DOUGLAS C. MONTGOMERY
GEORGE C. RUNGER
NORMA F. HUBELE

ENGINEERING STATISTICS

Engineering Statistics

Douglas C. Montgomery
George C. Runger
Norma Faris Hubele
Arizona State University

John Wiley & Sons, Inc.
New York Chichester Weinheim Brisbane Singapore Toronto

ACQUISITIONS EDITOR Wayne Anderson
MARKETING MANAGER Harper Mooy
SENIOR PRODUCTION MANAGER Lucille Buonocore
SENIOR PRODUCTION EDITOR Monique Calello
SENIOR DESIGNER Kevin Murphy
ILLUSTRATION COORDINATOR Jaime Perea
ILLUSTRATION STUDIO Wellington Studios
MANUFACTURING MANAGER Monique Calello

This book was set in 10/12 Times Roman by Monotype Composition Company and printed and bound by Quebecor Printing, Fairfield. The cover was printed by Phoenix Color Corporation.

This book is printed on acid-free paper. ∞

The paper in this book was manufactured by a mill whose forest management programs include sustained yield harvesting of its timberlands. Sustained yield harvesting principles ensure that the numbers of trees cut each year does not exceed the amount of new growth.

Library of Congress Cataloging-in-Publication Data

Engineering statistics / Douglas C. Montgomery, George C. Runger, Norma Faris Hubele.
p. cm.
"The book is based on a longer text Douglas C. Montgomery, George C. Runger. Applied statistics and probability for engineers. New York: Wiley, 1994"—Preface.
Includes bibliographical references and index.
ISBN 0-471-17026-7 (cloth : alk. paper)
1. Engineering—Statistical methods. 2. Probabilities.
I. Montgomery, Douglas C. II. Runger, George C. III. Hubele, Norma Faris.

TA340.E53 1997
519.5′024′62—dc21

97-39795
CIP
AC

Printed in the United States of America

10 9 8 7 6 5 4

To

Meredith, Neil, Colin, and Cheryl

Rebecca, Elisa, George, and Taylor

Norman and Michelle

Preface

Engineers play a significant role in the modern world. They are responsible for the design and development of most of the products that our society uses, as well as the manufacturing processes that make these products. Engineers are also involved in many aspects of the management of both industrial enterprises and business or service organizations. Fundamental training in engineering develops skills in problem formulation, analysis, and solution that are valuable in a wide range of settings.

Solving many types of engineering problems requires an appreciation of variability and some understanding of how to use both descriptive and analytical tools in dealing with variability. Statistics is the branch of applied mathematics that is concerned with variability and its impact on decision making. This is an introductory textbook for a first course in engineering statistics. While many of the topics we present are fundamental to the use of statistics in other disciplines, we have elected to focus on an engineering audience, because this approach will best meet the needs of engineering students by allowing them to concentrate on the applications of statistics to their disciplines. Consequently, our examples and exercises are engineering based, and in almost all cases, we have used a real problem setting or the data either from a published source or from our own consulting experience.

Engineers in all disciplines should take at least one course in statistics. Indeed, the Accreditation Board on Engineering and Technology is requiring that engineers learn about statistics and how to use statistical methodology effectively as part of their formal undergraduate training. Because of other program requirements, most engineering students will take only one statistics course. This book has been designed to serve as a text for the one-term statistics course for all engineering students.

ORGANIZATION OF THE BOOK

The book is based on a more comprehensive text [Montgomery, D. C. and Runger, G. C., *Applied Statistics and Probability for Engineers,* John Wiley & Sons, New York, 1994] that has been used by instructors in a one- or two-semester course. We have taken the key topics for a one-semester course from that book as the basis of this text. As a result of this condensation and revision, this book has a modest mathematical level. Engineering students who have completed one semester of calculus should have no difficulty reading nearly all of the text. Our intent is to give the student an understanding of statistical methodology and how it may be applied in the solution of engineering problems, not the mathematical theory of statistics.

Chapter 1 introduces the role of statistics and probability in engineering problem solving. Statistical thinking and the associated methods are illustrated and contrasted with other engineering modeling approaches within the context of the engineering problem-solving method. Highlights of the value of statistical methodologies are discussed using simple examples. Simple summary statistics are introduced.

Chapter 2 illustrates the useful information provided by simple summary and graphical displays. Computer procedures for analyzing large data sets are given. Data analysis methods such as histograms, stem-and-leaf plots, and frequency distributions are illustrated. Using these displays to obtain insight into the behavior of the data or underlying system is emphasized.

Chapter 3 introduces the concepts of a random variable and the probability distribution that describes the behavior of that random variable. We concentrate on the normal distribution, because of its fundamental role in the statistical tools that are frequently applied in engineering. We have tried to avoid using sophisticated mathematics and the event–sample space orientation traditionally used to present this material to engineering students. An in-depth understanding of probability is not necessary to understand how to use statistics for effective engineering problem solving. Other topics in this chapter include expected values, variances, probability plotting, correlation, and the central limit theorem.

Chapters 4 and 5 present the basic tools of statistical inference: point estimation, confidence intervals, and hypothesis testing. Techniques for a single sample are in Chapter 4, and two-sample inference techniques are in Chapter 5. Our presentation is distinctly applications oriented and stresses the simple comparative-experiment nature of these procedures. We want engineering students to become interested in how these methods can be used to solve real-world problems and to learn some aspects of the concepts behind them so that they can see how to apply them in other settings. We give a logical, heuristic development of the techniques, not a mathematically rigorous one.

Empirical model building is introduced in Chapter 6. Both simple and multiple linear regression models are presented, and the use of these models as approximations to mechanistic models is discussed. We show the student how to find the least squares estimates of the regression coefficients, perform the standard statistical tests and confidence intervals, and use the model residuals for assessing model adequacy. Although this chapter makes some modest use of matrix algebra, we emphasize the use of the computer for regression model fitting and analysis.

Chapter 7 formally introduces the design of engineering experiments, although much of Chapters 4 and 5 was the foundation for this topic. We emphasize the factorial design and, in particular, the case in which all the experimental factors are at two levels. Our practical experience indicates that if engineers know how to set up a factorial experiment with all factors at two levels, conduct the experiment properly, and correctly analyze the resulting data, they can successfully attack a large majority of the engineering experiments that they will encounter in the real world. Consequently, we have written this chapter to accomplish these objectives. We also introduce fractional factorial designs and response surface methods.

Statistical quality control is introduced in Chapter 8. The important topic of Shewhart control charts is emphasized. The $\overline{X}$ and R charts are presented, along with some simple control charting techniques for individuals and attribute data. We also discuss some aspects of estimating the capability of a process.

The students should be encouraged to work problems to master the subject matter. The book contains an ample number of problems of different levels of difficulty. The end-of-section exercises are intended to reinforce the concepts and techniques introduced in that section. These exercises are more structured than the end-of-chapter supplemental exercises, which generally require more formulation or conceptual thinking. We use the supplemental exercises as integrating problems to reinforce mastery of concepts, as opposed to analytical technique. The Team Exercises challenge the student to apply chapter methods and concepts to problems requiring data collection. As noted below, the use of statistics software in problem solution should be an integral part of the course.

USING THE BOOK

We strongly believe that an introductory course in statistics for undergraduate engineering students should be, first and foremost, an *applied course.* The primary emphasis should be on data description, inference (confidence intervals and tests), model building, designing engineering experiments, and statistical quality control, *because these are the techniques that they will need to know how to use as practicing engineers.* There is a tendency in teaching these courses to spend a great deal of time on probability and random variables (and, indeed, some engineers, such as industrial and electrical engineers, do need to know more about these subjects than students in other disciplines) and to emphasize the mathematically oriented aspects of the subject. This can turn an engineering statistics course into a "baby math-stat" course. This type of course can be fun to teach and much easier on the instructor, because it is almost always easier to teach theory than application, but it does not prepare the student for professional practice.

In our course taught at Arizona State University, students meet twice weekly, once in a large classroom and once in a small computer laboratory. Students are responsible for reading assignments, individual homework problems, and team projects. In-class team activities include designing experiments, generating data, and performing analyses. The supplemental problems and team exercises in this text are a good source for these activities. The intent is to provide an active learning environment with challenging problems that foster the development of skills for analysis and synthesis.

USING THE COMPUTER

In practice, engineers use computers to apply statistical methods in solving problems. Therefore, we strongly recommend that the computer be integrated into the course. Throughout the book, we have presented output from *Minitab* as typical examples of what can be done with modern computer software. In teaching, we have used *Statgraphics,* *Minitab*, *Excel*, and several other statistics packages or spreadsheets. We did not clutter the book with examples from many different packages, because *how* the instructor integrates the software into the class is ultimately more important than *which* package is used. All text data and the instructor manual are available in electronic form.

In our large-class meeting times, we have access to computer software. We show the student how the technique is implemented in the software as soon as it is discussed in class. We recommend this as a teaching format. Low-cost student versions of many popular software packages are available, and many institutions have statistics software available on a local area network, so access for the students is typically not a problem.

Computer software can be used to do many exercises in this text. Some exercises, however, have small computer icons in the margin. We highly recommend using software in these instances.

WEB SITE

Current supporting material for instructors and students is available at the Web site www.wiley.com/college/montES. We will use this site to communicate our latest information about innovations and recommendations for effectively using this text and we hope to elicit your feedback. In-depth case studies that illustrate an integration of several analysis methods will be posted as they are developed. Electronic versions of select data from the text are posted there for your convenience.

ACKNOWLEDGMENTS

We would like to express our appreciation for support in developing some of this text material from the Course and Curriculum Development Program of the Undergraduate Education Division of the National Science Foundation. We also thank Teri Reed Rhoads, Lora Zimmer, and Sharon Lewis for their work as Graduate assistants in the development of the course based on this text. We are thankful to Connie Borror for her thorough work on the instructor's volume. We appreciate the staff support and resources provided by the Department of Industrial and Management Systems Engineering at Arizona State University.

Several reviewers provided many helpful suggestions, including Dr. Hongshik Ahn, SUNY, Stony Brook; Dr. James Simpson, USAF Academy; Dr. John D. O'Neil, California Polytechnic University, Pomona; Dr. Charles Donaghey, University of Houston. We are also indebted to Dr. Smiley Cheng for permission to adapt many of the statistical tables from his excellent book (with Dr. James Fu), *Statistical Tables for Classroom and Exam Room*. John Wiley and Sons, Prentice Hall, the Biometrika Trustees, The American Statistical Association, the Institute of Mathematical Statistics, and the editors of *Biometrics* allowed us to use copyrighted material, for which we are grateful.

This project was supported, in part, by the
National Science Foundation
Opinions expressed are those of the authors and not necessarily those of the Foundation

Douglas C. Montgomery
George C. Runger
Norma Faris Hubele

Contents

1 The Role of Statistics in Engineering 1

1-1 The Engineering Method and Statistical Thinking 1
1-2 Collecting Engineering Data 10
1-3 Mechanistic and Empirical Models 12
1-4 Designing Experimental Investigations 16
1-5 Observing Processes over Time 19

2 Data Summary and Presentation 24

2-1 The Importance of Data Summary and Display 24
2-2 The Stem-and-Leaf Diagram 25
2-3 The Frequency Distribution and Histogram 30
2-4 The Box Plot 35
2-5 Time Sequence Plots 37

3 Random Variables and Probability Distributions 46

3-1 Introduction 47

3-2 Random Variables 49

3-3 Probability 51

3-4 Probability Density Function, Mean, and Variance of a Continuous Random Variable 56
- 3-4.1 Probability Density Function 56
- 3-4.2 Mean and Variance of a Continuous Random Variable 59

3-5 Normal Distribution 63

3-6 Probability Plots 75

3-7 Probability Mass Function, Mean, and Variance of a Discrete Random Variable 79
- 3-7.1 Probability Mass Function 80
- 3-7.2 Mean and Variance of a Discrete Random Variable 81

3-8 Binomial Distribution 84

3-9 Poisson Process 92
- 3-9.1 Poisson Distribution 92
- 3-9.2 Exponential Distribution 99

3-10 Normal Approximation to the Binomial and Poisson Distributions 105

3-11 Correlation and Independence 109
- 3-11.1 Correlation 109
- 3-11.2 Independence 113

3-12 Random Samples, Statistics, and the Central Limit Theorem 115

4 Decision Making for a Single Sample 131

4-1 Statistical Inference 132

4-2 Point Estimation 133

4-3 Hypothesis Testing 140
- 4-3.1 Statistical Hypotheses 140
- 4-3.2 Testing Statistical Hypotheses 142
- 4-3.3 One-Sided and Two-Sided Hypotheses 150
- 4-3.4 General Procedure for Hypothesis Testing 152

4-4 Inference on the Mean of a Population, Variance Known 154
- 4-4.1 Hypothesis Testing on the Mean 154
- 4-4.2 *P*-Values in Hypothesis Testing 157
- 4-4.3 Type II Error and Choice of Sample Size 158
- 4-4.4 Large-Sample Test 161
- 4-4.5 Some Practical Comments on Hypothesis Testing 161
- 4-4.6 Confidence Interval on the Mean 163
- 4-4.7 General Method for Deriving a Confidence Interval 169

4-5 Inference on the Mean of a Population, Variance Unknown 171
- 4-5.1 Hypothesis Testing on the Mean 172
- 4-5.2 *P*-Value for a *t*-Test 176
- 4-5.3 Computer Solution 177
- 4-5.4 Choice of Sample Size 178
- 4-5.5 Confidence Interval on the Mean 180

4-6 Inference on the Variance of a Normal Population 183
- 4-6.1 Hypothesis Testing on the Variance of a Normal Population 183
- 4-6.2 Confidence Interval on the Variance of a Normal Population 187

4-7 Inference on a Population Proportion 189
- 4-7.1 Hypothesis Testing on a Binomial Proportion 190
- 4-7.2 Type II Error and Choice of Sample Size 191

4-7.3 Confidence Interval on a Binomial Proportion 193

4-8 Summary Table of Inference Procedures for a Single Sample 198

4-9 Testing for Goodness of Fit 198

5 Decision Making for Two Samples 210

5-1 Introduction 211

5-2 Inference for a Difference in Means, Variances Known 211
- 5-2.1 Hypothesis Tests for a Difference in Means, Variances Known 212
- 5-2.2 Choice of Sample Size 214
- 5-2.3 Identifying Cause and Effect 215
- 5-2.4 Confidence Interval on a Difference in Means, Variances Known 216

5-3 Inference for the Difference in Means of Two Normal Distributions, Variances Unknown 221
- 5-3.1 Hypotheses Tests for the Difference in Means 221
- 5-3.2 Choice of Sample Size 227
- 5-3.3 Confidence Interval on the Difference in Means 228
- 5-3.4 Computer Solution 230

5-4 The Paired *t*-Test 235

5-5 Inferences on the Variances of Two Normal Populations 243
- 5-5.1 Tests of Hypotheses on the Ratio of Two Variances 243
- 5-5.2 Confidence Interval on the Ratio of Two Variances 247

5-6 Inference on Two Population Proportions 250
- 5-6.1 Large-Sample Test for H_0: $p_1 = p_2$ 250
- 5-6.2 The β-Error and Choice of Sample Size 252
- 5-6.3 A Confidence Interval for $p_1 - p_2$ 254

5-7 Summary Table for Inference Procedures for Two Samples 256

5-8 What If We Have More Than Two Means? 256
- 5-8.1 An Example 257
- 5-8.2 The Analysis of Variance 258

6 Building Empirical Models 275

6-1 Introduction to Empirical Models 275

6-2 Least Squares Estimation of the Parameters 283
- 6-2.1 Simple Linear Regression 283
- 6-2.2 Multiple Linear Regression 287

6-3 Properties of the Least Squares Estimators and Estimation of σ^2 297

6-4 Hypothesis Testing in Linear Regression 301
- 6-4.1 Test for Significance of Regression 301
- 6-4.2 Tests on Individual Regression Coefficients 304

6-5 Confidence Intervals in Linear Regression 308
- 6-5.1 Confidence Intervals on Individual Regression Coefficients 308
- 6-5.2 Confidence Interval on the Mean Response 309

6-6 Prediction of New Observations 313

6-7 Assessing the Adequacy of the Regression Model 317
- 6-7.1 Residual Analysis 317
- 6-7.2 The Coefficient of Multiple Determination 322
- 6-7.3 Influential Observations 324

7 Design of Engineering Experiments 333

7-1 The Strategy of Experimentation 333

7-2 Some Applications of Experimental Design Techniques 335

7-3 Factorial Experiments 339

7-4 2^k Factorial Design 344
- 7-4.1 2^2 Design 344
- 7-4.2 Analysis 347
- 7-4.3 Residual Analysis and Model Checking 351

7-5 2^k Design for $k \geq 3$ Factors 356

7-6 Single Replicate of a 2^k Design 364

7-7 Addition of Center Points to a 2^k Design 369

7-8 Fractional Replication of a 2^k Design 375
- 7-8.1 One-Half Fraction of a 2^k Design 376
- 7-8.2 Smaller Fractions: 2^{k-p} Fractional Factorial Design 383

7-9 Response Surface Methods and Designs 395
- 7-9.1 Method of Steepest Ascent 397
- 7-9.2 Analysis of a Second-Order Response Surface 400

7-10 Factorial Experiments with More Than Two Levels 410

8 Statistical Quality Control 426

8-1 Quality Improvement and Statistics 426

8-2 Statistical Quality Control 428

8-3 Statistical Process Control 428

8-4 Introduction to Control Charts 429
- 8-4.1 Basic Principles 429
- 8-4.2 Design of a Control Chart 433
- 8-4.3 Rational Subgroups 435
- 8-4.4 Analysis of Patterns on Control Charts 436

8-5 $\overline{X}$ and R control Charts 439

8-6 Control Charts for Individual Measurements 447

8-7 Process Capability 453

8-8 Attribute Control Charts 459
- 8-8.1 *P* Chart (Control Chart for Proportions) 459
- 8-8.2 *U* Chart (Control Chart for Defects per Unit) 462

8-9 Control Chart Performance 465

Appendices 473

A. Statistical Tables and Charts A-1

B. Bibliography B-1

C. Answers to Selected Problems C-1

Index I-1

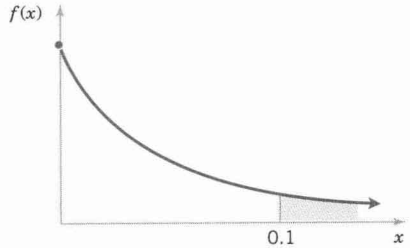

Figure 3-27 Probability for the exponential distribution in Example 3-27.

Determine the interval of time such that the probability that no log-on occurs in the interval is 0.90. The question asks for the length of time x such that $P(X > x) = 0.90$. Now,

$$P(X > x) = e^{-25x} = 0.90$$

Therefore, upon taking logarithms of both sides

$$x = 0.00421 \text{ hour} = 0.25 \text{ minute}$$

Furthermore, the mean time until the next log-on is

$$E(X) = 1/25 = 0.04 \text{ hour} = 2.4 \text{ minutes}$$

The standard deviation of the time until the next log-on is

$$\sigma_X = 1/25 \text{ hours} = 2.4 \text{ minutes}$$

In Example 3-27, the probability that there are no log-ons in a 6-minute interval is 0.082 regardless of the starting time of the interval. A Poisson process assumes that events occur uniformly throughout the interval of observation; that is, there is no clustering of events. If the log-ons are well modeled by a Poisson process, the probability that the first log-on after noon occurs after 12:06 P.M. is the same as the probability that the first log-on after 3:00 P.M. occurs after 3:06 P.M. And if someone logs on at 2:22 P.M., the probability the next log-on occurs after 2:28 P.M. is still 0.082.

Our starting point for observing the system does not matter. However, if there are high-use periods during the day, such as right after 8:00 A.M., followed by a period of low use, a Poisson process is not an appropriate model for log-ons and the distribution is not appropriate for computing probabilities. It might be reasonable to model each of the high- and low-use periods by a separate Poisson process, employing a larger value for λ during the high-use periods and a smaller value otherwise. Then, an exponential distribution with the correspondıng value of λ can be used to calculate log-on probabilities for the high- and low-use periods.

An even more interesting property of an exponential random variable is the **lack of**

memory property. In Example 3-27, suppose that there are no log-ons from 12:00 to 12:15; the probability that there are no log-ons from 12:15 to 12:21 is still 0.082. Because we have already been waiting for 15 minutes, we feel that we are "due." That is, the probability of a log-on in the next 6 minutes should be greater than 0.082. However, for an exponential distribution this is not true.

The lack of memory property is not that surprising when you consider the development of a Poisson process. In that development, we assumed that an interval could be partitioned into small intervals that were independent. These subintervals are similar to independent, Bernoulli trials that comprise a binomial process; knowledge of previous results does not affect the probabilities of events in future subintervals.

The exponential distribution is often used in reliability studies as the model for the time until failure of a device. For example, the lifetime of a semiconductor chip might be modeled as an exponential random variable with a mean of 40,000 hours. The lack of memory property of the exponential distribution implies that the device does not wear out. That is, regardless of how long the device has been operating, the probability of a failure in the next 1000 hours is the same as the probability of a failure in the first 1000 hours of operation. The lifetime of a device with failures caused by random shocks might be appropriately modeled as an exponential random variable. However, the lifetime of a device that suffers slow mechanical wear, such as bearing wear, is better modeled by a distribution that does not lack memory.

EXERCISES FOR SECTION 3-9.2

3-88. Suppose X has an exponential distribution with $\lambda = 2$. Determine the following.

(a) $P(X \leq 0)$
(b) $P(X \geq 2)$
(c) $P(X \leq 1)$
(d) $P(1 < X < 2)$
(e) Find the value of x such that $P(X < x) = 0.05$.

3-89. Suppose X has an exponential distribution with mean equal to 10. Determine the following.

(a) $P(X > 10)$
(b) $P(X > 20)$
(c) $P(X > 30)$
(d) Find the value of x such that $P(X < x) = 0.95$.

3-90. Suppose the counts recorded by a geiger counter follow a Poisson process with an average of two counts per minute.

(a) What is the probability that there are no counts in a 30-second interval?
(b) What is the probability that the first count occurs in less than 10 seconds?
(c) What is the probability that the first count occurs between 1 and 2 minutes after start-up?

3-91. Continuation of Exercise 3-90.

(a) What is the mean time between counts?
(b) What is the standard deviation of the time between counts?
(c) Determine x, such that the probability that at least one count occurs before time x minutes is 0.95.

3-92. The time between calls to a plumbing supply business is exponentially distributed with a mean time between calls of 15 minutes.

(a) What is the probability that there are no calls within a 30-minute interval?
(b) What is the probability that at least one call arrives within a 10-minute interval?

(c) What is the probability that the first call arrives within 5 and 10 minutes after opening?

(d) Determine the length of an interval of time such that the probability of at least one call in the interval is 0.90.

3-93. The time between arrivals of taxis at a busy intersection is exponentially distributed with a mean of 10 minutes.

(a) What is the probability that you wait longer than one hour for a taxi?

(b) Suppose you have already been waiting for one hour for a taxi, what is the probability that one arrives within the next 10 minutes?

3-94. Continuation of Exercise 3-93.

(a) Determine x such that the probability that you wait more than x minutes is 0.10.

(b) Determine x such that the probability that you wait less than x minutes is 0.90.

(c) Determine x such that the probability that you wait less than x minutes is 0.50.

3-95. The distance between major cracks in a highway follows an exponential distribution with a mean of 5 miles.

(a) What is the probability that there are no major cracks in a 10-mile stretch of the highway?

(b) What is the probability that there are two major cracks in a 10-mile stretch of the highway?

(c) What is the standard deviation of the distance between major cracks?

3-96. Continuation of Exercise 3-95.

(a) What is the probability that the first major crack occurs between 12 and 15 miles of the start of inspection?

(b) What is the probability that there are no major cracks in two separate 5-mile stretches of the highway?

(c) Given that there are no cracks in the first 5 miles inspected, what is the probability that there are no major cracks in the next 10 miles inspected?

3-97. The time between the arrival of electronic messages at your computer is exponentially distributed with a mean of two hours.

(a) What is the probability that you do not receive a message during a two-hour period?

(b) If you have not had a message in the last four hours, what is the probability that you do not receive a message in the next two hours?

(c) What is the expected time between your fifth and sixth message?

3-98. The time between arrivals of small aircraft at a county airport is exponentially distributed with a mean of one hour. What is the probability that more than three aircraft arrive within an hour?

3-99. Continuation of Exercise 3-98.

(a) If 30 separate one-hour intervals are chosen, what is the probability that no interval contains more than three arrivals?

(b) Determine the length of an interval of time (in hours) such that the probability that no arrivals occur during the interval is 0.10.

3-100. The time between calls to a corporate office is exponentially distributed with a mean of 10 minutes.

(a) What is the probability that there are more than three calls in one-half hour?

(b) What is the probability that there are no calls within one-half hour?

(c) Determine x such that the probability that there are no calls within x hours is 0.01.

(d) What is the probability that there are no calls within a two-hour interval?

(e) If four nonoverlapping one-half hour intervals are selected, what is the probability that none of these intervals contains any call?

3-10 NORMAL APPROXIMATION TO THE BINOMIAL AND POISSON DISTRIBUTIONS

Because a binomial random variable is a count from repeated independent trials, the central limit theorem can be applied. Consequently, it should not be surprising to use the normal distribution to approximate binomial probabilities for cases in which n is large. The following example illustrates that for many physical systems the binomial model is appropriate with an extremely large value for n. In these cases, it is difficult to calculate probabilities by using the binomial distribution. Fortunately, the normal approximation is most effective in these cases. An illustration is provided in Fig. 3-28.

EXAMPLE 3-28

In a digital communication channel, assume that the number of bits received in error can be modeled by a binomial random variable, and assume that the probability that a bit is received in error is 1×10^{-5}. If 16 million bits are transmitted, what is the probability that more than 150 errors occur?

Let the random variable X denote the number of errors. Then X is a binomial random variable and

$$\begin{aligned} P(X > 150) &= 1 - P(X \leq 150) \\ &= 1 - \sum_{x=0}^{150} \binom{16{,}000{,}000}{x} (10^{-5})^x (1 - 10^{-5})^{16{,}000{,}000 - x} \end{aligned}$$

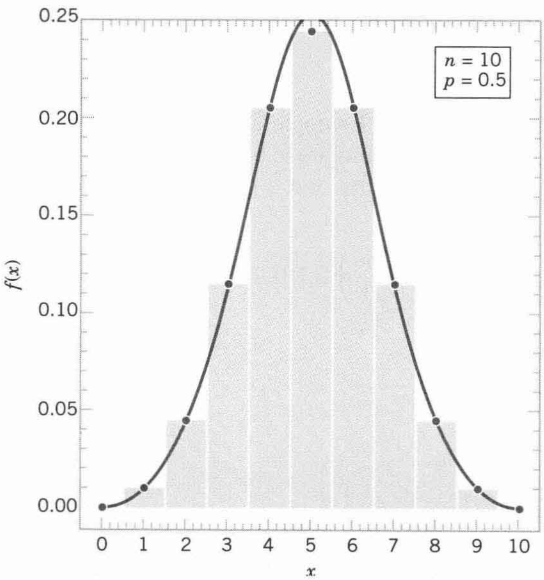

Figure 3-28 Normal approximation to the binomial distribution.

Clearly, the probability in Example 3-28 is difficult to compute. Fortunately, the normal distribution can be used to provide an excellent approximation in this example.

If X is a binomial random variable, then

$$Z = \frac{X - np}{\sqrt{np\,(1-p)}} \tag{3-14}$$

is approximately a standard normal random variable.

Recall that for a binomial variable X, $E(X) = np$ and $V(X) = np(1 - p)$. Consequently, the normal approximation is nothing more than the formula for standardizing the random variable X. Probabilities involving X can be approximated by using a standard normal random variable. The normal approximation to the binomial distribution is good *if n is large enough relative to p;* in particular, whenever $np > 5$ and $n(1 - p) > 5$. The digital communication problem in Example 3-27 is solved as follows

$$\begin{aligned} P(X > 150) &= P\left(\frac{X - 160}{\sqrt{160(1 - 10^{-5})}} > \frac{150 - 160}{\sqrt{160(1 - 10^{-5})}}\right) \\ &= P(Z > -0.79) = P(Z < 0.79) = 0.785 \end{aligned}$$

EXAMPLE 3-29

Again consider the transmission of bits in Example 3-28. To judge how well the normal approximation works, assume only $n = 50$ bits are to be transmitted and that the probability of an error is $p = 0.1$. The exact probability that 2 or less errors occur is

$$\begin{aligned} P(X \le 2) &= \binom{50}{0}0.9^{50} + \binom{50}{1}0.1(0.9^{49}) + \binom{50}{2}0.1^2(0.9^{48}) \\ &= 0.11 \end{aligned}$$

Based on the normal approximation

$$P(X \le 2) = P\left(\frac{X - 5}{2.12} < \frac{2 - 5}{2.12}\right) = P(Z < -1.415) = 0.08$$

For a sample as small as 50 bits, the normal approximation is reasonable.