an illustration, consider the case where $I \sim Beta(5, 10, 50(10)^3, 100(10)^3)$ and $P_r \sim Beta(5, 5, 100, 200)$. The expected effort $E(Eff_{SW})$ is

$$E(Eff_{SW}) = \xi\left(50(10)^3 + (100(10)^3 - 50(10)^3)\frac{5}{5+10}\right) = (66,666.67)\xi$$

where

$$\xi = \frac{1}{(100)^9} \frac{\Gamma(10)}{\Gamma(5)\Gamma(5)} \int_{100}^{200} y^{-1}(y-100)^4(200-y)^4\,dy = 0.0067358$$

Therefore

$$E(Eff_{SW}) = (66,666.67)(0.0067358) = 449.053 \text{ staff-months}$$

The value for ξ was determined by numerical integration.

A determination of $Var(Eff_{SW})$ completes this discussion. By definition

$$Var(Eff_{SW}) = E(Eff_{SW}^2) - \left[E(Eff_{SW})\right]^2$$

From equation 5-122

$$E(Eff_{SW}^2) = M_I(3) M_{P_r}(-1)$$

where

$$M_I(3) = \frac{1}{(b_1 - a_1)^{\alpha_1+\beta_1-1}} \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \int_{a_1}^{b_1} t^2(t-a_1)^{\alpha_1-1}(b_1-t)^{\beta_1-1}\,dt$$

$$M_{P_r}(-1) = \frac{1}{(b_2 - a_2)^{\alpha_2+\beta_2-1}} \frac{\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)} \int_{a_2}^{b_2} y^{-2}(y-a_2)^{\alpha_2-1}(b_2-y)^{\beta_2-1}\,dy$$

If $I \sim Beta(5,10,50(10)^3,100(10)^3)$ and $P_r \sim Beta(5,5,100,200)$, then a numerical integration of the above two integrals yields

$$M_I(3) = 4.47917(10)^9 \text{ and } M_{P_r}(-1) = 0.000045852$$

therefore

$$E(Eff_{SW}^2) = (4.47917(10)^9)(0.000045852) = 205378.9028 \text{ (staff-months)}^2$$

so,

$$Var(Eff_{SW}) = 3730.3 \text{ (staff-months)}^2$$

$$\sigma_{Eff_{SW}} = \sqrt{Var(Eff_{SW})} = 61.07 \text{ staff-months}$$

In summary, the effort mean and standard deviation (rounded) is

$$E(Eff_{SW}) = 449 \text{ staff-months}$$

$$\sigma_{Eff_{SW}} = 61 \text{ staff-months} \blacklozenge$$

Table 5-9. Transformation Formulas
Useful in Cost Uncertainty Analysis

Operation	Distribution	Transformation	A Cost Analysis Application
1. Multiplication of a random variable by a constant a $U = aX$	X is any distribution	$E(U) = aE(X)$ $Var(U) = a^2 Var(X)$	Many types of applications; a could represent a labor rate (dollars per staff-month) and X could represent an effort (in staff-months).
2. Addition of a constant a to a random variable $U = a + X$	X is any distribution	$E(U) = a + E(X)$ $Var(U) = Var(X)$	Many types of applications; a could represent a fixed cost, while X could represent a variable cost (whose precise value is uncertain).
3. Case A Sum of two uniform random variables $U = X_1 + X_2$	X_i's independent and $X_1 \sim Unif(a_1, b_1)$ $X_2 \sim Unif(a_2, b_2)$	$U \sim Trap((a_1 + a_2), (a_2 + b_1), (a_1 + b_2), (b_1 + b_2))$ if $b_1 - a_1 < b_2 - a_2$; $U \sim Trng((a_1 + a_2), m, (b_1 + b_2))$ if $b_1 - a_1 = b_2 - a_2$, where $m = \frac{1}{2}[(a_1 + a_2) + (b_1 + b_2)]$	Refer to exercise 8 in this chapter. Also refer to the discussion pertaining to figure 5-18.

Table 5-9. Transformation Formulas
Useful in Cost Uncertainty Analysis (Continued)

Operation	Distribution	Transformation	A Cost Analysis Application
3. Case B Sum of two normal random variables $U = X_1 + X_2$	X_i's independent and $X_1 \sim N(a_1, b_1)$ $X_2 \sim N(a_2, b_2)$	$U \sim N(a_1 + a_2, b_1 + b_2)$	Refer to chapter 6, section 6.2.2, for a further discussion.
4. Sum of n random variables $U = X_1 + X_2 + \ldots + X_n$	Central limit theorem conditions (theorem 5-10) are met.	U approaches a normal distribution as n becomes large, with $E(U) = E(X_1) + E(X_2) + \ldots + E(X_n)$ $Var(U) = Var(X_1) + Var(X_2) + \ldots + Var(X_n)$	Most common application is summing cost elements across a system's work breakdown structure (WBS). In this case, X_i represents the cost of the ith cost element in the WBS (refer to figure 1-3). Also, refer to chapter 6, section 6.2.2, for a further discussion.
5. The ratio of a single random variable $U = \frac{1}{X}$	$X \sim Unif(a, b)$	$f_U(u) = \frac{1}{b-a}\frac{1}{u^2} \quad \frac{1}{b} \le u \le \frac{1}{a}$	Refer to case discussion 5-4.

5. (Concluded)

The ratio of a single random variable $U = \frac{1}{X}$	$X \sim Unif(a, b)$	$E\left(\frac{1}{X}\right) = \frac{1}{b-a}\ln\left(\frac{b}{a}\right)$ $Var\left(\frac{1}{X}\right) = \frac{1}{ba} - \left(\frac{1}{b-a}\ln\left(\frac{b}{a}\right)\right)^2$	Refer to exercise 19 in this chapter.

6.

Product of two random variables $U = X_1 X_2$	Any distribution with X_i's *independent*; or $Cov(X_1, X_2) = 0$ $Cov(X_1^2, X_2^2) = 0$. In this case, let $E(X_1) = \mu_1,\ E(X_2) = \mu_2$ $Var(X_1) = \sigma_1^2,\ Var(X_2) = \sigma_2^2$	$E(U) = \mu_1\mu_2$ $Var(U) = (\sigma_1^2 + \mu_1^2)(\sigma_2^2 + \mu_2^2) - (\mu_1\mu_2)^2$	Refer to case discussion 5-3.

7.

Product of n random variables $U = X_1 X_2 \cdots X_{n-1} X_n$	Any distribution with X_i's *independent* $E(X_i) = \mu_i$ $Var(X_i) = \sigma_i^2$	U approaches the lognormal distribution as $n \to \infty$ $E(U) = \mu_1\mu_2 \cdots \mu_{n-1}\mu_n$ $Var(U) = \prod_{i=1}^{n}(\sigma_i^2 + \mu_i^2) - \prod_{i=1}^{n}\mu_i^2$	Refer to example 4-9.

Table 5-9. Transformation Formulas
Useful in Cost Uncertainty Analysis (Concluded)

Operation	Distribution	Transformation	A Cost Analysis Application
8. Case A (Uncorrelated) Ratio of two uncorrelated random variables $U = \frac{X_1}{X_2}$	Any distribution X_i's *independent* $E(X_1) = \mu_1,\ E(X_2) = \mu_2$ $Var(X_1) = \sigma_1^2,\ Var(X_2) = \sigma_2^2$ where $E\left(\frac{1}{X_2}\right)$ and $E\left(\frac{1}{X_2^2}\right)$ exist and are known.	$E(U) = \mu_1 E\left(\frac{1}{X_2}\right)$ $Var(U) = (\sigma_1^2 + \mu_1^2)E\left(\frac{1}{X_2^2}\right) - \left[\mu_1 E\left(\frac{1}{X_2}\right)\right]^2$	Refer to case discussion 5-4.
8. Case B (Correlated) [12] Ratio of two correlated random variables $U = \frac{X_1}{X_2}$	Any distribution X_i's *correlated* $E(X_1) = \mu_1,$ $E(X_2) = \mu_2\ (\mu_2 \neq 0)$ $Var(X_1) = \sigma_1^2,\ Var(X_2) = \sigma_2^2$ ρ is the correlation between X_1 and X_2	$E(U) \approx \frac{\mu_1}{\mu_2} + \frac{1}{(\mu_2)^2}\left(\sigma_2^2 \frac{\mu_1}{\mu_2} - \rho\sigma_1\sigma_2\right)$ $Var(U) \approx \sigma_2^2 \frac{(\mu_1)^2}{(\mu_2)^4} + \frac{\sigma_1^2}{(\mu_2)^2} - 2\rho\sigma_1\sigma_2 \frac{\mu_1}{(\mu_2)^3}$	Refer to case discussion 5-4, but with I and P_r treated as *correlated* random variables.

Exercises

1. In example 5-2, $Eff_{SysTest} = XY$ and X and Y have joint probability density function

$$f(x,y) = \begin{cases} \dfrac{1}{240} & 5 \le x \le 15, \quad 12 \le y \le 36 \\ 0 & \text{otherwise} \end{cases}$$

a) Sketch the event spaces associated with events A, B, and C where

$$A = \{Eff_{SysTest} \le 240\}$$

$$B = \{Eff_{SysTest} \le 240 \mid X \le 12\}$$

$$C = \{\{Eff_{SysTest} \le 240\} \cap \{X \le 12\} \cap \{Y \le 20\}\}$$

b) From part a) compute $P(A)$, $P(B)$, and $P(C)$.

2. In example 5-3, $Eff_{SW} = \dfrac{X}{Y}$ and X and Y have joint probability density function

$$f(x,y) = \begin{cases} \dfrac{1}{5(10^6)} & 50{,}000 \le x \le 100{,}000, \quad 100 \le y \le 200 \\ 0 & \text{otherwise} \end{cases}$$

Find

a) $P(Eff_{SW} \le 313)$

b) $P(Eff_{SW} \le 410 \mid X \le 70{,}000)$

c) $P(\{Eff_{SW} \le 410\} \cap \{X \le 70{,}000\} \cap \{Y \ge 150\})$

3. Suppose $f(x,y) = \begin{cases} \dfrac{1}{240} & 5 \le x \le 15, \quad 12 \le y \le 36 \\ 0 & \text{otherwise} \end{cases}$

Compute

a) $f_X(x)$ using equation 5-10

b) $f_Y(y)$ using equation 5-11

c) $P(X \le 10 | Y = 24)$

d) $P(Y > 24 | X = 10)$

e) Are X and Y dependent or independent random variables? Justify your answer.

4. a) If X and Y are random variables with means μ_X and μ_Y, show that

$$Cov(X,Y) = E(XY) - \mu_X \mu_Y$$

b) If X and Y are random variables, show that $Cov(X,Y) = Cov(Y,X)$; for any real numbers a, b, c, and d show that

$$Cov(aX + b, cY + d) = acCov(X,Y)$$

c) Show that $Cov(X,Y) = 0$ if X and Y are *independent* random variables.

d) Show that $Cov(X,X) = Var(X)$. Given this, show that $Corr(X,X) = 1$.

e) Show that $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$.

5. Suppose Y, X_1, and X_2 are independent random variables. If $Z = X_1 + X_2$ show that Y and Z are uncorrelated.

6. If $Y = X^{100}$ and $X \sim Unif(0,1)$ show that $\rho_{X,Y} = 0.24$.

7. Let the total cost of a system's prime mission equipment (PME) be denoted by $Cost_{PME}$. Let $Cost_{PME} = X_1 + X_2$ where X_1 is the total cost of the system's hardware and X_2 is the total cost of the system's software. Assume X_1 and X_2 are independent random variables. Suppose the cost to integrate and assemble the system's hardware and software is denoted by $Cost_{I\&A}$. If $Cost_{I\&A} = \frac{1}{10} X_1 + \frac{1}{5} X_2$

a) Determine a general formula for $Corr(Cost_{PME}, Cost_{I\&A})$.

b) Compute $Corr(Cost_{PME}, Cost_{I\&A})$ when $\sigma_{X_1} = \sigma_{X_2}$.

8. Let $Cost_{PMP}$ denote the total cost of a system's prime mission product (PMP). Let $Cost_{PMP} = Cost_{PME} + Cost_{I\&A}$, where $Cost_{PME} = X_1 + X_2$ and $Cost_{I\&A} = \frac{1}{10}X_1 + \frac{1}{5}X_2$. Let X_1 and X_2 denote the total costs (\$M) of the system's hardware and software. If X_1 and X_2 are independent random variables with $X_1 \sim Unif(5,10)$ and $X_2 \sim Unif(30,45)$ compute

a) $E(Cost_{PMP})$

b) $Var(Cost_{PMP})$

c) $F_{Cost_{PME}}(x_1 + x_2)$ (refer to table 5-9 for the distribution function of the sum of two independent uniformly distributed random variables).

d) Using $F_{Cost_{PME}}(x_1 + x_2)$, determine d such that $P(Cost_{PME} \le d) = 0.75$.

9. Suppose X_1, X_2, X_3 are the cost element costs of an electronic system. Let the system's total cost be given by $Cost_{Sys} = X_1 + X_2 + X_3$, where X_1, X_2, X_3 are given in the table below. Let X_1 and W be independent random variables.

Cost Element Name	Cost Element Cost X_i (\$M)	Distribution of X_i or the Applicable Functional Relationship
Prime Mission Product (PMP)	X_1	$N(12.5, 6.6)$
System Eng. & Prgm Mgt (SEPM)	X_2	$X_2 = \frac{1}{2}X_1$
System Test & Evaluation (STE)	X_3	$X_3 = \frac{1}{4}X_1 + \frac{1}{8}X_2 + W$, where $W \sim Unif(0.6, 1.0)$

a) Write a general formula for $E(Cost_{Sys})$ and compute its value.

b) Show that $Var(Cost_{Sys}) = \frac{841}{256} Var(X_1) + Var(W)$ from the expression

$$Var(Cost_{Sys}) = Var(X_1) + Var(X_2) + Var(X_3) + 2[Cov(X_1, X_2) + Cov(X_1, X_3) + Cov(X_2, X_3)]$$

c) Compute $Var(Cost_{Sys})$.

10. In case discussion 5-1, the K-S test revealed the normal distribution as a plausible model of the underlying distribution function for $Cost_{SSA}$. Use the K-S test on the data in table 5-4 to show the lognormal distribution is also a plausible model.

11. In example 5-7, X denoted the number of engineering staff required to test a new rocket propulsion system. The number of months Y required to design, conduct, and analyze the test was given by $Y = 2X + 3$. If X is uniformly distributed in the interval $5 \le x \le 15$, determine

a) $F_Y(y)$ b) $f_Y(y)$

12. In example 5-10, verify that $F_W(1750) = 0.75 = F_{Hours}(87.67)$.

13. Suppose the direct engineering hours to design a new communication satellite is given by $Hours = 4 + 2\sqrt{W}$, where W is the satellite's weight, in pounds. Suppose the uncertainty in the satellite's weight is captured by a triangular distribution; that is, $W \sim Trng(1000, 1500, 2000)$. Suppose the satellite design team assessed 1500 pounds to be the point estimate for weight; that is, $w_{PE} = 1500$.

a) Determine the cumulative distribution function of *Hours*.

b) Compute $P(Hours \le h_{PE})$, where $h_{PE} = 4 + 2\sqrt{w_{PE}}$.

c) Determine the probability density function of *Hours*.

14. Suppose the development effort Eff_{SW} for a software project is defined by $Eff_{SW} = c_1 I^{c_2}$. If $I \sim Trng(a,m,b)$ derive $F_{Eff_{SW}}(s)$, $f_{Eff_{SW}}(s)$, $E(Eff_{SW})$, $Var(Eff_{SW})$.

15. Suppose the development schedule for a software project is defined by $T_{SW} = k_1 (Eff_{SW})^{k_2}$, where $Eff_{SW} = c_1 I^{c_2}$. Answer the following:

    a) If $I \sim Unif(a,b)$ derive $F_{T_{SW}}(t)$, $f_{T_{SW}}(t)$, $E(T_{SW})$, $Var(T_{SW})$.

    b) If $I \sim Trng(a,m,b)$ derive $F_{T_{SW}}(t)$, $f_{T_{SW}}(t)$, $E(T_{SW})$, $Var(T_{SW})$.

16. In example 5-13, the effort (staff-months) to develop software for a new system was given by $Eff_{SW} = 2.8 I^{1.2}$. The development schedule (months) was given by $T_{SW} = 2.5(Eff_{SW})^{0.32}$. If $I \sim Unif(30,80)$, use theorem 5-11 to show the following:

    a) $F_{Eff_{SW}}(518) = F_{T_{SW}}(18.5) = 0.95$

    b) $F_{T_{SW}}(t) = \frac{1}{50}\left[\left(\frac{t}{3.48}\right)^{\frac{1}{0.384}} - 30\right] \quad 12.8 \le t \le 18.7$

17. The uncertainties in the amount of code to develop for the radar system in example 5-16, was represented by the independent random variables $I_1, I_2, I_3, \ldots, I_{14}$. Let $I_{Total} = I_1 + I_2 + I_3 + \ldots + I_{14}$, where each I is in thousands of delivered source instructions (KDSI). From the information in table 5-7, use the central limit theorem to determine the 0.25-fractile and the 0.75-fractile of $Eff_{SW} = 2.8(I_{Total})^{1.2}$.

18. Refer to example 5-2 and use theorem 5-12 to find the general formula for the probability density function of $Eff_{SysTest}$.

19. a) Let X and Y be independent random variables with $Z=(X+Y)^2$. Show that $E(Z)=M_X(3)+M_Y(3)+2M_X(2)M_Y(2)$.

b) Suppose $X \sim Unif(a,b)$. Use theorem 5-13 and the definition of the Mellin transform (equation 5-96) to show that

$$E\left(\frac{1}{X}\right)=\frac{1}{b-a}\ln\left(\frac{b}{a}\right) \text{ and } Var\left(\frac{1}{X}\right)=\frac{1}{ba}-\left(\frac{1}{b-a}\ln\left(\frac{b}{a}\right)\right)^2$$

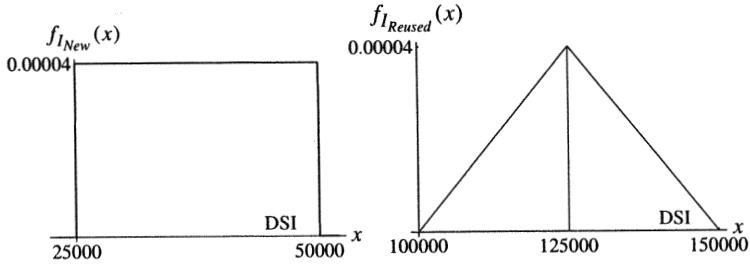
20. In example 5-19, a new software application was being developed that consisted of a mixture of new code I_{New} and reused code I_{Reused}. Suppose I_{New} and I_{Reused} are independent random variables with probability density functions given in example 5-19 (shown below for convenience). If the effort Eff_{SW} associated with developing the application is a function of the equivalent size I_{Equiv}, where

$$I_{Equiv}=I_{New}+I_{Reused}^{0.857}$$

and

$$Eff_{SW}=2.8\left(\frac{1}{1000}I_{Equiv}\right)^{1.2}$$

use the Mellin transform technique to approximate $E(Eff_{SW})$.



Hint: Use the first three terms of the binomial series expansion of $(I_{Equiv})^{1.2}$, given by

$$(I_{Equiv})^{1.2} \approx (I_{New})^{1.2}+1.2(I_{New})^{0.2}I_{Reused}^{0.857}+0.12(I_{New})^{-0.8}I_{Reused}^{1.714}$$

References

1. Lurie, P. M., and M. S. Goldberg. 1993. *A Handbook of Cost Risk Analysis Methods*, P-2734. Alexandria, Virginia: The Institute for Defense Analyses.

2. Quirin, W. L. 1978. *Probability and Statistics*. New York: Harper and Row, Publishers, Inc.

3. Garvey, P. R. 1990. A General Analytic Approach to System Cost Uncertainty Analysis, in W. R. Greer, Jr., and D. A. Nussbaum (eds.). *Cost Estimating and Analysis: Tools and Techniques*. pp. 161-181. New York: Springer-Verlag.

4. Law, A. M., and W. D. Kelton. 1991. *Simulation Modeling and Analysis*, 2nd ed. New York: McGraw-Hill, Inc.

5. Boehm, B. W. 1981. *Software Engineering Economics*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

6. Garvey, P. R., and F. D. Powell. 1989. Three Methods for Quantifying Software Development Effort Uncertainty, in B. W. Boehm (ed.). *Software Risk Management*. pp. 292-306. Washington, DC: IEEE Computer Society Press.

7. Mood, A. M., F. A. Graybill, and D. C. Boes. 1974. *Introduction to the Theory of Statistics*, 3rd ed. New York: McGraw-Hill, Inc.

8. Cramer, H. 1966. *Mathematical Methods of Statistics*. Princeton, New Jersey: Princeton University Press.

9. Epstein, B. 1948. Some Applications of the Mellin Transform in Statistics. *Annals of Mathematical Statistics*, Vol. 19, pp. 370-379.

10. Giffin, W. C. 1975. *Transform Techniques for Probability Modeling*. New York: Academic Press, Inc.

11. Conte, S. D., H. E. Dunsmore, and V. Y. Shen. 1986. *Software Engineering Metrics and Models*. Menlo Park, California: The Benjamin/Cummings Publishing Company, Inc.

12. Rice, J. A. 1995. *Mathematical Statistics and Data Analysis*, 2nd ed. Belmont, California: Duxbury Press.

Chapter 6

System Cost Uncertainty Analysis

A reasonable probability is
the only certainty.
Edgar Watson Howe
Country Town Sayings [1911]

Our wisdom and deliberation for the
most part follow the lead of chance.
Michel Eyquem de Montaigne
Essays [1580]

This chapter illustrates how key concepts developed thus far combine to produce the probability distribution of a system's total cost. Chapter 7 will extend this discussion to the joint and conditional distributions of a system's total cost and schedule. Chapter 6 begins with an introduction to the work breakdown structure, a primary method for determining a system's total cost.

6.1 Work Breakdown Structures

The work breakdown structure (WBS) is a framework for identifying all elements of cost that relate to the tasks and activities of developing, producing, deploying, sustaining, and disposing a system. Work breakdown structures are unique to the system under consideration. They are developed according to the specific requirements and functions the system has to perform. Work breakdown structures are defined for classes of systems. These classes include electronic systems, aircraft systems, surface vehicles, ship systems, and spacecraft systems [1,2].

Work breakdown structures are tiered by a hierarchy of cost elements. A typical electronic system WBS is illustrated in figure 6-1. Shown are four hierarchies, or indenture levels. The first level represents the entire system (e.g., the air traffic control radar system). The second level reflects the major cost elements of the system. In figure 6-1, these elements include prime

mission product (PMP), system engineering, program management, and system test and evaluation.

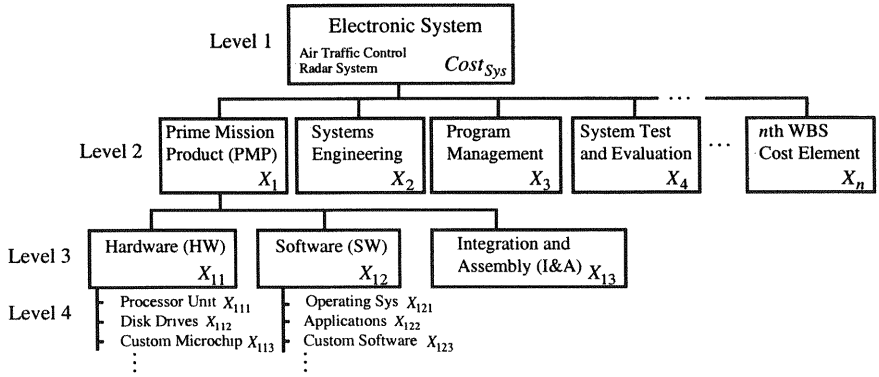


Figure 6-1. An Illustrative Electronic System WBS

The following defines each level 2 cost element.

- *Prime Mission Product (PMP)* — This element refers to the hardware and software used to accomplish the primary mission of the system. It includes the engineering effort and management activities associated with the system's individual hardware components and software functions, as well as the effort to integrate, assemble, test, and checkout the system's hardware and software.

- *Systems Engineering* — This element encompasses the overall engineering effort to define and deploy the system. It includes integrating the technical efforts of design engineering, specialty engineering (e.g., reliability engineering, security engineering), production engineering, and integrated test planning to produce an operational system.

- *Program Management* — This element includes all effort associated with the business and administrative management of the system. This includes cost, schedule, and performance measurement, as well as contract administration, data management, and customer/user liaison activities.

- *System Test and Evaluation* — This element includes all test engineering, test planning, and related technical efforts (test mockups, prototypes) to insure the deployed system has been tested against its requirements.

In figure 6-1, the PMP cost element is divided into its level 3 cost elements. At this level, the radar's hardware, software, and integration cost elements are defined. A further division of PMP into its level 4 cost elements is also shown in figure 6-1. Here, the individual cost elements of the system's hardware and software are defined. In practice, the number of levels specified in a system's WBS reflects the extent the system itself is defined. In most instances, cost elements are seldom specified below level 6 in a system's work breakdown structure.

Certain cost elements in a WBS qualify as *configuration items.* A configuration item is an aggregation of hardware or software that satisfies a particular end-use function of the system. A custom made microchip or developed software applications are typically designated as configuration items. This designation means the item is subject to *configuration management.* Configuration management is the process of documenting, monitoring, and controlling change to the configuration item's technical baseline. Cost elements placed under configuration management typically begin to appear at level 4 of a WBS.

The WBS is the definitive cost element structure of a system. It is the basis upon which the system's cost is determine (or modeled). From a WBS perspective a system's total cost (which we will denote by $Cost_{Sys}$) is a summation of cost element costs, summed across the levels of the WBS. In figure 6-1,

$$Cost_{Sys} = X_1 + X_2 + X_3 + X_4 + \ldots + X_n \qquad (6\text{-}1)$$

where the first term in equation 6-1, X_1, is

$$X_1 = X_{11} + X_{12} + X_{13} + \ldots + X_{1k} \qquad (6\text{-}2)$$

and k is the number of level 3 cost elements associated with X_1. Similarly,

$$X_{11} = X_{111} + X_{112} + X_{113} + \ldots + X_{11j} \qquad (6\text{-}3)$$

where j is the number of level 4 cost elements associated with X_{11}. The other

terms in equation 6-1, $X_2, X_3, X_4, \ldots X_n$, are defined in a similar manner. This layered sum of cost element costs is often referred to as the "roll-up" cost.

Cost elements of a work breakdown structure are specific to the system class. Cost elements* of a satellite system are illustrated in figure 6-2.

1 Satellite System
- 1.1 Launch Vehicle Segment
- 1.2 Space Segment
  - 1.2.1 Satellite Integration, Assembly, and Test
  - 1.2.2 Spacecraft Bus
    - 1.2.2.1 Spacecraft Bus Integration, Assembly, and Test
    - 1.2.2.2 Structures and Mechanical Assembly Subsystem
    - 1.2.2.3 Attitude Determination and Control Subsystem
    - 1.2.2.4 Thermal Control Subsystem
    - 1.2.2.5 Electrical Power Subsystem
    - 1.2.2.6 Telemetry and Communication Subsystem
    - 1.2.2.7 Propulsion Subsystem
  - 1.2.3 Payload
    - 1.2.3.1 Payload Hardware
    - 1.2.3.2 Payload Software
- 1.3 Command, Control, and Communications Segment
- 1.4 Systems Engineering and Program Management
- 1.5 System Test and Evaluation
- 1.6 Peculiar Support Equipment
- 1.7 Common Support Equipment
- 1.8 Operations and Support
- 1.9 Flight Support Operations
- 1.10 Program Office

Figure 6-2. Illustrative Spacecraft WBS

Notice the difference between these cost elements and those of the electronic system WBS, shown in figure 6-1. In the satellite system, its cost elements are grouped into segments. Within segments, these elements are divided into

---

* Cost element indenture levels are identified by numbering conventions that may or may not incorporate decimals. The convention used is a matter of presentation style.

levels. Levels can reflect subsystems, such as the spacecraft bus (platform) in figure 6-2. For context, the spacecraft bus elements are defined below.

- *Spacecraft Bus Integration, Assembly, and Test* — This element refers to all efforts associated with the cost of integrating, assembling, and testing the individual subsystems that constitute the spacecraft bus.

- *Structures and Mechanical Assembly Subsystem* — This element (subsystem) refers to the central frame of the spacecraft that provides support and mounting surfaces for all equipment. It includes deployment mechanisms, the solar array boom, experimental booms, antenna supports, and mechanical design equipment.

- *Attitude Determination and Control Subsystem* — This element (subsystem) measures and maintains the orientation of the space vehicle relative to an inertial or external reference. Attitude determination components include inertial measurement devices (e.g., gyroscopes, accelerometers), earth sensors, sun sensors, horizon sensors, and magnetometers. Attitude control adjusts and maintains the space vehicle's attitude and stabilization. Attitude control components include fuel lines, fuel tanks, thrusters, inertia wheels, and any associated electronics.

- *Thermal Control Subsystem* — This element (subsystem) maintains the temperature of the spacecraft and mission payload through heat transfer between space vehicle elements. Thermal control techniques may be passive or active. Passive techniques include special paint, mirrors, and insulation. Active techniques include heat pipes, louvers, and heaters.

- *Electrical Power Subsystem* (EPS) — This element (subsystem) generates, converts, regulates, stores, and distributes electrical power between major space vehicle subsystems. Two common types of EPS's are solar and electrochemical. Typical components of the EPS include solar array for power generation, batteries for power storage, as well as wiring harnesses, regulators, switching electronics, converters, and components for power conditioning and distribution.

- *Telemetry and Communication Subsystem* — This element (subsystem) measures the space vehicle's conditions (health and status), processes health and status data and mission data, stores and transmits data to ground receivers, as well as receives, processes, and initiates commands from ground controllers. This subsystem also maintains the track of the space vehicle; typical components include data processors, transmitters, receivers, antennas, decoders, amplifiers, and tape recorders.

- *Propulsion Subsystem* — This element (subsystem), also referred to as Apogee Kick Motor (AKM), provides reaction force for the final maneuver into orbit and for orbit changes. Typical components include solid rocket motor and explosive squibs, nozzle control mechanisms, thrust sensing and shut-down controls, as well as any required cabling, wiring, and plumbing.

As mentioned earlier, a system's work breakdown structure is tailored from general work breakdown structures specific to the system's class. The satellite system WBS in figure 6-2 was tailored from the general WBS for the Unmanned Space Vehicle Cost Model (USCM) [3]. The USCM WBS is presented in figure 6-3.

1 Space Vehicle
- 1.1 Integration, Assembly, & System Test (IA&T)
- 1.2 Spacecraft
  - 1.2.1 Structure, Interstage/Adapter
  - 1.2.2 Thermal Control
  - 1.2.3 Attitude Determination Control System (ADCS)
    - 1.2.3.1 Attitude Determination
    - 1.2.3.2 Reaction Control System
  - 1.2.4 Electrical Power Supply (EPS)
    - 1.2.4.1 Power Generation
    - 1.2.4.2 Power Storage
    - 1.2.4.3 Power Conditioning and Distribution (PCD)
  - 1.2.5 Telemetry, Tracking, and Command
    - 1.2.5.1 Transmitter
    - 1.2.5.2 Receiver/Exciter
    - 1.2.5.3 Transponder
    - 1.2.5.4 Digital Electronics (Signal/Data Processor)
    - 1.2.5.5 Analog Electronics
    - 1.2.5.6 Antennas
    - 1.2.5.7 RF Distribution
- 1.3 Communications Payload
  - 1.3.1 Transmitter
  - 1.3.2 Receiver/Exciter
  - 1.3.3 Transponder
  - 1.3.4 Digital Electronics (Signal/Data Processor)
  - 1.3.5 Analog Electronics
  - 1.3.6 Antennas
  - 1.3.7 RF Distribution
- 1.4 Program-Level
  - 1.4.1 Program Management
  - 1.4.2 Systems Engineering
  - 1.4.3 Data

2 Aerospace Ground Equipment

3 Launch and Orbital Operations and Support

Figure 6-3. Unmanned Spacecraft WBS [3]

Work breakdown structures can be quite complex. They may involve many segments and levels, as well as numerous cost elements. Because the WBS is the basis for deriving a system's cost, WBS's may also contain a variety of mathematical relationships. These relationships are traditionally known as cost estimating relationships (CERs).* Their primary purpose is to generate point estimate costs of various WBS cost elements. Table 6-1 illustrates some spacecraft-related CERs.

Table 6-1. Illustrative CERs for Spacecraft Cost Elements [4]
(Nonrecurring Development Costs, FY92 ($K))

Cost Element	Input Parameters	CER
Attitude Control-Attitude Determination	Z_1 = Dry Weight (kg)	$X_{AttDeterm} = 3330 Z_1^{0.46}$
Telemetry, Tracking, & Command	Z_1 = Weight (kg)	$X_{TT\&C} = 1955 + 199 Z_1$
Structure/Thermal	Z_1 = Weight (kg)	$X_{S/T} = 2640 + 416 Z_1^{0.66}$
Electrical Power Supply (EPS)	Z_1 = EPS Weight (kg) Z_2 = Beginning of Life Power (kg·watts)	$X_{EPS} = 5303 + 0.108 (Z_1 Z_2)^{0.97}$
Payload Communication Electronics	Z_1 = Weight (kg)	$X_{Comm} = 917 Z_1^{0.70}$

In summary, a work breakdown structure provides the framework for developing a system's cost. It further serves as the framework for an analysis of the system's cost uncertainty. The complexity of these analyses is dictated by the complexity of the WBS and its associated CERs.

The following illustrates how probability methods are applied to the problem of quantifying a system's cost uncertainty within the framework of the WBS. Case discussions are presented to link theory to practice.

---

* Most CERs are statistically derived from data on cost and technical characteristics. This book uses the term CER to include those that are logically based, as well as those developed by statistical methods.

6.2 An Analytical Framework

This section focuses on the application of probability methods for quantifying the uncertainty in a system's cost. The WBS will provide the analytical framework for quantifying this uncertainty, which is expressed as a probability distribution. Analytical methods from probability theory are stressed. Analytical methods provide insight into problem structure and subtleties not always apparent from empirically based methods, such as Monte Carlo simulation.*

6.2.1 Computing the System Cost Mean and Variance

From equation 6-1, we see that system cost, denoted by $Cost_{Sys}$, is a summation of work breakdown structure cost element costs. Illustrated in figure 6-4, we define $Cost_{Sys}$ as

$$Cost_{Sys} = X_1 + X_2 + X_3 + \ldots + X_n \tag{6-4}$$

If $X_1, X_2, X_3, \ldots, X_n$ are independent, then from theorem 5-7 and theorem 5-8

$$E(Cost_{Sys}) = E(X_1) + E(X_2) + E(X_3) + \ldots + E(X_n) \tag{6-5}$$

$$Var(Cost_{Sys}) = Var(X_1) + Var(X_2) + Var(X_3) + \ldots + Var(X_n) \tag{6-6}$$

If $X_1, X_2, X_3, \ldots, X_n$ are *not independent*, then

$$Var(Cost_{Sys}) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \rho_{X_i,X_j}\sigma_{X_i}\sigma_{X_j} \tag{6-7}$$

Equations 6-5 through 6-7 are the formal expressions for the mean and variance of $Cost_{Sys}$. The following case discussions illustrate how these expressions are used.

---

* Monte Carlo simulation is an empirical method often used for quantifying cost uncertainty. The concept underlying this method is discussed in section 6.3.

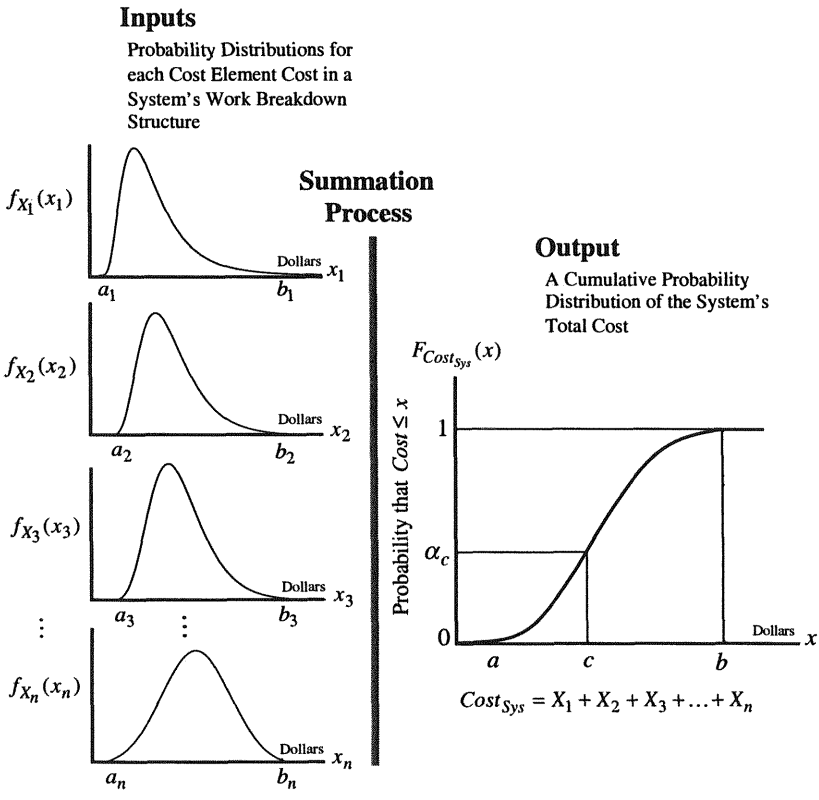


Figure 6-4. Cumulative Probability Distribution of $Cost_{Sys}$

Case Discussion 6-1 [5] Suppose the cost element costs $X_1, X_2, X_3, \ldots, X_{10}$ of an electronic system are given by the WBS in table 6-2. Let

$$Cost_{Sys} = X_1 + X_2 + X_3 + \ldots + X_{10}$$

Suppose the random variables $X_1, W, X_5, X_7, X_8, X_9$ (defined in table 6-2) are independent.

a) Compute $E(Cost_{Sys})$ and $Var(Cost_{Sys})$.

b) What distribution function approximates the distribution of $Cost_{Sys}$?

c) Find the value of $Cost_{Sys}$ that has a 5 percent chance of being exceeded.

Table 6-2. WBS for Case Discussion 6-1

Cost Element Name	Cost Element Cost X_i ($M)	Distribution of X_i or the Applicable Functional Relationship
Prime Mission Product (PMP)	X_1	$N(12.5, 6.6)$
System Engineering and Program Management (SEPM)	X_2	$X_2 = \frac{1}{2} X_1$
System Test & Evaluation (STE)	X_3	$X_3 = \frac{1}{4} X_1 + \frac{1}{8} X_2 + W$, where $W \sim Unif(0.6, 1.0)$
Data and Technical Orders	X_4	$X_4 = \frac{1}{10} X_1$
Site Survey and Activation	X_5	$Trng(5.1, 6.6, 12.1)$
Initial Spares	X_6	$X_6 = \frac{1}{10} X_1$
System Warranty	X_7	$Unif(0.9, 1.3)$
Early Prototype Phase	X_8	$Trng(1.0, 1.5, 2.4)$
Operations Support	X_9	$Trng(0.9, 1.2, 1.6)$
System Training	X_{10}	$X_{10} = \frac{1}{4} X_1$

a) It is given that

$$Cost_{Sys} = X_1 + X_2 + X_3 + \ldots + X_{10} \tag{6-8}$$

Using the relationships given in table 6-2, equation 6-8 is equivalent to

$$Cost_{Sys} = X_1 + \tfrac{1}{2} X_1 + \left(\tfrac{1}{4} X_1 + \tfrac{1}{8} X_2 + W\right) + \tfrac{1}{10} X_1 + X_5 + \tfrac{1}{10} X_1 + X_7 + X_8 + X_9 + \tfrac{1}{4} X_1$$

Combining the above terms yields

$$Cost_{Sys} = \tfrac{181}{80} X_1 + W + X_5 + X_7 + X_8 + X_9 \tag{6-9}$$

From theorem 5-7 (and equation 6-5)

$$E(Cost_{Sys}) = \tfrac{181}{80} E(X_1) + E(W) + E(X_5) + E(X_7) + E(X_8) + E(X_9) \tag{6-10}$$

From theorem 5-8 (and equation 6-6)

$$Var(Cost_{Sys}) = \left(\frac{181}{80}\right)^2 Var(X_1) + Var(W) + Var(X_5) + Var(X_7) + Var(X_8) + Var(X_9) \qquad (6\text{-}11)$$

since X_1, W, X_5, X_7, X_8, and X_9 are independent random variables. To compute the mean and variance of $Cost_{Sys}$ we need the means and variances of X_1, W, X_5, X_7, X_8, and X_9. Table 6-3 presents these statistics.

Table 6-3. Cost Statistics for X_1, W, X_5, X_7, X_8, and X_9

Cost Element Cost X_i ($M)	$E(X_i)$ ($M)	$Var(X_i)$ ($M)^2
X_1	12.500	6.6
W	0.800	0.16/12
X_5	7.933	40.75/18
X_7	1.100	0.16/12
X_8	1.633	1.51/18
X_9	1.233	0.37/18

The statistics in table 6-3 were determined by distribution-specific formulas given in chapter 4. For instance, since $X_1 \sim N(12.5, 6.6)$ we know from theorem 4-6 that $E(X_1) = 12.5$ and $Var(X_1) = 6.6$. Since W is a uniform distribution, from theorem 4-2

$$E(W) = \frac{0.6+1}{2} = 0.8 \quad \text{and} \quad Var(W) = \frac{(1-0.6)^2}{12} = \frac{0.16}{12} = 0.01\overline{333}$$

Since X_5 is a triangular distribution, from theorem 4-3

$$E(X_5) = \tfrac{1}{3}(5.1 + 6.6 + 12.1) = 7.933$$

$$Var(X_5) = \tfrac{1}{18}\left[(6.6-5.1)(6.6-12.1) + (12.1-5.1)^2\right] = 40.75/18$$

Substituting the data in table 6-3 into equations 6-10 and 6-11 we obtain

$$E(Cost_{Sys}) = 40.98 \text{ (\$M)} \tag{6-12}$$

$$Var(Cost_{Sys}) = 36.18 \text{ (\$M)}^2 \tag{6-13}$$

b) To approximate the distribution function of $Cost_{Sys}$, observe the following. First, the random variables X_1, W, X_5, X_7, X_8, and X_9 are independent. Hence, the central limit theorem will *affect* the shape of the distribution of $Cost_{Sys}$. Second, the random variables X_2, X_3, X_4, X_6, and X_{10} are highly correlated to X_1, which is $N(12.5, 6.6)$. It can be shown that

$$\rho_{X_v, X_1} = 1 \quad (v = 2,4,6,10) \text{ and } \rho_{X_3, X_1} = 0.9898$$

Thus, it is reasonable to conclude (for this case) the distribution function for $Cost_{Sys}$ is approximately normal — with mean and variance given by equations 6-12 and 6-13, respectively. The cumulative distribution function for $Cost_{Sys}$, assumed to be approximately normal, is shown in figure 6-5.

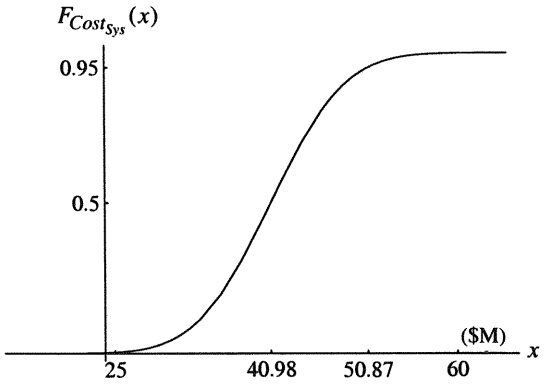


Figure 6-5. Assumed Normal CDF for $Cost_{Sys}$

c) In figure 6-5, note that $P(Cost_{Sys} \leq 50.87) = 0.95$. This means a value of 50.87 (\$M) for $Cost_{Sys}$ has only a 5 percent chance of being exceeded. To

arrive at this value, it is necessary to find x such that $P(Cost_{Sys} \leq x) = 0.95$. From equation 4-22

$$P\left(\frac{Cost_{Sys} - E(Cost_{Sys})}{\sigma} \leq \frac{x - E(Cost_{Sys})}{\sigma}\right) = 0.95$$

$$= P\left(Z \leq \frac{x - 40.98}{6.015}\right) = 0.95 \qquad (6\text{-}14)$$

Since $Cost_{Sys} \sim N(40.98, 36.18)$, from table A-1 (appendix A)

$$\frac{x - 40.98}{6.015} = 1.645$$

and $x = 50.87$. Thus, a value of 50.87 ($M) for $Cost_{Sys}$ has only a 5 percent chance of *being exceeded*. Equivalently, 50.87 ($M) is the 0.95-fractile (i.e., $x_{0.95} = 50.87$) of $Cost_{Sys}$. Furthermore, we can say the cost reserve (refer to chapter 1) needed for a 95 percent chance of *not exceeding* 50.87 ($M) is 9.9 ($M) above the expected cost of the system.

Further Considerations

Distribution Function of $Cost_{Sys}$ — In case discussion 6-1, it was assumed the distribution function for $Cost_{Sys}$ could be approximated by a normal distribution. How reasonable is this assumption? A series of 20 "points" is shown in figure 6-5a. These points reflect random statistical samples (values) of $Cost_{Sys}$, sampled by Monte Carlo simulation (explained in section 6.3). The curve implied by these "points" represents the simulated distribution function* for $Cost_{Sys}$. The curve given by the solid line in figure 6-5a is the assumed normal distribution for $Cost_{Sys}$ — as shown in figure 6-5. With this

* The simulated distribution is an empirically developed distribution. In establishing this distribution, no assumption is made that the distribution function for $Cost_{Sys}$ is normal.

in mind, observe in figure 6-5a how closely the simulated distribution for $Cost_{Sys}$ matches the assumed normal distribution.

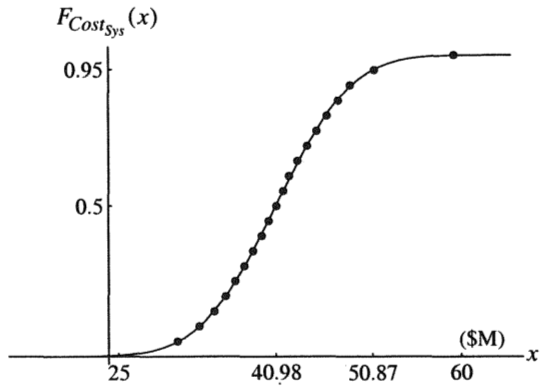


Figure 6-5a. Assumed Normal CDF for $Cost_{Sys}$ (defined by the solid line) vs the Simulated CDF (defined by the points)

The closeness with which these "points" fall along the curve given by the solid line, in figure 6-5a, *visually suggests* the reasonableness of the assumption that the distribution function for $Cost_{Sys}$ can be *approximated* by a normal. Although this is a practical conclusion, it is an informal one. A more formal conclusion could be derived from the Kolmogorov-Smirnov (K-S) test, illustrated in case discussion 5-1. This would reveal whether the normal distribution is a *statistically plausible model* for the underlying distribution function of $Cost_{Sys}$, in this case.

Correlation — In case discussion 6-1, a significant amount of correlation exists between certain pairs of cost element costs. In table 6-2, the five cost element costs X_2, X_3, X_4, X_6, and X_{10} were functionally related to X_1, the system's PMP cost. In particular, X_2, X_4, X_6, and X_{10} were linearly related to X_1 by the expression

$$X_v = a_v X_1 \tag{6-15}$$

where $v = 2,4,6,10$, $a_2 = 1/2$, $a_4 = a_6 = 1/10$, and $a_{10} = 1/4$. In table 6-2, cost element cost X_3 was a linear combination of X_1, X_2, and W; specifically,

$$X_3 = \tfrac{1}{4} X_1 + \tfrac{1}{8} X_2 + W \tag{6-16}$$

where X_1 and W were given to be independent random variables and $X_2 = \frac{1}{2} X_1$. The functional relationships given by equations 6-15 and 6-16 imply the following correlations:

$\rho_{X_v, X_1} = 1$ for $v = 2,4,6,10$

$\rho_{X_1, W} = 0$ from theorem 5-3

$\rho_{X_2, W} = \rho_{\frac{1}{2}X_1, W} = \rho_{X_1, W} = 0$ from theorems 5-6 and 5-3

$\rho_{X_3, X_1} = \rho_{\frac{5}{16}X_1 + W, X_1} = 0.9898$ from theorem 6-1 (see below)

$\rho_{X_3, X_2} = \rho_{X_3, \frac{1}{2}X_1} = \rho_{X_3, X_1} = 0.9898$ from theorem 5-6

$\rho_{X_3, W} = \rho_{\frac{5}{16}X_1 + W, W} = 0.1424$ from theorem 6-1

Theorem 6-1 If $Y = aX + Z$ where a is a real number and X and Z are *independent* random variables then

$$\rho_{Y,X} = a\frac{\sigma_X}{\sigma_Y} \text{ and } \rho_{Y,Z} = \frac{\sigma_Z}{\sigma_Y}$$

A proof of this theorem can be developed from equation 5-29 (chapter 5).

The correlations listed above reveal the degree correlation exists between pairs of cost element costs in this WBS (table 6-2). The existence of these correlations is hard to notice when $Cost_{Sys}$ is expressed in the form

$$Cost_{Sys} = \tfrac{181}{80} X_1 + W + X_5 + X_7 + X_8 + X_9$$

In the above, $Cost_{Sys}$ is now written as the sum of six *independent* random variables instead of the sum of ten random variables (equation 6-8).

Capturing the combined effect of these correlations on the distribution function of $Cost_{Sys}$ is accounted for by the coefficient $\frac{181}{80}$. This case discussion illustrates how correlation can exist in a WBS, by virtue of the functional relationships defined among the cost element costs. Functional relationships such as those in this WBS (table 6-2) are *very* common in cost analysis. Although these relationships are primarily defined for developing the point estimate of $Cost_{Sys}$, such relationships come along with implied correlations. Cost analysts *must* be aware of this implication so as not to inadvertently *induce* correlation (or consider it absent) when it is already present. This concludes case discussion 6-1.♦

Many cost elements in case discussion 6-1 were a function of a single random variable. Thus, computing $E(Cost_{Sys})$ and $Var(Cost_{Sys})$ was "relatively" straightforward. More complex relationships are given in case discussion 6-2. Case discussion 6-2 illustrates the computation of $E(Cost_{Sys})$ and $Var(Cost_{Sys})$ when cost elements are functions of two or more random variables. In addition, it will be seen how a program's schedule can be incorporated into cost estimating relationships. Case discussion 6-2 is the last in this chapter. It lays the groundwork for studying cost-schedule probability tradeoffs and will be revisited in chapter 7.

Case Discussion 6-2 Suppose the government is acquiring a new digital information system. The system consists of three large screen displays for "situation rooms," forty-seven display workstations, two support processors, and a suite of electronic communications equipment. Suppose the system requires new software to be developed for the large screen display, as well as for the display workstation. The work breakdown structure for this system is given in figure 6-6. Cost element data for this WBS are provided in table 6-4. Additional information about these data follows.

Additional Information

Figure 6-6 presents the system's WBS. Table 6-4 presents the cost element data associated with this WBS.

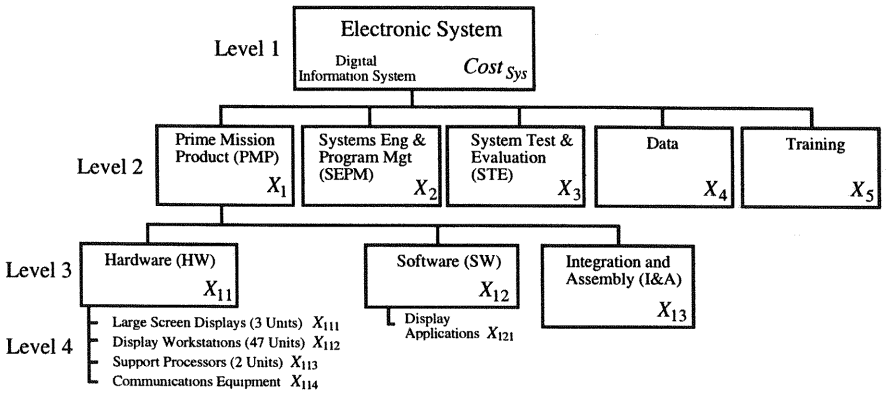


Figure 6-6. Case Discussion 6-2 Work Breakdown Structure

Table 6-4. Cost Element Data for Case Discussion 6-2

WBS Cost Element Cost ($K)	Functional Relationship (if applicable)	Distribution (if applicable)	Distributions of Random Variables (if applicable)
X_{111}		$Unif(700,750)$	
X_{112}		$Unif(3200,4000)$	
X_{113}		$Unif(200,250)$	
X_{114}		$Unif(350,380)$	
X_{121}	$\ell_{r_{SW}}\left(2.8I^{1.2}\right)$		$\ell_{r_{SW}} \sim Unif(10,15)$ $I \sim Trng(80,100,150)$
X_{13}	$0.05\left(X_{121}+\sum_{s=1}^{4} X_{11s}\right)$		
X_2	$\left(\ell_{r_{SEPM}}\right)\cdot$ $(SL_{SEPM})(PrgmSched)$		$\ell_{r_{SEPM}} \sim Unif(20,25)$ $SL_{SEPM} \sim Trng(12,15,25)$ $PrgmSched \sim N(33.36,1.94)$

Table 6-4. Cost Element Data for Case Discussion 6-2
(Concluded)

WBS Cost Element Cost ($K)	Functional Relationship (if applicable)	Distribution (if applicable)	Distributions of Random Variables (if applicable)
X_3	$\left(\ell_{r_{STE}}\right) \cdot (SL_{STE})(PrgmSched)$		$\ell_{r_{STE}} \sim Unif(15,20)$ $SL_{STE} \sim Unif(4,7)$ $PrgmSched \sim N(33.36, 1.94)$
X_4	$0.05\left(X_{13} + X_{121} + \sum_{s=1}^{4} X_{11s}\right)$		
X_5	$0.02\left(X_{13} + X_{121} + \sum_{s=1}^{4} X_{11s}\right)$		

In figure 6-6, the total cost of the digital information system is

$$Cost_{Sys} = X_1 + X_2 + X_3 + X_4 + X_5 \qquad (6\text{-}17)$$

Furthermore, assume in table 6-4 that X_{111}, X_{112}, X_{113}, X_{114}, $\ell_{r_{SW}}$, I, $\ell_{r_{SEPM}}$, SL_{SEPM}, $PrgmSched$, $\ell_{r_{STE}}$, and SL_{STE} are *independent random variables*. In table 6-4, we have the following random variable definitions.

- $\ell_{r_{SW}}$, $\ell_{r_{SEPM}}$, and $\ell_{r_{STE}}$ are labor rates for software (SW) development, systems engineering and program management (SEPM), and system test and evaluation (STE), respectively; the units are in ($K) per staff-month.
- I denotes the number of delivered source instructions (DSI) to be developed. The units are in thousands (K); that is, I is expressed in terms of KDSI (as discussed in chapter 5).
- SL_{SEPM} and SL_{STE} represents staff-levels (i.e., the number of persons) for the SEPM and STE activities, respectively.
- $PrgmSched$ denotes the total months to complete the development of the digital information system.

From the information given in this case discussion,

a) Determine $E(Cost_{Sys})$ and $Var(Cost_{Sys})$.

b) Discuss correlations implied by the relationships in table 6-4.

c) What distribution function(s) approximate $F_{Cost_{Sys}}(x)$?

Preliminaries — Before beginning part a), a simplified expression for $Cost_{Sys}$ will be developed. Recall from equation 6-17, the system's total cost is given by

$$Cost_{Sys} = X_1 + X_2 + X_3 + X_4 + X_5$$

This can be written as

$$Cost_{Sys} = Cost_{PMP} + X_2 + X_3 + X_4 + X_5 \qquad (6\text{-}18)$$

where

$$Cost_{PMP} = X_1 = X_{11} + X_{12} + X_{13} \qquad (6\text{-}19)$$

From figure 6-6 and equation 6-19

$$\begin{aligned} Cost_{PMP} = X_1 &= X_{11} + X_{12} + X_{13} \\ &= (X_{111} + X_{112} + X_{113} + X_{114}) + (X_{121}) + X_{13} \end{aligned}$$

From table 6-4

$$X_{13} = 0.05\left(X_{121} + \sum_{s=1}^{4} X_{11s}\right)$$

Combining these relationships

$$Cost_{PMP} = 1.05(X_{111} + X_{112} + X_{113} + X_{114} + X_{121}) \qquad (6\text{-}19a)$$

$$Cost_{PMP} = 1.05(X_{11} + X_{12}) \qquad (6\text{-}20)$$

In electronic systems, the sum $(X_{11} + X_{12})$ is known as the *prime mission equipment* (PME) cost, that is, PME is the total cost of just the system's hardware and software. Thus, equation 6-20 is equivalent to

$$Cost_{PMP} = 1.05 Cost_{PME} \qquad (6\text{-}21)$$

Equation 6-21 will be used later in this case discussion. Returning to equation 6-18 we had

$$Cost_{Sys} = Cost_{PMP} + (X_2 + X_3) + (X_4 + X_5) \tag{6-22}$$

From table 6-4, X_4 and X_5 can be written as

$$X_4 = 0.05\left(X_{13} + X_{121} + \sum_{s=1}^{4} X_{11s}\right) = 0.05X_1 = 0.05Cost_{PMP}$$

$$X_5 = 0.02\left(X_{13} + X_{121} + \sum_{s=1}^{4} X_{11s}\right) = 0.02X_1 = 0.02Cost_{PMP}$$

This simplifies $Cost_{Sys}$ (equation 6-22) to

$$Cost_{Sys} = 1.07Cost_{PMP} + (X_2 + X_3)$$

$$Cost_{Sys} = 1.07Cost_{PMP} + Q \tag{6-23}$$

where $Q = (X_2 + X_3)$. We will now work with equation 6-23 to determine the mean and variance of $Cost_{Sys}$.

Part a) From theorem 5-7, $E(Cost_{Sys})$ is

$$E(Cost_{Sys}) = 1.07E(Cost_{PMP}) + E(Q) \tag{6-24}$$

It can be shown, in this case, that $Cov(Cost_{PMP}, Q) = 0$. Therefore, from theorem 5-8, $Var(Cost_{Sys})$ is

$$Var(Cost_{Sys}) = (1.07)^2 Var(Cost_{PMP}) + Var(Q) \tag{6-25}$$

To compute $E(Cost_{Sys})$ and $Var(Cost_{Sys})$, it is necessary to determine the means and variances of $Cost_{PMP}$ and Q. Because these computations are lengthy, part a) is separated into three sections a.1), a.2), and a.3). They are defined as follows:

a.1) Focuses on computing the mean and variance of $Cost_{PMP}$.

a.2) Focuses on computing the mean and variance of Q.

a.3) Combines a.1) and a.2) to determine the mean and variance of $Cost_{Sys}$.

a.1) Mean and Variance of $Cost_{PMP}$

To compute $E(Cost_{PMP})$ and $Var(Cost_{PMP})$, recall from equation 6-21

$$Cost_{PMP} = 1.05 Cost_{PME} \tag{6-26}$$

where

$$Cost_{PME} = X_{11} + X_{12} = (X_{111} + X_{112} + X_{113} + X_{114}) + X_{121} \tag{6-27}$$

From equations 6-26 and 6-27

$$\begin{aligned} E(Cost_{PMP}) &= 1.05 E(Cost_{PME}) \\ &= 1.05 E((X_{111} + X_{112} + X_{113} + X_{114}) + X_{121}) \\ &= 1.05[E(X_{111}) + E(X_{112}) + E(X_{113}) + E(X_{114}) + E(X_{121})] \end{aligned} \tag{6-28}$$

Since it is assumed (refer to figure 6-6), in this case discussion, X_{111}, X_{112}, X_{113}, X_{114}, and X_{121} are independent random variables, we can write

$$\begin{aligned} Var(Cost_{PMP}) &= 1.05^2 Var(Cost_{PME}) \\ &= 1.05^2 Var((X_{111} + X_{112} + X_{113} + X_{114}) + X_{121}) \\ &= 1.05^2 [Var(X_{111}) + Var(X_{112}) + Var(X_{113}) + Var(X_{114}) + Var(X_{121})] \end{aligned} \tag{6-29}$$

From table 6-4, $X_{111} \sim Unif(700, 750)$; therefore, from theorem 4-2

$$E(X_{111}) = \frac{700 + 750}{2} = 725 \text{ and } Var(X_{111}) = \frac{(750 - 700)^2}{12} = 208.333$$

Similarly, for X_{112}, X_{113}, and X_{114}

$$E(X_{112}) = \frac{3200 + 4000}{2} = 3600 \text{ and } Var(X_{112}) = \frac{(4000 - 3200)^2}{12} = 53333.333$$

$$E(X_{113}) = \frac{200 + 250}{2} = 225 \text{ and } Var(X_{113}) = \frac{(250 - 200)^2}{12} = 208.333$$

$$E(X_{114}) = \frac{350 + 380}{2} = 365 \text{ and } Var(X_{114}) = \frac{(380 - 350)^2}{12} = 75$$

To complete the calculation of $E(Cost_{PMP})$ and $Var(Cost_{PMP})$ it is necessary to compute mean and variance of X_{121} (the cost to develop the display software). Two methods from chapter 5 will show ways this can be done.

Method 1 — Transformation of Variables Approach

In this method, transformation formulas developed in chapter 5, specifically those summarized in table 5-5, are used. From table 6-4, software cost, denoted by X_{121}, is

$$X_{121} = \ell_{r_{SW}}\left(2.8I^{1.2}\right) \tag{6-30}$$

It was given the random variables $\ell_{r_{SW}}$ and I are independent. From theorem 5-5

$$E(X_{121}) = E(\ell_{r_{SW}})E\left(2.8I^{1.2}\right)$$

Since $\ell_{r_{SW}} \sim Unif(10,15)$, from theorem 4-2 $E(\ell_{r_{SW}}) = 12.5$. Therefore,

$$E(X_{121}) = 12.5\left[E\left(2.8I^{1.2}\right)\right] \tag{6-31}$$

Recall if $Eff_{SW} = c_1 I^{c_2}$, and $I \sim Trng(a,m,b)$, then from equation 5-76 (table 5-5)

$$\begin{aligned} E(Eff_{SW}) = {} & c_1 \frac{2}{b-a}\frac{1}{m-a}\left[\frac{m^{c_2+2}-a^{c_2+2}}{c_2+2} + \frac{a^{c_2+2}-am^{c_2+1}}{c_2+1}\right] \\ & + c_1 \frac{2}{b-a}\frac{1}{m-b}\left[\frac{b^{c_2+2}-m^{c_2+2}}{c_2+2} + \frac{bm^{c_2+1}-b^{c_2+2}}{c_2+1}\right] \end{aligned} \tag{6-32}$$

Relating equation 6-32 to this case, $c_1 = 2.8$, $c_2 = 1.2$, $a = 80$, $m = 100$, and $b = 150$. Substituting these values into equation 6-32 yields $E(2.8I^{1.2}) = 790.23$. Therefore,

$$\boxed{E(X_{121}) = 12.5[790.23] = 9877.875} \tag{6-33}$$

We next compute $Var(X_{121})$. From theorem 3-10 and the above results

$$Var(X_{121}) = E\left(X_{121}^2\right) - [E(X_{121})]^2 = E\left(X_{121}^2\right) - [9877.875]^2 \tag{6-34}$$

To determine $Var(X_{121})$, it remains to determine $E\left(X_{121}^2\right)$ in equation 6-34. Now,

$$E\left(X_{121}^2\right) = E\left(\ell_{r_{SW}}^2 \cdot \left(2.8I^{1.2}\right)^2\right) = E\left(\ell_{r_{SW}}^2\left(7.84I^{2.4}\right)\right) \tag{6-35}$$

Since the random variables $\ell_{r_{SW}}$ and I are independent

$$E\left(X_{121}^2\right) = E\left(\ell_{r_{SW}}^2\right)E\left(7.84I^{2.4}\right) \tag{6-36}$$

We will take the following approach to compute $E\left(\ell_{r_{SW}}^2\right)$. Since

$$Var\left(\ell_{r_{SW}}\right) = E\left(\ell_{r_{SW}}^2\right) - \left[E\left(\ell_{r_{SW}}\right)\right]^2$$

We have

$$E\left(\ell_{r_{SW}}^2\right) = Var\left(\ell_{r_{SW}}\right) + \left[E\left(\ell_{r_{SW}}\right)\right]^2 \tag{6-37}$$

Since $\ell_{r_{SW}} \sim Unif(10,15)$, from theorem 4-2

$$E(\ell_{r_{SW}}) = 12.5 \text{ and } Var\left(\ell_{r_{SW}}\right) = \frac{(15-10)^2}{12} = \frac{25}{12}$$

Substituting these values into equation 6-37 yields $E\left(\ell_{r_{SW}}^2\right) = 158\frac{1}{3}$. Therefore, equation 6-36 becomes

$$E\left(X_{121}^2\right) = 158\tfrac{1}{3}\, E\left(7.84 I^{2.4}\right) \tag{6-38}$$

To compute $E\left(7.84 I^{2.4}\right)$ equation 6-32 will be used again with $c_1 = 7.84$, $c_2 = 2.4$, and $a = 80$, $m = 100$, and $b = 150$. Substituting these values into equation 6-32 yields $E\left(7.84 I^{2.4}\right) = 640626.866$. Therefore,

$$E\left(X_{121}^2\right) = 158\tfrac{1}{3}(640626.866) = 101432587.1 \tag{6-39}$$

Hence, equation 6-34 becomes

$$\boxed{Var(X_{121}) = 101432587.1 - [9877.875]^2 = 3860172.585} \tag{6-40}$$

$$\text{and } \sigma_{X_{121}} = \sqrt{Var\left(X_{121}\right)} = 1964.732$$

Method 2 — Mellin Transform Approach

In this method, the Mellin transform (refer to section 5.5) is used to illustrate an alternative approach to computing $E(X_{121})$ and $Var(X_{121})$. Recall that

$$X_{121} = \ell_{r_{SW}}\left(2.8 I^{1.2}\right) \tag{6-41}$$

From theorem 5-13, the Mellin transform of X_{121} is

$$M_{X_{121}}(s) = M_{\ell_{r_{SW}}}(s)(2.8)^{s-1} M_I(1.2s - 1.2 + 1) \tag{6-42}$$

From equation 5-98

$$E(X_{121}) = M_{X_{121}}(2) = M_{\ell_{r_{SW}}}(2)(2.8)^{2-1} M_I(1.2(2) - 1.2 + 1)$$

$$E(X_{121}) = 2.8 M_{\ell_{r_{SW}}}(2) M_I(2.2) \tag{6-43}$$

Since $\ell_{rsw} \sim Unif(10,15)$, from table 5-8 (equation 5-108)

$$M_{\ell_{rsw}}(2) = \frac{1}{2}\frac{1}{(15-10)}\left(15^2 - 10^2\right) = 12.5$$

Since $I \sim Trng(80,100,150)$, from table 5-8 (equation 5-109) with $s = 2.2$, $a = 80$, $m = 100$, and $b = 150$ we have

$$M_I(2.2) = 282.225$$

Therefore

$$\boxed{E(X_{121}) = 2.8(12.5)(282.225) = 9877.875} \tag{6-44}$$

To compute $Var(X_{121})$ we have

$$Var(X_{121}) = E(X_{121}^2) - \left[E(X_{121})\right]^2 = E(X_{121}^2) - [9877.875]^2 \tag{6-45}$$

From equation 5-99

$$\begin{aligned} E(X_{121}^2) &= M_{X_{121}}(3) \\ &= M_{\ell_{rsw}}(3)(2.8)^{3-1} M_I(1.2(3) - 1.2 + 1) \\ &= (2.8)^2 M_{\ell_{rsw}}(3) M_I(3.4) \end{aligned} \tag{6-46}$$

where

$$M_{\ell_{rsw}}(3) = \frac{1}{3}\frac{1}{(15-10)}\left(15^3 - 10^3\right) = 158\tfrac{1}{3}$$

and $M_I(3.4) = 81712.61045$

Substituting these values into equation 6-46 yields $E(X_{121}^2) = 101432587.1$. Therefore

$$\boxed{Var(X_{121}) = 101432587.1 - [9877.875]^2 = 3860172.585} \tag{6-47}$$

and $\sigma_{X_{121}} = \sqrt{Var(X_{121})} = 1964.732$ ♦

All the information needed to complete the computation of $E(Cost_{PMP})$ and $Var(Cost_{PMP})$ is available. From equation 6-28, recall that

$$E(Cost_{PMP}) = 1.05E(Cost_{PME})$$

$$= 1.05\left[E(X_{111}) + E(X_{112}) + E(X_{113}) + E(X_{114}) + E(X_{121})\right] \tag{6-48}$$

Substituting the expected value computations developed in the above discussions into equation 6-48 yields

$$E(Cost_{PMP}) = 1.05[725 + 3600 + 225 + 365 + 9877.875]$$
$$= 15532.52 \text{ (\$K)} \quad (6\text{-}49)$$

From equation 6-29

$$Var(Cost_{PMP}) = 1.05^2 Var(Cost_{PME})$$
$$= 1.05^2 \begin{bmatrix} Var(X_{111}) + Var(X_{112}) \\ + Var(X_{113}) + Var(X_{114}) + Var(X_{121}) \end{bmatrix} \quad (6\text{-}50)$$

Substituting the variance computations developed in the above discussions into equation 6-50 yields

$$Var(Cost_{PMP}) = 1.05^2 \begin{bmatrix} 208.333 + 53333.333 \\ + 208.333 + 75 + 3860172.585 \end{bmatrix} = 4315182.336 \text{ (\$K)}^2 \quad (6\text{-}51)$$

$$\text{and } \sigma_{Cost_{PMP}} = \sqrt{Var(Cost_{PMP})} = 2077.3 \text{ (\$K)}$$

a.2) Mean and Variance of Q

The above discussion presented the mean and variance of the system's prime mission product cost. To complete the computation of $E(Cost_{Sys})$ and $Var(Cost_{Sys})$, defined by equations 6-24 and 6-25, the values of $E(Q)$ and $Var(Q)$, where $Q = X_2 + X_3$, must be determined.

From table 6-4, observe that X_2 and X_3 are *not independent* random variables. They are both a function of the random variable *PrgmSched*. From theorem 5-7, $E(Q)$ is the sum of the means of X_2 and X_3 regardless of whether or not the two random variables are independent. Hence,

$$E(Q) = E(X_2 + X_3) = E(X_2) + E(X_3) \quad (6\text{-}52)$$

However, because X_2 and X_3 are not independent, $Var(Q)$ is *not* just the sum of their respective variances. Applying theorem 5-8 to this particular case,

$$Var(Q) = Var(X_2) + Var(X_3) + 2\rho_{X_2,X_3}\sigma_{X_2}\sigma_{X_3} \tag{6-53}$$

The following presents the computations for the means and variances of X_2 and X_3, as well as ρ_{X_2,X_3}, their correlation coefficient.

Mean and Variance of X_2

From the WBS in figure 6-6, recall that the cost of systems engineering and program management (SEPM) is denoted by X_2. From table 6-4, X_2 is a function of three random variables; specifically,

$$X_2 = \ell_{r_{SEPM}}(SL_{SEPM})(PrgmSched) \tag{6-54}$$

Given $\ell_{r_{SEPM}}$, SL_{SEPM}, and $PrgmSched$ are independent random variables

$$E(X_2) = E\left(\ell_{r_{SEPM}}\right)E\left(SL_{SEPM}\right)E(PrgmSched) \tag{6-55}$$

From the distribution functions for $\ell_{r_{SEPM}}$, SL_{SEPM}, and $PrgmSched$ in table 6-4, it can be shown that

$$\boxed{E(X_2) = (22.5)(17\tfrac{1}{3})(33.36) = 13010.4} \tag{6-56}$$

The variance of X_2 is

$$Var(X_2) = E(X_2^2) - [E(X_2)]^2$$

which is equivalent to

$$Var(X_2) = E\left(\ell^2_{r_{SEPM}}\, SL^2_{SEPM}\, PrgmSched^2\right) - [13010.4]^2 \tag{6-57}$$

where $E(X_2^2) = E\left(\ell^2_{r_{SEPM}}\, SL^2_{SEPM}\, PrgmSched^2\right)$. To compute $Var(X_2)$, it remains to determine $E\left(\ell^2_{r_{SEPM}}\, SL^2_{SEPM}\, PrgmSched^2\right)$. Again, since $\ell_{r_{SEPM}}$, SL_{SEPM}, and $PrgmSched$ are independent

$$E(X_2^2) = E\left(\ell^2_{r_{SEPM}}\, SL^2_{SEPM}\, PrgmSched^2\right) = E\left(\ell^2_{r_{SEPM}}\right)E\left(SL^2_{SEPM}\right)E\left(PrgmSched^2\right) \tag{6-58}$$

Similar to the previous calculations involving $\ell_{r_{SW}}$, it is left to the reader to show that

$$E\left(\ell^2_{r_{SEPM}}\right) = 508\tfrac{1}{3} \tag{6-59}$$

since $\ell_{r_{SEPM}} \sim Unif(20,25)$.

From table 6-4, the distribution function for SEPM staff-level is triangular, specifically $SL_{SEPM} \sim Trng(12,15,25)$. To determine $E\left(SL_{SEPM}^2\right)$ the relationship

$$E\left(SL_{SEPM}^2\right) = Var(SL_{SEPM}) + \left[E(SL_{SEPM})\right]^2 \tag{6-60}$$

is used. From theorem 4-3, it can be shown that

$$Var(SL_{SEPM}) = 7.7222 \text{ and } E(SL_{SEPM}) = 17\tfrac{1}{3}$$

Therefore,

$$E\left(SL_{SEPM}^2\right) = 7.7222 + \left[17\tfrac{1}{3}\right]^2 = 308.166 \tag{6-61}$$

The last term in equation 6-58 is $E\left(PrgmSched^2\right)$. To compute this expected value, note that

$$E\left(PrgmSched^2\right) = Var(PrgmSched) + \left[E(PrgmSched)\right]^2 \tag{6-62}$$

From table 6-4, $PrgmSched \sim N(33.36,1.94)$. Therefore

$$E\left(PrgmSched^2\right) = 1.94 + [33.36]^2 = 1114.829 \tag{6-63}$$

The expected value of each term in equation 6-58 has now been determined. Thus,

$$E\left(\ell_{r_{SEPM}}^2 SL_{SEPM}^2 PrgmSched^2\right) = (508\tfrac{1}{3})(308.166)(1114.829)$$

$$= 174639133.4 \tag{6-64}$$

Combining the above results $Var(X_2)$ is

$$\boxed{Var(X_2) = 174639133.4 - [13010.4]^2 = 5368625.24} \tag{6-65}$$

$$\text{and } \sigma_{X_2} = \sqrt{Var(X_2)} = 2317.03$$

Mean and Variance of X_3

From the WBS in figure 6-6, recall the cost of system test and evaluation (STE) is denoted by X_3. From table 6-4, X_3 is a function of three independent random variables; specifically,

$$X_3 = \ell_{r_{STE}}(SL_{STE})(PrgmSched) \tag{6-66}$$

The same approach to determine the mean and variance of the cost of SEPM can be used to determine the mean and variance of the cost of STE. For this reason, it is left to the reader to verify the following:

$$\boxed{E(X_3) = 3210.9} \tag{6-67}$$

$$E\left(X_3^2\right) = E\left(\ell_{r_{STE}}^2\right)E\left(SL_{STE}^2\right)E\left(PrgmSched^2\right) \tag{6-68}$$

$$= (308\tfrac{1}{3})(31)(1114.829) = 10655907.19$$

Since $\ell_{r_{STE}} \sim Unif\,(15,20)$ — Since $SL_{STE} \sim Unif\,(4,7)$

With

$$Var(X_3) = E(X_3^2) - \left[E(X_3)\right]^2 \tag{6-69}$$

Substituting the results from equations 6-67 and 6-68 into equation 6-69 yields

$$\boxed{Var(X_3) = 346028.38} \tag{6-70}$$

and $\sigma_{X_3} = \sqrt{Var(X_3)} = 588.242$

Correlation Between X_2 and X_3

By definition (equation 5-29), the correlation between X_2 and X_3 is

$$\rho_{X_2,X_3} = \frac{Cov(X_2,X_3)}{\sigma_{X_2}\sigma_{X_3}} = \frac{E(X_2X_3) - E(X_2)E(X_3)}{\sigma_{X_2}\sigma_{X_3}} \tag{6-71}$$

From table 6-4, it was given that

$$X_2 = \ell_{r_{SEPM}} SL_{SEPM}\, PrgmSched \tag{6-72}$$

$$X_3 = \ell_{r_{STE}} SL_{STE}\, PrgmSched \tag{6-73}$$

All the terms in equation 6-71, except for $E(X_2X_3)$, have been determined from the above computations. The term $E(X_2X_3)$ is

$$E(X_2X_3) = E\left(\ell_{r_{SEPM}} SL_{SEPM}\, PrgmSched \cdot \ell_{r_{STE}} SL_{STE}\, PrgmSched\right)$$

$$= E\left(\ell_{r_{SEPM}} \ell_{r_{STE}} SL_{SEPM} SL_{STE}\, PrgmSched^2\right) \tag{6-74}$$

Since $\ell_{r_{SEPM}}$, $\ell_{r_{STE}}$, SL_{SEPM}, SL_{STE}, and $PrgmSched$ were given to be independent random variables, equation 6-74 can be written as

$$E(X_2X_3) = E(\ell_{r_{SEPM}})E(\ell_{r_{STE}})E(SL_{SEPM})E(SL_{STE})E(PrgmSched^2) \tag{6-75}$$

It can be determined that

$$E(X_2X_3) = (22.5)(17.5)(17\tfrac{1}{3})(5.5)(1114.829) = 41847893.59 \qquad (6\text{-}76)$$

In equations 6-75 and 6-76, the term $E(PrgmSched^2) = 1114.829$ comes from equation 6-63. Substituting the result from equation 6-76 into equation 6-71 yields

$$\rho_{X_2,X_3} = \frac{41847893.59 - (13010.4)(3210.9)}{(2317.03)(588.242)} = 0.0534 \qquad (6\text{-}77)$$

All the terms necessary to complete the computation of $E(Cost_{Sys})$ and $Var(Cost_{Sys})$ have now been determined.

a.3) Mean and Variance of $Cost_{Sys}$

From equation 6-24

$$\begin{aligned} E(Cost_{Sys}) &= 1.07E(Cost_{PMP}) + E(Q) \\ &= 1.07E(Cost_{PMP}) + E(X_2 + X_3) \\ &= 1.07E(Cost_{PMP}) + E(X_2) + E(X_3) \\ &= 1.07(15532.52) + 13010.4 + 3210.9 \\ &= 32841.1 \text{ (\$K)} \end{aligned} \qquad (6\text{-}78)$$

From equation 6-25

$$\begin{aligned} Var(Cost_{Sys}) &= (1.07)^2 Var(Cost_{PMP}) + Var(Q) \\ &= (1.07)^2 Var(Cost_{PMP}) + Var(X_2 + X_3) \\ &= (1.07)^2 Var(Cost_{PMP}) + Var(X_2) + Var(X_3) + 2\rho_{X_2,X_3}\sigma_{X_2}\sigma_{X_3} \\ &= (1.07)^2(4315182.336) + 5368625.24 + 346028.38 + 2(0.0534)(2317.03)(588.242) \\ &= 10800671.5 \text{ (\$K)}^2 \end{aligned} \qquad (6\text{-}79)$$

which implies

$$\sigma_{Cost_{Sys}} = \sqrt{Var(Cost_{Sys})} = 3286.44 \text{ (\$K)} \qquad (6\text{-}80)$$

In summary, the mean cost of the digital information system is 32.8 (\$M) and the standard deviation is 3.3 (\$M). This concludes part a) of this case discussion.

Part b) Some Implied Correlations

This section discusses the correlations implied by some of the cost relationships in this WBS. The correlation between cost element costs X_i, for $i = 1,\ldots,5$, is best explored from the relationships given in table 6-4. From equation 6-21, we have

$$Cost_{PMP} = 1.05\, Cost_{PME}$$

Since $Cost_{PMP}$ is a linear function of $Cost_{PME}$ (with positive slope) the correlation between $Cost_{PMP}$ and $Cost_{PME}$ is unity. In table 6-4, we are also given that

$$X_{13} = 0.05\left(X_{121} + \sum_{s=1}^{4} X_{11s} \right) = 0.05 Cost_{PME}$$

Thus, the correlation between X_{13} (the integration and assembly cost) and $Cost_{PME}$ is unity. There also exists perfect correlation between $Cost_{PMP}$ and other cost element costs in this WBS. From table 6-4 and the *Preliminaries* section of this case discussion, we can write

$$X_4 = 0.05 Cost_{PMP} \text{ and } X_5 = 0.02 Cost_{PMP}$$

Thus, there are implied correlations between X_4 and $Cost_{PMP}$ and X_5 and $Cost_{PMP}$ because of these functional (mathematical) relationships. Here, the correlation between the cost of Data, denoted by X_4, and $Cost_{PMP}$ is unity. Similarly, the correlation between the cost of Training, denoted by X_5, and $Cost_{PMP}$ is unity. These relationships illustrate "logical" or "factor-based" cost relationships, which are common in electronic systems cost analyses.

Lastly, there is another important correlation in this case discussion. Notice the costs of SEPM and STE, denoted by X_2 and X_3, are functions of *PrgmSched* — the system's development schedule. As a result, a positive correlation exists between $Cost_{Sys}$ and *PrgmSched*. The following presents a derivation of this correlation.

Correlation Between $Cost_{Sys}$ and *PrgmSched*

From equation 6-23, recall that

$$Cost_{Sys} = 1.07 Cost_{PMP} + (X_2 + X_3) = 1.07 Cost_{PMP} + Q \qquad (6\text{-}81)$$

To simplify notation, let $C \equiv Cost_{Sys}$ and $P \equiv PrgmSched$. The correlation between the system's total cost C and its development schedule P will be determined. By definition, this correlation is

$$\rho_{C,P} = \frac{E(CP) - E(C)E(P)}{\sigma_C \sigma_P} \qquad (6\text{-}82)$$

where

$E(C) = 32841.1$ (from equation 6-78)

$E(P) = 33.36$ (seen in table 6-4)

$\sigma_C = 3286.44$ (from equation 6-80)

$\sigma_P = \sqrt{1.94} = 1.39283$ (seen from table 6-4)

To determine $\rho_{C,P}$ we need $E(CP)$. Multiplying equation 6-81 by P, we can write

$$E(CP) = E\left[(1.07 Cost_{PMP} + Q)P\right]$$

$$= 1.07 E(Cost_{PMP} P) + E(QP)$$

It can be shown, in this case, that $Cov(Cost_{PMP}, P) = 0$. Therefore, from theorem 5-1

$$E(Cost_{PMP} P) - E(Cost_{PMP})E(P) = 0 \Rightarrow E(Cost_{PMP} P) = E(Cost_{PMP})E(P)$$

Thus,

$$E(CP) = 1.07 E(Cost_{PMP})E(P) + E(QP)$$

$$E(CP) = 1.07(15532.52)(33.36) + E(QP) \qquad (6\text{-}83)$$

To complete the computation of $E(CP)$ it remains to determine $E(QP)$. Given the specifics of this case discussion, the random variables Q and P are *not* independent so $E(QP) \neq E(Q)E(P)$. The computation of $E(QP)$ proceeds as follows:

$$E(QP) = E\left[(X_2 + X_3)P\right] = E\left[\ell_{r_{SEPM}} SL_{SEPM} P^2 + \ell_{r_{STE}} SL_{STE} P^2\right]$$

$$= E\left[\ell_{r_{SEPM}} SL_{SEPM} P^2\right] + E\left[\ell_{r_{STE}} SL_{STE} P^2\right] \qquad (6\text{-}84)$$

Since the random variables $\ell_{r_{SEPM}}$, $\ell_{r_{STE}}$, SL_{SEPM}, SL_{STE}, and P were given to be independent, equation 6-84 can be written as

$$E(QP) = \left(E\left[\ell_{r_{SEPM}}\right]E[SL_{SEPM}] + E\left[\ell_{r_{STE}}\right]E[SL_{STE}]\right)E\left[P^2\right] \qquad (6\text{-}85)$$

$$= \left(22.5(17\tfrac{1}{3}) + 17.5(5.5)\right)1114.829$$

$$= 542085.6013$$

Therefore

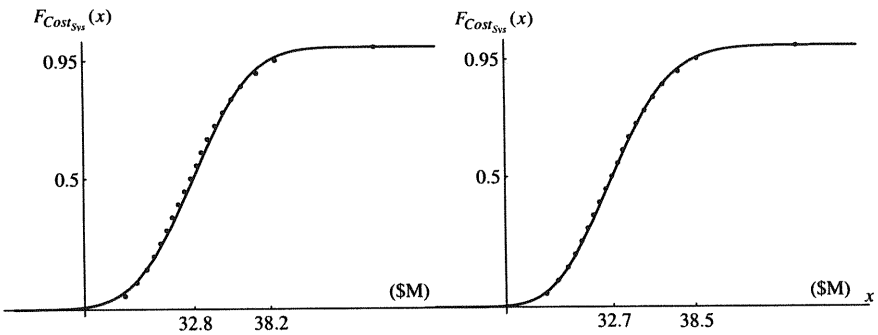
$$E(CP) = 1.07(15532.52)(33.36) + 542085.6013 = 1096522.009$$

and

$$\rho_{C,P} = \frac{E(CP) - E(C)E(P)}{\sigma_C \sigma_P} = \frac{1096522.009 - (32841.1)(33.36)}{(3286.44)(1.39283)} = 0.206 \qquad (6\text{-}86)$$

Part c) Distribution Function Approximation to $F_{Cost_{Sys}}(x)$

Figure 6-7 presents distributions that approximate the cumulative distribution function of the system's total cost. The curves defined by the two solid lines reflect two assumed theoretical distributions. They are a normal distribution (the left picture in figure 6-7) and a lognormal distribution (the right picture in figure 6-7), each with mean 32.8 ($M) and standard deviation 3.3 ($M).



Normal CDF vs the Simulated CDF LogNormal CDF vs the Simulated CDF

Figure 6-7. Assumed Theoretical CDFs vs the Simulated CDF for $Cost_{Sys}$

A third distribution is shown in figure 6-7 by a series of 20 "points". These points reflect random statistical samples (values) of $Cost_{Sys}$, sampled by Monte Carlo simulation (explained in section 6.3). In figure 6-7, the curve implied by these "points" is the simulated distribution function for $Cost_{Sys}$. Observe in figure 6-7 how closely this simulated distribution matches the assumed normal distribution, as well as the assumed lognormal distribution for $Cost_{Sys}$. The closeness with which these "points" fall along the two curves (each defined by the solid lines in figure 6-7) *visually suggests* the reasonableness of the assumption that the distribution function for $Cost_{Sys}$ can be approximated by a normal or by a lognormal. Although this is a practical conclusion, it is an informal one. A more formal conclusion could be derived from the Kolmogorov-Smirnov (K-S) test, illustrated in case discussion 5-1. This would reveal whether the normal and the lognormal distributions are *statistically plausible models* for the underlying distribution function of $Cost_{Sys}$, in this case.

6.2.2 Approximating the Distribution Function of System Cost

This section provides *guidance* for approximating the distribution function of a system's total cost. Some of this guidance reflects mathematical theory; some of it reflects observations from numerous project applications.

In the examples and case discussions presented in this book, the normal distribution often approximates the distribution function of a system's total cost. There are many reasons for this. Primary among them is $Cost_{Sys}$ (a system's total cost) is a summation of WBS cost element costs. Within the WBS, it is typical to have a mixture of independent and correlated cost element costs. The greater the number of independent cost element costs, the more it is that the distribution function of $Cost_{Sys}$ is approximately normal. Why is this? It is essentially the phenomenon described by the central limit

theorem (theorem 5-10). Seen in this book, the central limit theorem is very powerful. It does not take many independent cost element costs for the distribution of $Cost_{Sys}$ to move towards normality. Such a move is evidenced when 1) a sufficient number of independent cost element costs are summed and 2) no cost element's cost distribution has a much larger standard deviation than the standard deviations of the other cost element cost distributions. When conditions in the WBS result in $Cost_{Sys}$ being positively skewed (i.e., a non-normal distribution function), then the lognormal often [6,7]* approximates the distribution function of $Cost_{Sys}$.

What drives the distribution of $Cost_{Sys}$ to be normal or to be skewed? To address this, cost relationships that frequently occur in a system's WBS are examined. The electronic system is used to provide a context for the discussion. Work breakdown structures associated with other system classes (e.g., spacecraft systems) can also exhibit properties similar to those discussed below.

From the electronic system WBS in figure 6-8, $Cost_{Sys}$ is defined by

$$Cost_{Sys} = X_1 + X_2 + X_3 + X_4 + \ldots + X_n \tag{6-87}$$

where $X_1, X_2, X_3, X_4, \ldots, X_n$ denote the n costs of the system's level 2 cost elements (refer to equation 6-1). These elements include (but are not limited to) the system's prime mission product (PMP), as well as the system's systems engineering, program management, and system test. Referring to figure 6-8, equation 6-87 can also be written as

$$Cost_{Sys} = Cost_{PMP} + \sum_{i=2}^{n} X_i \tag{6-88}$$

where $Cost_{PMP} = X_1$.

---

* Many practitioners [8-11] have empirically shown the beta distribution also well approximates the distribution of $Cost_{Sys}$.

In the cost analysis of electronic systems, the distribution function of $Cost_{Sys}$ is often *observed* to be approximately normal. Situations specific to cost analysis contribute to this observation. The following cases describe the most common of these situations. In each case, the distribution functions for $Cost_{PMP}, X_2, X_3, X_4, \ldots, X_n$ are assumed to be "well-behaved" (e.g., unimodal, continuous).

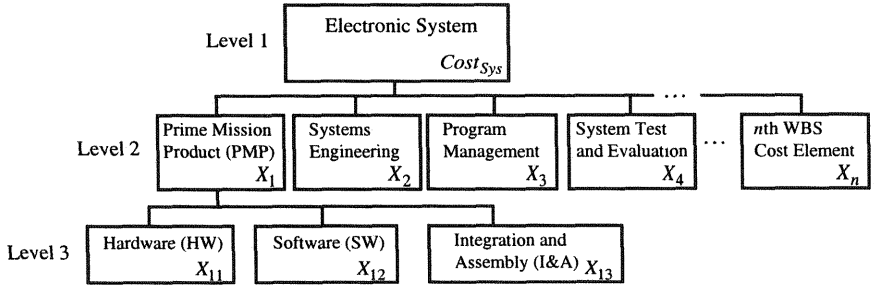


Figure 6-8. An Electronic System WBS

Case A

If (in equation 6-88) the distribution function of $Cost_{PMP}$ *is normal and* $X_2, X_3, X_4, \ldots, X_n$ *are linear functions of* $Cost_{PMP}$*, such as* $X_i = a_i Cost_{PMP}$ *where* $a_i \geq 0$ $(i = 2, \ldots, n)$*, then the distribution function of* $Cost_{Sys}$ *is normal with mean*

$$(1 + a_2 + a_3 + \ldots + a_n) E(Cost_{PMP})$$

and variance

$$(1 + a_2 + a_3 + \ldots + a_n)^2 Var(Cost_{PMP})$$

Case A is a direct consequence of the following proposition.

Proposition 6-1 If X is a normal random variable and $Y = aX$, where a is a constant, then the distribution function for Y is normal with mean $aE(X)$ and variance $a^2 Var(X)$.

Case B

If (in equation 6-88) $Cost_{PMP}$ and $X_2, X_3, X_4, \ldots, X_n$ are independent random variables and each are normally distributed, then the distribution function of $Cost_{Sys}$ is normal with

$$\text{mean } E(Cost_{PMP}) + \sum_{i=2}^{n} E(X_i) \text{ and variance } Var(Cost_{PMP}) + \sum_{i=2}^{n} Var(X_i)$$

Case B is a direct consequence of the following proposition.

Proposition 6-2 If $X_1, X_2, X_3, \ldots, X_k$ are independent *normally* distributed random variables and $Y = X_1 + X_2 + X_3 + \ldots + X_k$, then, regardless of the size of k, the distribution function of Y is normal with mean $\sum_{i=1}^{k} E(X_i)$ and variance $\sum_{i=1}^{k} Var(X_i)$.

Case C

Suppose (in equation 6-88) $Cost_{PMP}$, $X_2, X_3, X_4, \ldots, X_n$ are independent random variables. Furthermore, suppose $Cost_{PMP}$, $X_2, X_3, X_4, \ldots, X_n$ are <u>not</u> necessarily each normally distributed. If the number of cost element costs in the sequence $Cost_{PMP}$, $X_2, X_3, X_4, \ldots, X_n$ is sufficiently large with none dominating in standard deviation, then (by the central limit theorem) the distribution function of $Cost_{Sys}$ is approximately normal with

$$\text{mean } E(Cost_{PMP}) + \sum_{i=2}^{n} E(X_i) \text{ and variance } Var(Cost_{PMP}) + \sum_{i=2}^{n} Var(X_i)$$

The above three cases stem from mathematical theory. The next two cases originate from observations. They are not intended to be rigorous findings; rather, they reflect results often seen in practice.

Case D

Suppose (in equation 6-88) $Cost_{PMP}$ is normal and $Cost_{PMP}, X_2, X_3, X_4, \ldots, X_n$ are independent random variables. Furthermore, suppose $X_2, X_3, X_4, \ldots, X_n$ are not necessarily each normally distributed. If the number of cost element costs in the sequence $X_2, X_3, X_4, \ldots, X_n$ is

sufficiently large with no X_i $(i=2,\ldots,n)$ *dominating in standard deviation, then the distribution function of* $Cost_{Sys}$ *is approximately normal with*

$$\textit{mean}\quad E(Cost_{PMP})+\sum_{i=2}^{n}E(X_i)\ \textit{and variance}\ Var(Cost_{PMP})+\sum_{i=2}^{n}Var(X_i)$$

Case D stems from the influences of the central limit theorem and proposition 6-2. To see this, recall from equation 6-88 $Cost_{Sys}$ is given by

$$Cost_{Sys}=Cost_{PMP}+\sum_{i=2}^{n}X_i$$

If the distribution function for $Cost_{PMP}$ is normal and the distribution function of the sum $\sum_{i=2}^{n}X_i$ is *approximately* normal (by the central limit theorem), then $Cost_{Sys}$ is *approximately* the sum of two normally distributed random variables. In case D, $Cost_{PMP}$ and $\sum_{i=2}^{n}X_i$ are independent. Thus, from proposition 6-2, the distribution function of $Cost_{Sys}$ is *approximately* normal.

Case E

Suppose (in equation 6-88) $Cost_{PMP}$ *is normal. Suppose the sequence* $X_2,X_3,X_4,\ldots,X_n$ *contains some cost element costs correlated to* $Cost_{PMP}$ *(with correlation coefficient* ρ_{Cost_{PMP},X_i}*) and some that are uncorrelated to* $Cost_{PMP}$*. Suppose* $X_2,X_3,X_4,\ldots,X_n$ *are mutually independent random variables. If the number of* X_i*'s* $(i\geq 2)$ *uncorrelated to* $Cost_{PMP}$ *is sufficiently large, with none of the* X_i*'s (correlated or uncorrelated to* $Cost_{PMP}$*) dominating in standard deviation, then the distribution function of* $Cost_{Sys}$ *is approximately normal with*

$$\textit{mean}\ E(Cost_{PMP})+\sum_{i=2}^{n}E(X_i)$$

$$\textit{and variance}\ Var(Cost_{PMP})+\sum_{i=2}^{n}Var(X_i)+2\sum_{i=2}^{n}\rho_{Cost_{PMP},X_i}\sigma_{Cost_{PMP}}\sigma_{X_i}$$

In all but case C, the distribution function for $Cost_{PMP}$ was given to be normal. This is common in electronic systems. The normality of $Cost_{PMP}$ is primarily driven by the central limit theorem, where $Cost_{PMP}$ typically reflects the sum of many *independent* hardware and software costs.

6.2.2.1 The Normality of $Cost_{PMP}$

In electronic systems (refer to the WBS in figure 6-8) $Cost_{PMP}$ is defined as the sum of three cost element costs; specifically,

$$Cost_{PMP} = X_1 = X_{11} + X_{12} + X_{13} \tag{6-89}$$

Equation 6-89 can also be written as

$$Cost_{PMP} = Cost_{PME} + X_{13} \tag{6-90}$$

where $Cost_{PME}$ is the system's prime mission equipment cost. It represents the total cost of the system's hardware and software; that is,

$$Cost_{PME} = X_{11} + X_{12} \tag{6-91}$$

The normality of $Cost_{PMP}$ will be discussed by examining distribution functions that frequently characterize X_{11}, X_{12}, and X_{13}.

Distribution Function of Hardware Cost

Typically, a system's total hardware cost X_{11} is the sum of the individual hardware item costs. Referring to figure 6-8, suppose

$$X_{11} = X_{111} + X_{112} + X_{113} + \ldots + X_{11j} \tag{6-92}$$

where X_{11i} $(i = 1,2,\ldots,j)$ are independent random variables representing the costs of the individual hardware items. Under appropriate conditions, the distribution function of X_{11} can be approximately normal by the central limit theorem (theorem 5-10); that is, $X_{11} \sim N\big(E(X_{11}), Var(X_{11})\big)$ with

$$E(X_{11}) = E(X_{111}) + E(X_{112}) + E(X_{113}) + \ldots + E(X_{11j})$$

$$Var(X_{11}) = Var(X_{111}) + Var(X_{112}) + Var(X_{113}) + \ldots + Var(X_{11j})$$

If the distribution functions for X_{11i} $(i = 1,2,\ldots,j)$ are well behaved, then the approximation (in most cases) is good for small j (e.g., not less than or equal to $j = 5$ hardware items). The more asymmetric (skewed) the distribution functions are for X_{11i} $(i = 1,2,\ldots,j)$, the larger j must be for X_{11} to become approximately normal.

In practice, it is *very* common to see the normal distribution approximate X_{11}, particularly in systems designed around the use of *commercial hardware items*. The uncertainty in the cost of such items tends to vary independently and cost analysts often describe these uncertainties by distribution functions that are well behaved.

The cost distribution functions of hardware items that require *custom development* may be *asymmetric*. In practice, this asymmetry typically reflects a positive skew. The presence of asymmetry in the distribution functions for X_{11i} $(i = 1,2,\ldots,j)$ *will* affect how well (or how quickly) the normal distribution approximates X_{11}. If j (in equation 6-92) is sufficiently large and the asymmetry is isolated to just a few hardware items whose cost standard deviations contribute only a small amount to the standard deviation of X_{11}, then the distribution of X_{11} may still be approximately normal. If X_{11} is the sum of just a few asymmetric distributions (i.e., j is small), then the distribution of X_{11} may indeed be non-normal. In such circumstances, the lognormal (or beta distribution) might well approximate the distribution function of X_{11}. It is a good exercise for the reader to study this further. After reading section 6.3, use the Monte Carlo simulation technique to study the reasonableness of certain distribution function approximations of X_{11}. Do this using various symmetric and asymmetric distributions for the costs of the hardware items X_{11i} $(i = 1,2,\ldots,j)$.

Distribution Function of Software Cost

Can the distribution function of software cost also be approximated by the normal distribution? The answer depends on *how* software cost is determined. Cost analysts sometimes determine software cost according to the equation

$$X_{12} = \ell_{r_{SW}} \left[c_1 (I_{X_{121}})^{c_2} + c_1 (I_{X_{122}})^{c_2} + \ldots + c_1 (I_{X_{12k}})^{c_2} \right] \qquad (6\text{-}93)$$

where $I_{X_{12i}} (i = 1,2,\ldots,k)$ is the number of thousands of delivered source instructions (KDSI) to be developed for the ith software function in the system and c_1, c_2, and $\ell_{r_{SW}}$ are constants (discussed in section 5.4.2). Equation 6-93 is traditionally applied in cases where the individual software functions are independently developed. Such functions would have minimal-to-no interdependencies. They would integrate and execute in the system in a highly modular fashion. Under this formulation, if $\ell_{r_{SW}}$ is a constant, k is sufficiently large, and $I_{X_{121}}, I_{X_{122}}, \ldots, I_{X_{12k}}$ are independent random variables, then, by the central limit theorem, the distribution function of X_{12} will be approximately normal. This result is dependent on the way X_{12} is *mathematically defined*. Other definitions for X_{12} *may* yield distribution functions for X_{12} that are skewed. Two such definitions are given by equations 6-94 and 6-95.

$$X_{12} = \ell_{r_{SW}} \frac{I}{P_r} \qquad (6\text{-}94)$$

$$X_{12} = \ell_{r_{SW}} \left[\begin{array}{l} \overbrace{c_1 (I_{X_{121}})^{c_2} + c_1 (I_{X_{122}})^{c_2} + \ldots + c_1 (I_{X_{12m}})^{c_2}}^{\text{Software functions that have independent development efforts}} \\ \underbrace{+ c_1 (I_{X_{12(m+1)}} + I_{X_{12(m+2)}} + \ldots + I_{X_{12(m+k)}})^{c_2}}_{\text{Software functions that have dependent development efforts}} \end{array} \right] \qquad (6\text{-}95)$$

Equation 6-94 (refer to chapter 5) *might* be used when software cost is based on the total size I (in DSI) of the software to be developed and its development productivity rate P_r (i.e., DSI per staff-month). Here, I and P_r may or may not be independent random variables. Equation 6-95 is traditionally applied when a combination of independently developed software functions (the first part of equation 6-95) and a set of software functions that share functionality (the last part of equation 6-95) characterize the system.

In the definitions for X_{12} (given by equations 6-93, 6-94, and 6-95) it would be reasonable to consider $\ell_{r_{SW}}$ a random variable instead of a constant. This consideration also affects whether the distribution function of X_{12} can be approximated by a normal distribution. The reader is encouraged to explore these questions further, using the Monte Carlo simulation technique discussed in section 6.3.

Distribution Function of Integration and Assembly (I&A)

Similar to the above discussion, the distribution function for X_{13} — the cost to integrate, assemble, and checkout the system's hardware and software (known in the cost analysis community as I&A) is also driven by how X_{13} is mathematically defined. The following approaches are commonly used to define X_{13}.

Approach 1 — Cost Factor

Cost analysts often define X_{13} as a scalar multiple of $Cost_{PME}$, that is,

$$X_{13} = aCost_{PME} \tag{6-96}$$

where $a > 0$. For electronic systems, a typical value for a is 0.05. If $Cost_{PME}$ is normally distributed, then from proposition 6-1

$$X_{13} \sim N\left(aE(Cost_{PME}), a^2 Var(Cost_{PME})\right) \tag{6-97}$$

Under this approach, the correlation between X_{13} and $Cost_{PME}$ is unity.

Approach 2 — Level of Effort

Another way cost analysts define X_{13} is by a level-of-effort formulation; that is,

$$X_{13} = \ell_{r_{X_{13}}} SL_{X_{13}} T_{X_{13}} \tag{6-98}$$

where $\ell_{r_{X_{13}}}$ is a labor rate (e.g., dollars per staff-month), $SL_{X_{13}}$ is the staff-level (i.e., the number of persons) needed for I&A, and $T_{X_{13}}$ is the number of months needed for I&A activities. From chapter 5 (table 5-9), if n is sufficiently large, then the distribution function of a product of n-independent random variables is approximately lognormal. If $\ell_{r_{X_{13}}}$, $SL_{X_{13}}$, and $T_{X_{13}}$ are independent then X_{13} is the product of three independent random variables. Are three independent random variables enough for the distribution function of X_{13} to be well approximated by the lognormal? After reading section 6.3, use the Monte Carlo simulation technique to explore this question.

To summarize, conditions can occur in the WBS that drive the distribution functions for X_{11}, X_{12}, X_{13} to be normal (or approximately normal). Recall $Cost_{PMP}$ is defined by

$$Cost_{PMP} = X_{11} + X_{12} + X_{13} = Cost_{PME} + X_{13} \tag{6-99}$$

where

$$Cost_{PME} = X_{11} + X_{12} \tag{6-100}$$

If X_{11} and X_{12} are independent normal random variables, then the distribution function for $Cost_{PME}$ is normal with mean

$$E(X_{11}) + E(X_{12})$$

and variance

$$Var(X_{11}) + Var(X_{12})$$

Furthermore, if $Cost_{PME}$ is normally distributed and X_{13} is defined by approach 1; that is, $X_{13} = aCost_{PME}$ $a > 0$ then $Cost_{PMP}$ is normally distributed (by proposition 6-1) with mean

$$(1+a)\left[E(X_{11}) + E(X_{12})\right]$$

and variance

$$(1+a)^2\left[Var(X_{11})+Var(X_{12})\right]$$

Even if X_{13} is not normal, which is certainly possible in approach 2, the distribution function of $Cost_{PMP}$ may still be approximately normal. However, this depends on the extent the distribution of $Cost_{PME}$ influences the overall distribution of $Cost_{PMP}$. If $Cost_{PME}$ is normal with standard deviation *significantly* larger than the standard deviation of X_{13} and X_{13} is independent of $Cost_{PME}$, it is possible that the normal distribution approximates the distribution of $Cost_{PMP}$. Again, it is a worthwhile exercise for the reader to explore cases when this is (and is not) true.

From these discussions, it is seen how frequently the distribution function for $Cost_{Sys}$ can become approximately normal. This is *not* to argue that $Cost_{Sys}$ is always normally distributed. Rather, it is to encourage cost analysts to *study the mathematical relationships* they define in a work breakdown structure to see whether analytical approximations to the distribution function of $Cost_{Sys}$ can be argued. Where possible, analytical forms of the distribution function (e.g., the normal, the lognormal, the beta) of $Cost_{Sys}$ are desirable. Such forms reveal much information about the "cost-behavior" in a system's work breakdown structure. They offer analysts and decision-makers insight about this behavior, so potential areas for cost-reductions and tradeoffs might be easily seen.

6.3 Monte Carlo Simulation

Throughout the many examples and case discussions presented in this book, analytical techniques have been used to develop (or approximate) the probability distribution of a system's cost. As previously stressed, analytical solutions to these types of problems are recommended. However, at times there are limitations when using analytical techniques. A system's work breakdown structure cost model can contain cost estimating relationships too

complex for strict analytical study. In such circumstances, a technique known as the Monte Carlo method is frequently used. This section provides an introduction to this method.

The Monte Carlo method falls into a class of techniques known as simulation. Simulation has varying definitions among practitioners. For instance, Winston [12] defines *simulation* as a technique that imitates the operation of a real-world system as it evolves over time. Rubinstein [13] offers a definition close to the context of this book:

> "Simulation is a numerical technique for conducting experiments on a digital computer, which involves certain types of mathematical and logical models that describe the behavior of a business or economic system (or some component thereof) over extended periods of real time."

With easy access to powerful microcomputers and applications software (such as electronic spreadsheets), simulation is a widely used problem-solving technique in management science and operations research.

The Monte Carlo method involves the generation of random variables from known, or assumed, probability distributions. The process of generating random variables from such distributions is known as *random variate generation* or *Monte Carlo sampling*. Simulations driven by Monte Carlo sampling are known as *Monte Carlo simulations*. Mentioned in the first chapter, one of the earliest applications of Monte Carlo simulation to cost analysis problems was at the RAND Corporation [14]. Since then, Monte Carlo simulation became (and remains) a popular approach for studying cost uncertainty, as well as in evaluating the cost-effectiveness of a system's design alternatives.

For cost uncertainty analysis, Monte Carlo simulation can be used to develop the empirical distribution of a system's cost. In concert with Rubinstein's definition, the WBS serves as the mathematical/logical cost model

of the system within which to conduct the simulation. In this context, the steps in a Monte Carlo simulation are as follows:

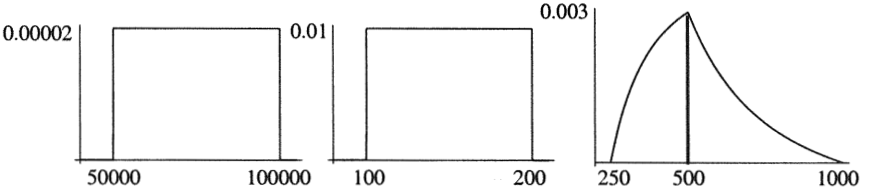
- For each random variable defined in the system's WBS, randomly select (sample) a value from its distribution function, which is known (or assumed).
- Once a set of feasible values for each random variable has been established, combine these values according to the mathematical relationships specified across the WBS (such as the relationships given in case discussions 6-1 and 6-2). This process produces a single value for the system's total cost.
- Repeat the above two steps n-times (e.g., ten-thousand times). This produces n-values each representing a possible (i.e., feasible) value for the system's total cost.
- Develop a frequency distribution from these n-values. This distribution is the simulated (i.e., empirical) distribution of total system cost.

In cost uncertainty analysis, Monte Carlo simulations are generally static simulations. *Static simulations* are those used to study the behavior of a system (or model) at a specific point in time. In contrast, *dynamic simulations* are those used to study such behavior as it changes over time.

To illustrate the concept of Monte Carlo sampling, consider the problem of determining the mean effort (in staff-months) to develop a software application. For discussion purposes, assume effort Eff_{SW} (refer to chapter 5) is given by

$$Eff_{SW} = \frac{I}{P_r} \tag{6-101}$$

where the distribution functions for I and P_r are given in figure 6-9.



I	P_r		Eff_{SW}
53,423	113		472.77
97,346	182		534.87
78,321	142	$Eff_{SW} = \frac{I}{P_r}$	551.56
92,837	187		496.45
51,837	158		328.08
93,283	163		572.29
73,176	109		671.34
83,710	151		554.37
67,194	169		397.60
66,512	102		652.08

Arithmetic Average 523.14 staff-months
Expected Value 519.86 staff-months
(see example 5-17)

Figure 6-9. Monte Carlo Sampling — 10 Random Samples Drawn from the Distribution Functions for I and P_r

In the Monte Carlo method, samples for I and P_r are randomly drawn from their distribution functions. These samples are Monte Carlo samples. For each sample (value) of I and P_r, a value for Eff_{SW} is computed according to equation 6-101. This process of sampling I and P_r and computing the associated Eff_{SW} is repeated thousands of times. From the many sampled values of Eff_{SW}, a simulated (empirical) probability distribution of Eff_{SW} is determined. In addition, various statistical measures such as the mean of Eff_{SW} can be computed from these sampled values. In figure 6-9, ten random samples of I and P_r are shown along with the associated values of Eff_{SW}. From these samples an average value of Eff_{SW} is computed. After

only ten Monte Carlo samples, this average is close to the computed expected value of Eff_{SW} (refer to example 5-17).

A way to randomly sample values from a given distribution function is essential to the Monte Carlo method. There are a number of well-established techniques for randomly sampling values. One method is the inverse transform method, which is presented in the following section. For a full discussion of random variate generation techniques, as well as the general topic of modeling and simulation, the reader is directed to Rubinstein [13] and Law and Kelton [15].

The Inverse Transform Method

The inverse transform method (ITM) is a popular technique for generating random variates from continuous distributions. It is a relatively straightforward method for distribution functions that exist in closed form, such as the uniform or triangular distributions (see chapter 4). Alternative random variate generation techniques, such as those described in Law and Kelton [15], are recommended for working with distribution functions that are not in closed form. The following illustrates the ITM.

Suppose a set of random variates for the size of a software application must be generated, where the distribution function for size (expressed as delivered source instructions I) is given by

$$I \sim Unif(50000,100000)$$

From equation 4-5 (chapter 4), the cumulative distribution function for I is

$$F_I(t) = \frac{t - 50000}{50000} \quad 50,000 \le t \le 100,000 \qquad (6\text{-}102)$$

To apply the ITM a random number η, where $0 \le \eta \le 1$, is generated. Next, a value for t that satisfies $\eta = F_I(t)$ is found. Repeating this process for various η produces Monte Carlo samples that stem from the given distribution function. In this case, Monte Carlo samples of I whose underlying distribution function is equation 6-102

are generated. For example, if a random number generator (discussed next) produces $\eta = 0.06846$, then the value of t such that

$$0.06846 = \frac{t - 50000}{50000}$$

is 53423, which is the first value of I shown in figure 6-9. Generalizing further, the above expression can be solved for any η; this yields

$$t = 50{,}000(\eta + 1) \qquad (6\text{-}103)$$

Equation 6-103 is known as the *random variate generator* for I. In particular, notice if $\eta = 0$, $\eta = \frac{1}{2}$, and $\eta = 1$, then equation 6-103 generates $t = 50{,}000$, $t = 75{,}000$ (which is the median of I), and $t = 100{,}000$, respectively. Thus, for any random number η the random variate generator given by equation 6-103 will produce Monte Carlo samples whose underlying distribution function is precisely that given by equation 6-102.

Essential to random variate generators is the generation of random numbers identified in the above discussion by η. In general, *random numbers* are independent random variables uniformly distributed over the unit interval. In Monte Carlo sampling, independent random samples are drawn from the *standard uniform distribution*, defined by equation 6-104.

$$f_X(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \qquad (6\text{-}104)$$

The statistical literature offers a number of algorithms for generating random numbers. One such generator, commonly available in many present-day software applications, is given by the recursive relationship

$$x_{i+1} = (ax_i + c)(\text{mod}\, m) \quad (i = 0, 1, 2, \ldots) \qquad (6\text{-}105)$$

where a (the *multiplier*), c (the *increment*), and m (the *modulus*) are nonnegative integers. Generators that produce random numbers by equation 6-105 are known as *linear congruential generators* [13, 15]. They produce a sequence of integers between 0 and $m - 1$. Equation 6-105 is equivalent to

$$x_{i+1} = ax_i + c - m\kappa_i \tag{6-106}$$

where $\kappa_i = [(ax_i + c)/m]$ is the largest *integer* less than or equal to $(ax_i + c)/m$. For each x_i $(i \geq 1)$, the associated random number between 0 and 1 is generated by $\eta_{i+1} = (x_{i+1})/m$. For example, suppose $a = 75$, $c = 50$, $m = 5000$, and $x_0 = 20$. The term x_0 is known as the initial value or seed. It is assigned arbitrarily to the random number generator. Using equation 6-106, the first two random numbers, η_1 and η_2, associated with the sequence of integers $x_1, x_2, \ldots, x_{4999}$ are

$$x_1 = 75(20) + 50 - 5000\kappa_0 = 1550 - 5000(0) = 1550$$

$$x_2 = 75(1550) + 50 - 5000\kappa_1 = 116300 - 5000(23) = 1300$$

where

$$\kappa_0 = [(75(20) + 50)/5000] = 0$$

$$\kappa_1 = [(75(1550) + 50)/5000] = [23.26] = 23$$

Thus, $$\eta_1 = \frac{1550}{5000} = 0.310 \text{ and } \eta_2 = \frac{1300}{5000} = 0.260$$

In a strict sense, random numbers generated by recursive relationships are not "purely random." Because they are produced by a deterministic procedure, with results that can be replicated, such numbers are considered "pseudorandom." In practice, the values of a, c, m, and x_0 are selected in a way to create a sequence of x_i's such that their corresponding η_i's appear to be statistically independent uniformly distributed random variates in the unit interval.

The Question of Sample Size in Monte Carlo Simulations

In Monte Carlo simulations, a question frequently asked is "*How many trials (the sample size) are necessary to have confidence in the outputs of the simulation*?" Morgan and Henrion [16] provide a guideline for determining sample size as a function of the precision desired in the outputs of a Monte Carlo simulation. Specifically, formulas are presented to address the question:

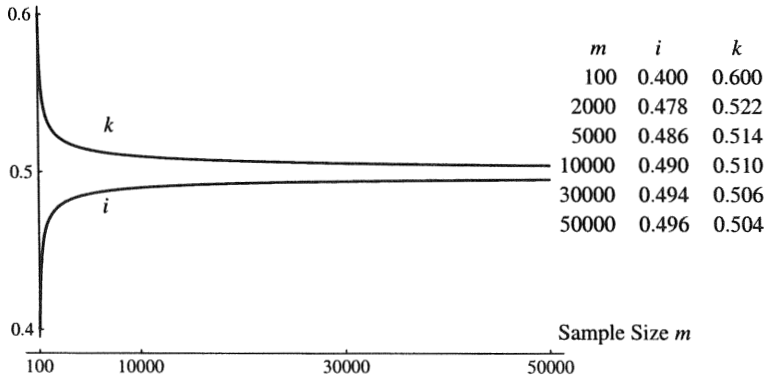
"*What sample size is needed so that, with probability α, a true fractile of the underlying distribution falls between a pair of fractiles estimated from the Monte Carlo sample*?"

Morgan-Henrion Guideline [16] Define m as the sample size and let x_p be the p-fractile of X (the underlying distribution); that is, $P(X \le x_p) = p$. Let c satisfy the probability $P(-c \le Z \le c) = \alpha$, where $Z \sim N(0,1)$. Then, the pair of fractiles $(\hat{x}_i, \hat{x}_k)$ estimated from a Monte Carlo sample with

$$i = \frac{mp - c\sqrt{mp(1-p)}}{m} = p - c\sqrt{\frac{p(1-p)}{m}} \tag{6-107}$$

$$k = \frac{mp + c\sqrt{mp(1-p)}}{m} = p + c\sqrt{\frac{p(1-p)}{m}} \tag{6-108}$$

contains x_p with probability α. For different sample sizes m, figure 6-10 illustrates with probability 0.95 ($c \approx 2$) the values of i and k such that the true median of the distribution falls between $(\hat{x}_i, \hat{x}_k)$. The lower and upper curves in figure 6-10 are generated from equations 6-107 and 6-108. As the sample size increases, the difference between the lower and upper curves decreases dramatically. With 100 samples you can be 95 percent confident the true median $x_{0.50}$ falls between the estimated fractiles $\hat{x}_{0.40}$ and $\hat{x}_{0.60}$. Increasing that sample size by a factor of 100 ($m = 10{,}000$) brings the same degree of confidence to within $\hat{x}_{0.49}$ to $\hat{x}_{0.51}$. As a guideline, 10,000 trials (Monte Carlo samples) should be sufficient to meet the precision requirements for most Monte Carlo simulations, particularly those conducted for cost uncertainty analyses.



m	i	k
100	0.400	0.600
2000	0.478	0.522
5000	0.486	0.514
10000	0.490	0.510
30000	0.494	0.506
50000	0.496	0.504

Figure 6-10. Sample Size for Monte Carlo Simulations

Exercises

Exercises 1 through 4 refer to case discussion 6-1.

1. Review case discussion 6-1 and verify the computations that led to $E(Cost_{Sys})$ and $Var(Cost_{Sys})$.

2. Prove theorem 6-1.

3. Referring to case discussion 6-1, use theorem 6-1 to show that
   a) $\rho_{X_3,X_1} = 0.9898$ b) $\rho_{X_3,W} = 0.1424$

4. The coordinates listed below are the twenty points shown in figure 6-5a. They are values for $(x, F_{Cost_{Sys}}(x))$ determined by Monte Carlo simulation. The simulation was run with a sample size of $n = 5000$.

   (31.01,0.05), (33.225,0.10), (34.76,0.15), (35.885,0.20), (36.849,0.25), (37.785,0.30), (38.67,0.35), (39.563,0.40), (40.272,0.45), (41.069,0.50), (41.728,0.55), (42.326,0.60), (43.191,0.65), (44.183,0.70), (45.151,0.75), (46.208,0.80), (47.368,0.85), (48.548,0.90), (51.028,0.95), (59.235,1)

Using the values above for $(x, F_{Cost_{Sys}}(x))$, apply the K-S test (chapter 5) to show $Cost_{Sys} \sim N(40.98, 36.18)$ is a statistically plausible model for the distribution function of $Cost_{Sys}$.

Exercises 5 through 9 refer to case discussion 6-2.

5. Review case discussion 6-2 and verify the computations that led to $E(Cost_{Sys})$ and $Var(Cost_{Sys})$.

6. Referring to table 6-4 and equation 6-19a, show that

   a) $Cov(Cost_{PMP}, Q) = 0$, where $Q = X_2 + X_3$

   b) $Cov(Cost_{PMP}, P) = 0$, where $P = PrgmSched$

7. Use the Mellin transform technique to verify, in case discussion 6-2, the mean and variance of the cost of STE, which was denoted by X_3.

8. Review case discussion 6-2 and verify the computations that led to the correlation between $Cost_{Sys}$ and $PrgmSched$.

9. The coordinates listed below are the twenty points shown in figure 6-7. They are values for $(x, F_{Cost_{Sys}}(x))$ determined by Monte Carlo simulation. The simulation was run with a sample size of $n = 5000$.

   (27.88,0.05), (28.72,0.10), (29.44,0.15), (29.97,0.20), (30.45,0.25), (30.9,0.30), (31.3,0.35), (31.74,0.40), (32.2,0.45), (32.64,0.50), (33.07,0.55), (33.43,0.60), (33.87,0.65), (34.41,0.70), (34.99,0.75), (35.6,0.80), (36.29,0.85), (37.39,0.90), (38.72,0.95), (45.71,1)

   Using the values above for $(x, F_{Cost_{Sys}}(x))$, apply the K-S test to show that a normal distribution and a lognormal distribution, each with mean 32.8 ($M) and standard deviation 3.3 ($M), are statistically plausible models for the distribution functions of $Cost_{Sys}$.

10. Use the *Inverse Transform Method* (section 6.3) to develop a random

number generator that produces triangularly distributed random variables.

References

1. United States Department of Defense. 1993. *Work Breakdown Structures for Defense Materiel Items*, MIL-STD-881B.

2. Blanchard, B. S., and W. J. Fabrycky. 1990. *Systems Engineering and Analysis*, 2nd ed. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

3. United States Air Force. 1994. *Unmanned Spacecraft Cost Model* (USCM), 7th ed. Los Angeles Air Force Base, California.

4. Larson, W. J., and J. R. Wertz (eds.). 1995. *Space Mission Analysis and Design*, 2nd ed. Norwell, Massachusetts: Kluwer Academic Press.

5. Garvey, P. R. 1990. A General Analytic Approach to System Cost Uncertainty Analysis, in W. R. Greer, Jr., and D. A. Nussbaum (eds.). *Cost Estimating and Analysis: Tools and Techniques*. pp. 161-181. New York: Springer-Verlag.

6. Garvey, P. R. 1996 (Spring). Modeling Cost and Schedule Uncertainties – A Work Breakdown Structure Perspective. *Military Operations Research*, V2, N1, pp. 37-43.

7. Abramson, R. L., and P. H. Young. 1997 (Spring). FRISKEM–Formal Risk Evaluation Methodology. *The Journal of Cost Analysis*, pp. 29-38.

8. Sobel, S. 1965. *A Computerized Technique to Express Uncertainty in Advanced System Cost Estimates*, ESD-TR-65-79. Bedford, Massachusetts: The MITRE Corporation.

9. McNichols, G. R. 1984 (Spring). The State of the Art of Cost Uncertainty Analysis. *Journal of Cost Analysis*, Vol. 1, No. 1, pp. 149-174.

10. Neimeier, H. 1994. Analytic Uncertainty Modeling. *International System Dynamics Conference Proceedings*. Stirling, Scotland.

11. Black R. L., and J. J. Wilder. 1982. Probabilistic Cost Approximations When Inputs Are Dependent. *Proceedings of the 1982 International Society of Parametric Analysts (ISPA) Conference*.

12. Winston, W. L. 1994. *Operations Research — Applications and Algorithms*. Belmont, California: Duxbury Press.

13. Rubinstein, R. Y. 1981. *Simulation and the Monte Carlo Method.* New York: John Wiley & Sons, Inc.

14. Dienemann, P. F. 1966. *Estimating Uncertainty Using Monte Carlo Techniques*, RM-4854-PR. Santa Monica, California: The RAND Corporation.

15. Law, A. M., and W. D. Kelton. 1991. *Simulation Modeling and Analysis*, 2nd ed. New York: McGraw-Hill, Inc.

16. Morgan, M. G., and M. Henrion. 1990. *Uncertainty: A Guide to Dealing With Uncertainty in Quantitative Risk and Policy Analysis.* New York: Cambridge University Press.

Chapter 7

Modeling Cost and Schedule Uncertainties — An Application of Joint Probability Theory

All uncertainty is fruitful...so long as it is accompanied by the wish to understand.
Antonio Machado
Juan De Mairena (1943)

Life is the art of drawing sufficient conclusions from insufficient premises.
Samuel Butler
Lord, What is Man?
Note-Books (1912)

7.1 Introduction

When cost uncertainty analyses are presented to decision-makers, questions often asked are "*What is the chance the system can be delivered within cost and schedule*?" "*How likely might the point estimate cost be exceeded for a given schedule*?" "*How are cost reserve recommendations affected by schedule risk*?" During the past thirty years, techniques from univariate probability theory have been widely applied to provide insight into $P(Cost \leq x_1)$ and $P(Schedule \leq x_2)$. Although it has long been recognized that a system's cost and schedule are correlated, little has been applied from multivariate probability theory to study joint cost-schedule distributions. A multivariate probability model would provide analysts and decision-makers visibility into joint and conditional cost-schedule probabilities, such as

$$P(Cost \leq x_1 \text{ and } Schedule \leq x_2)$$

and

$$P(Cost \leq x_1 \mid Schedule = x_2)$$

This chapter introduces modeling cost and schedule uncertainties by joint probability distributions. A family of joint distributions [1] has been developed for this purpose. This family consists of the classical bivariate normal and two lesser known joint distributions, the bivariate normal-lognormal and the bivariate lognormal. Experiences with Monte Carlo

simulations suggest these distributions are plausible models for computing joint and conditional cost-schedule probabilities. Appendixes B and C summarize key statistical formulas associated with the bivariate normal-lognormal and bivariate lognormal distributions. Formulas for the bivariate normal distribution are well known and are summarized in this chapter.

7.2 Joint Probability Models for Cost-Schedule

Mentioned above, decision-makers often require understanding how uncertainties between a system's cost and schedule interact. A decision-maker might bet on a "high-risk" schedule in hopes of keeping the system's cost within requirements. On the other hand, the decision-maker may be willing to assume "more cost" for a schedule with a small chance of being exceeded. This is a common tradeoff faced by decision-makers on systems engineering projects. This is illustrated in figure 7-1.

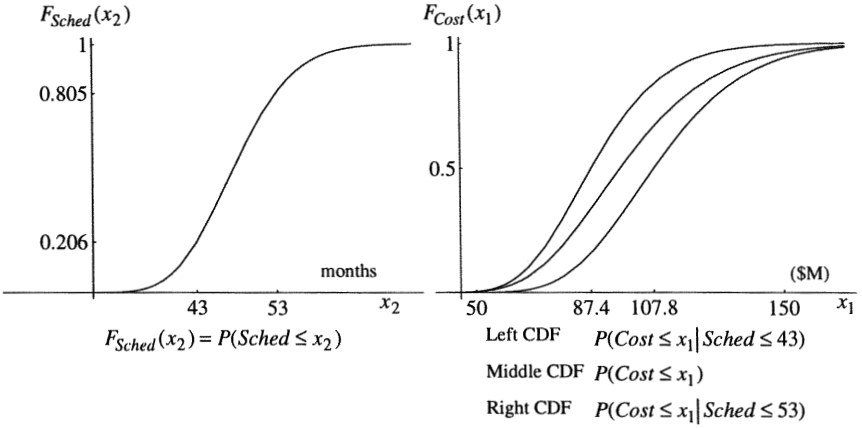


Figure 7-1. Illustrative Distributions for a System's Cost and Schedule

Suppose the cumulative distribution functions for a system's cost and schedule are shown in figure 7-1. The cumulative distribution function for schedule (the left-side of figure 7-1) indicates a 20 percent chance of delivering the system within 43 months. However, there is slightly better than

an 80 percent chance of doing so in 53 months. Given this information, a decision-maker might ask, *What is the cost tradeoff given these two possible schedule outcomes?* To answer this question, we need the distribution function of the system's cost *conditioned* on schedule. Three cumulative distribution functions for the system's cost are shown on the right-side of figure 7-1. The left CDF is the cost distribution *conditioned* on a schedule of 43 months. The right CDF is the cost distribution *conditioned* on a schedule of 53 months. The middle CDF is the overall cost distribution conditioned across the *entire* schedule distribution (i.e., not conditioned on a *specific* schedule outcome). The difference between the conditional median cost (107.8 ($M)) given a schedule of 53 months and the conditional median cost (87.4 ($M)) given a "high-risk" schedule of 43 months is 20.4 ($M).* In the context of figure 7-1, this difference in cost is certainly significant for any cost-schedule tradeoffs under consideration. This discussion highlights how joint probability models can be used to analyze cost-schedule interactions and reveal important tradeoffs between them.

The following presents a family of bivariate probability distributions for modeling cost-schedule uncertainty. This family of distributions are candidate theoretical models that may be assumed by an analyst, when joint or conditional cost-schedule probabilities are needed. These distributions have key features desirable for cost analysis. First, they can directly incorporate correlation between cost and schedule on a given system. Second, we will see that their marginal distributions are either both normal, normal and lognormal, or both lognormal. Shown throughout this book, marginal distributions such as these are frequently observed in Monte Carlo simulations [2,3] of system cost and schedule.

---

* Example 7-4 will discuss figure 7-1 further and show how these conditional median costs are determined.

7.2.1 The Bivariate Normal

This section presents the classical bivariate normal distribution and summarizes its major characteristics. An important feature of this distribution is its marginal distributions, which are both univariate normal.

In cost analysis, normal distributions can arise when a system's cost is the sum of many independent WBS cost element costs. Normal distributions can also occur in schedule analyses. For instance, a system's schedule is approximately normal if it is the sum of many independent activities in a schedule network. If normal distributions characterize a system's cost and schedule, then the bivariate normal could serve as an *assumed** model of their joint distribution.

Mathematical Definition

Suppose X_1 and X_2 are two random variables defined on $-\infty < x_1 < \infty$ and $-\infty < x_2 < \infty$. Let

$$E(X_1) = \mu_{X_1} = \mu_1 \tag{7-1}$$

$$E(X_2) = \mu_{X_2} = \mu_2 \tag{7-2}$$

$$Var(X_1) = \sigma^2_{X_1} = \sigma^2_1 \tag{7-3}$$

$$Var(X_2) = \sigma^2_{X_2} = \sigma^2_2 \tag{7-4}$$

The pair of random variables

$$(X_1, X_2) \sim Bivariate\, N((\mu_1, \mu_2), (\sigma^2_1, \sigma^2_2, \rho_{1,2})) \tag{7-5}$$

has a bivariate normal distribution if

---

* In general, the true joint distribution of (X_1, X_2) cannot be *uniquely* determined from the marginal distributions of X_1 and X_2. Only when random variables are *independent* can their joint distribution be obtained from their marginal distributions. From chapter 5 (section 5.1.2) recall that two random variables X_1 and X_2 are independent *if and only if* $f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$.

$$f_{X_1,X_2}(x_1,x_2)=\frac{1}{(2\pi)\sigma_1\sigma_2\sqrt{1-\rho_{1,2}^2}}e^{-\frac{1}{2}w} \tag{7-6}$$

where

$$w=\frac{1}{1-\rho_{1,2}^2}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2-2\rho_{1,2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}$$

for $-\infty < x_1 < \infty$ and $-\infty < x_2 < \infty$. The terms μ_i and σ_i^2 $(i=1,2)$ in the above expression are given by equations 7-1 through 7-4. The correlation term $\rho_{1,2}$ in equation 7-6 is

$$\rho_{1,2}=\rho_{X_1,X_2} \tag{7-7}$$

The admissible values for $\rho_{1,2}$ are given by the interval

$$-1<\rho_{1,2}<1$$

If two continuous random variables X_1 and X_2 have a bivariate normal distribution, then

$$P(a_1 \le X_1 \le b_1 \text{ and } a_2 \le X_2 \le b_2)=\int_{a_2}^{b_2}\int_{a_1}^{b_1} f_{X_1,X_2}(x_1,x_2)\,dx_1\,dx_2 \tag{7-8}$$

where $f_{X_1,X_2}(x_1,x_2)$ is given by equation 7-6.

Marginal and Conditional Distributions

A characteristic of the bivariate normal distribution is the distribution of X_1 and the distribution of X_2 are each univariate normal. These are the marginal distributions. They are given by

$$f_1(x_1)=\frac{1}{\sqrt{2\pi}\,\sigma_1}e^{-\frac{1}{2}\left[(x_1-\mu_1)^2/\sigma_1^2\right]} \tag{7-9}$$

$$f_2(x_2)=\frac{1}{\sqrt{2\pi}\,\sigma_2}e^{-\frac{1}{2}\left[(x_2-\mu_2)^2/\sigma_2^2\right]} \tag{7-10}$$

Important tradeoffs in cost analysis often involve assessing the impact a given set of schedules has on the likelihood that system cost will not exceed a required threshold. To make these assessments, the conditional probability distribution is needed. Conditional distributions provide probabilities of the type $P(X_1 \le a | X_2 = b)$. If two continuous random variables X_1 and X_2 have a bivariate normal distribution, then the conditional probability density function of X_1 given $X_2 = x_2$, denoted by $f_{X_1|x_2}(x_1)$, is normally distributed. That is,

$$X_1 | x_2 \sim N(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(x_2 - \mu_2), \sigma_1^2(1-\rho_{1,2}^2)) \qquad (7\text{-}11)$$

Similarly

$$X_2 | x_1 \sim N(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1 - \mu_1), \sigma_2^2(1-\rho_{1,2}^2)) \qquad (7\text{-}12)$$

From equations 7-11 and 7-12, the conditional means and variances of the bivariate normal distribution are

$$E(X_1|x_2) = \mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(x_2 - \mu_2) \qquad (7\text{-}13)$$

$$E(X_2|x_1) = \mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1 - \mu_1) \qquad (7\text{-}14)$$

$$Var(X_1|x_2) = \sigma_1^2(1-\rho_{1,2}^2) \qquad (7\text{-}15)$$

$$Var(X_2|x_1) = \sigma_2^2(1-\rho_{1,2}^2) \qquad (7\text{-}16)$$

Views of the Bivariate Normal

Figures 7-2 and 7-3 provide views of a bivariate normal density function. These figures are plots of

$$(X_1, X_2) \sim Bivariate\, N((100,48),(625,36,0.5))$$

Figure 7-2 is a surface view of this function, which has a "hill-like"

appearance. The marginal distributions of X_1 and X_2, viewed from the sides of the surface, are both univariate normal.

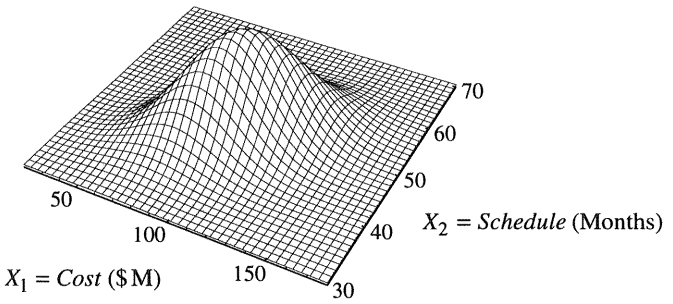


Figure 7-2. A Bivariate Normal Density
$(X_1, X_2) \sim Bivariate\, N((100,48),(625,36,0.5))$

The peak of the bivariate normal density function occurs at $x_1 = \mu_1$ and $x_2 = \mu_2$. In particular,

$$f_{X_1,X_2}(\mu_1,\mu_2) = \frac{1}{(2\pi)\sigma_1\sigma_2\sqrt{1-\rho_{1,2}^2}}$$

Another way to view the bivariate normal is to look at its topography, also known as its *contours*. Contours of constant probability density h are produced by finding x_1 and x_2 such that $h = f_{X_1,X_2}(x_1,x_2)$. In general, contours of the *bivariate normal* are ellipses concentric at (μ_1,μ_2). Figure 7-3 illustrates a set of contours for the bivariate normal density specified in figure 7-2. The innermost ellipse corresponds to $h = 0.001$, the middle ellipse corresponds to $h = 0.0005$, and the outer ellipse corresponds to $h = 0.0001$. The contour associated with the peak of the bivariate normal is given by the single point (μ_1,μ_2).

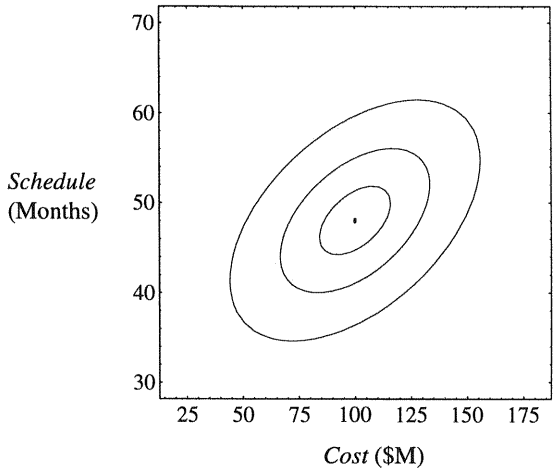


Figure 7-3. Contours of a Bivariate Normal Density
$(X_1, X_2) \sim Bivariate\, N((100,48),(625,36,0.5))$

Example 7-1 Prove that the function given by equation 7-6 is indeed a joint probability density function.

Solution To prove this, it is necessary to show

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)\,dx_2\,dx_1 = 1 \qquad (7\text{-}17)$$

With some algebra, the density function $f_{X_1,X_2}(x_1,x_2)$ (equation 7-6) can be factored as

$$f_{X_1,X_2}(x_1,x_2) = \left\{\frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x_1-\mu_1)^2/2\sigma_1^2}\right\} Q(x_1,x_2)$$

where

$$Q(x_1,x_2) = \left\{\frac{1}{\sqrt{2\pi}(\sigma_2\sqrt{1-\rho_{1,2}^2})} e^{-(x_2-b)^2/2\sigma_2^2(1-\rho_{1,2}^2)}\right\}$$

and $b = \mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1 - \mu_1)$. Substituting this factorization of $f_{X_1,X_2}(x_1, x_2)$ into equation 7-17 yields

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x_1-\mu_1)^2/2\sigma_1^2} \left\{ \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}(\sigma_2\sqrt{1-\rho_{1,2}^2})} e^{-(x_2-b)^2/2\sigma_2^2(1-\rho_{1,2}^2)} dx_2 \right\} dx_1$$

The right-most integrand in the expression above is the probability density function of a $N(b, \sigma_2^2(1-\rho_{1,2}^2))$ random variable, which by definition has integral equal to unity. Similarly, the left-most integrand in the expression above is the probability density of a $N(\mu_1, \sigma_1^2)$ random variable. Therefore,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\, dx_2\, dx_1 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x_1-\mu_1)^2/2\sigma_1^2} dx_1 = 1$$

Example 7-2 Suppose the joint probability density function of a system's cost and schedule is a bivariate normal given by

$$(X_1, X_2) \sim \textit{Bivariate}\, N((100, 48), (625, 36, 0.5))$$

where X_1 is the random variable that denotes the system's cost ($M) and X_2 is the random variable that denotes the system's schedule (months). Determine the median cost of the system conditioned on a schedule of 53 months.

Solution Following the notation specific to expression (7-5)

$$(X_1, X_2) \sim \textit{Bivariate}\, N((100, 48), (625, 36, 0.5))$$

implies $\mu_1 = 100$, $\mu_2 = 48$, $\sigma_1^2 = 625$, $\sigma_2^2 = 36$, and $\rho_{1,2} = 0.5$. The median system cost conditioned on a schedule of 53 months is found by computing $Med(X_1 | x_2 = 53)$. From expression 7-11, the conditional distribution of $X_1 | x_2$ is

$$X_1|x_2 \sim N(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(x_2 - \mu_2), \sigma_1^2(1-\rho_{1,2}^2))$$

Given the parameters $\mu_1 = 100$, $\mu_2 = 48$, $\sigma_1^2 = 625$, $\sigma_2^2 = 36$, and $\rho_{1,2} = 0.5$

$$X_1|x_2 \sim N(100 + 2.0833(x_2 - 48), 625(1-(0.5)^2))$$

and $$X_1|53 \sim N(110.42, 468.75)$$

Since the conditional distribution of system cost $X_1|x_2$ is normal,

$$Med(X_1|53) = E(X_1|53) = 110.42 \text{ (\$M)}$$

Figure 7-4 depicts the cumulative conditional cost distribution of $X_1|53$. The "point" shown along the distribution is aligned to $Med(X_1|53)$.

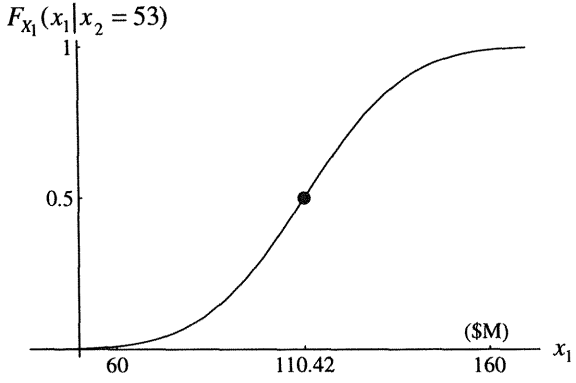


Figure 7-4. Cumulative Conditional Cost Distribution

$(X_1, X_2) \sim Bivariate\ N((100,48),(625,36,0.5))$

7.2.2 The Bivariate Normal-LogNormal

This section presents the bivariate normal-lognormal distribution and summarizes its major characteristics. An important feature of this distribution is its marginal distributions. One is normal and the other is lognormal.

In cost analysis, it is common for the distribution functions of a system's cost and schedule to be normal and lognormal, respectively, In particular, a

system's schedule is often observed (from Monte Carlo simulations) to be lognormal if it is the sum of many positively correlated schedule activities in an overall schedule network. Thus, if normal and lognormal distributions characterize a system's cost and schedule (or vice versa), then the bivariate normal-lognormal could serve as an *assumed* model of their joint distribution.

Mathematical Definition

Suppose $Y_1 = X_1$ and $Y_2 = \ln X_2$ are two random variables where X_1 and X_2 are defined on $-\infty < x_1 < \infty$ and $0 < x_2 < \infty$. If Y_1 and Y_2 each have a normal distribution, then the mean and variance of Y_i $(i = 1,2)$ are

$$E(Y_1) = \mu_{Y_1} = \mu_{X_1} = \mu_1 \tag{7-18}$$

$$Var(Y_1) = \sigma_{Y_1}^2 = \sigma_{X_1}^2 = \sigma_1^2 \tag{7-19}$$

$$E(Y_2) = \mu_{Y_2} = \mu_2 = \frac{1}{2}\ln\left[\frac{(\mu_{X_2})^4}{(\mu_{X_2})^2 + \sigma_{X_2}^2}\right] \tag{7-20}$$

$$Var(Y_2) = \sigma_{Y_2}^2 = \sigma_2^2 = \ln\left[\frac{(\mu_{X_2})^2 + \sigma_{X_2}^2}{(\mu_{X_2})^2}\right] \tag{7-21}$$

The pair of random variables

$$(X_1, X_2) \sim \textit{Bivariate NLogN}((\mu_1, \mu_2), (\sigma_1^2, \sigma_2^2, \rho_{1,2})) \tag{7-22}$$

has a bivariate normal-lognormal distribution if

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{(2\pi)\sigma_1\sigma_2\sqrt{1-\rho_{1,2}^2}\, x_2} e^{-\frac{1}{2}w} \tag{7-23}$$

where

$$w = \frac{1}{1-\rho_{1,2}^2}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho_{1,2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{\ln x_2-\mu_2}{\sigma_2}\right)+\left(\frac{\ln x_2-\mu_2}{\sigma_2}\right)^2\right\}$$

for $-\infty < x_1 < \infty$ and $0 < x_2 < \infty$. The terms μ_i and σ_i^2 $(i=1,2)$ in the above expression are specifically given by equations 7-18 through 7-21. The correlation term $\rho_{1,2}$ in equation 7-23 (derived in appendix B) is

$$\rho_{1,2} = \rho_{Y_1,Y_2} = \rho_{X_1,\ln X_2} = \rho_{X_1,X_2}\frac{(e^{\sigma_2^2}-1)^{1/2}}{\sigma_2} \qquad (7\text{-}24)$$

The admissible values for $\rho_{1,2}$ are given by the interval $-1 < \rho_{1,2} < 1$. Therefore, admissible values for ρ_{X_1,X_2} (in equation 7-24) are *restricted* to the interval

$$\frac{-\sigma_2}{\sqrt{e^{\sigma_2^2}-1}} < \rho_{X_1,X_2} < \frac{\sigma_2}{\sqrt{e^{\sigma_2^2}-1}} \qquad (7\text{-}25)$$

If two continuous random variables X_1 and X_2 have a bivariate normal-lognormal distribution, then

$$P(a_1 \le X_1 \le b_1 \text{ and } a_2 \le X_2 \le b_2) = \int_{a_2}^{b_2}\int_{a_1}^{b_1} f_{X_1,X_2}(x_1,x_2)\,dx_1\,dx_2 \qquad (7\text{-}26)$$

where $f_{X_1,X_2}(x_1,x_2)$ is given by equation 7-23.

Marginal and Conditional Distributions

For the bivariate normal-lognormal distribution given by equation 7-23, the distribution of X_1 is normal and the distribution of X_2 is lognormal. These are the marginal distributions. They are given by

$$f_1(x_1) = \frac{1}{\sqrt{2\pi}\,\sigma_1}e^{-\frac{1}{2}\left[(x_1-\mu_1)^2/\sigma_1^2\right]} \qquad (7\text{-}27)$$

$$f_2(x_2) = \frac{1}{\sqrt{2\pi}\,\sigma_2\,x_2} e^{-\frac{1}{2}\left[(\ln x_2 - \mu_2)^2/\sigma_2^2\right]} \tag{7-28}$$

The conditional distributions of the bivariate normal-lognormal distribution are normal and lognormal. In particular,

$$X_1|x_2 \sim N(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2), \sigma_1^2(1-\rho_{1,2}^2)) \tag{7-29}$$

and

$$X_2|x_1 \sim LogN(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1 - \mu_1), \sigma_2^2(1-\rho_{1,2}^2)) \tag{7-30}$$

From these conditional distributions it can be readily shown (left for the reader) that

$$E(X_1|x_2) = \mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2) \tag{7-31}$$

$$E(X_2|x_1) = e^{\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1-\mu_1) + \frac{1}{2}\sigma_2^2(1-\rho_{1,2}^2)} \tag{7-32}$$

and

$$Var(X_1|x_2) = \sigma_1^2(1-\rho_{1,2}^2) \tag{7-33}$$

$$Var(X_2|x_1) = e^{2(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1-\mu_1))} e^z(e^z - 1) \tag{7-34}$$

where $z = \sigma_2^2(1-\rho_{1,2}^2)$.

Views of the Bivariate Normal-LogNormal

Figures 7-5 and 7-6 provide views of a bivariate normal-lognormal density function. These figures are plots of

$$(X_1, X_2) \sim Bivariate\ NLogN((100, 3.86345), (625, 0.0155042, 0.501944))$$

Figure 7-5 is a surface view of the function, which has a "hill-like" appearance. The marginal distributions of X_1 and X_2, when viewed from the sides of the surface, are univariate normal and univariate lognormal,

respectively. A topographic view of a bivariate normal-lognormal density function in figure 7-5 is shown in figure 7-6.

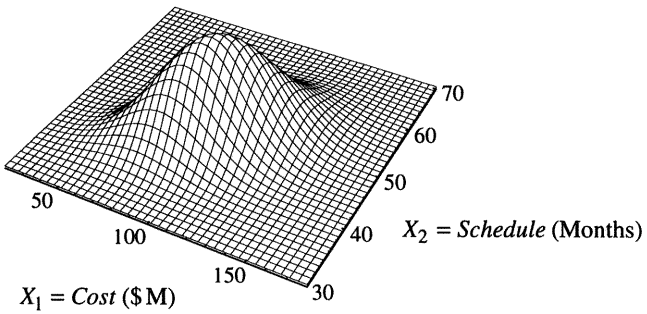


Figure 7-5. A Bivariate Normal-LogNormal Density

$(X_1, X_2) \sim$ *Bivariate NLogN*((100,3.86345),(625,0.0155042,0.501944))

In figure 7-6, the innermost contour corresponds to $h = 0.001$, the middle contour corresponds to $h = 0.0005$, and the outer contour corresponds to $h = 0.0001$. The point $(\mu_{X_1}, \mu_{X_2}) = (100, 48)$, shown in figure 7-6, stems from

$$(X_1, X_2) \sim \textit{Bivariate NLogN}((100, 3.86345), (625, 0.0155042, 0.501944))$$

This is seen in the following example.

Example 7-3 Assume the joint probability density function of a system's cost X_1 and schedule X_2 is bivariate normal-lognormal with density function given by equation 7-23. Suppose X_1 has mean 100 ($M) and variance 625 ($M)^2. Suppose X_2 has mean 48 (months) and variance 36 (months)^2. If the correlation between the system's cost and schedule is

$$\rho_{X_1, X_2} = 0.5$$

determine the median system cost conditioned on a schedule of 53 months.

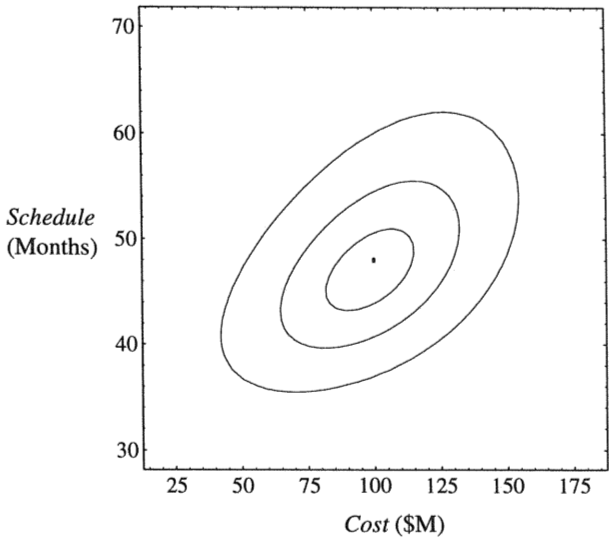


Figure 7-6. Contours of a Bivariate Normal-LogNormal Density

$(X_1, X_2) \sim$ *Bivariate NLogN*((100,3.86345),(625,0.0155042,0.501944))

Solution First, determine the five parameters that specify the bivariate normal-lognormal defined by expression 7-22. Since $\mu_{X_1} = 100$, $\sigma^2_{X_1} = 625$, $\mu_{X_2} = 48$, $\sigma^2_{X_2} = 36$, equations 7-18 through 7-21 give

$$E(Y_1) = \mu_{Y_1} = \mu_{X_1} = \mu_1 = 100 \tag{7-35}$$

$$Var(Y_1) = \sigma^2_{Y_1} = \sigma^2_{X_1} = \sigma^2_1 = 625 \tag{7-36}$$

$$E(Y_2) = \mu_{Y_2} = \mu_2 = \frac{1}{2}\ln\left[\frac{(\mu_{X_2})^4}{(\mu_{X_2})^2 + \sigma^2_{X_2}}\right] = 3.86345 \tag{7-37}$$

$$Var(Y_2) = \sigma^2_{Y_2} = \sigma^2_2 = \ln\left[\frac{(\mu_{X_2})^2 + \sigma^2_{X_2}}{(\mu_{X_2})^2}\right] = 0.0155042 \tag{7-38}$$

$$\rho_{1,2} = \rho_{Y_1,Y_2} = \rho_{X_1,\ln X_2} = \rho_{X_1,X_2} \frac{(e^{\sigma_2^2} - 1)^{1/2}}{\sigma_2} = 0.501944 \qquad (7\text{-}39)$$

From 7-25, the interval for the correlation between X_1 and X_2, in this example, is restricted to

$$-0.996126 < \rho_{X_1,X_2} < 0.996126$$

Thus, the correlation given between the system's cost and schedule is admissible since $-0.996126 < 0.5 < 0.996126$. From the above computations, the parameters of the bivariate normal-lognormal distribution are

$$(X_1, X_2) \sim \textit{Bivariate NLogN}((100, 3.86345), (625, 0.0155042, 0.501944)) \quad (7\text{-}40)$$

The median system cost conditioned on a schedule of 53 months is found by computing $Med(X_1 | x_2 = 53)$. From expression 7-29, the conditional distribution of $X_1 | x_2$ is

$$X_1 | x_2 \sim N(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2), \sigma_1^2(1 - \rho_{1,2}^2))$$

From equation 7-35 through 7-39

$$X_1 | x_2 \sim N(100 + 100.8(\ln x_2 - 3.86345), 625(1 - (0.501944)^2))$$

and $X_1 | 53 \sim N(110.8, 467.5)$

Since the conditional distribution of system cost $X_1 | x_2$ is normal

$$Med(X_1 | 53) = E(X_1 | 53) = 110.8 \text{ (\$M)}$$

Figure 7-7 depicts the cumulative conditional cost distribution of $X_1 | 53$. The "point" shown along the distribution is aligned to $Med(X_1 | 53)$.

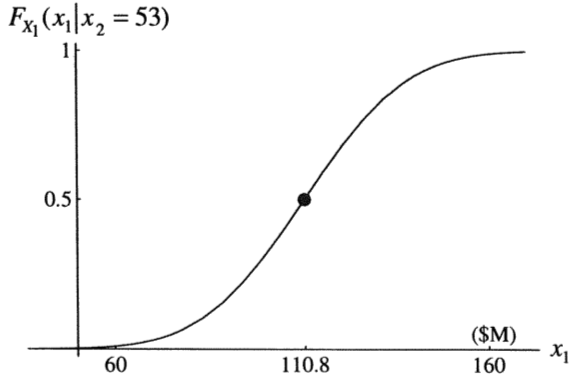


Figure 7-7. Cumulative Conditional Cost Distribution

$(X_1, X_2) \sim$ *Bivariate NLogN*((100,3.86345),(625,0.0155042,0.501944))

7.2.3 The Bivariate LogNormal

This section presents the bivariate lognormal and summarizes its major characteristics. From a practical perspective, if the distribution functions of a system's cost and schedule are lognormal, then the bivariate lognormal could serve as an *assumed* model of their joint distribution. However, it again must be emphasized that this is indeed an assumption. In general, the true joint distribution of a pair of random variables (X_1, X_2) cannot be *uniquely* determined from the marginal distributions of X_1 and X_2. Only when random variables are *independent* can their joint distribution be obtained from their marginal distributions.

Mathematical Definition

Suppose $Y_1 = \ln X_1$ and $Y_2 = \ln X_2$ are two random variables where X_1 and X_2 are defined on $0 < x_1 < \infty$ and $0 < x_2 < \infty$. If Y_1 and Y_2 each have a normal distribution, then the mean and variance of Y_i $(i = 1,2)$ are

$$E(Y_i) = \mu_{Y_i} = \mu_i = \frac{1}{2}\ln\left[\frac{(\mu_{X_i})^4}{(\mu_{X_i})^2 + \sigma_{X_i}^2}\right] \tag{7-41}$$

$$Var(Y_i) = \sigma_{Y_i}^2 = \sigma_i^2 = \ln\left[\frac{(\mu_{X_i})^2 + \sigma_{X_i}^2}{(\mu_{X_i})^2}\right] \tag{7-42}$$

The pair of random variables

$$(X_1, X_2) \sim \text{Bivariate } LogN((\mu_1, \mu_2), (\sigma_1^2, \sigma_2^2, \rho_{1,2})) \tag{7-43}$$

has a bivariate lognormal distribution if

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{(2\pi)\sigma_1\sigma_2\sqrt{1-\rho_{1,2}^2}\,x_1x_2}e^{-\frac{1}{2}w} \tag{7-44}$$

where

$$w = \frac{1}{1-\rho_{1,2}^2}\left\{\left(\frac{\ln x_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho_{1,2}\left(\frac{\ln x_1 - \mu_1}{\sigma_1}\right)\left(\frac{\ln x_2 - \mu_2}{\sigma_2}\right) + \left(\frac{\ln x_2 - \mu_2}{\sigma_2}\right)^2\right\}$$

for $0 < x_1 < \infty$ and $0 < x_2 < \infty$. The terms μ_i and σ_i^2 $(i = 1,2)$ in the above expression are given by equation 7-41 and equation 7-42. The correlation term $\rho_{1,2}$ in equation 7-44 (derived in appendix C) is

$$\rho_{1,2} = \frac{1}{\sigma_1\sigma_2}\ln\left[1 + \rho_{X_1,X_2}\sqrt{e^{\sigma_1^2} - 1}\sqrt{e^{\sigma_2^2} - 1}\right] \tag{7-45}$$

The admissible values for $\rho_{1,2}$ are given by the interval $-1 < \rho_{1,2} < 1$. From this, it can be shown that admissible values for ρ_{X_1,X_2} (in equation 7-45) are *restricted* to the interval

$$\frac{e^{-\sigma_1\sigma_2} - 1}{\sqrt{e^{\sigma_1^2} - 1}\sqrt{e^{\sigma_2^2} - 1}} < \rho_{X_1,X_2} < \frac{e^{\sigma_1\sigma_2} - 1}{\sqrt{e^{\sigma_1^2} - 1}\sqrt{e^{\sigma_2^2} - 1}} \tag{7-46}$$

If two continuous random variables X_1 and X_2 have a bivariate lognormal distribution, then

$$P(a_1 \le X_1 \le b_1 \text{ and } a_2 \le X_2 \le b_2) = \int_{a_2}^{b_2}\int_{a_1}^{b_1} f_{X_1,X_2}(x_1,x_2)\,dx_1\,dx_2 \quad (7\text{-}47)$$

where $f_{X_1,X_2}(x_1,x_2)$ is given by equation 7-44.

Marginal and Conditional Distributions

For the bivariate lognormal distribution (given by equation 7-44), the distribution of X_1 is lognormal and the distribution of X_2 is lognormal. The marginal distributions are given by

$$f_1(x_1) = \frac{1}{\sqrt{2\pi}\,\sigma_1 x_1} e^{-\frac{1}{2}\left[(\ln x_1 - \mu_1)^2/\sigma_1^2\right]} \quad (7\text{-}48)$$

$$f_2(x_2) = \frac{1}{\sqrt{2\pi}\,\sigma_2 x_2} e^{-\frac{1}{2}\left[(\ln x_2 - \mu_2)^2/\sigma_2^2\right]} \quad (7\text{-}49)$$

The conditional distributions of the bivariate lognormal distribution are both lognormal. In particular,

$$X_1 | x_2 \sim LogN(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2),\ \sigma_1^2(1-\rho_{1,2}^2)) \quad (7\text{-}50)$$

and

$$X_2 | x_1 \sim LogN(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(\ln x_1 - \mu_1),\ \sigma_2^2(1-\rho_{1,2}^2)) \quad (7\text{-}51)$$

From these conditional distributions it can be readily shown (left for the reader) that

$$E(X_1|x_2) = x_2^{\frac{\sigma_1}{\sigma_2}\rho_{1,2}} e^{\mu_1 - \frac{\sigma_1}{\sigma_2}\rho_{1,2}\mu_2 + \frac{1}{2}\sigma_1^2(1-\rho_{1,2}^2)} \quad (7\text{-}52)$$

$$E(X_2|x_1) = x_1^{\frac{\sigma_2}{\sigma_1}\rho_{1,2}} e^{\mu_2 - \frac{\sigma_2}{\sigma_1}\rho_{1,2}\mu_1 + \frac{1}{2}\sigma_2^2(1-\rho_{1,2}^2)} \quad (7\text{-}53)$$

and $$Var(X_1|x_2) = x_2^{2\frac{\sigma_1}{\sigma_2}\rho_{1,2}} e^{2(\mu_1 - \frac{\sigma_1}{\sigma_2}\rho_{1,2}\mu_2)} e^{z^o}(e^{z^o} - 1) \qquad (7\text{-}54)$$

$$Var(X_2|x_1) = x_1^{2\frac{\sigma_2}{\sigma_1}\rho_{1,2}} e^{2(\mu_2 - \frac{\sigma_2}{\sigma_1}\rho_{1,2}\mu_1)} e^{z}(e^{z} - 1) \qquad (7\text{-}55)$$

where $z^o = \sigma_1^2(1-\rho_{1,2}^2)$ and $z = \sigma_2^2(1-\rho_{1,2}^2)$.

Views of the Bivariate LogNormal

Figures 7-8 and 7-9 provide views of a bivariate lognormal density function. These figures are plots of

$(X_1, X_2) \sim Bivariate\ LogN((4.57486, 3.86345), (0.0606246, 0.0155042, 0.505708))$

Figure 7-8 is a surface view of the function, which has a "hill-like" appearance. The marginal distributions of X_1 and X_2, viewed from the sides of the surface, are both univariate lognormal.

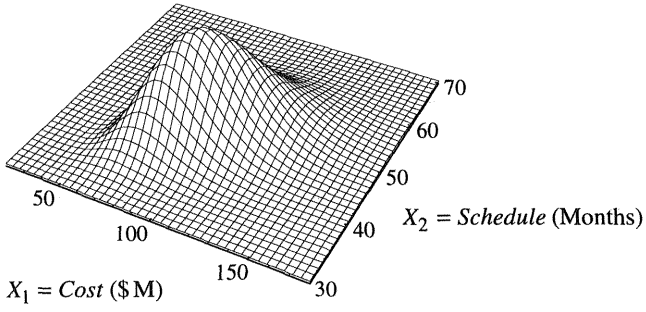


Figure 7-8. A Bivariate LogNormal Density

$(X_1, X_2) \sim Bivariate\ LogN((4.57486, 3.86345), (0.0606246, 0.0155042, 0.505708))$

A topographic view of a bivariate lognormal density function in figure 7-8 is shown in figure 7-9. In figure 7-9, the innermost contour corresponds to $h = 0.001$, the middle contour corresponds to $h = 0.0005$, and the outer

contour corresponds to $h = 0.0001$. The point $(\mu_{X_1}, \mu_{X_2}) = (100, 48)$, shown in figure 7-9, stems from

$(X_1, X_2) \sim$ *Bivariate LogN*$((4.57486, 3.86345), (0.0606246, 0.0155042, 0.505708))$

This is seen in the following example.

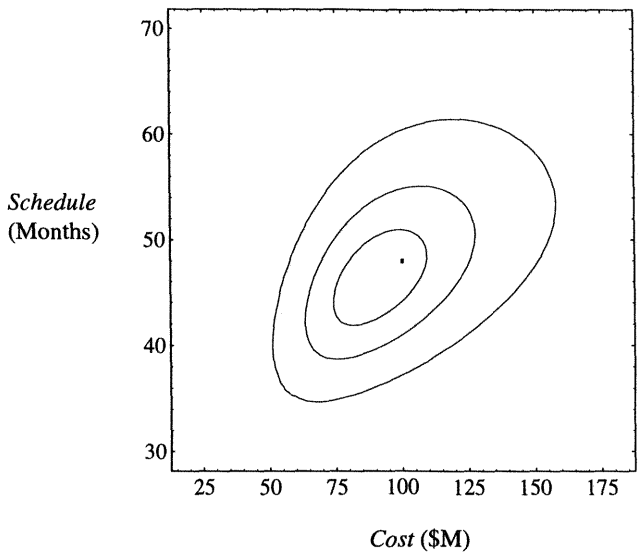


Figure 7-9. Contours of a Bivariate LogNormal Density

$(X_1, X_2) \sim$ *Bivariate LogN*$((4.57486, 3.86345), (0.0606246, 0.0155042, 0.505708))$

Example 7-4 Assume the joint probability density function of a system's cost X_1 and schedule X_2 is bivariate lognormal with density function given by equation 7-44. Suppose X_1 has mean 100 ($M) and variance 625 ($M)^2. Suppose X_2 has mean 48 (months) and variance 36 (months)^2. Let cost and schedule have a correlation of 0.5. Show that the difference between the median system cost conditioned on a schedule with a 20 percent chance of being achieved and the median system cost conditioned on a schedule with an 80 percent chance of being achieved is 20.4 ($M).

Solution It is given that $\mu_{X_1} = 100$, $\sigma^2_{X_1} = 625$, $\mu_{X_2} = 48$, $\sigma^2_{X_2} = 36$, and $\rho_{X_1,X_2} = 0.5$. From equations 7-41, 7-42, and 7-45 the parameters of the bivariate lognormal, given in expression 7-43, are

$$\mu_1 = \frac{1}{2}\ln\left[\frac{(\mu_{X_1})^4}{(\mu_{X_1})^2 + \sigma^2_{X_1}}\right] = 4.57486 \qquad \sigma_1^2 = \ln\left[\frac{(\mu_{X_1})^2 + \sigma^2_{X_1}}{(\mu_{X_1})^2}\right] = 0.0606246$$

$$\mu_2 = \frac{1}{2}\ln\left[\frac{(\mu_{X_2})^4}{(\mu_{X_2})^2 + \sigma^2_{X_2}}\right] = 3.86345 \qquad \sigma_2^2 = \ln\left[\frac{(\mu_{X_2})^2 + \sigma^2_{X_2}}{(\mu_{X_2})^2}\right] = 0.0155042$$

$$\rho_{1,2} = \frac{1}{\sigma_1\sigma_2}\ln\left[1 + \rho_{X_1,X_2}\sqrt{e^{\sigma_1^2} - 1}\sqrt{e^{\sigma_2^2} - 1}\right] = 0.50578$$

From expression 7-50, the cost distribution X_1 conditioned on a schedule of x_2 months is

$$X_1 | x_2 \sim LogN(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2), \sigma_1^2(1 - \rho_{1,2}^2))$$

so

$$X_1 | x_2 \sim LogN(4.57486 + (\ln x_2 - 3.86345), 0.0451204) \qquad (7\text{-}56)$$

Figure 7-1 illustrates the cumulative distribution functions associated with this example. The schedule distribution is shown on the left-side of figure 7-1. Since X_2 is lognormal with mean 48 (months) and variance 36 (months)^2, $X_2 \sim LogN(3.86345, 0.0155042)$. It is left to the reader to show (chapter 4, section 4.4) the value of x_2 such that $P(X_2 \le x_2) = 0.20$ is 43 months (rounded). Similarly, the value of x_2 such that $P(X_2 \le x_2) = 0.80$ is 53 months (rounded). From expression 7-56, the conditional cost distribution given $x_2 = 43$ months is

$$X_1 | 43 \sim LogN(4.47, 0.045)$$

Likewise, the conditional cost distribution given $x_2 = 53$ months is

$$X_1 | 53 \sim LogN(4.68, 0.045)$$

Since $X_1 | x_2$ is lognormal, we know from equation 4-39 (chapter 4) that

$$Med(X_1 | 43) = e^{4.47} = 87.4 \text{ (\$M)}$$

$$Med(X_1 | 53) = e^{4.68} = 107.8 \text{ (\$M)}$$

Therefore, the difference between the median system cost conditioned on a schedule with a 20 percent chance of being achieved and the median system cost conditioned on a schedule with an 80 percent chance of being achieved is 20.4 (\$M).

7.2.4 Case Discussion

In case discussion 7-1, we determine the cost of the digital information system (discussed in case discussion 6-2) that has a 5 percent chance of being exceeded but is conditioned on a development schedule that has a 5 percent chance of being exceeded.

Case Discussion 7-1 In case discussion 6-2 (chapter 6), the random variable $Cost_{Sys}$ denoted the total cost (\$K) of a digital information system and the random variable *PrgmSched* represented its development duration (in months). Suppose the joint probability density function of $Cost_{Sys}$ and *PrgmSched* is bivariate normal. Let b be the number of months such that $P(PrgmSched \le b) = 0.95$, where *PrgmSched* is normally distributed with $E(PrgmSched) = 33.36$ (months) and $Var(PrgmSched) = 1.94$ (months)^2. Determine a such that $P(Cost_{Sys} \le a | PrgmSched = b) = 0.95$.

To determine a, we first find b such that $P(PrgmSched \le b) = 0.95$. This probability can be written as $P(PrgmSched \le b) = P(Z \le v)$, where

$$\upsilon = \frac{b - E(PrgmSched)}{\sigma_{PrgmSched}} = \frac{b - 33.36}{1.39283}$$

From table A-1, we have $P(Z \le \upsilon) = 0.95$ if $\upsilon = \frac{b - 33.36}{1.39283} = 1.645 \Rightarrow b = 35.65$.

Now, it remains to determine a such that

$$P(Cost_{Sys} \le a \mid PrgmSched = 35.65) = 0.95$$

Since the joint probability density function of $Cost_{Sys}$ and $PrgmSched$ is given to be bivariate normal, from expression 7-11, the distribution of $Cost_{Sys}$ conditioned on $PrgmSched$ is

$$Cost_{Sys} \mid PrgmSched = x_2 \sim N(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(x_2 - \mu_2), \sigma_1^2(1 - \rho_{1,2}^2))$$

From case discussion 6-2 (chapter 6)

$$\mu_1 = E(Cost_{Sys}) = 32841.1 \qquad \sigma_1^2 = Var(Cost_{Sys}) = 10800698.3$$

$$\mu_2 = E(PrgmSched) = 33.36 \qquad \sigma_2^2 = Var(PrgmSched) = 1.94$$

and $\rho_{1,2} = \rho_{Cost_{Sys},PrgmSched} = 0.206$

Therefore,

$$Cost_{Sys} \mid PrgmSched = x_2 \sim N(32841.1 + 486.06(x_2 - \mu_2), 10342359.87)$$

At $x_2 = 35.65$ we have

$$Cost_{Sys} \mid PrgmSched = 35.65 \sim N(33954.18, 10342359.87)$$

The density function of $Cost_{Sys}$ conditioned on a system schedule of 35.65 months is normal, with mean 33954.18 ($K) and variance 10342359.87 ($K)^2. To find a such that $P(Cost_{Sys} \le a \mid PrgmSched = 35.65) = 0.95$, let

$$P(Cost_{Sys} \le a \mid PrgmSched = 35.65) = P(Z \le \varphi)$$

where $\varphi = \frac{a - 33954.18}{\sqrt{10342359.87}}$. From table A-1, $P(Z \le \varphi) = 0.95$ if

$$\varphi = \frac{a - 33954.18}{\sqrt{10342359.87}} = 1.645$$

This implies that $a = 39244.4$. Thus, the cost of the digital information system that has only a 5 percent chance of being exceeded, when conditioned on a schedule having the same chance of being exceeded, is 39244.4 ($K).

7.3 Summary

The family of distributions described in this chapter provides an analytical basis for computing joint and conditional cost-schedule probabilities. They are mathematical models that might be hypothesized for capturing the joint interactions between a system's cost and schedule.

Seen throughout this chapter, a parameter required by these models is the correlation between cost and schedule.* This can be a difficult value to determine. One approach is the direct computation of the correlation as illustrated in case discussion 6-2 (refer to equation 6-86). However, in some instances this might not be analytically possible or practical. Another approach is to obtain an estimate of the correlation, from sample values generated by Monte Carlo simulation. This is a reasonable method that can be done regardless of the complexity of the cost-schedule estimation relationships. Subjective assessments might be used. However, care must be taken to specify an *admissible correlation* for the particular pair of random variables. Furthermore, there may already exist an implied correlation by virtue of how the cost-schedule estimation relationships are mathematically defined (refer to case discussion 6-2). Subjectively specifying a correlation when one is already present (only its magnitude is unknown or yet to be

* Because these models treat cost and schedule as correlated random variables, it is important to recognize that *they do not capture causal impacts* that schedule compression or extension has on cost.

determined) is double counting correlation. Such a situation invalidates the mathematical integrity of the cost uncertainty analysis.

In summary, systems engineering typically takes place in environments of limited funds and challenging schedules. It is incumbent upon engineers and analysts to continually assess affordability relative to the chance of jointly meeting cost and schedule, or meeting cost for a given feasible schedule, against specific tradeoffs in system requirements, acquisition strategies, and post-development support. The distributions described in this chapter are one way such assessments may be made.

Exercises

1. Suppose the mean cost and mean schedule of a program is 100 ($M) and 48 months, respectively. Furthermore, suppose the program's cost and schedule variances are 625 (months)^2 and 36 (months)^2, respectively. If the correlation between the program's cost and schedule is 0.5, find x_1 such that

   a) $P(Cost \le x_1 \,|\, x_2 = 53\,\text{months}) = 0.95$ if program cost and schedule have a bivariate normal distribution.

   b) $P(Cost \le x_1 \,|\, x_2 = 53\,\text{months}) = 0.95$ if program cost and schedule have a bivariate normal-lognormal distribution.

   c) $P(Cost \le x_1 \,|\, x_2 = 53\,\text{months}) = 0.95$ if program cost and schedule have a bivariate lognormal distribution.

2. Suppose $(X_1, X_2) \sim Bivariate\ NLogN((\mu_1, \mu_2), (\sigma_1^2, \sigma_2^2, \rho_{1,2}))$ where μ_1, μ_2, σ_1^2, σ_2^2, and $\rho_{1,2}$ are defined in section 7.2.2. If $\mu_{X_2} = \sqrt{e}$ and $\sigma_{X_2}^2 = e(e-1)$ show that $-\frac{1}{\sqrt{e-1}} < \rho_{X_1,X_2} < \frac{1}{\sqrt{e-1}}$.

3. Suppose $(X_1, X_2) \sim$ *Bivariate LogN*$((\mu_1, \mu_2), (\sigma_1^2, \sigma_2^2, \rho_{1,2}))$ where μ_1, μ_2, σ_1^2, σ_2^2, and $\rho_{1,2}$ are defined in section 7.2.3. If $\mu_{X_1} = \mu_{X_2} = \sqrt{e}$ and $\sigma_{X_1}^2 = \sigma_{X_2}^2 = e(e-1)$ show that $-\frac{1}{e} < \rho_{X_1, X_2} < 1$.

4. Assume the joint probability density function of program cost X_1 and schedule X_2 is bivariate normal-lognormal with density function given by equation 7-23. Suppose X_1 has mean 100 (\$M) and variance 625 (\$M)^2. Suppose X_2 has mean 48 (months) and variance 36 (months)^2. Let program cost and schedule have a correlation of 0.5. Compute the difference between the median program cost conditioned on a schedule with a 50 percent chance of being achieved, and the median program cost conditioned on a schedule with a 95 percent chance of being achieved.

5. Show that the functions given by equations 7-23 and 7-44 are each joint probability density functions.

6. If $(X_1, X_2) \sim$ *Bivariate NLogN*$((\mu_1, \mu_2), (\sigma_1^2, \sigma_2^2, \rho_{1,2}))$, where μ_1, μ_2, σ_1^2, σ_2^2, and $\rho_{1,2}$ are defined in section 7.2.2, show that

   a) $E(X_1 | x_2) = \mu_1 + \frac{\sigma_1}{\sigma_2} \rho_{1,2} (\ln x_2 - \mu_2)$

   b) $E(X_2 | x_1) = e^{\mu_2 + \frac{\sigma_2}{\sigma_1} \rho_{1,2} (x_1 - \mu_1) + \frac{1}{2} \sigma_2^2 (1 - \rho_{1,2}^2)}$

   c) $Var(X_1 | x_2) = \sigma_1^2 (1 - \rho_{1,2}^2)$

   d) $Var(X_2 | x_1) = e^{2(\mu_2 + \frac{\sigma_2}{\sigma_1} \rho_{1,2} (x_1 - \mu_1))} e^z (e^z - 1)$ where $z = \sigma_2^2 (1 - \rho_{1,2}^2)$

7. If $(X_1, X_2) \sim$ *Bivariate LogN*$((\mu_1, \mu_2), (\sigma_1^2, \sigma_2^2, \rho_{1,2}))$, where μ_1, μ_2, σ_1^2, σ_2^2, and $\rho_{1,2}$ are defined in section 7.2.3, show that

a) $E(X_1|x_2) = x_2^{\frac{\sigma_1}{\sigma_2}\rho_{1,2}} e^{\mu_1 - \frac{\sigma_1}{\sigma_2}\rho_{1,2}\mu_2 + \frac{1}{2}\sigma_1^2(1-\rho_{1,2}^2)}$

b) $E(X_2|x_1) = x_1^{\frac{\sigma_2}{\sigma_1}\rho_{1,2}} e^{\mu_2 - \frac{\sigma_2}{\sigma_1}\rho_{1,2}\mu_1 + \frac{1}{2}\sigma_2^2(1-\rho_{1,2}^2)}$

c) $Var(X_1|x_2) = x_2^{2\frac{\sigma_1}{\sigma_2}\rho_{1,2}} e^{2(\mu_1 - \frac{\sigma_1}{\sigma_2}\rho_{1,2}\mu_2)} e^{z^o}(e^{z^o} - 1)$

d) $Var(X_2|x_1) = x_1^{2\frac{\sigma_2}{\sigma_1}\rho_{1,2}} e^{2(\mu_2 - \frac{\sigma_2}{\sigma_1}\rho_{1,2}\mu_1)} e^{z}(e^{z} - 1)$

where $z^o = \sigma_1^2(1-\rho_{1,2}^2)$ and $z = \sigma_2^2(1-\rho_{1,2}^2)$.

References

1. Garvey, P. R. 1996 (Spring). Modeling Cost and Schedule Uncertainties – A Work Breakdown Structure Perspective. *Military Operations Research*, V2, N1, pp. 37-43.

2. Garvey, P. R., and A. E. Taub. 1997 (Spring). A Joint Probability Model for Cost and Schedule Uncertainties. *The Journal of Cost Analysis*, pp. 3-27.

3. Abramson, R. L., and P. H. Young. 1997 (Spring). FRISKEM–Formal Risk Evaluation Methodology. *The Journal of Cost Analysis*, pp. 29-38.

This Page Intentionally Left Blank

Epilogue

Considerations and Recommended Practices

One thorn of experience is worth a whole wilderness of warning.
James Russell Lowell (1819-1891)
Shakespeare Once More

And long experience made him sage.
John Gay (1688-1732)
Fables. Part I. The Shepherd and the Philosopher

The following provides a set of considerations and recommended practices when performing cost uncertainty analyses. They reflect the author's insights and experiences in developing, refining, and applying many of the techniques presented in this book.

Treating Cost as a Random Variable — The cost of a future system can be significantly affected by uncertainty. The existence of uncertainty implies the existence of a range of possible costs. How can a decision-maker be shown the chance a particular cost in the range of possible costs will be realized? The probability distribution is a recommended approach for providing this insight. Probability distributions result when independent variables (e.g., weight, power-output, staff-level) used to derive a system's cost randomly assume values across ranges of possible values. For instance, the cost of a satellite might be derived on the basis of a range of possible weight values, with each value randomly occurring. This approach treats cost as a random variable. It is a recognition that values for these variables (such as weight) are not typically known with sufficient precision to perfectly predict cost, *at a time when such predictions are needed.* This point is further articulated by the author's long-time colleague S. A. Book*...

---

* Book, S. A. 1997. Cost Risk Analysis – A Tutorial. *Risk Management Symposium Proceedings*. Los Angeles, California: The Aerospace Corporation.

"The mathematical vehicle for working with a range of possible costs is the probability distribution, with cost itself viewed as a "random variable". Such terminology does not imply, of course, that costs are "random" (though well they may be!) but rather that they are composed of a large number of very small pieces, whose individual contributions to the whole we do not have the ability to investigate in a degree of detail sufficient to calculate the total cost precisely. It is much more efficient for us to recognize that virtually all components of cost are simply "uncertain" and to find some way to assign probabilities to various possible ranges of costs. An analogue is the situation in coin tossing where, in theory, if we knew all the physics involved and solved all the differential equations, we could predict with certainty whether a coin would fall "heads" or "tails". However, the combination of influences acting on the coin are too complicated to understand in sufficient detail to calculate the physical parameters of the coin's motion. So we do the next best thing: we bet that the uncertainties will probably average out in such a way that the coin will fall "heads" half the time and "tails" the other half. It is much more efficient to consider the deterministic physical process of coin tossing to be a "random" statistical process and to assign probabilities of 0.50 to each of the two possible outcomes, heads or tails."

Risk versus Uncertainty — In this book we make a distinction between the terms risk and uncertainty. *Risk* is the chance of loss or injury. In a situation that includes favorable and unfavorable events, risk is the probability an unfavorable event occurs. *Uncertainty* is the indefiniteness about the outcome of a situation. We analyze uncertainty *for the purpose of measuring risk.* In systems engineering, the analysis might focus on measuring the risk of: failing to achieve performance objectives, overrunning the budgeted cost, or delivering the system too late to meet user needs.

Subjective Probability Assessments — Probability theory is a well-established formalism for quantifying uncertainty. Introduced in chapter 2, its application to real-world systems engineering problems often involves the use of subjective probabilities.

Subjective probabilities are those assigned to events on the basis of personal judgment. They are measures of a person's degree-of-belief that an event will occur. Subjective probabilities are associated with one-time, nonrepeatable events — those whose probabilities cannot be objectively determined from a sample space of outcomes developed by repeated trials, or experimentation. Subjective probabilities must be consistent with the axioms of probability (refer to chapter 2). For instance, if an engineer assigns a probability of 0.70 to the event "*the number of gates for the new processor chip* ***will not exceed*** *12000*," then it must follow the chip *will exceed* 12000 gates with probability 0.30. Subjective probabilities are *conditional* on the state of the person's knowledge, which changes with time.

To be credible, subjective probabilities should *only* be assigned to events by subject matter experts — persons with significant experience with events similar to the one under consideration. Instead of assigning a single subjective probability to an event, subject experts often find it easier to describe a function that depicts a distribution of probabilities. Such a distribution is sometimes called a *subjective probability distribution*. Subjective probability distributions are governed by the same mathematical properties of probability distributions associated with discrete or continuous random variables (described in chapter 3). Subjective probability distributions are most common in cost uncertainty analysis, particularly on the input-side of the process (refer to figure 1-3 and the case discussions in chapter 6). Because of their nature, subjective probability distributions can be thought of as "belief functions." They describe a subject expert's belief in the distribution of probabilities for an event under consideration. Probability theory provides the mathematical formalism with which we operate (add, subtract, multiply, and divide) on these belief functions.

Correlation — Correlation is a necessary consideration in cost uncertainty analysis. It can exist between the costs of work breakdown structure (WBS) cost elements.

Correlation can also exist between the cost of a cost element and the variables (e.g., weight, schedule) that define its cost.

Statistical theory offers a number of ways to measure correlation. Two popular measures are Pearson's product-moment correlation and Spearman's rank correlation. Subtleties concerning these measures must be understood to avoid errors in a cost uncertainty analysis. Pearson's product-moment correlation measures *linearity* between two random variables. Spearman's rank correlation measures their *monotonicity*. Thus, these two measures of correlation *can be very different*. This is illustrated in figure 5-10 (chapter 5). Furthermore, the variance of a sum of random variables is a function of Pearson's product-moment correlation, *not* Spearman's rank correlation. Thus, from a WBS perspective, Pearson's product-moment correlation is the *only correct* measure of correlation to use when computing the variance of a sum of cost element costs.

In cost uncertainty analysis, care must be taken if it is necessary to subjectively specify Pearson correlations. Pearson correlations can be restricted to a *subinterval* of -1 to +1 for random variables characterized by certain types of distribution functions. This is illustrated in chapter 7. Thus, the Pearson correlation between any two random variables cannot be assigned a value in a completely arbitrary way. If it *is* necessary to subjectively specify Pearson correlations, the reader should review the recently published work of Lurie-Golberg.*

In practice, it is recommended that analysts express associations within the WBS through functional relationships (cost equations), as illustrated in case discussions 6-1 and 6-2. This allows the Pearson correlations *implied by these relationships* to be captured in the overall analysis. Pearson correlations that originate from logically defined functional relationships are more easily defended in cost reviews than those made on the basis of subjective assessments.

---

* Lurie, P. M., and M. S. Goldberg. 1998. An Approximate Method for Sampling Correlated Random Variables from Partially-Specified Distributions. *Management Science*, Vol. 44, No. 2, pp. 203-218.

Capturing Cost-Schedule Uncertainties — Decision-makers require understanding how uncertainties between a system's cost and schedule interact. A decision-maker might bet on a "high-risk" schedule in hopes of keeping the system's cost within requirements. On the other hand, the decision-maker may be willing to assume "more cost" for a schedule with a small chance of being exceeded. This is a common tradeoff faced by decision-makers on systems engineering projects. The family of distributions in chapter 7 provides an analytical basis for computing this tradeoff, using joint and conditional cost-schedule probabilities. This family is a set of mathematical models that might be *hypothesized* for capturing the joint interactions between cost and schedule.

A parameter required by these models is the correlation between cost and schedule.* Direct computation is one approach for determining this parameter, as illustrated in case discussion 6-2. However, in some instances this might not be analytically possible or practical. Another approach is to obtain an estimate of the correlation from sample values generated by Monte Carlo simulation. This is a reasonable method that can be done regardless of the complexity of the cost-schedule estimation relationships. Subjective assessments might be used. However, care must again be taken to specify an *admissible correlation.* Furthermore, there may already exist an implied correlation by virtue of how the cost-schedule estimation relationships are mathematically defined (refer to case discussion 6-2). Subjectively specifying a correlation when one is already present (only its magnitude is unknown) is *double counting correlation.* Such a situation invalidates the mathematical integrity of the cost uncertainty analysis.

Approximating the Distribution Function of a System's Total Cost — Cost analysts are encouraged to *study the mathematical relationships* they define in a system's work breakdown structure, to see whether analytical approximations to the distribution function of $Cost_{Sys}$ (a system's total cost) can be argued. Analytical

---

* Because these models treat cost and schedule as correlated random variables, it is important to recognize that *they do not capture causal impacts* that schedule compression or extension has on cost.

approximations can reveal much information about the "cost-behavior" in a system's WBS. Chapter 6 (section 6.2.2) presented five cases when the normal distribution approximates the distribution function of a system's total cost. There are many reasons for this approximation. Primary among them is that $Cost_{Sys}$ is a summation of WBS cost element costs. Seen in the chapter 6 case discussions, it is typical to have a mixture of independent and correlated cost element costs within a system's WBS. Because of the central limit theorem (theorem 5-10, chapter 5), the greater the number of independent cost element costs the more it is that the distribution function of $Cost_{Sys}$ is approximately normal. The central limit theorem is very powerful. It does not take many independent cost element costs for the distribution function of $Cost_{Sys}$ to move towards normality. Such a move is evidenced when (1) a sufficient number of independent cost element costs are summed and (2) when no cost element's cost distribution has a much larger standard deviation than the standard deviations of the other cost element cost distributions. When conditions in the WBS result in $Cost_{Sys}$ being positively skewed (i.e., a non-normal distribution function), then the lognormal often approximates the distribution function of $Cost_{Sys}$.

Monte Carlo simulation is another approach for developing an empirical approximation to the distribution function of $Cost_{Sys}$. The Monte Carlo method, discussed in section 6.3, is often needed when a system's WBS contains cost estimating relationships too complex for strict analytical study. In Monte Carlo simulations, a question frequently asked is "*How many trials are necessary to have confidence in the output of the simulation*?" As a guideline, 10,000 trials (Monte Carlo samples) should be sufficient to meet the precision requirements for most Monte Carlo simulations; particularly those for cost uncertainty analyses.

Benefits of Cost Uncertainty Analysis — Cost uncertainty analysis provides decision-makers many benefits and important insights. These include:

Establishing a Cost and Schedule Risk Baseline — Baseline probability distributions of a system's cost and schedule can be developed for a given system configuration, acquisition strategy, and cost-schedule estimation approach. This baseline provides decision-makers visibility into potentially high-payoff areas for risk reduction initiatives. Baseline distributions assist in determining a system's cost and schedule that simultaneously have a specified probability of not being exceeded (chapter 7). They can also provide decision-makers an assessment of the likelihood of achieving a budgeted (or proposed) cost and schedule, or cost for a given feasible schedule.

Determining Cost Reserve — Cost uncertainty analysis provides a basis for determining cost reserve as a function of the uncertainties specific to a system. The analysis provides the direct link between the amount of cost reserve to recommend and the probability that a system's cost will not exceed a prescribed (or desired) magnitude (refer to figure 1-6, chapter 1). An analysis should be conducted to verify the recommended cost reserve covers fortuitous events (e.g., unplanned code growth, unplanned schedule delays) deemed possible by the system's engineering team. Finally, it is sometimes necessary to allocate cost reserve dollars into the cost elements of a system's work breakdown structure. The reader is directed to the Book-Young algorithm* as an approach for making this allocation.

Conducting Risk Reduction Tradeoff Analyses — Cost uncertainty analyses can be conducted to study the payoff of implementing risk reduction initiatives (e.g., rapid prototyping) on lessening a system's cost and schedule risks. Furthermore, families of probability distribution functions can be generated to compare the cost and cost risk impacts of alternative system requirements, schedule uncertainties, and competing system configurations or acquisition strategies.

---

* Book, S. A. 1997. Cost Risk Analysis – A Tutorial. *Risk Management Symposium Proceedings*. Los Angeles, California: The Aerospace Corporation.

Documenting the Cost Uncertainty Analysis — The validity and meaningfulness of a cost uncertainty analysis relies on the engineering team's experience, judgment, and knowledge of the system's uncertainties. Formulating and documenting a supporting rationale that summarizes the team's collective insights into these uncertainties is *the critical part of the process*. Without a well documented rationale the credibility of the analysis can be easily questioned. The details of the analysis methodology are important and should also be documented. The methodology *must be technically sound* and offer value-added problem structure, analyses, and insights otherwise not visible. Decisions that successfully eliminate uncertainty, or reduce it to acceptable levels, are ultimately driven by human judgment. This at best is aided by, not directed by, the methods presented in this book.

Some Additional Reading

1. Cooper, D. F., and C. B. Chapman. 1987. *Risk Analysis for Large Projects — Models, Methods, & Cases*. Chichester, United Kingdom: John Wiley & Sons Ltd.

2. Vose, D. 1996. *Quantitative Risk Analysis: A Guide to Monte Carlo Simulation Modelling*. Chichester, United Kingdom: John Wiley & Sons Ltd.

Appendix A

Statistical Tables and Related Integrals

A.1 Table A-1 presents values of the cumulative distribution function of the standard normal distribution. These values are denoted by $F_Z(z)$, which is given by

$$F_Z(z) = P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy \qquad \text{(A-1)}$$

Example: a) What is $P(Z \le 0.33)$? b) What is $P(Z \le -0.33)$?

a) From equation A-1 $F_Z(0.33) = P(Z \le 0.33)$; from table A-1 $F_Z(0.33) = 0.6293$

b) Since $Z \sim N(0,1)$ we have $P(Z \le -z) = P(Z > z) = 1 - P(Z \le z)$; therefore, in this example, $F_Z(-0.33) = P(Z \le -0.33) = P(Z > 0.33) = 1 - P(Z \le 0.33) = 1 - 0.6293 = 0.3707$

Table A-1. Percentiles of the Standard Normal Distribution
(the 3-digit columns are z, the 8-digit columns are $F_Z(z)$)

0.00	0.5000000	0.21	0.5831661	0.42	0.6627572	0.63	0.7356528
0.01	0.5039894	0.22	0.5870644	0.43	0.6664021	0.64	0.7389138
0.02	0.5079784	0.23	0.5909541	0.44	0.6700314	0.65	0.7421540
0.03	0.5119665	0.24	0.5948348	0.45	0.6736448	0.66	0.7453732
0.04	0.5159535	0.25	0.5987063	0.46	0.6772419	0.67	0.7485712
0.05	0.5199389	0.26	0.6025681	0.47	0.6808225	0.68	0.7517478
0.06	0.5239223	0.27	0.6064198	0.48	0.6843863	0.69	0.7549030
0.07	0.5279032	0.28	0.6102612	0.49	0.6879331	0.70	0.7580364
0.08	0.5318814	0.29	0.6140918	0.50	0.6914625	0.71	0.7611480
0.09	0.5358565	0.30	0.6179114	0.51	0.6949743	0.72	0.7642376
0.10	0.5398279	0.31	0.6217195	0.52	0.6984682	0.73	0.7673050
0.11	0.5437954	0.32	0.6255158	0.53	0.7019441	0.74	0.7703501
0.12	0.5477585	0.33	0.6293000	0.54	0.7054015	0.75	0.7733727
0.13	0.5517168	0.34	0.6330717	0.55	0.7088403	0.76	0.7763728
0.14	0.5556700	0.35	0.6368306	0.56	0.7122603	0.77	0.7793501
0.15	0.5596177	0.36	0.6405764	0.57	0.7156612	0.78	0.7823046
0.16	0.5635595	0.37	0.6443087	0.58	0.7190427	0.79	0.7852362
0.17	0.5674949	0.38	0.6480272	0.59	0.7224047	0.80	0.7881447
0.18	0.5714237	0.39	0.6517317	0.60	0.7257469	0.81	0.7910300
0.19	0.5753454	0.40	0.6554217	0.61	0.7290692	0.82	0.7938920
0.20	0.5792597	0.41	0.6590970	0.62	0.7323712	0.83	0.7967307

Table A-1. Percentiles of the Standard Normal Distribution (Concluded)
(the 3-digit columns are z, the 8-digit columns are $F_Z(z)$)

0.84	0.7995459	1.05	0.8531409	1.26	0.8961653	1.47	0.9292191
0.85	0.8023375	1.06	0.8554277	1.27	0.8979576	1.48	0.9305633
0.86	0.8051055	1.07	0.8576903	1.28	0.8997274	1.49	0.9318879
0.87	0.8078498	1.08	0.8599289	1.29	0.9014746	1.50	0.9331928
0.88	0.8105704	1.09	0.8621434	1.30	0.9031995	1.51	0.9344783
0.89	0.8132671	1.10	0.8643339	1.31	0.9049020	1.52	0.9357445
0.90	0.8159399	1.11	0.8665004	1.32	0.9065824	1.53	0.9369916
0.91	0.8185888	1.12	0.8686431	1.33	0.9082408	1.54	0.9382198
0.92	0.8212136	1.13	0.8707618	1.34	0.9098773	1.55	0.9394292
0.93	0.8238145	1.14	0.8728568	1.35	0.9114919	1.56	0.9406200
0.94	0.8263912	1.15	0.8749280	1.36	0.9130850	1.57	0.9417924
0.95	0.8289439	1.16	0.8769755	1.37	0.9146565	1.58	0.9429466
0.96	0.8314724	1.17	0.8789995	1.38	0.9162066	1.59	0.9440826
0.97	0.8339768	1.18	0.8809998	1.39	0.9177355	1.60	0.9452007
0.98	0.8364569	1.19	0.8829767	1.40	0.9192433	1.61	0.9463011
0.99	0.8389129	1.20	0.8849303	1.41	0.9207301	1.62	0.9473839
1.00	0.8413447	1.21	0.8868605	1.42	0.9221961	1.63	0.9484493
1.01	0.8437523	1.22	0.8887675	1.43	0.9236414	1.64	0.9494974
1.02	0.8461358	1.23	0.8906514	1.44	0.9250663	1.65	0.9505285
1.03	0.8484950	1.24	0.8925122	1.45	0.9264707	1.66	0.9515428
1.04	0.8508300	1.25	0.8943502	1.46	0.9278549	1.67	0.9525403
1.68	0.9535214	1.89	0.9706211	2.10	0.9821356	2.31	0.9895559
1.69	0.9544861	1.90	0.9712835	2.11	0.9825709	2.32	0.9898296
1.70	0.9554346	1.91	0.9719335	2.12	0.9829970	2.33	0.9900969
1.71	0.9563671	1.92	0.9725711	2.13	0.9834143	2.40	0.9918025
1.72	0.9572838	1.93	0.9731967	2.14	0.9838227	2.50	0.9937903
1.73	0.9581849	1.94	0.9738102	2.15	0.9842224	2.60	0.9953388
1.74	0.9590705	1.95	0.9744120	2.16	0.9846137	2.70	0.9965330
1.75	0.9599409	1.96	0.9750022	2.17	0.9849966	2.80	0.9974448
1.76	0.9607961	1.97	0.9755809	2.18	0.9853713	2.90	0.9981341
1.77	0.9616365	1.98	0.9761483	2.19	0.9857379	3.00	0.9986500
1.78	0.9624621	1.99	0.9767046	2.20	0.9860966	3.10	0.9990323
1.79	0.9632731	2.00	0.9772499	2.21	0.9864475	3.20	0.9993128
1.80	0.9640697	2.01	0.9777845	2.22	0.9867907	3.30	0.9995165
1.81	0.9648522	2.02	0.9783084	2.23	0.9871263	3.40	0.9996630
1.82	0.9656206	2.03	0.9788218	2.24	0.9874546	3.50	0.9997673
1.83	0.9663751	2.04	0.9793249	2.25	0.9877756	3.60	0.9998409
1.84	0.9671159	2.05	0.9798179	2.26	0.9880894	3.70	0.9998922
1.85	0.9678433	2.06	0.9803008	2.27	0.9883962	3.80	0.9999276
1.86	0.9685573	2.07	0.9807739	2.28	0.9886962	3.90	0.9999519
1.87	0.9692582	2.08	0.9812373	2.29	0.9889894	4.00	0.9999683
1.88	0.9699460	2.09	0.9816912	2.30	0.9892759	5.00	0.9999997

A.2 Table A-2 is used for the Kolmogorov-Smirnov goodness of fit test. The values in table A-2 apply *only when all the parameters of the hypothesized distribution are known*, that is, none of the distribution's parameters are estimated (or derived) from the sample data. The reader is directed to Law and Kelton* and Stephens** for an expanded discussion of table A-2. In table A-2, D is the Kolmogorov-Smirnov test statistic defined as

$$D = \max_x \left| F_X(x) - \hat{F}_X(x) \right|$$

This statistic measures the largest vertical distance between the hypothesized cumulative distribution function $F_X(x)$ and the empirical (observed) cumulative distribution function $\hat{F}_X(x)$, developed from the sample data.

Table A-2. Modified Critical Values for the Kolmogorov-Smirnov Test Statistic

(Applicable when the parameters of the hypothesized distribution $F_X(x)$ are known and not estimated from the sample data)

Let n denote the number of samples. If

$$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) D > c_{1-\alpha}$$

reject the claim that the observed values come from the hypothesized distribution; otherwise accept it.

α	$1-\alpha$	$c_{1-\alpha}$
0.010	0.990	1.628
0.025	0.975	1.480
0.050	0.950	1.358
0.100	0.900	1.224
0.150	0.850	1.138

* Law, A. M., and W. D. Kelton. 1991. *Simulation Modeling and Analysis*, 2nd ed. New York: McGraw-Hill, Inc.

** Stephens, M. A. 1974. EDF Statistics for Goodness of Fit and Some Comparisons. *J. Am. Statist. Assoc.*, 69, pp. 730-737.

A.3 Integrals Related to the Normal Probability Density Function

The following integrals are often useful in proofs and computations involving the normal probability density function. In each integral, a is a real number and b is a positive real number. The first integral is the integral of the normal probability density function. The second integral is the mean of a normally distributed random variable. The third integral is the second moment of a normally distributed random variable, with mean a and variance b^2.

$$1. \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,b} e^{-\frac{(x-a)^2}{2b^2}} dx = 1 \qquad \text{(A-2)}$$

$$2. \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\,b} e^{-\frac{(x-a)^2}{2b^2}} dx = a \qquad \text{(A-3)}$$

$$3. \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}\,b} e^{-\frac{(x-a)^2}{2b^2}} dx = a^2 + b^2 \qquad \text{(A-4)}$$

A.4 Sums of Independent Uniform Random Variables

Suppose the random variable U is defined as the sum of n uniformly distributed *independent* random variables, that is

$$U = U_1 + U_2 + U_3 + \ldots + U_n$$

where $U_i \sim Unif(0,1)$ for $i = 1,2,3,\ldots,n$. Let $f_U(u)$ denote the probability density function of U. From theorem 5-12 (chapter 5) a general expression for $f_U(u)$ can be developed. A convenient form of this expression is given below [Cramer, 1966].

$$f_U(u) = \frac{1}{(n-1)!}\left[u^{n-1} - \binom{n}{1}(u-1)^{n-1} + \binom{n}{2}(u-2)^{n-1} - \cdots\right]$$

In the expression above, $0 < u < n$ and the summation is continued as long as the arguments u, $(u-1)$, $(u-2)$, $\cdots$ are positive.* From the central limit theorem, as n increases the distribution function of U will approach a normal distribution with mean $\frac{n}{2}$ and variance $\frac{n}{12}$. This is illustrated in figure A-1.

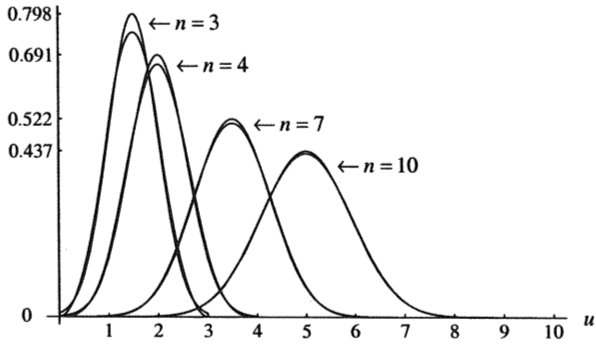


Figure A-1. Probability Density Functions for Sums of Uniform Independent Random Variables

* Cramer, H. 1966. *Mathematical Methods of Statistics*. pp. 245. Princeton, New Jersey: Princeton University Press.

Figure A-1 shows pairs of PDFs plotted for $n = 3, 4, 7$, and 10. The left-most pair show plots of $f_{Normal}(u)$ and $f_U(u)$, respectively, for $n = 3$; specifically,

$$f_{Normal}(u) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\frac{3}{12}}} e^{-\frac{1}{2}\left[\frac{(u-\frac{3}{2})^2}{\frac{3}{12}}\right]}$$

and

$$f_U(u) = \begin{cases} \frac{1}{2}u^2 & 0 < u < 1 \\ \frac{1}{2}(u^2 - 3(u-1)^2) & 1 < u < 2 \\ \frac{1}{2}(u^2 - 3(u-1)^2 + 3(u-2)^2) & 2 < u < 3 \end{cases}$$

The second pair of PDFs (from the left) show plots of $f_{Normal}(u)$ and $f_U(u)$, respectively, for $n = 4$; specifically,

$$f_{Normal}(u) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\frac{4}{12}}} e^{-\frac{1}{2}\left[\frac{(u-\frac{4}{2})^2}{\frac{4}{12}}\right]}$$

and

$$f_U(u) = \begin{cases} \frac{1}{6}u^3 & 0 < u < 1 \\ \frac{1}{6}(u^3 - 4(u-1)^3) & 1 < u < 2 \\ \frac{1}{6}(u^3 - 4(u-1)^3 + 6(u-2)^3) & 2 < u < 3 \\ \frac{1}{6}(u^3 - 4(u-1)^3 + 6(u-2)^3 - 4(u-3)^3) & 3 < u < 4 \end{cases}$$

A similar convention holds for the two remaining pairs of PDFs plotted in figure A-1. The values shown along the vertical axis, in figure A-1, correspond to values for $f_{Normal}(u)$.

Table A-3 compares the cumulative probabilities derived from each PDF pair in figure A-1. In table A-3, the columns labeled $F_U(u)$ and $F_{Normal}(u)$ are defined as follows:

$$F_U(u) = \int_0^u f_U(t)\,dt$$

and

$$F_{Normal}(u) = \int_{-\infty}^{u} f_{Normal}(t)\,dt$$

where

$$f_{Normal}(t) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\frac{n}{12}}} e^{-\frac{1}{2}\left[\frac{(t-\frac{n}{2})^2}{\frac{n}{12}}\right]}$$

for $n = 3$, 4, 7, and 10.

Table A-3. Sums of Independent Uniform Random Variables — Cumulative Probability

$$U = U_1 + U_2 + U_3 + \ldots + U_n$$

$$U_i \sim Unif(0,1) \quad i = 1,2,3,\ldots,n$$

$n=3$	$F_U(u)$	$F_{Normal}(u)$	$n=4$	$F_U(u)$	$F_{Normal}(u)$
$0<u<1$	0.16666667	0.158655	$0<u<1$	0.041666667	0.0416323
$0<u<2$	0.83333334	0.841345	$0<u<2$	0.499999997	0.5
$0<u<3$	1	0.99865	$0<u<3$	0.958333327	0.958368
			$0<u<4$	1	0.999734

$n=7$	$F_U(u)$	$F_{Normal}(u)$	$n=10$	$F_U(u)$	$F_{Normal}(u)$
$0<u<1$	0.0001984127	0.000531557	$0<u<1$	0.00000027557	0.00000588567
$0<u<2$	0.0240079367	0.0247673	$0<u<2$	0.00027943121	0.0005075
$0<u<3$	0.2603174567	0.256345	$0<u<3$	0.01346285321	0.0142299
$0<u<4$	0.7396825367	0.743655	$0<u<4$	0.13890156321	0.136661
$0<u<5$	0.9759920567	0.975233	$0<u<5$	0.49999999321	0.5
$0<u<6$	0.9998015807	0.999468	$0<u<6$	0.86109842321	0.863339
$0<u<7$	1	0.999998	$0<u<7$	0.98653713321	0.98577
			$0<u<8$	0.99972055521	0.999492
			$0<u<9$	0.99999971085	0.999994
			$0<u<10$	1	0.999999978398

This Page Intentionally Left Blank

Appendix B

The Bivariate Normal-LogNormal Distribution

Let $Y_1 = X_1$ and $Y_2 = \ln X_2$ where X_1 and X_2 are random variables defined on $-\infty < x_1 < \infty$ and $0 < x_2 < \infty$. If Y_1 and Y_2 each have a normal distribution then

$$E(Y_1) = \mu_{Y_1} = \mu_{X_1} = \mu_1 \qquad Var(Y_1) = \sigma^2_{Y_1} = \sigma^2_{X_1} = \sigma^2_1$$

$$E(Y_2) = \mu_{Y_2} = \mu_2 = \frac{1}{2}\ln\left[\frac{(\mu_{X_2})^4}{(\mu_{X_2})^2 + \sigma^2_{X_2}}\right] \quad Var(Y_2) = \sigma^2_{Y_2} = \sigma^2_2 = \ln\left[\frac{(\mu_{X_2})^2 + \sigma^2_{X_2}}{(\mu_{X_2})^2}\right]$$

The pair of random variables

$$(X_1, X_2) \sim Bivariate\ NLogN((\mu_1, \mu_2), (\sigma^2_1, \sigma^2_2, \rho_{1,2}))$$

has a bivariate normal-lognormal distribution if

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{(2\pi)\sigma_1\sigma_2\sqrt{1-\rho^2_{1,2}}\,x_2} e^{-\frac{1}{2}w}$$

where

$$-1 < \rho_{1,2} = \rho_{Y_1,Y_2} = \rho_{X_1,\ln X_2} < 1$$

and

$$w = \frac{1}{1-\rho^2_{1,2}}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho_{1,2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{\ln x_2-\mu_2}{\sigma_2}\right) + \left(\frac{\ln x_2-\mu_2}{\sigma_2}\right)^2\right\}$$

Theorem B-1 If $(X_1, X_2) \sim Bivariate\ NLogN((\mu_1, \mu_2), (\sigma^2_1, \sigma^2_2, \rho_{1,2}))$ then

$$\rho_{1,2} = \rho_{X_1,X_2}\frac{(e^{\sigma^2_2} - 1)^{1/2}}{\sigma_2}$$

Proof:

By definition

$$\rho_{X_1,X_2} = \frac{Cov(X_1,X_2)}{\sigma_{X_1}\sigma_{X_2}} = \frac{\sigma_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}}$$

where

$$\sigma_{X_1X_2} = \int_0^\infty \int_{-\infty}^\infty (x_1-\mu_1)(x_2-\mu_2) f_{X_1,X_2}(x_1,x_2)dx_1dx_2$$

and $\sigma_{X_1} = \sigma_1$. Since X_2 is lognormal

$$\sigma_{X_2} = (e^{2\mu_2+\sigma_2^2}(e^{\sigma_2^2}-1))^{1/2} = E(X_2)(e^{\sigma_2^2}-1)^{1/2}$$

Thus, $$\rho_{X_1,X_2} = \frac{\sigma_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}} = \frac{\sigma_{X_1X_2}}{\sigma_1 E(X_2)(e^{\sigma_2^2}-1)^{1/2}}$$

To compute $\sigma_{X_1X_2}$, let $t_1 = \dfrac{x_1-\mu_1}{\sigma_1}$ and $t_2 = \dfrac{\ln x_2 - \mu_2}{\sigma_2}$; therefore,

$$\sigma_{X_1X_2} = \frac{1}{2\pi\sqrt{1-\rho_{1,2}^2}} \int_{-\infty}^\infty \int_{-\infty}^\infty (\sigma_1 t_1)(e^{\mu_2+\sigma_2 t_2}-\mu_2) e^{-\frac{1}{2(1-\rho_{1,2}^2)}(t_1^2-2\rho_{1,2}t_1t_2+t_2^2)} dt_1dt_2$$

$$= \frac{1}{2\pi\sqrt{1-\rho_{1,2}^2}} \int_{-\infty}^\infty (\sigma_1 t_1)[I_1-\mu_2 I_2]dt_1$$

where

$$I_1 = \int_{-\infty}^\infty e^{\mu_2+\sigma_2 t_2} e^{-\frac{1}{2(1-\rho_{1,2}^2)}(t_1^2-2\rho_{1,2}t_1t_2+t_2^2)} dt_2$$

and

$$I_2 = \int_{-\infty}^\infty e^{-\frac{1}{2(1-\rho_{1,2}^2)}(t_1^2-2\rho_{1,2}t_1t_2+t_2^2)} dt_2$$

To determine I_1, note the integrand can be written as

$$I_1 = e^{\mu_2} \int_{-\infty}^\infty e^{-\frac{1}{2(1-\rho_{1,2}^2)}(t_1^2-2[\rho_{1,2}t_1+(1-\rho_{1,2}^2)\sigma_2]t_2+t_2^2)} dt_2$$

Letting

$$A = A(t_1) = \rho_{1,2}t_1 + (1-\rho_{1,2}^2)\sigma_2$$

and noting that

$$t_2^2 - 2At_2 = (t_2 - A)^2 - A^2$$

we can write

$$I_1 = e^{\mu_2} e^{-\frac{1}{2(1-\rho_{1,2}^2)}t_1^2} e^{\frac{1}{2(1-\rho_{1,2}^2)}A^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho_{1,2}^2)}(t_2-A)^2} dt_2$$

$$I_1 = e^{\mu_2} e^{-\frac{1}{2(1-\rho_{1,2}^2)}t_1^2} e^{\frac{1}{2(1-\rho_{1,2}^2)}A^2} \sqrt{2\pi}\sqrt{(1-\rho_{1,2}^2)}$$

To determine I_2, note the integrand can be written as

$$I_2 = e^{-\frac{1}{2(1-\rho_{1,2}^2)}t_1^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho_{1,2}^2)}(t_2^2 - 2\rho_{1,2}t_1t_2)} dt_2$$

Letting $B = B(t_1) = \rho_{1,2}t_1$ and noting that $t_2^2 - 2Bt_2 = (t_2 - B)^2 - B^2$ we have

$$I_2 = e^{-\frac{1}{2(1-\rho_{1,2}^2)}t_1^2} e^{\frac{1}{2(1-\rho_{1,2}^2)}B^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho_{1,2}^2)}(t_2-B)^2} dt_2$$

$$I_2 = e^{-\frac{1}{2(1-\rho_{1,2}^2)}t_1^2} e^{\frac{1}{2(1-\rho_{1,2}^2)}B^2} \sqrt{2\pi}\sqrt{(1-\rho_{1,2}^2)}$$

Thus,

$$I_1 - \mu_2 I_2 = e^{\frac{-t_1^2}{2(1-\rho_{1,2}^2)}} \sqrt{2\pi}\sqrt{(1-\rho_{1,2}^2)} \left[e^{\mu_2} e^{\frac{A^2}{2(1-\rho_{1,2}^2)}} - \mu_2 e^{\frac{B^2}{2(1-\rho_{1,2}^2)}} \right]$$

and

$$\sigma_{X_1X_2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma_1 t_1) e^{\frac{-t_1^2}{2(1-\rho_{1,2}^2)}} \left[e^{\mu_2} e^{\frac{A^2}{2(1-\rho_{1,2}^2)}} - \mu_2 e^{\frac{B^2}{2(1-\rho_{1,2}^2)}} \right] dt_1$$

$$\sigma_{X_1X_2} = \frac{1}{\sqrt{2\pi}}\left[e^{\mu_2}\sigma_1\int_{-\infty}^{\infty} t_1 e^{\frac{-(t_1^2-A^2)}{2(1-\rho_{1,2}^2)}}dt_1 - \mu_2\sigma_1\int_{-\infty}^{\infty} t_1 e^{\frac{-(t_1^2-B^2)}{2(1-\rho_{1,2}^2)}}dt_1\right]$$

$$\sigma_{X_1X_2} = \frac{1}{\sqrt{2\pi}}\left[e^{\mu_2}\sigma_1\int_{-\infty}^{\infty} t_1 e^{-\frac{1}{2}(t_1-\rho_{1,2}\sigma_2)^2+\frac{1}{2}\sigma_2^2}dt_1 - \mu_2\sigma_1\int_{-\infty}^{\infty} t_1 e^{-t_1^2/2}dt_1\right]$$

$$\sigma_{X_1X_2} = \frac{1}{\sqrt{2\pi}}\left[e^{\mu_2}\sigma_1 e^{\frac{1}{2}\sigma_2^2}\int_{-\infty}^{\infty} t_1 e^{-\frac{1}{2}(t_1-\rho_{1,2}\sigma_2)^2}dt_1 - \mu_2\sigma_1\cdot 0\right]$$

$$\sigma_{X_1X_2} = \frac{1}{\sqrt{2\pi}}\left[e^{\mu_2+\sigma_2^2/2}\sigma_1\rho_{1,2}\sigma_2\sqrt{2\pi}\right] = E(X_2)\rho_{1,2}\sigma_1\sigma_2$$

Hence,

$$\rho_{X_1,X_2} = \frac{\sigma_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}} = \frac{E(X_2)\rho_{1,2}\sigma_1\sigma_2}{\sigma_1\left[e^{2\mu_2+\sigma_2^2}(e^{\sigma_2^2}-1)\right]^{1/2}} = \frac{E(X_2)\rho_{1,2}\sigma_1\sigma_2}{\sigma_1\left[E(X_2)(e^{\sigma_2^2}-1)^{1/2}\right]}$$

Thus,

$$\rho_{1,2} = \rho_{X_1,X_2}\frac{(e^{\sigma_2^2}-1)^{1/2}}{\sigma_2} \tag{B-1}$$

Theorem B-2 If $(X_1, X_2) \sim$ *Bivariate NLogN*$((\mu_1,\mu_2),(\sigma_1^2,\sigma_2^2,\rho_{1,2}))$, then

$$f_1(x_1) = \frac{1}{\sqrt{2\pi}\,\sigma_1}e^{-\frac{1}{2}\left[(x_1-\mu_1)^2/\sigma_1^2\right]}$$

and

$$f_2(x_2) = \frac{1}{\sqrt{2\pi}\,\sigma_2 x_2}e^{-\frac{1}{2}\left[(\ln x_2-\mu_2)^2/\sigma_2^2\right]}$$

Proof:

By definition

$$f_1(x_1) = \int_0^\infty f_{X_1,X_2}(x_1,x_2)\,dx_2$$

$$f_2(x_2) = \int_{-\infty}^\infty f_{X_1,X_2}(x_1,x_2)\,dx_1$$

The density function $f_{X_1,X_2}(x_1,x_2)$ can be factored as

$$f_{X_1,X_2}(x_1,x_2) = \left\{\frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x_1-\mu_1)^2/2\sigma_1^2}\right\} Q(x_1,x_2) \qquad \text{(B-2)}$$

where

$$Q(x_1,x_2) = \left\{\frac{1}{\sqrt{2\pi}(\sigma_2\sqrt{1-\rho_{1,2}^2})\,x_2} e^{-(\ln x_2 - b)^2/2\sigma_2^2(1-\rho_{1,2}^2)}\right\}$$

and

$$b = \mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1-\mu_1).$$

Therefore,

$$\begin{aligned} f_1(x_1) &= \int_0^\infty \left\{\frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x_1-\mu_1)^2/2\sigma_1^2}\right\} Q(x_1,x_2)\,dx_2 \\ &= \left\{\frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x_1-\mu_1)^2/2\sigma_1^2}\right\} \int_0^\infty Q(x_1,x_2)\,dx_2 \\ &= \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\left[(x_1-\mu_1)^2/\sigma_1^2\right]} \end{aligned}$$

since the integrand is the density function of a $LogN(b,\sigma_2^2(1-\rho_{1,2}^2))$ random variable. To compute $f_2(x_2)$, the density function $f_{X_1,X_2}(x_1,x_2)$ is factored as

$$f_{X_1,X_2}(x_1,x_2) = Q^*(x_1,x_2)\left\{\frac{1}{\sqrt{2\pi}\sigma_2}\frac{1}{x_2} e^{-(\ln x_2-\mu_2)^2/2\sigma_2^2}\right\} \qquad \text{(B-3)}$$

where

$$Q^*(x_1,x_2)=\left\{\frac{1}{\sqrt{2\pi}(\sigma_1\sqrt{1-\rho_{1,2}^2})}e^{-(x_1-b^*)^2/2\sigma_1^2(1-\rho_{1,2}^2)}\right\}$$

and

$$b^*=\mu_1+\frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2-\mu_2)$$

Therefore,

$$\begin{aligned}f_2(x_2)&=\int_{-\infty}^{\infty}\left\{\frac{1}{\sqrt{2\pi}\sigma_2}\frac{1}{x_2}e^{-(\ln x_2-\mu_2)^2/2\sigma_2^2}\right\}Q^*(x_1,x_2)dx_1\\&=\left\{\frac{1}{\sqrt{2\pi}\sigma_2}\frac{1}{x_2}e^{-(\ln x_2-\mu_2)^2/2\sigma_2^2}\right\}\int_{-\infty}^{\infty}Q^*(x_1,x_2)dx_1\\&=\frac{1}{\sqrt{2\pi}\,\sigma_2 x_2}e^{-\frac{1}{2}\left[(\ln x_2-\mu_2)^2/\sigma_2^2\right]}\end{aligned}$$

since the integrand is the density function of a $N(b^*,\sigma_1^2(1-\rho_{1,2}^2))$ random variable.

Theorem B-3 If $(X_1,X_2)\sim$ *Bivariate NLogN*$((\mu_1,\mu_2),(\sigma_1^2,\sigma_2^2,\rho_{1,2}))$, then

$$X_1|x_2\sim N(\mu_1+\frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2-\mu_2),\sigma_1^2(1-\rho_{1,2}^2))$$

$$X_2|x_1\sim LogN(\mu_2+\frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1-\mu_1),\sigma_2^2(1-\rho_{1,2}^2))$$

Proof:

By definition,

$$f_{X_1|x_2}(x_1)=\frac{f_{X_1,X_2}(x_1,x_2)}{f_2(x_2)}=\frac{\left\{\frac{1}{\sqrt{2\pi}\,\sigma_2 x_2}e^{-\frac{1}{2}\left[(\ln x_2-\mu_2)^2/\sigma_2^2\right]}\right\}Q^*(x_1,x_2)}{\frac{1}{\sqrt{2\pi}\,\sigma_2 x_2}e^{-\frac{1}{2}\left[(\ln x_2-\mu_2)^2/\sigma_2^2\right]}}$$

$$f_{X_1|x_2}(x_1)=Q^*(x_1,x_2)$$

Thus, from equation B-3

$$X_1|x_2 \sim N(b^*, \sigma_1^2(1-\rho_{1,2}^2))$$

where

$$b^* = \mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2)$$

Similarly,

$$f_{X_2|x_1}(x_2) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_1(x_1)} = \frac{\left\{\frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\left[(x_1-\mu_1)^2/\sigma_1^2\right]}\right\} Q(x_1,x_2)}{\frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\left[(x_1-\mu_1)^2/\sigma_1^2\right]}}$$

$$f_{X_2|x_1}(x_2) = Q(x_1,x_2)$$

Thus, from equation B-2

$$X_2|x_1 \sim LogN(b, \sigma_2^2(1-\rho_{1,2}^2))$$

where

$$b = \mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1-\mu_1)$$

Theorem B-4 If $(X_1, X_2) \sim$ *Bivariate NLogN*$((\mu_1,\mu_2),(\sigma_1^2,\sigma_2^2,\rho_{1,2}))$, then

$$E(X_2|x_1) = e^{\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1-\mu_1) + \frac{1}{2}\sigma_2^2(1-\rho_{1,2}^2)}$$

$$Var(X_2|x_1) = e^{2\left(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1-\mu_1)\right)} e^z(e^z-1)$$

$$E(X_1|x_2) = \mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2)$$

$$Var(X_1|x_2) = \sigma_1^2(1-\rho_{1,2}^2)$$

where $z = \sigma_2^2(1-\rho_{1,2}^2)$.

Proof:

Theorem B-3 proved that

$$X_2|x_1 \sim LogN(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1 - \mu_1), \sigma_2^2(1-\rho_{1,2}^2))$$

Therefore,

$$E(X_2|x_1) = e^{\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1-\mu_1) + \frac{1}{2}\sigma_2^2(1-\rho_{1,2}^2)}$$

$$Var(X_2|x_1) = e^{2(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1-\mu_1))} e^z(e^z - 1)$$

where $z = \sigma_2^2(1-\rho_{1,2}^2)$.

Theorem B-3 also proved that

$$X_1|x_2 \sim N(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2), \sigma_1^2(1-\rho_{1,2}^2))$$

Therefore, it follows immediately from the properties of the normal distribution that

$$E(X_1|x_2) = \mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2)$$

$$Var(X_1|x_2) = \sigma_1^2(1-\rho_{1,2}^2)$$

Theorem B-5 If $(X_1, X_2) \sim$ *Bivariate NLogN*$((\mu_1, \mu_2), (\sigma_1^2, \sigma_2^2, \rho_{1,2}))$, then

$$Median(X_2|x_1) = e^{\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1-\mu_1)}$$

$$Mode(X_2|x_1) = e^{\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1-\mu_1) - \sigma_2^2(1-\rho_{1,2}^2)}$$

$$Median(X_1|x_2) = E(X_1|x_2)$$

$$Mode(X_1|x_2) = E(X_1|x_2)$$

Proof:

Since $X_2|x_1$ is lognormally distributed,

$$Median(X_2|x_1) = e^{\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1-\mu_1)}$$

and

$$Mode(X_2|x_1) = e^{\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(x_1-\mu_1) - \sigma_2^2(1-\rho_{1,2}^2)}$$

Since $X_1|x_2$ is normally distributed, it follows immediately that

$$Median(X_1|x_2) = E(X_1|x_2)$$

$$Mode(X_1|x_2) = E(X_1|x_2)$$

Property B-1 If $(X_1, X_2) \sim$ *Bivariate NLogN*$((\mu_1,\mu_2),(\sigma_1^2,\sigma_2^2,\rho_{1,2}))$, then

$$E(X_1|Median(X_2|\mu_1)) = \mu_1$$

Proof:

From theorem B-4, it was established that

$$E(X_1|x_2) = \mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2)$$

From theorem B-5,

$$Median(X_2|x_1 = \mu_1) = e^{\mu_2}$$

It follows that

$$\begin{aligned} E(X_1|Median(X_2|\mu_1)) = E(X_1|e^{\mu_2}) &= \mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln e^{\mu_2} - \mu_2) \\ &= \mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\mu_2 - \mu_2) \\ &= \mu_1 \end{aligned}$$

This Page Intentionally Left Blank

Appendix C

The Bivariate LogNormal Distribution

Let $Y_1 = \ln X_1$ and $Y_2 = \ln X_2$ where X_1 and X_2 are random variables defined on $0 < x_1 < \infty$ and $0 < x_2 < \infty$. If Y_1 and Y_2 each have a normal distribution then

$$E(Y_1) = \mu_{Y_1} = \mu_1 = \frac{1}{2}\ln\left[\frac{(\mu_{X_1})^4}{(\mu_{X_1})^2 + \sigma_{X_1}^2}\right] \quad Var(Y_1) = \sigma_{Y_1}^2 = \sigma_1^2 = \ln\left[\frac{(\mu_{X_1})^2 + \sigma_{X_1}^2}{(\mu_{X_1})^2}\right]$$

$$E(Y_2) = \mu_{Y_2} = \mu_2 = \frac{1}{2}\ln\left[\frac{(\mu_{X_2})^4}{(\mu_{X_2})^2 + \sigma_{X_2}^2}\right] \quad Var(Y_2) = \sigma_{Y_2}^2 = \sigma_2^2 = \ln\left[\frac{(\mu_{X_2})^2 + \sigma_{X_2}^2}{(\mu_{X_2})^2}\right]$$

The pair of random variables

$$(X_1, X_2) \sim \textit{Bivariate LogN}((\mu_1, \mu_2), (\sigma_1^2, \sigma_2^2, \rho_{1,2}))$$

has a bivariate lognormal distribution if

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{(2\pi)\sigma_1\sigma_2\sqrt{1-\rho_{1,2}^2}\, x_1 x_2} e^{-\frac{1}{2}w}$$

where

$$-1 < \rho_{1,2} = \rho_{Y_1,Y_2} = \rho_{\ln X_1, \ln X_2} < 1$$

and

$$w = \frac{1}{1-\rho_{1,2}^2}\left\{\left(\frac{\ln x_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho_{1,2}\left(\frac{\ln x_1 - \mu_1}{\sigma_1}\right)\left(\frac{\ln x_2 - \mu_2}{\sigma_2}\right) + \left(\frac{\ln x_2 - \mu_2}{\sigma_2}\right)^2\right\}$$

Theorem C-1 If $(X_1, X_2) \sim$ *Bivariate LogN*$((\mu_1, \mu_2),(\sigma_1^2, \sigma_2^2, \rho_{1,2}))$ then

$$\rho_{X_1,X_2} = \frac{e^{\rho_{1,2}\sigma_1\sigma_2} - 1}{\sqrt{e^{\sigma_1^2} - 1}\sqrt{e^{\sigma_2^2} - 1}}$$

Proof:

By definition,

$$\rho_{X_1,X_2} = \frac{Cov(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}} = \frac{E(X_1X_2) - E(X_1)E(X_2)}{\sigma_{X_1}\sigma_{X_2}} \qquad \text{(C-1)}$$

Since $Y_1 = \ln X_1$ and $Y_2 = \ln X_2$,

$$E(X_1X_2) = E(e^{Y_1}e^{Y_2}) = E(e^{Y_1+Y_2})$$

Since $Y_i \sim N(\mu_i, \sigma_i^2)$ (for $i = 1,2$), the expectation $E(e^{Y_1+Y_2})$ is a special evaluation of the moment generating function* of a bivariate normal, which is

$$M(t_1,t_2) = E(e^{t_1Y_1+t_2Y_2}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1y_1+t_2y_2} f(y_1,y_2)dy_1dy_2$$

$$= e^{(\mu_1t_1+\mu_2t_2)+\frac{1}{2}(\sigma_1^2t_1^2+2\rho_{Y_1,Y_2}\sigma_1\sigma_2t_1t_2+\sigma_2^2t_2^2)}$$

for some real t_1 and t_2. Therefore,

$$E(X_1X_2) = E(e^{Y_1}e^{Y_2}) = E(e^{Y_1+Y_2}) = e^{(\mu_1+\mu_2)+\frac{1}{2}(\sigma_1^2+\sigma_2^2+2\rho_{Y_1,Y_2}\sigma_1\sigma_2)}$$

To determine the remaining terms in equation (C-1), for $r \geq 0$ the moments of X_1 and X_2 are

$$E(X_i^r) = e^{r\mu_i+\frac{1}{2}r^2\sigma_i^2} \qquad \text{(C-2)}$$

Thus,

$$E(X_1) = e^{\mu_1+\frac{1}{2}\sigma_1^2}$$

$$E(X_2) = e^{\mu_2+\frac{1}{2}\sigma_2^2}$$

---

* Refer to Ross, S. 1994. *A First Course in Probability*, 4th ed. New York: Macmillan College Publishing Company.

and

$$\sigma_{X_1}^2 = Var(X_1) = E(X_1^2) - (E(X_1))^2 = e^{2\mu_1+2\sigma_1^2} - (e^{\mu_1+\frac{1}{2}\sigma_1^2})^2$$
$$= e^{2\mu_1+2\sigma_1^2} - e^{2\mu_1+\sigma_1^2}$$

$$\sigma_{X_2}^2 = Var(X_2) = E(X_2^2) - (E(X_2))^2 = e^{2\mu_2+2\sigma_2^2} - (e^{\mu_2+\frac{1}{2}\sigma_2^2})^2$$
$$= e^{2\mu_2+2\sigma_2^2} - e^{2\mu_2+\sigma_2^2}$$

Substituting into equation C-1,

$$\rho_{X_1,X_2} = \frac{E(X_1X_2) - E(X_1)E(X_2)}{\sigma_{X_1}\sigma_{X_2}}$$

$$\rho_{X_1,X_2} = \frac{e^{(\mu_1+\mu_2)+\frac{1}{2}(\sigma_1^2+2\rho_{Y_1,Y_2}\sigma_1\sigma_2+\sigma_2^2)} - (e^{\mu_1+\frac{1}{2}\sigma_1^2})(e^{\mu_2+\frac{1}{2}\sigma_2^2})}{\sqrt{e^{2\mu_1+2\sigma_1^2} - e^{2\mu_1+\sigma_1^2}}\sqrt{e^{2\mu_2+2\sigma_2^2} - e^{2\mu_2+\sigma_2^2}}}$$

This can be factored as:

$$\rho_{X_1,X_2} = \frac{e^{(\mu_1+\mu_2)+\frac{1}{2}(\sigma_1^2+\sigma_2^2)}(e^{\rho_{1,2}\sigma_1\sigma_2} - 1)}{e^{(\mu_1+\mu_2)+\frac{1}{2}(\sigma_1^2+\sigma_2^2)}\sqrt{e^{\sigma_1^2} - 1}\sqrt{e^{\sigma_2^2} - 1}}$$

Thus,

$$\rho_{X_1,X_2} = \frac{e^{\rho_{1,2}\sigma_1\sigma_2} - 1}{\sqrt{e^{\sigma_1^2} - 1}\sqrt{e^{\sigma_2^2} - 1}} \qquad \text{(C-3)}$$

Theorem C-2 If $(X_1, X_2) \sim$ *Bivariate LogN*$((\mu_1,\mu_2),(\sigma_1^2,\sigma_2^2,\rho_{1,2}))$, then

$$f_1(x_1) = \frac{1}{\sqrt{2\pi}\,\sigma_1 x_1} e^{-\frac{1}{2}\left[(\ln x_1 - \mu_1)^2/\sigma_1^2\right]}$$

and

$$f_2(x_2) = \frac{1}{\sqrt{2\pi}\,\sigma_2 x_2} e^{-\frac{1}{2}\left[(\ln x_2 - \mu_2)^2/\sigma_2^2\right]}$$

Proof:

By definition,

$$f_1(x_1) = \int_0^\infty f_{X_1,X_2}(x_1,x_2)\,dx_2$$

$$f_2(x_2) = \int_0^\infty f_{X_1,X_2}(x_1,x_2)\,dx_1$$

The density function $f_{X_1,X_2}(x_1,x_2)$ can be factored as

$$f_{X_1,X_2}(x_1,x_2) = \left\{\frac{1}{\sqrt{2\pi}\sigma_1}\frac{1}{x_1}e^{-(\ln x_1-\mu_1)^2/2\sigma_1^2}\right\}Q(x_1,x_2) \qquad \text{(C-4)}$$

where

$$Q(x_1,x_2) = \left\{\frac{1}{\sqrt{2\pi}(\sigma_2\sqrt{1-\rho_{1,2}^2})\,x_2}e^{-(\ln x_2-b)^2/2\sigma_2^2(1-\rho_{1,2}^2)}\right\}$$

and

$$b = \mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(\ln x_1 - \mu_1)$$

Therefore,

$$\begin{aligned} f_1(x_1) &= \int_0^\infty \left\{\frac{1}{\sqrt{2\pi}\sigma_1}\frac{1}{x_1}e^{-(\ln x_1-\mu_1)^2/2\sigma_1^2}\right\}Q(x_1,x_2)\,dx_2 \\ &= \left\{\frac{1}{\sqrt{2\pi}\sigma_1}\frac{1}{x_1}e^{-(\ln x_1-\mu_1)^2/2\sigma_1^2}\right\}\int_0^\infty Q(x_1,x_2)\,dx_2 \\ &= \frac{1}{\sqrt{2\pi}\,\sigma_1 x_1}e^{-\frac{1}{2}\left[(\ln x_1-\mu_1)^2/\sigma_1^2\right]} \end{aligned}$$

since the integrand is the probability density function of a $LogN(b,\sigma_2^2(1-\rho_{1,2}^2))$ random variable.

To compute $f_2(x_2)$, the density function $f_{X_1,X_2}(x_1,x_2)$ is factored as

$$f_{X_1,X_2}(x_1,x_2) = Q^*(x_1,x_2)\left\{\frac{1}{\sqrt{2\pi}\sigma_2}\frac{1}{x_2}e^{-(\ln x_2-\mu_2)^2/2\sigma_2^2}\right\} \qquad \text{(C-5)}$$

where

$$Q^*(x_1,x_2)=\left\{\frac{1}{\sqrt{2\pi}(\sigma_1\sqrt{1-\rho_{1,2}^2})\,x_1}e^{-(\ln x_1-b^*)^2/2\sigma_1^2(1-\rho_{1,2}^2)}\right\}$$

and

$$b^*=\mu_1+\frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2-\mu_2)$$

Therefore,

$$f_2(x_2)=\int_0^\infty\left\{\frac{1}{\sqrt{2\pi}\sigma_2}\frac{1}{x_2}e^{-(\ln x_2-\mu_2)^2/2\sigma_2^2}\right\}Q^*(x_1,x_2)\,dx_1$$

$$=\left\{\frac{1}{\sqrt{2\pi}\sigma_2}\frac{1}{x_2}e^{-(\ln x_2-\mu_2)^2/2\sigma_2^2}\right\}\int_0^\infty Q^*(x_1,x_2)\,dx_1$$

$$=\frac{1}{\sqrt{2\pi}\sigma_2}\frac{1}{x_2}e^{-(\ln x_2-\mu_2)^2/2\sigma_2^2}$$

since the integrand is the probability density function of a $LogN(b^*,\sigma_1^2(1-\rho_{1,2}^2))$ random variable.

Theorem C-3 If $(X_1,X_2)\sim$ *Bivariate* $LogN((\mu_1,\mu_2),(\sigma_1^2,\sigma_2^2,\rho_{1,2}))$, then

$$X_1|x_2\sim LogN(\mu_1+\frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2-\mu_2),\sigma_1^2(1-\rho_{1,2}^2))$$

$$X_2|x_1\sim LogN(\mu_2+\frac{\sigma_2}{\sigma_1}\rho_{1,2}(\ln x_1-\mu_1),\sigma_2^2(1-\rho_{1,2}^2))$$

Proof:

By definition,

$$f_{X_1|x_2}(x_1)=\frac{f_{X_1,X_2}(x_1,x_2)}{f_2(x_2)}=\frac{\left\{\frac{1}{\sqrt{2\pi}\,\sigma_2\,x_2}e^{-\frac{1}{2}\left[(\ln x_2-\mu_2)^2/\sigma_2^2\right]}\right\}Q^*(x_1,x_2)}{\frac{1}{\sqrt{2\pi}\,\sigma_2\,x_2}e^{-\frac{1}{2}\left[(\ln x_2-\mu_2)^2/\sigma_2^2\right]}}$$

$$f_{X_1|x_2}(x_1) = Q^*(x_1, x_2)$$

Thus, from equation C-5,

$$X_1|x_2 \sim LogN(b^*, \sigma_1^2(1-\rho_{1,2}^2))$$

where

$$b^* = \mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2)$$

Similarly,

$$f_{X_2|x_1}(x_2) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_1(x_1)} = \frac{\left\{\frac{1}{\sqrt{2\pi}\,\sigma_1 x_1} e^{-\frac{1}{2}\left[(\ln x_1 - \mu_1)^2/\sigma_1^2\right]}\right\} Q(x_1,x_2)}{\frac{1}{\sqrt{2\pi}\,\sigma_1 x_1} e^{-\frac{1}{2}\left[(\ln x_1 - \mu_1)^2/\sigma_1^2\right]}}$$

$$f_{X_2|x_1}(x_2) = Q(x_1, x_2)$$

Thus, from equation C-4,

$$X_2|x_1 \sim LogN(b, \sigma_2^2(1-\rho_{1,2}^2))$$

where

$$b = \mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(\ln x_1 - \mu_1)$$

Theorem C-4 If $(X_1, X_2) \sim$ *Bivariate* $LogN((\mu_1,\mu_2),(\sigma_1^2,\sigma_2^2,\rho_{1,2}))$, then

$$E(X_2|x_1) = x_1^{\frac{\sigma_2}{\sigma_1}\rho_{1,2}} e^{\mu_2 - \frac{\sigma_2}{\sigma_1}\rho_{1,2}\mu_1 + \frac{1}{2}\sigma_2^2(1-\rho_{1,2}^2)}$$

$$Var(X_2|x_1) = x_1^{2\frac{\sigma_2}{\sigma_1}\rho_{1,2}} e^{2(\mu_2 - \frac{\sigma_2}{\sigma_1}\rho_{1,2}\mu_1)} e^z(e^z - 1)$$

$$E(X_1|x_2) = x_2^{\frac{\sigma_1}{\sigma_2}\rho_{1,2}} e^{\mu_1 - \frac{\sigma_1}{\sigma_2}\rho_{1,2}\mu_2 + \frac{1}{2}\sigma_1^2(1-\rho_{1,2}^2)}$$

$$Var(X_1|x_2) = x_2^{2\frac{\sigma_1}{\sigma_2}\rho_{1,2}} e^{2(\mu_1 - \frac{\sigma_1}{\sigma_2}\rho_{1,2}\mu_2)} e^{z^*}(e^{z^*} - 1)$$

where

$$z = \sigma_2^2(1-\rho_{1,2}^2) \quad \text{and} \quad z^* = \sigma_1^2(1-\rho_{1,2}^2)$$

Proof:

Theorem C-3 proved that

$$X_2|x_1 \sim LogN(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(\ln x_1 - \mu_1), \sigma_2^2(1-\rho_{1,2}^2))$$

Therefore,

$$E(X_2|x_1) = e^{\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(\ln x_1 - \mu_1) + \frac{1}{2}\sigma_2^2(1-\rho_{1,2}^2)}$$
$$= x_1^{\frac{\sigma_2}{\sigma_1}\rho_{1,2}} e^{\mu_2 - \frac{\sigma_2}{\sigma_1}\rho_{1,2}\mu_1 + \frac{1}{2}\sigma_2^2(1-\rho_{1,2}^2)}$$

and

$$Var(X_2|x_1) = e^{2(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(\ln x_1 - \mu_1))} e^{\sigma_2^2(1-\rho_{1,2}^2)}(e^{\sigma_2^2(1-\rho_{1,2}^2)} - 1)$$
$$= x_1^{2\frac{\sigma_2}{\sigma_1}\rho_{1,2}} e^{2(\mu_2 - \frac{\sigma_2}{\sigma_1}\rho_{1,2}\mu_1)} e^z(e^z - 1)$$

Theorem C-3 also proved that

$$X_1|x_2 \sim LogN(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2), \sigma_1^2(1-\rho_{1,2}^2))$$

Therefore,

$$E(X_1|x_2) = e^{\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2) + \frac{1}{2}\sigma_1^2(1-\rho_{1,2}^2)}$$
$$= x_2^{\frac{\sigma_1}{\sigma_2}\rho_{1,2}} e^{\mu_1 - \frac{\sigma_1}{\sigma_2}\rho_{1,2}\mu_2 + \frac{1}{2}\sigma_1^2(1-\rho_{1,2}^2)}$$

and

$$Var(X_1|x_2) = e^{2(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2))} e^{\sigma_1^2(1-\rho_{1,2}^2)}(e^{\sigma_1^2(1-\rho_{1,2}^2)} - 1)$$

$$= x_2^{2\frac{\sigma_1}{\sigma_2}\rho_{1,2}} e^{2(\mu_1 - \frac{\sigma_1}{\sigma_2}\rho_{1,2}\mu_2)} e^{z^*}(e^{z^*} - 1)$$

Theorem C-5 If $(X_1, X_2) \sim$ *Bivariate LogN*$((\mu_1, \mu_2), (\sigma_1^2, \sigma_2^2, \rho_{1,2}))$, then

$$Median(X_2|x_1) = x_1^{\frac{\sigma_2}{\sigma_1}\rho_{1,2}} e^{\mu_2 - \frac{\sigma_2}{\sigma_1}\rho_{1,2}\mu_1}$$

$$Mode(X_2|x_1) = x_1^{\frac{\sigma_2}{\sigma_1}\rho_{1,2}} e^{\mu_2 - \frac{\sigma_2}{\sigma_1}\rho_{1,2}\mu_1 - \sigma_2^2(1-\rho_{1,2}^2)}$$

$$Median(X_1|x_2) = x_2^{\frac{\sigma_1}{\sigma_2}\rho_{1,2}} e^{\mu_1 - \frac{\sigma_1}{\sigma_2}\rho_{1,2}\mu_2}$$

$$Mode(X_1|x_2) = x_2^{\frac{\sigma_1}{\sigma_2}\rho_{1,2}} e^{\mu_1 - \frac{\sigma_1}{\sigma_2}\rho_{1,2}\mu_2 - \sigma_1^2(1-\rho_{1,2}^2)}$$

Proof:

From theorem C-3, it follows that

$$Median(X_2|x_1) = e^{\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(\ln x_1 - \mu_1)} = x_1^{\frac{\sigma_2}{\sigma_1}\rho_{1,2}} e^{\mu_2 - \frac{\sigma_2}{\sigma_1}\rho_{1,2}\mu_1}$$

$$Mode(X_2|x_1) = e^{\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{1,2}(\ln x_1 - \mu_1) - \sigma_2^2(1-\rho_{1,2}^2)} = x_1^{\frac{\sigma_2}{\sigma_1}\rho_{1,2}} e^{\mu_2 - \frac{\sigma_2}{\sigma_1}\rho_{1,2}\mu_1 - \sigma_2^2(1-\rho_{1,2}^2)}$$

$$Median(X_1|x_2) = e^{\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2)} = x_2^{\frac{\sigma_1}{\sigma_2}\rho_{1,2}} e^{\mu_1 - \frac{\sigma_1}{\sigma_2}\rho_{1,2}\mu_2}$$

$$Mode(X_1|x_2) = e^{\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{1,2}(\ln x_2 - \mu_2) - \sigma_1^2(1-\rho_{1,2}^2)} = x_2^{\frac{\sigma_1}{\sigma_2}\rho_{1,2}} e^{\mu_1 - \frac{\sigma_1}{\sigma_2}\rho_{1,2}\mu_2 - \sigma_1^2(1-\rho_{1,2}^2)}$$

Property C-1 If $(X_1, X_2) \sim$ *Bivariate LogN*$((\mu_1, \mu_2), (\sigma_1^2, \sigma_2^2, \rho_{1,2}))$, then the conditional coefficients of dispersion are

$$D_{F_{X_1|x_2}} = \frac{\left[Var(X_1|x_2)\right]^{1/2}}{E(X_1|x_2)} = \sqrt{(e^{z^*} - 1)}$$

$$D_{F_{X_2|x_1}} = \frac{\left[Var(X_2|x_1)\right]^{1/2}}{E(X_2|x_1)} = \sqrt{(e^z - 1)}$$

where $F_{X_1|x_2}$ and $F_{X_2|x_1}$ are the cumulative distributions of $f_{X_1|x_2}$ and $f_{X_2|x_1}$ This property is stated without proof. It is a direct algebraic consequence of theorem C-4.

This Page Intentionally Left Blank

This Page Intentionally Left Blank