

Population • Sample • Variable • Parameter
DAVID M. LEVINE • DAVID F. STEPHAN

- Descriptive • Inferential • Experimental
Data • Charts • Tables • Graphs • Models
- Statistics for finance, quality, marketing, science...or anything else
 - Easy explanations, real-world examples
 - Step-by-step instructions for Microsoft Excel and TI-83/84 calculators and downloadable practice files
- Mode • 3rd Quartile • $P(A)$ • Z Score

Even You Can Learn Statistics

*A Guide for
Everyone Who Has
Ever Been Afraid
Of Statistics*



Even You Can Learn Statistics

This page intentionally left blank

Even You Can Learn Statistics

**A Guide for Everyone Who Has Ever
Been Afraid of Statistics**

David M. Levine, Ph.D.

David F. Stephan



PEARSON PRENTICE HALL

An Imprint of PEARSON EDUCATION

Upper Saddle River, NJ • New York • London • San Francisco • Toronto

Sydney • Tokyo • Singapore • Hong Kong


Cape Town • Madrid • Paris • Milan • Munich • Amsterdam

www.ft-ph.com

Library of Congress Catalog-in-Publication: 2004107420

Executive Editor: *Jim Boyd*
Editorial Assistant: *Richard Winkler*
Marketing Manager: *Martin Litkowski*
International Marketing Manager: *Tim Galligan*
Managing Editor: *Gina Kanouse*
Project Editor: *Kayla Dugger*
Design Manager: *Sandra Schroeder*
Cover Designers: *Alan Clements and Gary Adair*
Composition and Interior Design: *Argosy and Jake McFarland*
Manufacturing Buyer: *Dan Uhrig*

© 2005 Pearson Education, Inc.

 Publishing as Pearson Prentice Hall
Upper Saddle River, NJ 07458

Prentice Hall offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales. For more information, please contact: U.S. Corporate and Government Sales, 1-800-382-3419, corpsales@pearsontechgroup.com. For sales outside of the U.S., please contact: International Sales, 1-317-581-3793, international@pearsontechgroup.com.

Company and product names mentioned herein are the trademarks or registered trademarks of their respective owners.

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

Printed in the United States of America
1st Printing

ISBN 0-13-146757-3

Pearson Education Ltd.
Pearson Education Australia Pty., Limited
Pearson Education Singapore, Pte. Ltd.
Pearson Education North Asia Ltd.
Pearson Education Canada, Ltd.
Pearson Educación de Mexico, S.A. de C.V.
Pearson Education—Japan
Pearson Education Malaysia, Pte. Ltd.

To our wives

Marilyn and Mary

To our children

Sharyn and Mark

And to our parents

in loving memory, Lee, Reuben, and Francis

in honor, Ruth

This page intentionally left blank

Table of Contents

Introduction	xvii
 Chapter 1 Fundamentals of Statistics	 1
1.1 The Five Basic Words of Statistics	2
Population	2
Sample	2
Parameter	2
Statistic	3
Variable	3
1.2 The Branches of Statistics	4
Descriptive Statistics	4
Inferential Statistics	5
1.3 Sources of Data	5
Published Sources	5
Experiments	6
Surveys	6
1.4 Sampling Concepts	7
Sampling	7
Probability Sampling	7
Simple Random Sampling	7
Frame	8
1.5 Sample Selection Methods	8
Sampling With Replacement	8
Sampling Without Replacement	8
One-Minute Summary	10
Test Yourself	11
Answers to Test Yourself Questions	14
References	14
 Chapter 2 Presenting Data in Charts and Tables	 17
2.1 Presenting Categorical Data	17
The Summary Table	17
The Bar Chart	18
The Pie Chart	19
The Pareto Diagram	20

Two-Way Cross-Classification Tables	22
2.2 Presenting Numerical Data	24
The Frequency and Percentage Distribution	24
Histogram	25
The Dot Scale Diagram	27
The Time-Series Plot	28
The Scatter Plot	28
2.3 Misusing Graphs	30
One-Minute Summary	32
Test Yourself	32
Answers to Test Yourself Questions	35
References	36
 Chapter 3 Descriptive Statistics for Numerical Variables	 37
3.1 Measures of Central Tendency	37
The Mean	37
The Median	40
The Mode	41
Quartiles	41
3.2 Measures of Variation	45
The Range	45
The Variance and the Standard Deviation	46
Standard (Z) Scores	49
3.3 Shape of Distributions	50
Symmetrical Shape	50
Left-Skewed Shape	50
Right-Skewed Shape	51
The Box-and-Whisker Plot	52
Important Equations	56
One-Minute Summary	56
Test Yourself	57
Answers to Test Yourself Questions	59
References	60
 Chapter 4 Probability	 61
4.1 Getting Started with Probability	61
Event	61
Elementary Event	62

Random Variable	62
Probability	62
Collectively Exhaustive Events	64
4.2 Some Rules of Probability	64
4.3 Assigning Probabilities	67
Classical Approach	67
Empirical Approach	67
Subjective Approach	68
One-Minute Summary	68
Test Yourself	68
Answers to Test Yourself Questions	70
References	70

Chapter 5 Probability Distributions 73

5.1 Probability Distributions for Discrete Variables	73
Discrete Probability Distribution	73
The Expected Value of a Random Variable	75
Standard Deviation of a Random Variable (σ)	76
5.2 The Binomial and Poisson Probability Distributions	79
The Binomial Distribution	79
The Poisson Distribution	83
5.3 Continuous Probability Distributions and the Normal Distribution	87
Normal Distribution	87
Using Standard Deviation Units	89
Finding the Z Value from the Area Under the Normal Curve	91
5.4 The Normal Probability Plot	94
Important Equations	96
One-Minute Summary	96
Test Yourself	97
Answers to Test Yourself Questions	101
References	102

Chapter 6 Sampling Distributions and Confidence Intervals 103

6.1 Sampling Distributions	104
Sampling Distribution	104
Sampling Distribution of the Mean and the Central Limit Theorem	104

Sampling Distribution of the Proportion	107
What You Need to Know About Sampling Distributions	107
6.2 Sampling Error and Confidence Intervals	107
Sampling Error	109
Confidence Interval Estimate	109
6.3 Confidence Interval Estimate for the Mean Using the t Distribution (σ Unknown)	111
t Distribution	112
6.4 Confidence Interval Estimation for the Proportion	116
Important Equations	119
One-Minute Summary	119
Test Yourself	119
Answers to Test Yourself Questions	122
References	122
 Chapter 7 Fundamentals of Hypothesis Testing	 125
7.1 The Null and Alternative Hypotheses	125
Null Hypothesis	126
Alternative Hypothesis	126
7.2 Hypothesis Testing Issues	127
Test Statistic	127
Practical Significance Versus Statistical Significance	128
7.3 Decision-Making Risks	129
Type I Error	129
Type II Error	129
Risk Trade-Off	130
7.4 Performing Hypothesis Testing	130
The p -Value Approach to Hypothesis Testing	131
p -Value	131
7.5 Types of Hypothesis Tests	132
Number of Groups	132
Relationship Stated in Alternative Hypothesis H_1	132
Type of Variable	132
One-Minute Summary	133
Test Yourself	133
Answers to Test Yourself Questions	135
References	135

Chapter 8 Hypothesis Testing: Z and t Tests	137
8.1 Testing for the Difference Between Two Proportions	137
8.2 Testing for the Difference Between the Means of Two Independent Groups	143
Pooled-Variance t Test	143
Pooled-Variance t Test Assumptions	148
8.3 The Paired t Test	150
Important Equations	155
One-Minute Summary	156
Test Yourself	156
Answers to Test Yourself Questions	157
References	158
 Chapter 9 Hypothesis Testing: Chi-Square Tests and the One-Way Analysis of Variance (ANOVA)	 159
9.1 Chi-Square Test for Two-Way Tables	159
9.2 One-Way Analysis of Variance (ANOVA): Testing for the Differences Among the Means of More Than Two Groups	166
One-Way ANOVA	166
The Three Variances of ANOVA	168
ANOVA Summary Table	170
One-Way ANOVA Assumptions	174
Important Equations	174
One-Minute Summary	175
Test Yourself	175
Answers to Test Yourself Questions	178
References	178
 Chapter 10 Regression Analysis	 181
10.1 Basics of Regression Analysis	182
Simple Linear Regression	182
10.2 Determining the Simple Linear Regression Equation	183
Y intercept	183
Slope	183
Least-Squares Method	184
Regression Model Prediction	187
10.3 Measures of Variation	191
Regression Sum of Squares (SSR)	191

Error Sum of Squares (SSE)	191
Total Sum of Squares (SST)	192
The Coefficient of Determination	193
The Coefficient of Correlation	194
Standard Error of the Estimate	194
10.4 Regression Assumptions	195
10.5 Residual Analysis	196
Residual	196
Evaluating the Assumptions	197
10.6 Inferences About the Slope	197
t Test for the Slope	198
Confidence Interval Estimate of the Slope (β_1)	200
10.7 Common Mistakes Using Regression Analysis	201
Important Equations	203
One-Minute Summary	205
Test Yourself	205
Answers to Test Yourself Questions	207
References	208

Chapter 11 Quality and Six Sigma Management Applications of Statistics **209**

11.1 Total Quality Management	209
11.2 Six Sigma Management	211
Six Sigma	211
The Six Sigma DMAIC Model	211
11.3 Control Charts	212
Special or Assignable Causes of Variation	212
Chance or Common Causes of Variation	213
Control Limits	213
The p Chart	214
11.4 The Parable of the Red Bead Experiment: Understanding Process Variability	219
Deming's Red Bead Experiment	220
11.5 Variables Control Charts for the Mean and Range	221
Important Equations	226
One-Minute Summary	227
Test Yourself	227
Answers to Test Yourself Questions	229
References	230

Appendix A TI Statistical Calculator Settings and Microsoft Excel Settings	231
A.1 TI Statistical Calculator Settings	231
“Ready State” Assumptions	231
Menu Selections	231
Statistical Function Entries by Menus	232
Primary Key Legend Convention	232
Mode Settings	232
Calculator Clearing and Reset	232
Data Storage	232
A.2 Microsoft Excel Settings	233
 Appendix B Review of Arithmetic and Algebra	 235
Assessment Quiz	235
Part 1	235
Part 2	236
Symbols	238
Addition	238
Subtraction	239
Multiplication	239
Division	240
Fractions	241
Exponents and Square Roots	242
Equations	243
Answers to Quiz	244
Part 1	244
Part 2	244
 Appendix C Statistical Tables	 245
C.1 The Cumulative Standardized Normal Distribution	246
C.2 Critical Values of t	248
C.3 Critical Values of χ^2	252
C.4 Critical Values of F	254
C.5 Control Chart Factors	262

Appendix D Using Microsoft Excel Wizards	263
D.1 Using the Chart Wizard	263
Choosing the Best Chart Options	264
D.2 Using the PivotTable Wizard	265
D.3 Using the Data Analysis Tools	267
D.4 Simple Linear Regression	267
Glossary	269
Index	277

Acknowledgements

This book would not have been produced without the helpful feedback of: Mark Berenson, Montclair State University; Howard Gitlow, University of Miami; Tim Krehbiel, Miami University; and Russ Hall.

We especially want to thank the staff at Financial Times/Pearson: Jim Boyd, for helping us make this book a reality, Kayla Dugger for her proofreading, Keith Cline for his copyediting, and Gina Kanouse for her work in the production of this text.

We have sought to make the content of this book as clear, accurate, and error-free as possible. We invite you to make suggestions or ask questions about the content if you think we have fallen short of our goals in any way. Please e-mail your comments to david_levine@baruch.cuny.edu.

About the Authors

Together, David M. Levine and David F. Stephan have more than 50 years teaching experience at the college level.

David M. Levine is Professor Emeritus of Statistics and Computer Information Systems at Baruch College (CUNY). A statistics education innovator who has co-authored several best-selling textbooks, including *Statistics for Managers Using Microsoft Excel*, *Basic Business Statistics: Concepts and Applications*, *Business Statistics: A First Course*, and *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*, he has recently finished two books on quality: *Quality Management* and *Six Sigma for Green Belts and Champions*.

David F. Stephan, an instructional designer and lecturer who pioneered the teaching of spreadsheet applications more than 20 years ago and now focuses on developing materials that make the Excel statistical functions more accessible to users, is a frequent coauthor of David Levine.

Introduction

The Even You Can Learn Statistics Owners Manual

In today's world, knowing how to apply statistics is more important than ever. *Even You Can Learn Statistics: The Easy to Use Guide for Everyone Who Has Ever Been Afraid of Statistics* will teach you the basic concepts that provide that understanding. You will also learn the most commonly used statistical methods and be able to practice those methods using a statistical calculator or a spreadsheet program. Please read the rest of this introduction so that you can become familiar with the distinctive features of this book. Be sure to visit the Web site for this book (www.prenticehall.com/youcanlearnstatistics), which contains free downloads and other material to support your learning.

Mathematics Is Always Optional!

Never mastered higher mathematics—or generally fearful of math? Not to worry, because in *Even You Can Learn Statistics*, you will find that every concept is explained in plain English, without the use of higher mathematics or mathematical symbols. Interested in the mathematical foundations behind statistics? *Even You Can Learn Statistics* includes **EQUATION BLACKBOARDS**, stand-alone sections that present the equations behind statistical methods and complement the main material. Either way, you can learn statistics.

Learning with the Concept-Interpretation Approach

Even You Can Learn Statistics uses a **Concept-Interpretation** approach to help you learn statistics. For each important statistical concept, you will first find a

CONCEPT, a plain-language definition that uses no complicated mathematical terms, followed by an

INTERPRETATION that fully explains the concept and its importance to statistics. When necessary, these sections also review the misconceptions and the errors people make when trying to apply the concept.

For simpler concepts, an **EXAMPLES** section lists real-life examples or applications of the statistical concepts. For more advanced concepts, **WORKED-OUT PROBLEMS** provide a complete solution to a statistical problem—including actual spreadsheet and calculator results—that illustrate how you can apply the concept to your own problems.

Practicing Statistics While You Learn Statistics

To enhance your learning of statistics, you should always review the **WORKED-OUT PROBLEMS**. If you want to practice what you have just learned, you can use the optional **CALCULATOR KEYS** and **SPREADSHEET SOLUTION** sections, which help you apply a statistical calculator or spreadsheet program to statistical analyses.

CALCULATOR KEYS sections give you the keystroke-by-keystroke steps to perform statistical analysis on a Texas Instruments statistical calculator from the TI-83 or TI-84 families, including TI-83 Plus and TI-84 Plus models. (You can adapt many sections for use with other TI statistical calculators—such as any model from the TI-86, TI-89, or Voyage 200 families—that have different keypads and arrangements of statistical functions.)

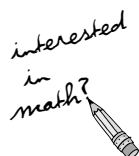
SPREADSHEET SOLUTION sections provide instructions for using the statistical capabilities of Microsoft Excel and identify files that you can download from the *Even You Can Learn Statistics* Web site that contain complete spreadsheets that you can use as models for your own problem solving.

If you plan to use either of these sections, review Appendix A for the conventions, software settings, and assumptions used for these sections.

In-Chapter Aids



As you read a chapter, look for Important Point icons that highlight key explanations. Download the data files from the Web site for this book (www.prenticehall.com/youcanlearnstatistics) so that you may examine the data under study in the Worked-out Problems. Even if you do not plan to use a calculator or a spreadsheet, look at the actual examples of their outputs to become familiar with how statistical results are reported.



Interested in Math? Then look for this icon throughout the book. And if you are not interested in math, remember that all of the passages with this icon can be skipped without losing any comprehension of the statistical methods presented.

End-of-Chapter Features

At the end of most chapters of *Even You Can Learn Statistics*, you will find these features that you can review to reinforce your learning.

Important Equations

A list of the important equations discussed in the chapter. Even if you are not interested in the mathematics of the statistical methods and have skipped the EQUATION BLACKBOARDS in the book, you can use these lists for reference and later study.

One-Minute Summary

A quick review of the significant topics of a chapter in outline form. When appropriate, the summaries also help guide you to make the right decisions about applying statistics to the data you seek to analyze.

Test Yourself

Explore how much you have retained with a set of questions that enable you to review and test yourself (with answers provided) on the concepts presented in a chapter.

Summary

Even You Can Learn Statistics can help you whether you are studying statistics as part of a formal course or just brushing up on your knowledge of statistics for a specific analysis. Be sure to visit the Web site for this book (www.prenticehall.com/youcanlearnstatistics). You are also invited to contact the authors via e-mail at david_levine@baruch.cuny.edu if you have any questions about this book.

This page intentionally left blank



Fundamentals of Statistics

1.1 The Five Basic Words of Statistics

1.2 The Branches of Statistics

1.3 Sources of Data

1.4 Sampling Concepts

1.5 Sample Selection Methods

One-Minute Summary

Test Yourself

Every day, you encounter numerical information that describes or analyzes some aspect of the world you live in. For example, here are some news items that appeared in the pages of *The New York Times* during a one-month period:

- Between 1969 and 2001, the rate of forearm fractures rose 52% for girls and 32% for boys, with the largest increases among children in early puberty, according to a recent Mayo Clinic study.
- Across the New York metropolitan area, the median sales price of a single-family home has risen by 75% since 1998, an increase of more than \$140,000.
- A study that explored the relationship between the price of a book and the number of copies of a book sold found that raising prices by 1% reduced sales by 4% at BN.com, but reduced sales by only 0.5% at Amazon.com.

Such stories as these would not be possible to understand without **statistics**, the branch of mathematics that consists of methods of processing and analyzing data to better support rational decision-making processes. Using statistics to better understand the world means more than just producing a new set of numerical information—you must *interpret* the results by reflecting on the significance and the importance of the results to the decision-making

process you face. Interpretation also means knowing when to ignore results, either because they are misleading, are produced by incorrect methods, or just restate the obvious, as this news story “reported” by the comedian David Letterman illustrates:

USA Today has come out with a new survey. Apparently, 3 out of every 4 people make up 75% of the population.

As newer technologies allow people to process and analyze ever-increasing amounts of data, statistics plays an increasingly important part of many decision-making processes today. Reading this chapter will help you understand the fundamentals of statistics and introduce you to concepts that are used throughout this book.

1.1 The Five Basic Words of Statistics



The five words *population*, *sample*, *parameter*, *statistic* (singular), and *variable* form the basic vocabulary of statistics. You cannot learn much about statistics unless you first learn the meanings of these five words.

Population

CONCEPT All the members of a group about which you want to draw a conclusion.

EXAMPLES All U.S. citizens who are currently registered to vote, all patients treated at a particular hospital last year, the entire daily output of a cereal factory's production line.

Sample

CONCEPT The part of the population selected for analysis.

EXAMPLES The registered voters selected to participate in a recent survey concerning their intention to vote in the next election, the patients selected to fill out a patient-satisfaction questionnaire, 100 boxes of cereal selected from a factory's production line.

Parameter

CONCEPT A numerical measure that describes a characteristic of a population.

EXAMPLES The percentage of all registered voters who intend to vote in the next election, the percentage of all patients who are very satisfied with the care they received, the average weight of all the cereal boxes produced on a factory's production line on a particular day.

Statistic

CONCEPT A numerical measure that describes a characteristic of a sample.

EXAMPLES The percentage in a sample of registered voters who intend to vote in the next election, the percentage in a sample of patients who are very satisfied with the care they received, the average weight of a sample of cereal boxes produced on a factory’s production line on a particular day.

INTERPRETATION Calculating statistics for a sample is the most common activity, because collecting population data is impractical for most actual decision-making situations.

Variable

CONCEPT A characteristic of an item or an individual that will be analyzed using statistics.

EXAMPLES Gender, the household income of the citizens who voted in the last presidential election, the publishing category (hardcover, trade paperback, mass-market paperback, textbook) of a book, the number of varieties of a brand of cereal.

INTERPRETATION All the variables taken together form the data of an analysis. Although you may have heard people saying that they are analyzing their data, they are, more precisely, analyzing their variables.

You should distinguish between a variable, such as gender, and its **value** for an individual, such as male. An **observation** is all the values for an individual item in the sample. For example, a survey might contain two variables, gender and age. The first observation might be male, 40. The second observation might be female, 45. The third observation might be female, 55. A **variable** is sometimes known as a column of data because of the convention of entering each observation as a unique row in a table of data. (Likewise, you may hear some refer to an observation as a row of data.)

Variables can be divided into the following types:

	Categorical Variables	Numerical Variables
Concept	The values of these variables are selected from an established list of categories.	The values of these variables involve a counted or measured value.
Subtypes	None.	Discrete values are counts of things. Continuous values are measures, and any value can theoretically occur, limited only by the precision of the measuring process.

(continues)

	Categorical Variables	Numerical Variables
Examples	<p>Gender, a variable that has the categories male and female.</p> <p>Academic major, a variable that might have the categories English, Math, Science, and History, among others.</p>	<p>The number of previous presidential elections in which a citizen voted, a discrete numerical variable.</p> <p>The household income of a citizen who voted, a continuous variable.</p>



All variables should have an **operational definition**—that is, a universally-accepted meaning that is clear to all associated with an analysis. Without operational definitions, confusion can occur. A famous example of such confusion was the tallying of votes in Florida during the 2000 U.S. presidential election in which, at various times, nine different definitions of a valid ballot were used. (A later analysis¹ determined that three of these definitions, including one pursued by Al Gore, led to margins of victory for George Bush that ranged from 225 to 493 votes and that the six others, including one pursued by George Bush, led to margins of victory for Al Gore that ranged from 42 to 171 votes.)

1.2 The Branches of Statistics

Two branches, *descriptive statistics* and *inferential statistics*, comprise the field of statistics.

Descriptive Statistics

CONCEPT The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

EXAMPLES The average age of citizens who voted for the winning candidate in the last presidential election, the average length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory's production line.

INTERPRETATION You are most likely to be familiar with this branch of statistics, because many examples arise in everyday life. Descriptive statistics forms the basis for analysis and discussion in such diverse fields as securities

¹ J. Calmes and E. P. Foldessy, "In Election Review, Bush Wins with No Supreme Court Help," *Wall Street Journal*, November 12, 2001, A1, A14

trading, the social sciences, government, the health sciences, and professional sports. A general familiarity and widespread availability of descriptive methods in many calculating devices and business software can often make using this branch of statistics seem deceptively easy. (Chapters 2 and 3 warn you of the common pitfalls of using descriptive methods.)

Inferential Statistics

CONCEPT The branch of statistics that analyzes sample data to draw conclusions about a population.

EXAMPLE A survey that sampled 2,001 full- or part-time workers ages 50 to 70, conducted by the American Association of Retired Persons (AARP), discovered that 70% of those polled planned to work past the traditional mid-60s retirement age. By using methods discussed in Section 6.4, this statistic could be used to draw conclusions about the population of all workers ages 50 to 70.

INTERPRETATION When you use inferential statistics, you start with a hypothesis and look to see whether the data are consistent with that hypothesis. Inferential statistical methods can be easily misapplied or misconstrued, and many inferential methods require the use of a calculator or computer. (A full explanation of common inferential methods appears in Chapters 6 through 9.)

1.3 Sources of Data

All statistical analysis begins by identifying the source of the data. Among the important sources of data are *published sources*, *experiments*, and *surveys*.

Published Sources

CONCEPT Data available in print or in electronic form, including data found on Internet Web sites. Primary data sources are those published by the individual or group that collected the data. Secondary data sources are those compiled from primary sources.

EXAMPLES Many U.S. federal agencies, including the Census Bureau, publish primary data sources that are available at the Web site www.fedstats.gov. Business news sections of daily newspapers commonly publish secondary source data compiled by business organizations and government agencies.

INTERPRETATION You should always consider the possible bias of the publisher and whether the data contain all the necessary and relevant variables

when using published sources. Remember, too, that *anyone* can publish data on the Internet.

Experiments

CONCEPT A process that studies the effect on a variable of varying the value(s) of another variable or variables, while keeping all other things equal. A typical experiment contains both a treatment group and a control group. The treatment group consists of those individuals or things that receive the treatment(s) being studied. The control group consists of those individuals or things that do not receive the treatment(s) being studied.

EXAMPLE Pharmaceutical companies use experimental studies to determine whether a new drug is effective. A group of patients who have many similar characteristics is divided into two subgroups. Members of one group, the treatment group, receive the new drug. Members of the other group, the control group, receive a **placebo**, a substance that has no medical effect. After a time period, statistics about each group are compared.

INTERPRETATION Proper experiments are either single-blind or double-blind. A study is a single-blind experiment if only the researcher conducting the study knows the identities of the members of the treatment and control groups. If neither the researcher nor study participants know who is in the treatment group and who is in the control group, the study is a double-blind experiment.

When conducting experiments that involve placebos, researchers also have to consider the **placebo effect**—that is, whether people in the control group will improve because they believe that they are getting a real substance that is intended to produce a positive result. When a control group shows as much improvement as the treatment group, a researcher can conclude that the placebo effect is a significant factor in the improvements of both groups.

Surveys

CONCEPT A process that uses questionnaires or similar means to gather values for the responses from a set of participants.

EXAMPLES The decennial U.S. census mail-in form, a poll of likely voters, a Web site instant poll or “question of the day.”

INTERPRETATION Surveys are either informal, open to anyone who wishes to participate; targeted, directed toward a specific group of individuals; or include people chosen at random. The type of survey affects how the data collected can be used and interpreted.

1.4 Sampling Concepts

Sampling

CONCEPT The process by which members of a population are selected for a sample.

EXAMPLES Choosing every fifth voter who leaves a polling place to interview, drawing playing cards randomly from a deck, polling every tenth visitor who views a certain Web site today.

INTERPRETATION The method by which sampling occurs, the identification of all items in a population, and the techniques used to select individual observations all affect sampling.

Probability Sampling

CONCEPT A sampling process that takes into consideration the chance of occurrence of each item being selected. Probability sampling increases your chances that the sample will be representative of the population.

EXAMPLES The registered voters selected to participate in a recent survey concerning their intention to vote in the next election, the patients selected to fill out a patient-satisfaction questionnaire, 100 boxes of cereal selected from a factory's production line.

INTERPRETATION You should use probability sampling whenever possible, because only this type of sampling allows you to apply inferential statistical methods to the data you collect. In contrast, you should use nonprobability sampling, in which the chance of occurrence of each item being selected is not known, to obtain rough approximations of results at low cost or for small-scale, initial, or pilot studies that will later be followed up by a more rigorous analysis. Surveys and polls that invite the public to call in or answer questions on a Web page are examples of nonprobability sampling.

Simple Random Sampling

CONCEPT The probability sampling process in which every individual or item from a population has the same chance of selection as every other individual or item. Every possible sample of a certain size has the same chance of being selected as every other sample that has that size.

EXAMPLES Selecting a playing card from a shuffled deck, generating a number by throwing a pair of perfect dice, or using a statistical device such as a table of random numbers.

INTERPRETATION Simple random sampling forms the basis for other random sampling techniques. The word random in the phrase *random sampling* may confuse you if you think that random implies the unexpected or the

unanticipated, as the word often does in everyday usage (as in random acts of kindness). However, in statistics, *random* implies no repeating patterns—that is, in a given sequence, a given pattern is equally likely (or unlikely) as another. From this sense of equal chance (and not unexpected or unanticipated) comes the term *random sampling*.

Frame

CONCEPT The list of all items in the population from which samples will be selected.

EXAMPLES Voter registration lists, municipal real estate records, customer or human resource databases, directories.

INTERPRETATION Frames influence the results of an analysis, and using two different frames can lead to different conclusions. You should always be careful to make sure your frame completely represents a population; otherwise any sample selected will be biased, and the results generated by analyses of that sample will be inaccurate.

1.5 Sample Selection Methods

Proper sampling can be done with or without replacement.

Sampling With Replacement

CONCEPT A sampling method in which each selected item is returned to the frame from which it was selected so that it has the same probability of being selected again.

EXAMPLE Selecting entries from a fishbowl and returning each entry to the fishbowl after it is drawn.

Sampling Without Replacement

CONCEPT A sampling method in which each selected item is not returned to the frame from which it was selected. Using this technique, an item can be selected no more than one time.

EXAMPLES Selecting numbers in state lottery games, selecting cards from a deck of cards during games of chance such as Blackjack.

INTERPRETATION Sampling without replacement means that an item can be selected no more than one time. You should choose sampling without

replacement over sampling with replacement, because statisticians generally consider the former to produce more desirable samples.

Other, more complex, sampling methods are also used in survey sampling. In a stratified sample, the items in the frame are first subdivided into separate subpopulations, or **strata**, and a simple random sample is conducted within each of the strata. In a **cluster sample**, the items in the frame are divided into several *clusters* so that each cluster is representative of the entire population. A random sampling of clusters is then taken, and all the items in each selected cluster or a sample from each cluster are then studied.



calculator keys

Entering Data

You can choose one of two ways to enter data values for a variable.

When entering one short list of values for a single variable:

Press [2nd][()] and enter the values separated by commas. (Press [·] to type a comma.) When you finish entering values, press [2nd][)]][STO ▶] and enter the name of the variable in which to store the values. For example, to store values in variable L1, press [2nd][1]. Press [ENTER] to complete the data entry. Your calculator will display the values separated by spaces and your screen will look similar to this:

```
{11,31,17,13,28}
→L1
{11 31 17 13 28}
```

When entering the values for several variables, or many values for a single variable:

Press [STAT]. Select 1:Edit and press [ENTER]. Use the cursor keys to move the cursor to the column of the variable for which you want to enter data. (If you have just cleared your RAM memory, the cursor will be in the column for variable L1.) Enter the first data value and press [ENTER]. Repeat until all values have been entered. Your screen will look similar to this:

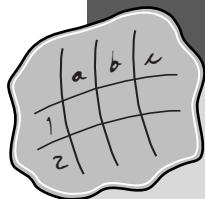
(continues)

L1	L2	L3
11	-----	-----
31		
17		
13		
28		
L1(5)=		

You can enter the data values for a second variable by using the cursor keys to move to the column of another variable. To delete values previously entered into a column, move the cursor to the name of variable and press [CLEAR][ENTER].

When you have finished entering all values, press [2nd][MODE] to quit and return to the main display.

If you have a connection cable and the TI Connect software, you can also enter values for a variable using the TI Data Editor application.



spreadsheet solution

Entering Data

Select **File** → **New**. Select **Blank Workbook** from the task pane. (If using an older version of Excel, select the **Workbook** icon in the New dialog box.) Click cell A1. Enter a name for variable in this cell and press [ENTER]. Type the first data value and press [ENTER]. Repeat until all values have been entered. Notice that every time you press [ENTER] the worksheet entry automatically advances down one row.

When you have finished entering data, select **File** → **Save As**, type a filename, and click the **Save** button to save your data.

One-Minute Summary

To understand statistics, you must first master the basic vocabulary presented in this chapter. You have also been introduced to data collection, the various sources of data, sampling methods, as well as the types of variables used in statistical analysis. The remaining chapters of this book focus on four important reasons for learning statistics:

- To present and describe information (Chapters 2 and 3)
- To draw conclusions about populations based only on sample results (Chapters 4 through 9)
- To obtain reliable forecasts (Chapter 10)
- To improve processes (Chapter 11)

Test Yourself

1. The portion of the population that is selected for analysis is called:
 - (a) a sample
 - (b) a frame
 - (c) a parameter
 - (d) a statistic
2. A summary measure that is computed from only a sample of the population is called:
 - (a) a parameter
 - (b) a population
 - (c) a discrete variable
 - (d) a statistic
3. The height of an individual is an example of a:
 - (a) discrete variable
 - (b) continuous variable
 - (c) categorical variable
 - (d) constant
4. The body style of an automobile (sedan, coupe, wagon, etc.) is an example of a:
 - (a) discrete variable
 - (b) continuous variable
 - (c) categorical variable
 - (d) constant
5. The number of credit cards in a person's wallet is an example of a:
 - (a) discrete variable
 - (b) continuous variable
 - (c) categorical variable
 - (d) constant
6. Statistical inference occurs when you:
 - (a) compute descriptive statistics from a sample
 - (b) take a complete census of a population
 - (c) present a graph of data
 - (d) take the results of a sample and draw conclusions about a population

7. The human resources director of a large corporation wants to develop a dental benefits package and decides to select 100 employees from a list of all 5,000 workers in order to study their preferences for the various components of a potential package. All the employees in the corporation constitute the _____.
 - (a) sample
 - (b) population
 - (c) statistic
 - (d) parameter
8. The human resources director of a large corporation wants to develop a dental benefits package and decides to select 100 employees from a list of all 5,000 workers in order to study their preferences for the various components of a potential package. The 100 employees who will participate in this study constitute the _____.
 - (a) sample
 - (b) population
 - (c) statistic
 - (d) parameter
9. Those methods involving the collection, presentation, and characterization of a set of data in order to properly describe the various features of that set of data are called:
 - (a) statistical inference
 - (b) the scientific method
 - (c) sampling
 - (d) descriptive statistics
10. Based on the results of a poll of 500 registered voters, the conclusion that the Republican candidate for U.S. president will win the upcoming election is an example of:
 - (a) inferential statistics
 - (b) descriptive statistics
 - (c) a parameter
 - (d) a statistic
11. A summary measure that is computed to describe a characteristic of an entire population is called:
 - (a) a parameter
 - (b) a population
 - (c) a discrete variable
 - (d) a statistic

12. You were working on a project to look at the value of the American dollar as compared to the English pound. You accessed an Internet site where you obtained this information for the past 50 years. Which method of data collection were you using?
 - (a) Published sources
 - (b) Experimentation
 - (c) Surveying
13. Which of the following is a discrete variable?
 - (a) The favorite flavor of ice cream of students at your local elementary school
 - (b) The time it takes for a certain student to walk to your local elementary school
 - (c) The distance between the home of a certain student and the local elementary school
 - (d) The number of teachers employed at your local elementary school
14. Which of the following is a continuous variable?
 - (a) The eye color of children eating at a fast-food chain
 - (b) The number of employees of a branch of a fast-food chain
 - (c) The temperature at which a hamburger is cooked at a branch of a fast-food chain
 - (d) The number of hamburgers sold in a day at a branch of a fast-food chain
15. The number of cars that arrive per hour at a parking lot is an example of:
 - (a) a categorical variable
 - (b) a discrete variable
 - (c) a continuous variable
 - (d) a statistic
16. The possible responses to the question “How long have you been living at your current residence?” are values from a continuous variable.
 - (a) True
 - (b) False
17. The possible responses to the question “How many times in the past three months have you visited a museum?” are values from a discrete variable.
 - (a) True
 - (b) False
18. An insurance company evaluates many variables about a person before deciding on an appropriate rate for automobile insurance. The number of accidents a person has had in the past three years is an example of a _____ variable.

19. An insurance company evaluates many variables about a person before deciding on an appropriate rate for automobile insurance. The distance a person drives in a day is an example of a _____ variable.
20. An insurance company evaluates many variables about a person before deciding on an appropriate rate for automobile insurance. A person's marital status is an example of a _____ variable.

Answers to Test Yourself Questions

1. a
2. d
3. b
4. c
5. a
6. d
7. b
8. a
9. d
10. a
11. a
12. a
13. d
14. c
15. b
16. a
17. a
18. discrete
19. continuous
20. categorical

References

1. Berenson, M. L., D. M. Levine, and T. C. Krehbiel. *Basic Business Statistics: Concepts and Applications, Ninth Edition*. Upper Saddle River, NJ: Prentice Hall, 2004.
2. Cochran, W. G. *Sampling Techniques, Third Edition*. New York: Wiley, 1977.

3. Gitlow, H. S., and D. M. Levine. *Six Sigma for Green Belts and Champions*. Upper Saddle River, NJ: Financial Times – Prentice Hall, 2005.
4. Levine, D. M., T. C. Krehbiel, and M. L. Berenson. *Business Statistics: A First Course, Third Edition*. Upper Saddle River, NJ: Prentice Hall, 2003.
5. Levine, D. M., D. Stephan, T. C. Krehbiel, and M. L. Berenson. *Statistics for Managers Using Microsoft Excel, Fourth Edition*. Upper Saddle River, NJ: Prentice Hall, 2005.
6. Levine, D. M., P. P. Ramsey, and R. K. Smidt, *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Upper Saddle River, NJ: Prentice Hall, 2001.
7. Microsoft Excel 2002. Redmond, WA: Microsoft Corporation, 2001.
8. Sincich, T., D. M. Levine, and D. Stephan, *Practical Statistics by Example Using Microsoft Excel and Minitab, Second Edition*. Upper Saddle River, NJ: Prentice Hall, 2002.

This page intentionally left blank



Presenting Data in Charts and Tables

2.1 Presenting Categorical Data

2.2 Presenting Numerical Data

2.3 Misusing Graphs

One-Minute Summary

Test Yourself

Presenting information effectively has become a must in a world that some say faces an information overload. Charts and tables are effective ways of presenting the categorical and numerical data of statistics. Even more important to the study of statistics, you must properly arrange and present categorical and numerical data in order to best apply the statistical methods described later in this book. Reading this chapter will help you learn to select and to develop appropriate tables and charts for both types of data.

2.1 Presenting Categorical Data

You present categorical data by sorting responses by categories. The count, amount, or percentage (part of the whole) of responses by category is then placed into a *summary table* or into one of several forms of charts.

The Summary Table

CONCEPT A two-column table in which the names of the categories are listed in the first column and the count, amount, or percentage of responses are listed in a second column. Sometimes additional columns present the same data in two or more ways (for example, as counts *and* percentages).

EXAMPLE*Blood Donation Behavior by Americans*

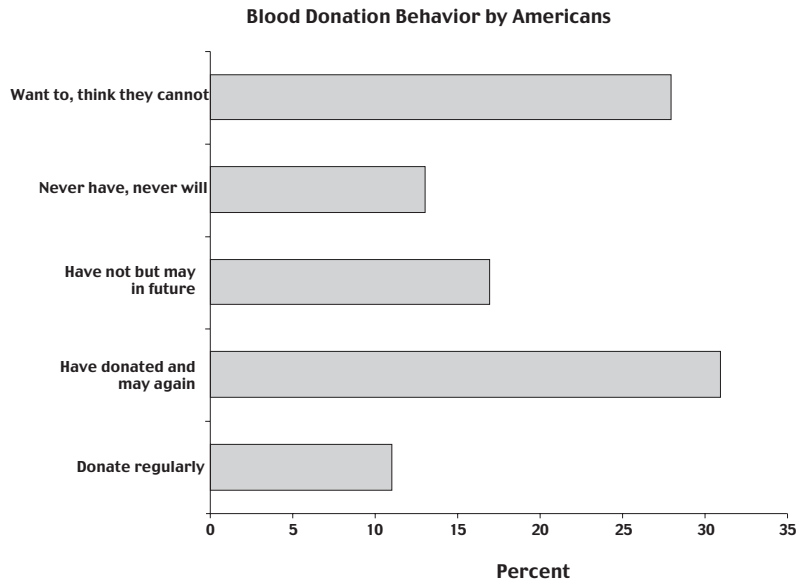
Behavior	Percentage (%)
Donate regularly	11
Have donated and may again	31
Have not but may in future	17
Never have, never will	13
Want to, but think they cannot	28

This table summarizes the results of a blood donation behavior survey that was conducted during the American Red Cross Save a Life Tour¹.

INTERPRETATION Summary tables enable you to see the big picture about a set of data. In the example, you can conclude that there seems a good opportunity to increase the percentage of people giving blood donations because only 11% donate regularly, and an equally small group (13%) say they will never donate.

The Bar Chart

CONCEPT A chart containing rectangles (“bars”) in which the length of each bar represents the count, amount, or percentage of responses of one category.

EXAMPLE

¹ USA Today Snapshots, “35,000 Blood Donations Needed Daily,” *USA Today*, August 20, 2003, p. 1.

This percentage bar chart presents the data of the summary table discussed in the previous example.

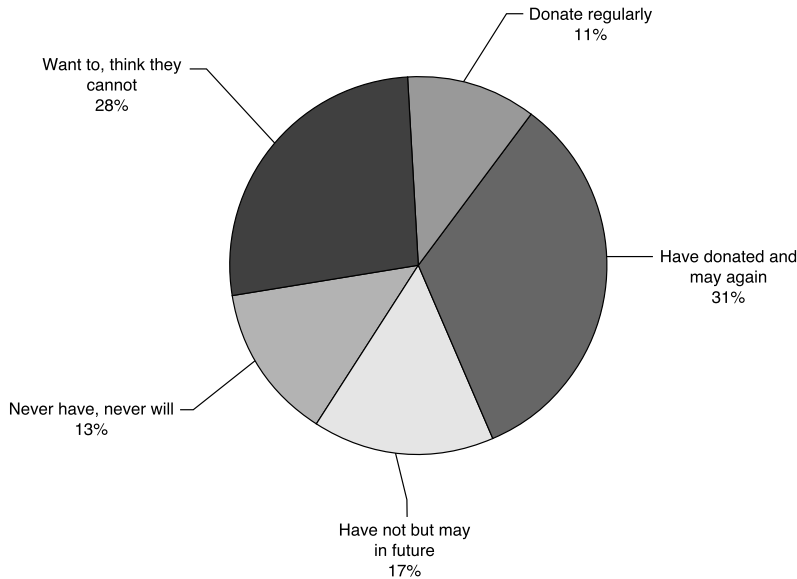
INTERPRETATION A bar chart better makes the point that the category “have donated and may donate again” is the single largest category for this example. For most people, scanning a bar chart is easier than scanning a column of numbers in which the numbers are unordered, as they are in the blood donation summary table.

The Pie Chart

CONCEPT A circle chart in which wedge-shaped areas—pie slices—represent the count, amount, or percentage of each category and the entire circle (“pie”) represents the total.

EXAMPLE

Blood Donation Behavior by Americans



This pie chart presents the data of the summary table discussed in the previous two examples.

INTERPRETATION By ordering the categories carefully, you can make the point that a majority of those surveyed either want to donate but think they cannot or are people who have donated in the past and may do so again in the future.

Although you will probably produce most of your pie charts using computers or calculators, you can also produce a pie chart using a protractor to divide up a drawn circle. To produce a pie chart in this way, first calculate percentages for each category. Then multiply each percentage by 360, the number of degrees in a circle, to get the number of degrees for the arc (part of circle) portion of each category's pie slice. (For example, for the "have donated blood regularly" category, multiply 11% by 360 degrees to get 39.6 degrees.) Mark the endpoints of this arc on the circle using the protractor, and draw lines from the endpoints to the center of the circle. (You may want to draw your circle using a compass so that the center of the circle can be easily identified.)



spreadsheet solution

Bar and Pie Charts

Download and open the **Chapter 2 Bar.xls** and **Chapter 2 Pie.xls** Excel files to see an example of a Microsoft Excel bar and pie chart. You can experiment with each chart by entering your own values in column B.

Microsoft Excel includes a Chart Wizard that assists you to produce custom charts. To learn more about this wizard, see Appendix D.1.

The Pareto Diagram

CONCEPT A special type of bar chart in which the counts, amounts, or percentages of each category are presented in descending order left to right, along with a superimposed plotted line that represents a running cumulative percentage.

EXAMPLE

Computer Keyboards Defects for a Three-Month Period

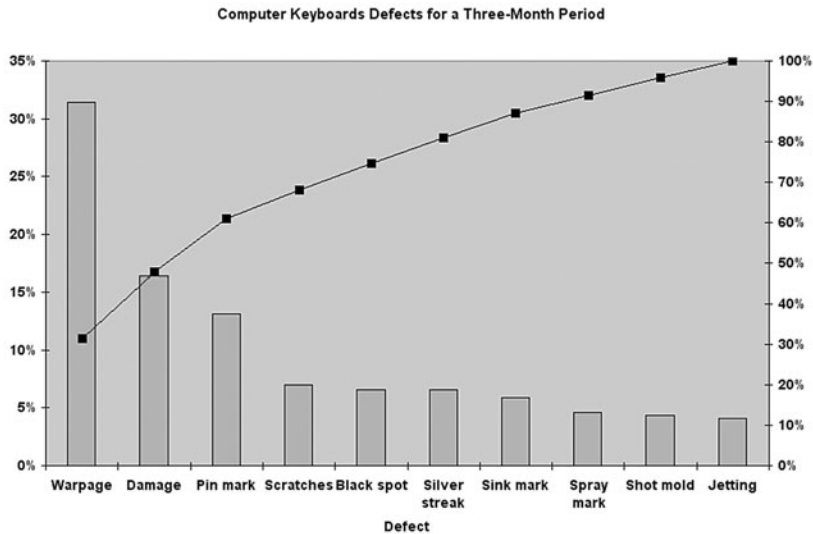
Defect	Frequency	Percentage
Black spot	413	6.53
Damage	1,039	16.43
Jetting	258	4.08
Pin mark	834	13.19
Scratches	442	6.99

Defect	Frequency	Percentage
Shot mold	275	4.35
Silver streak	413	6.53
Sink mark	371	5.87
Spray mark	292	4.62
Warpage	1,987	31.42
Total	6,324	100.01*

(Keyboard)

* Total percentage equals 100.1 due to rounding.

Source: U. H. Acharya and C. Mahesh, "Winning Back the Customer's Confidence: A Case Study on the Application of Design of Experiments to an Injection-Molding Process," *Quality Engineering*, 11, 1999, 357–363.



This Pareto diagram uses the data of the table that immediately precedes it to highlight the causes of computer keyboard defects manufactured during a three-month period.

INTERPRETATION When there are many categories, Pareto diagrams enable you to focus on the most important categories by visually separating the “vital few” from the “trivial many” categories. For the keyboard defects data, the Pareto diagram highlights that two categories, warpage and damage, account for nearly one-half of all defects, and that those two combined with the pin mark category account for more than 60% of all defects.



spreadsheet solution

Pareto Diagrams

Open the **Chapter 2 Pareto.xls** Excel file to see an example of a Microsoft Excel Pareto diagram chart. You can experiment by typing your own set of values—in descending order—in column B, rows 2 through 11. (Do not alter the entries in row 12 or columns C and D.) To produce approximations of custom Pareto diagrams, you can select the **Line – Column on 2 Axes custom** chart type in the Chart Wizard Step 1 dialog box and select the columns for the frequency and percentage data in Step 2.

Two-Way Cross-Classification Tables

CONCEPT A multicolumn table that presents the count or percentage of responses to two categorical variables. In two-way tables, the categories of one of the variables form the rows of the table, while the categories of the second variable form the columns. Cross-classification tables are also known as cross-tabulation tables.

EXAMPLES

Counts of Particles Found Cross-Classified by Wafer Condition

		Wafer Condition		Total
		Good	Bad	
Particles Found	Yes	14	36	50
	No	320	80	400
Total		334	116	450

This two-way cross-classification table summarizes the results of a manufacturing plant study that investigated whether particles found on silicon wafers affected the condition of a wafer. Equivalent tables, showing row percentage, column percentage, and overall total percentage, follow.

Row Percentages Table

		Wafer Condition		Total
		Good	Bad	
Particles Found	Yes	28.0	72.0	100.0
	No	80.0	20.0	100.0
Total		74.2	25.8	100.0

Column Percentages Table

		Wafer Condition		
		Good	Bad	Total
Particles Found	Yes	4.2	31.0	11.1
	No	95.8	69.0	88.9
Total		100.0	100.0	100.0

		Wafer Condition		
		Good	Bad	Total
Particles Found	Yes	3.1	8.0	11.1
	No	71.1	17.8	88.9
Total		74.2	25.8	100.0

Overall Total Percentages Table

INTERPRETATION The simplest two-way table has but two rows and two columns in its inner part (that is, excluding the total).

Column Variable

		1	2	Total
Row Variable	1	Count or percentage for row 1, column 1	Count or percentage for row 1, column 2	Total for row 1
	2	Count or percentage for row 2, column 1	Count or percentage for row 2, column 2	Total for row 2
Total		Total for column 1	Total for column 2	Overall total

Each cell in the inner part of the table represents the count or percentage of a pairing, or cross-classifying, of categories from each variable. Sometimes additional rows and columns present the percentages of the overall total, the percentages of the row total, and the percentages of the column total for each row and column combination.

Two-way tables can reveal what combination of values is most prevalent in data. In the example, the tables reveal that bad wafers are much more likely to have particles than the good wafers. Because the number of good and bad wafers was unequal in this example, you can best see this pattern in the Row Percentage table. That table shows that nearly three-quarters of the wafers that had particles were bad, but only 20% of wafers that did not have particles were bad.



spreadsheet solution

Two-Way Tables

Download and open the **Chapter 2 Two-Way.xls** Excel file to see an example of how you can use Microsoft Excel to create a two-way table.

Microsoft Excel includes a PivotTable Wizard that assists you to produce custom summary tables from sample data. To learn more about this wizard, see Appendix D.2.

2.2 Presenting Numerical Data

You can choose to present numerical data in either table or chart form. If you choose to use a table, you must first establish groups that represent different ranges of values and then place each value into the appropriate group. If you choose a chart, you can either plot each value directly on a chart or you can plot the contents of a table.

Many times you will want to do both, and this section reviews the commonly used frequency and percentage distribution tables and histogram and dot scale diagram charts. You present numerical data by either organizing the responses into groups that contain specific ranges of values or producing charts that represent each response as a point or a symbol. The frequency and percentage distributions, the histogram, and the dot scale diagram are among the many tables and charts that enable you to accomplish these tasks.

The Frequency and Percentage Distribution

CONCEPT A three-column table of grouped numerical data that contains the names of each group in the first column, the counts (frequencies) of each group in the second, and the percentages of each group in the third. This table may also appear as a two-column table that presents either the frequencies or the percentages.

EXAMPLE

Frequency and Percentage Distribution for the Viscosity of a Chemical

Viscosity	Frequency	Percentage
12.0 to under 13	2	1.67%
13.0 to under 14	14	11.67%
14.0 to under 15	45	37.50%

Viscosity	Frequency	Percentage
15.0 to under 16	39	32.50%
16.0 to under 17	17	14.17%
17.0 to under 18	2	1.67%
18.0 to under 19	1	0.83%

This frequency and percentage distribution presents viscosity (friction, as in automobile oil) measurements taken from 120 manufacturing batches, ordered from lowest to highest viscosity (shown below).

Viscosities from 120 Manufacturing Batches

12.6	12.8	13.0	13.1	13.3	13.3	13.4	13.5	13.6	13.7
13.7	13.7	13.8	13.8	13.9	13.9	14.0	14.0	14.0	14.1
14.1	14.1	14.2	14.2	14.2	14.3	14.3	14.3	14.3	14.3
14.3	14.4	14.4	14.4	14.4	14.4	14.4	14.4	14.4	14.5
14.5	14.5	14.5	14.5	14.5	14.6	14.6	14.6	14.7	14.7
14.8	14.8	14.8	14.8	14.9	14.9	14.9	14.9	14.9	14.9
14.9	15.0	15.0	15.0	15.0	15.1	15.1	15.1	15.1	15.2
15.2	15.2	15.2	15.2	15.2	15.2	15.2	15.3	15.3	15.3
15.3	15.3	15.4	15.4	15.4	15.4	15.5	15.5	15.6	15.6
15.6	15.6	15.6	15.7	15.7	15.7	15.8	15.8	15.9	15.9
16.0	16.0	16.0	16.0	16.1	16.1	16.1	16.2	16.3	16.4
16.4	16.5	16.5	16.6	16.8	16.9	16.9	17.0	17.6	18.6

(Chemical)

Source: Holmes and Mergen, "Parabolic Control Limits for the Exponentially Weighted Moving Average Control Charts," 1992, *Quality Engineering* 4(4): 487–495.

INTERPRETATION Frequency and percentage distributions enable you to quickly determine differences among the many values. In this example, you can quickly see that most of the viscosities are between 14.0 and 16.0, and that there are very few viscosities that are either below 13.0 or above 18.0.

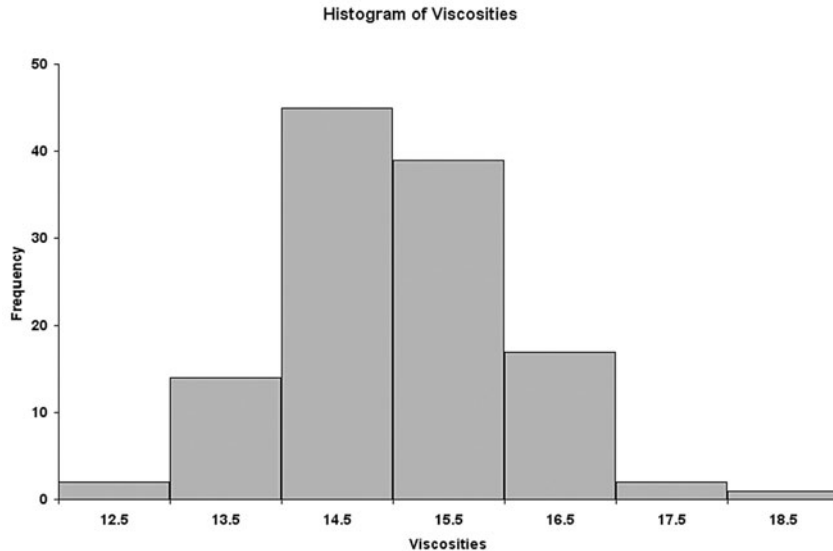
Care should be taken in forming groups for distributions, because the ranges of the group will affect how you perceive the data. For example, had the viscosity data been grouped into only two groups, below 15 and 15-and-above, you would see no pattern to the data.

Histogram

CONCEPT A special bar chart for grouped numerical data in which the frequencies or percentages of each group of numerical data are represented as

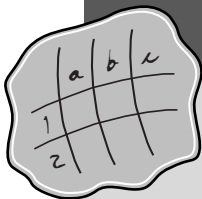
individual bars on the vertical Y-axis and the variable is plotted on the horizontal X-axis. In a histogram, there are no gaps between adjacent bars as there would be in a bar chart of categorical data.

EXAMPLE



This histogram presents the viscosity data of the previous example. The values below the bars (12.5, 13.5, 14.5, 15.5, 16.5, 17.5, and 18.5) are mid-points, the approximate middle value for each group of data. As with the frequency and percentage distributions, you can quickly see that very few viscosities are either below 13 or above 18.

INTERPRETATION Histograms reveal the overall shape of the frequencies for the groups. Histograms are considered symmetric if each side is an approximate mirror image of the other side. (The histogram of the example is an approximately symmetric histogram.)



spreadsheet solution

Frequency Distributions and Histograms

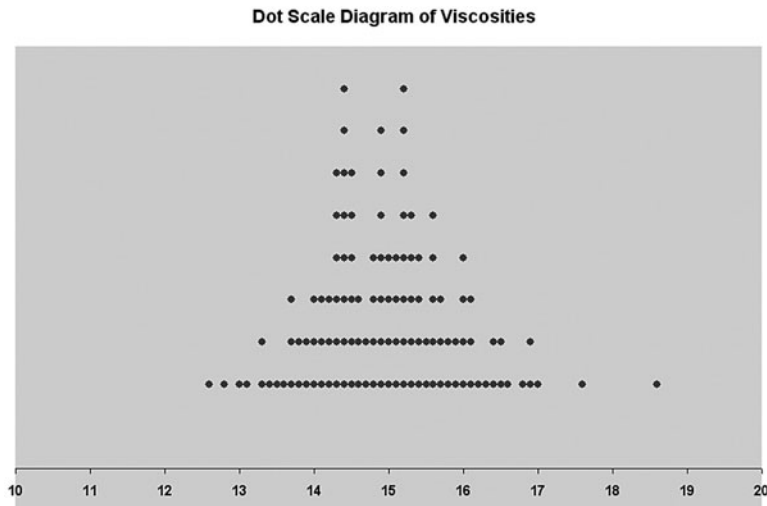
Download and open the **Chapter 2 Histogram.xls** Excel file to see an example of a frequency distribution and histogram for the viscosity data. You can experiment by typing your own set of values in column B, rows 2 through 8. (Do not alter the entries in row 12 or columns C and D.)

You can also produce approximations of histograms either using the Chart Wizard or the Data Analysis Histogram procedure.

The Dot Scale Diagram

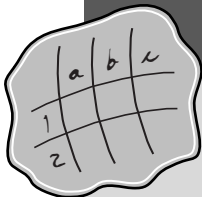
CONCEPT A chart in which each response is represented as a dot above a horizontal line that extends through the range of all values. Should two or more response values be identical, the dots for these responses are stacked (placed vertically) above each other.

EXAMPLE



This dot scale diagram presents the viscosity data for 120 batches of a chemical that is the basis for the previous two examples.

INTERPRETATION By avoiding the grouping of data, the dot scale diagram avoids the misleading patterns that histograms sometime produce. This dot scale diagram shows that the viscosity data is not as nicely symmetrical as the histogram would suggest. Note, however, that both charts clearly show a concentration of values in the center of the distribution between 14 and 16, and also show that very few batches have viscosities below 13 or above 18.



spreadsheet solution

Dot Scale Diagrams

Open the **Chapter 2 Dot Scale.xls** Excel file to see an example of a dot-scale diagram for the viscosity data. You

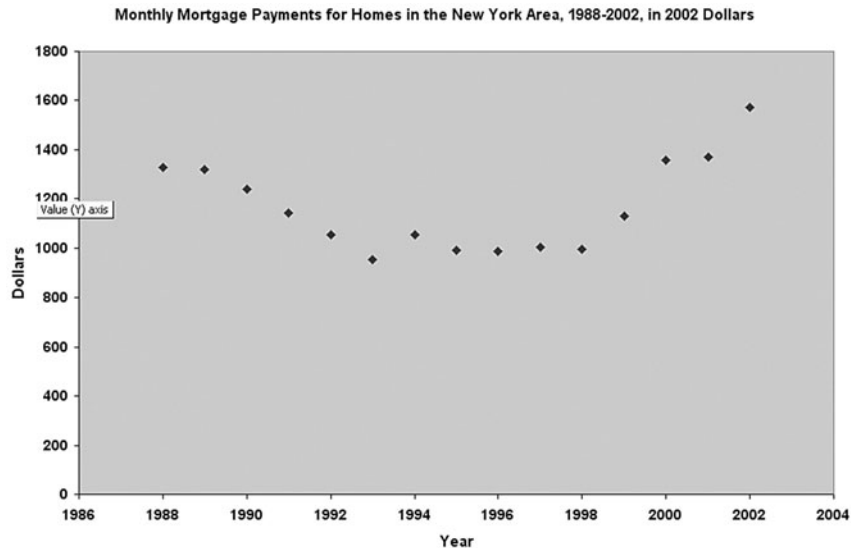
(continues)

can experiment by opening the smaller **Chapter 2 Dot Scale Practice.xls** file and typing your own data values in column A, rows 2 through 18.

The Time-Series Plot

CONCEPT A chart in which each point represents the value of a numerical variable at a specific time. By convention, the X-axis (the horizontal axis) always represents units of time, and the Y-axis (the vertical axis) always represents units of the variable.

EXAMPLE

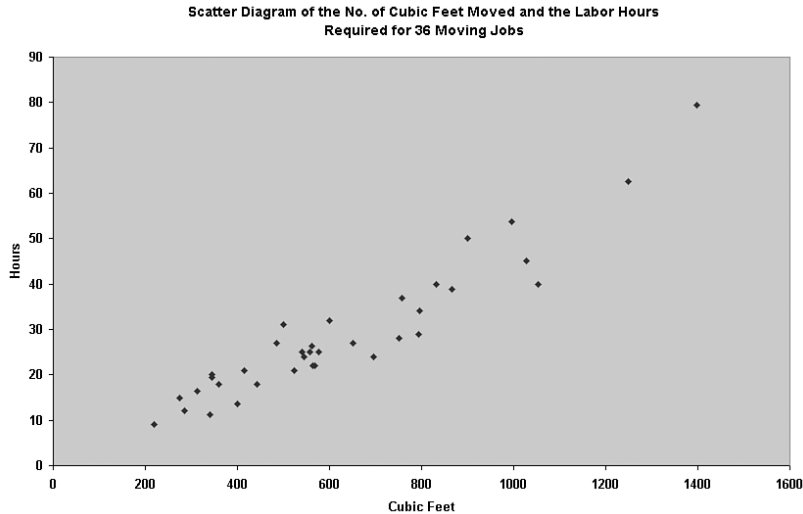


This time-series plot uses the monthly mortgage payments from a file of housing-related data for the years 1988 through 2002. (**NYHOUSING**)

INTERPRETATION Time-series plots can reveal patterns over time, patterns that can be quite hard to see when looking at a long list of numerical values. In this example, the plot reveals that monthly mortgage payments (when considered in constant dollars) dropped steadily in the late 1980s and early 1990s, only to level off and start rising again since 2000.

The Scatter Plot

CONCEPT A chart that plots the values of two numerical variables for each response. In a scatter plot, the X-axis (the horizontal axis) always represents units of one variable, and the Y-axis (the vertical axis) always represents units of the second variable.

EXAMPLE

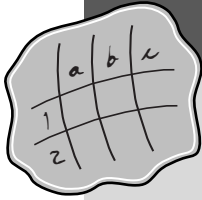
This scatter plot shows the number of cubic feet moved and the labor hours required for 36 moving jobs in which the travel and transport time is negligible. **(MOVING)**

Labor Hours and Cubic Feet Moved for 36 Moving Jobs

Cubic feet	24	13.5	26.25	25	9	20	22	11.25
Labor hours	545	400	562	540	220	344	569	340
Cubic feet	50	12	38.75	40	19.5	18	28	27
Labor hours	900	285	865	831	344	360	750	650
Cubic feet	21	15	25	45	29	21	22	16.5
Labor hours	415	275	557	1028	793	523	564	312
Cubic feet	37	32	34	25	34	25	31	24
Labor hours	757	600	796	557	796	577	500	695
Cubic feet	40	27	18	62.5	53.75	79.5		
Labor hours	1054	486	442	1249	995	1397		

INTERPRETATION Scatter plots help reveal patterns in the relationship between two numerical variables. The scatter plot for this example reveals a strong positive linear (straight line) relationship between the number of cubic feet moved and the number of labor hours required. Based on these data, a manager at the urban moving company could conclude that number of cubic feet being moved in a specific job will be a useful predictor of the

number of labor hours that will be needed. Using one numerical variable to predict the value of another is more fully discussed in Chapter 10.



spreadsheet solution

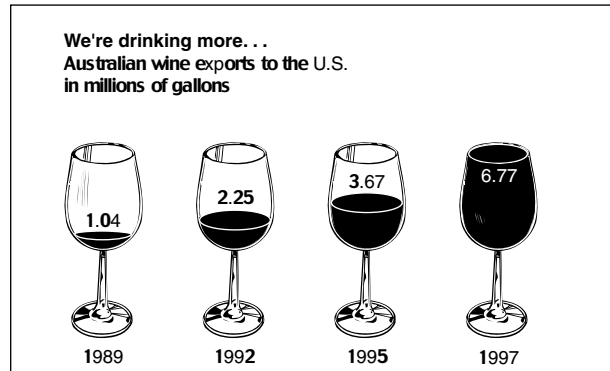
Scatter Plots

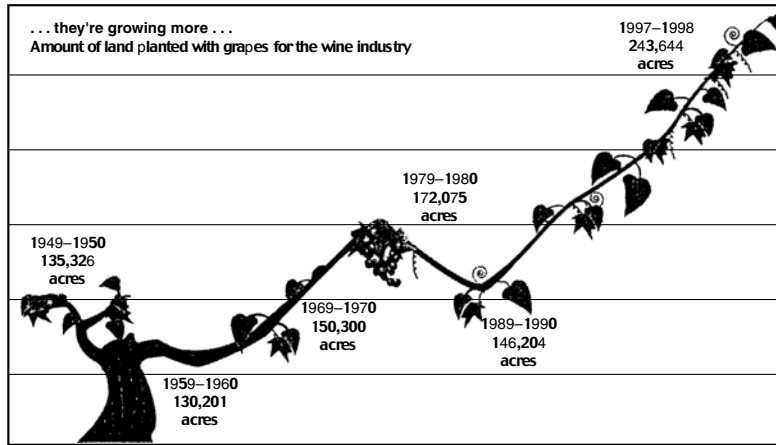
Download and open the **Chapter 2 Scatter.xls** Excel file to see an example of a scatter plot for the moving company data. You can experiment by typing your own data values in column A, rows 2 through 37. To produce custom scatter plots, you can select the XY (Scatter) *standard* chart type in the Chart Wizard Step 1 dialog box.

2.3 Misusing Graphs

Good graphical displays, such as those presented in this chapter, reveal what the data are conveying. Unfortunately, many graphs that you will see in the news media or in formal reports either are incorrect, misleading, or are so unnecessarily complicated that they never should be used.

EXAMPLE 1: Australian Wine Exports to the United States.



EXAMPLE 2: Amount of Land Planted with Grapes for the Wine Industry.

INTERPRETATION Using pictorial symbols, instead of bars or pies, always obscures the data and may create a false impression in the mind of the reader, especially if the pictorial symbols are representations of three-dimensional objects. In Example 1, the wine glass symbol fails to communicate that the 1997 data (6.77 million gallons) is almost twice the 1995 data (3.67 million gallons), nor does it accurately reflect that the 1992 data (2.25 million gallons) is a bit more than twice the 1.04 million gallons for 1989.

Example 2 combines the inaccuracy of using a picture (grape vine) instead of a standard shape with the error of having unlabeled and improperly scaled axes. A missing X-axis prevents the reader from immediately seeing that the 1997–1998 value is misplaced; by the scale of the graph, that data point should be closer to the rest of the data. A missing Y-axis prevents the reader from getting a better sense of the rate of change in land planted through the years. There are other problems as well; can you spot at least one more? (Hint: Compare the 1949–1950 data to the 1969–1970 data.)

When producing your own graphs, consider the following guidelines:

- Avoid unnecessary decorations or illustration around the borders or in the background.
- Avoid the use of fancy pictorial symbols to represent data values.
- In two-dimensional graphs, always include a scale for each axis.
- When charting non-negative values, the scale on the vertical axis should begin at zero.
- Always label every axis.
- Always supply a title.
- Always choose the simplest graph that can present your data.

One-Minute Summary

To choose an appropriate table or chart type, your starting point is always to determine whether your data are categorical or numerical.

If your data are categorical:

- Determine whether you have one or two variables to present.
- If one variable, use a summary table and/or bar chart, pie chart, or Pareto diagram.
- If two variables, use a two-way cross-classification table.

If your data are numerical:

- Determine whether you have one or two variables to present.
- If one variable, use a frequency and percentage distribution, histogram, or a dot scale diagram.
- If two variables, determine whether the time order of the data is important.
 - If yes, use a time-series plot.
 - If no, use a scatter plot.

Test Yourself

1. Which of the following graphical presentations is not appropriate for categorical data?
 - (a) Pareto diagram
 - (b) scatter plot
 - (c) bar chart
 - (d) pie chart
2. Which of the following graphical presentations is not appropriate for numerical data?
 - (a) histogram
 - (b) pie chart
 - (c) dot-scale diagram
 - (d) scatter diagram
3. A type of histogram in which the categories are plotted in the descending rank order of the magnitude of their frequencies is called a:
 - (a) bar chart
 - (b) pie chart
 - (c) scatter plot
 - (d) Pareto diagram

4. One of the advantages of a pie chart is that it shows that the total of all the categories of the pie adds to 100%.
 - (a) True
 - (b) False
5. The basic principle behind the _____ is the ability to separate the vital few categories from the trivial many categories.
 - (a) scatter plot
 - (b) dot scale diagram
 - (c) Pareto diagram
 - (d) pie chart
6. When studying the simultaneous responses to two categorical questions, you should set up a:
 - (a) histogram
 - (b) pie chart
 - (c) scatter plot
 - (d) cross-classification table
7. In a cross-classification table, the number of rows and columns:
 - (a) must always be the same
 - (b) must always be 2
 - (c) must add to 100%
 - (d) None of the above
8. Histograms are used for numerical data, whereas bar charts are suitable for categorical data.
 - (a) True
 - (b) False
9. A department store in a small town monitors customer complaints and organizes these complaints into six distinct categories. Over the past year, the company has received 534 complaints. One possible graphical method for representing these data is a Pareto diagram.
 - (a) True
 - (b) False
10. A department store in a small town monitors customer complaints and organizes these complaints into six distinct categories. Over the past year, the company has received 534 complaints. One possible graphical method for representing these data is a scatter plot.
 - (a) True
 - (b) False
11. A computer company collected information on the age of their customers. The youngest customer was 12, and the oldest was 72. To study the distribution of the age of its customers, it should use a pie chart.
 - (a) True
 - (b) False

12. A computer company collected information on the age of their customers. The youngest customer was 12, and the oldest was 72. To study the distribution of the age of its customers, it can use a histogram.
 - (a) True
 - (b) False
13. A financial services company wants to collect information on the weekly number of transactions. To study the weekly transactions, it can use a pie chart.
 - (a) True
 - (b) False
14. A financial services company wants to collect information on the weekly number of transactions. To study the weekly transactions, it can use a time-series plot.
 - (a) True
 - (b) False
15. A professor wants to study the relationship between the number of hours a student studied for an exam and the exam score achieved. The professor can use a time-series plot.
 - (a) True
 - (b) False
16. A professor wants to study the relationship between the number of hours a student studied for an exam and the exam score achieved. The professor can use a bar chart.
 - (a) True
 - (b) False
17. A professor wants to study the relationship between the number of hours a student studied for an exam and the exam score achieved. The professor can use a scatter plot.
 - (a) True
 - (b) False
18. If you wanted to compare the percentage of items that are in a particular category as compared to other categories, you should use a pie chart, not a bar chart.
 - (a) True
 - (b) False
19. To evaluate two categorical variables at the same time, a _____ should be developed.
20. A _____ is a vertical bar chart in which the rectangular bars are constructed at the boundaries of each class interval.
21. A _____ chart should be used when you are primarily concerned with the percentage of the total that is in each category.

- 22. A _____ chart should be used when you are primarily concerned with comparing the percentages in different categories.
- 23. A _____ should be used when you are studying a pattern between two numerical variables.
- 24. A _____ should be used to study the distribution of a numerical variable.
- 25. You have measured your pulse rate daily for 30 days. A _____ plot should be used to study the pulse rate.

Answers to Test Yourself Questions

- 1. b
- 2. b
- 3. d
- 4. a
- 5. c
- 6. d
- 7. d
- 8. a
- 9. a
- 10. b
- 11. b
- 12. a
- 13. b
- 14. a
- 15. b
- 16. b
- 17. a
- 18. b
- 19. cross-classification table
- 20. histogram
- 21. pie chart
- 22. bar chart
- 23. scatter plot
- 24. histogram or dot scale diagram
- 25. time-series plot

References

1. Beninger, J. M., and D. L. Robyn. 1978. "Quantitative Graphics in Statistics." *The American Statistician* 32: 1–11.
2. Berenson, M. L., D. M. Levine, and T. C. Krehbiel. *Basic Business Statistics: Concepts and Applications, Ninth Edition*. Upper Saddle River, NJ: Prentice Hall, 2004.
3. Gitlow, H. S., and D. M. Levine. *Six Sigma for Green Belts and Champions*. Upper Saddle River, NJ: Financial Times - Prentice Hall, 2005.
4. Levine, D. M., T. C. Krehbiel, and M. L. Berenson. *Business Statistics: A First Course, Third Edition*. Upper Saddle River, NJ: Prentice Hall, 2003.
5. Levine, D. M., D. Stephan, T. C. Krehbiel, and M. L. Berenson. *Statistics for Managers using Microsoft Excel, Fourth Edition*. Upper Saddle River, NJ: Prentice Hall, 2005.
6. Levine, D. M., P. P. Ramsey, and R. K. Smidt, *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Upper Saddle River, NJ: Prentice Hall, 2001.
7. Microsoft Excel 2002. Redmond, WA: Microsoft Corporation, 2001.
8. Sincich, T., D. M. Levine, and D. Stephan. *Practical Statistics by Example Using Microsoft Excel and Minitab, Second Edition*. Upper Saddle River, NJ: Prentice Hall, 2002.



Descriptive Statistics for Numerical Variables

3.1 Measures of Central Tendency

3.2 Measures of Variation

3.3 Shape of Distributions

Important Equations

One-Minute Summary

Test Yourself

Among the most important summarizing activities of descriptive statistics are those statistical methods that help measure properties of a numerical variable. Reading this chapter will allow you to learn about some of the methods used to identify the properties of central tendency, variation, and shape.

3.1 Measures of Central Tendency

Because the data values of most numerical variables show a tendency to group around a specific value, statisticians use a set of methods, collectively known as **measures of central tendency**, to help identify the general properties of that data. Three commonly used measures are the *arithmetic mean*, also known simply as the mean or average, the *median*, and the *mode*. You can calculate these measures as either sample statistics or population parameters.

The Mean

CONCEPT A number equal to the sum of the data values for a variable, divided by the number of data values that were summed.

EXAMPLES Many sports statistics (including baseball batting averages and football yards per reception), average SAT score for incoming freshmen at a college, average age of the workers in a company, average waiting times at a bank.

important point



INTERPRETATION The mean represents a “balance point” in a set of data values, similar to a fulcrum on a seesaw. As the only measure of central tendency that uses all the data values in a sample or population, the mean has one great weakness: individual extreme values can distort the balance point. Therefore, you should be wary of using this measure if the data you are trying to describe contains extreme values, as the second Worked-out Problem below illustrates.

WORKED-OUT PROBLEM 1 Although many people sometimes find themselves running late as they get ready to go to work, few measure the actual time it takes to get ready in the morning. Suppose you want to determine the typical time (in minutes) that elapses between your alarm clock’s programmed wake-up time and the time you leave your home for work. You decide to measure actual times for ten consecutive working days and record the following times:

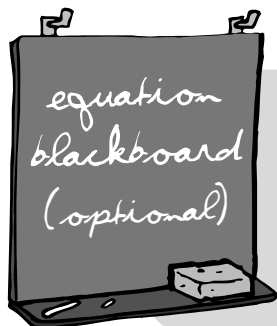
Day	1	2	3	4	5	6	7	8	9	10
Time	39	29	43	52	39	44	40	31	44	35

(Times)

To compute the mean time, first compute the sum of all the data values, $39 + 29 + 43 + 52 + 39 + 44 + 40 + 31 + 44 + 35$, which is 396. Then take this sum of 396 and divide by 10, the number of data values. The result, 39.6 minutes, is the mean time to get ready.

WORKED-OUT PROBLEM 2 Consider the same problem but imagine that on day 4 an exceptional occurrence such as oversleeping caused you to leave your home 50 minutes later than you had recorded for that day. That would make the time for day 4 102 minutes, the sum of all times, 446 minutes, and the mean (446 divided by 10), 44.6 minutes.


Note how one extreme value has dramatically changed the mean. Instead of being a number “somewhere” in the middle of the 10 get-ready times, the new mean of 44.6 minutes is greater than 9 of the 10 get-ready times. In this case, the mean fails as a measure of “central tendency.”



The worked-out problems calculate the mean of a sample of get-ready times. You need three symbols to write the equation for the mean calculation:

- an uppercase italic X with a horizontal line above it, \bar{X} , pronounced as “X bar,” that represents the number that is the mean of a sample.

interested
in
math?



- a subscripted uppercase italic X (for example, X_1) that represents one of the data values being summed. Because the problem contains 10 data values, there are 10 values, the first one labeled X_1 , the last one labeled X_{10} .
- a lowercase italic n , that represents the number of data values that were summed in this sample, a concept also known as the **sample size**. You pronounce n as “sample size” to avoid confusion with the symbol N that represents (and is pronounced as) the population size.

Using these symbols creates the following equation:

$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9 + X_{10}}{n}$$

By using an ellipsis (...), you can abbreviate the equation as:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_{10}}{n}$$

Using the insight that the value of the last subscript will always be equal to the value of n , you can generalize the formula as:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

By using the uppercase Greek letter sigma, Σ , a standard symbol that represents the summing of values, you can further simplify the formula as:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

$$\bar{X} = \frac{\Sigma X}{n} \quad \text{or more explicitly as:}$$

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

in which i represents a placeholder for a subscript and the $i = 1$ and n below and above the sigma represent the range of the subscripts used.

The Median

CONCEPT The middle value in a set of data values for a variable when the data values have been ordered from lowest to highest value. When the number of data values to be summarized is even, you perform a special calculation to determine the median (see Interpretation on page 40) because data sets with an even number of values have no natural middle value.

EXAMPLES Many economic statistics such as median household income for a region; many marketing statistics such as the median age for buying a consumer product; in education, the established middle point for many standardized tests.

INTERPRETATION The median splits the set of ranked data values into equal-in-numbers parts. Extreme values do not affect the median, making the median a good alternative to the mean when such values occur.

When the number of data values to be summarized is even, the median is calculated by taking the mean of the two values closest to the middle, when all values are ranked from lowest to highest. For example, if there were 6 ranked values, you would calculate the mean of the third and fourth values; and if there were 10 ranked values, you would calculate the mean of the fifth and sixth values.

When you are calculating the median for a very large number of values, you may not be able to quickly identify the middle value (when the number of data values is odd) or the middle two values (when the number of data values is even). To quickly determine the middle position, add 1 to the number of data values and divide that sum by 2. For example, for 127 values, divide 128 by 2 to get 64 and determine that the median is the 64th ranked value. For 70 values, divide 71 by 2 to get 35.5 and calculate the mean of the 35th and 36th ranked values—the two values closest the middle—to determine the median.

important point



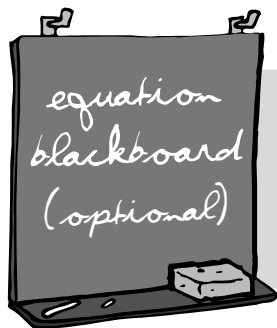
WORKED-OUT PROBLEM 1 You are asked to calculate the median age of a group of employees whose individual ages are 47, 23, 34, 22, and 27. You calculate the median by first ranking the values from lowest to highest: 22, 23, 27, 34, and 47. Because there are 5 values, the natural middle is the third ranked value, 27, making the median 27. This means that half of the workers are 27 years old or less and half the workers are 27 years old or more.

WORKED-OUT PROBLEM 2 You are asked to calculate the median for the original set of 10 get-ready times from Worked-out Problem 1 on page 38 (that determined the mean). As an ordered list lowest to highest, those values (shown with their ordered position) are as follows.

Time	29	31	35	39	39	40	43	44	44	52
Ordered Position	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th

Because there is an even number of data values, 10, you need to calculate the mean of the two values closest to the middle—that is, the fifth and sixth val-

ues, 39 and 40. The mean of 39 and 40 is 39.5, making the median for the set of 10 times 39.5 minutes.



Using the n symbol previously defined on page 39, you can define the median as:

$$\text{Median} = \frac{n+1}{2} \text{th ranked value}$$

interested
in
math?

The Mode

CONCEPT The value (or values) in a set of data values for a variable that appears most frequently.

EXAMPLES The most common score on an exam, the most likely income, the commuting time that occurs most often.

INTERPRETATION Similar to the median, extreme values do not affect the mode; unlike the median, however, the mode can vary much more from sample to sample than the median (or mean).

Some sets of data values have no mode—all the unique values appear the same number of times. Other sets of data values can have more than one mode, such as the get-ready times on page 38 in which two modes occur, 39 minutes and 44 minutes, because each of these values appears twice and all other values appear once.

Quartiles

CONCEPT The three values that split a set of ranked data values for a variable into four equal parts—quarters, or quartiles. The **first quartile**, Q_1 , is the value such that 25.0% of the ranked data values are smaller and 75.0% are larger. The **second quartile**, Q_2 , is another name for the **median**, which, as discussed on page 40, splits the ranked values into two equal parts. The **third quartile**, Q_3 , is the value such that 75.0% of the ranked values are smaller and 25.0% are larger.

EXAMPLE Many math and reading standardized tests for children report results in terms of quartiles.

INTERPRETATION Quartiles help summarize large sets of data values by allowing you to identify the 25th, 50th, and 75th percentiles. If you scored in

the third quartile on a standardized test, your score was in the top 25% of all scores. If your score *was equal to* the third quartile, the 75th percentile, then 25% of all scores were higher and 75% were lower. If you did exceptionally well, and learned that that your score was reported as the 99th percentile, you would know that your score was in the top 1% of all scores (and therefore greater than 99% of all scores).

To quickly determine the first quartile, add 1 to the number of data values and divide that sum by 4. For example, for 11 values, add 1 to 11 to get 12 and divide 12 by 4 to get 3 and determine that the first quartile is the third ranked value. To quickly determine the third quartile, add 1 to the number of data values and divide that sum by 4 and then multiply the quotient by 3. For the same example, you would multiply the quotient 3 by 3 to get 9 and determine that the third quartile is the ninth ranked value). (Use the instructions for calculating the median to determine the second quartile.)

When the result of this arithmetic is not a whole number, select the ranked positions immediately below and above the number calculated. For example, for 10 values, the result (for the first quartile) would be 2.75 ($10 + 1$ is 11, $11/4$ is 2.75), and you would select the second and third ranked values. With these values, do the following:

1. Multiply the larger ranked value by the decimal fraction of the original result (0.75 in the example.)
2. Multiply the smaller ranked value by 1 minus the decimal fraction of the original result (0.25 for the example, because $1 - 0.75$ is 0.25).
3. Add the two products to determine the quartile value.

Special case: Should the two ranked values selected be the same number, then the quartile is that number and you can skip the previous two multiplications and one addition.

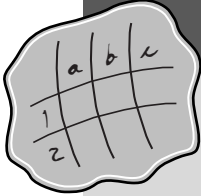
WORKED-OUT PROBLEM 1 You are asked to determine the first quartile for the ranked get-ready times first shown on page 40 and reproduced here.

Time	29	31	35	39	39	40	43	44	44	52
Ranked Value	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th

You first add 1 to 10, the number of values, and divide by 4 to get 2.75 to identify the second and third ranked values, 31 and 35. You multiply 35, the larger value, by the decimal fraction 0.75 to get 26.25. You multiply 31, the smaller value, by the decimal fraction 0.25 to get 7.75, and then add 26.25 and 7.75 to produce 34, the first quartile value, indicating that 25% of the get-ready times are 34 minutes or below and that the other 75% are 34 minutes or above.

WORKED-OUT PROBLEM 2 You are asked to determine the third quartile for the ranked get-ready times. Add 1 to 10 to get 11, divide by 4 to get 2.75, and multiply by 3 to get 8.25 to identify the 8th and 9th ordered values, 44

and 44. By the special case described on page 42, the third quartile is 44. Had the 9th value been 48, you would have multiplied 48 by 0.25 to get 12 and multiplied 44 by 0.75 to get 33 and then added 12 and 33 to get 45, the third quartile value.



spreadsheet solution

Descriptive Statistics

Download and open the **Chapter 3 Descriptive.xls** Excel file to see examples of formulas that use the Average, Median, and Mode functions that calculate the mean, median, and mode of a set of ranked values. To produce a table of descriptive statistics about a variable, similar to the one shown below, use the Data Analysis Descriptive Statistics procedure with the Summary Statistics output option.

	A	B
1	<i>Get-Ready Time</i>	
2		
3	Mean	39.6
4	Standard Error	2.140612581
5	Median	39.5
6	Mode	39
7	Standard Deviation	6.769211344
8	Sample Variance	45.82222222
9	Kurtosis	0.137510253
10	Skewness	0.085756573
11	Range	23
12	Minimum	29
13	Maximum	52
14	Sum	396
15	Count	10

Some of the additional descriptive statistics included in this worksheet are discussed in Sections 3.2 and 3.3.

WORKED-OUT PROBLEM 3 You want to undertake a study that compares the cost for a restaurant meal in a major city to the cost of a similar meal in the suburbs outside the city. You collect data about the cost of a meal per

person from a sample of 50 city restaurants and 50 suburban restaurants and arrange the 100 numbers in two ranked sets as follows:

City Cost Data

14	22	23	25	26	27	30	31	31	32	
33	34	34	35	35	35	36	36	37	37	
38	38	38	39	39	39	39	40	41	42	
43	44	44	44	44	45	45	48	48	49	
50	50	50	50	51	51	53	53	56	63	(City)

Suburban Cost Data

23	23	24	24	25	25	26	26	26	26	
27	27	28	28	29	29	29	30	30	30	
30	31	31	32	32	32	33	33	34	34	
36	37	37	37	38	38	38	38	38	38	
39	39	41	43	44	44	48	51	51	55	(Suburban)

Due to the many data values involved, you decide to use Microsoft Excel to calculate the mean and median of the two groups of cost data. You enter the city cost data into column A of a blank worksheet and enter the suburban cost data into column B of the same worksheet. You use the Data Analysis Descriptive Statistics procedure, specifying the input range as A1:B50, and generate these results:

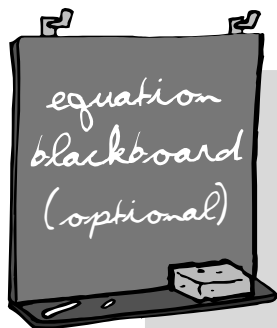
	A	B	C	D
1	City		Suburban	
2				
3	Mean	39.74	Mean	33.74
4	Standard Error	1.365107456	Standard Error	1.091641714
5	Median	39	Median	32
6	Mode	39	Mode	38
7	Standard Deviation	9.652767394	Standard Deviation	7.719072589
8	Sample Variance	93.17591837	Sample Variance	59.58408163
9	Kurtosis	0.199787977	Kurtosis	0.307227068
10	Skewness	-0.184147951	Skewness	0.827684018
11	Range	49	Range	32
12	Minimum	14	Minimum	23
13	Maximum	63	Maximum	55
14	Sum	1987	Sum	1687
15	Count	50	Count	50

From the results, you note the following:

- The mean cost of city meals, \$39.74, is higher than the mean cost of suburban meals, \$33.74.

- The median cost of a city meal, \$39, is also higher than the median suburban cost, \$32.
- The first and third quartiles for city meals (\$34 and \$48) are also higher than their suburban counterparts (\$28 and \$38).

From the mean, median, and quartiles, you can conclude that the cost of a restaurant meal per person seems higher for restaurants in the city than in the suburbs of that city.



interested
in
math?

Using the equation for the median developed earlier,

$$\text{Median} = \frac{n+1}{2} \text{ ranked value}$$

you can express the first quartile Q_1 as $Q_1 = \frac{n+1}{4} \text{th}$ ranked value

and the third quartile Q_3 , as

$$Q_3 = \frac{3(n+1)}{4} \text{th ranked value.}$$

3.2 Measures of Variation

A second important property that describes a set of numerical data is variation. **Variation** is the amount of **dispersion**, or “spread,” in the data. Four frequently used measures of variation are the *range*, the *variance*, the *standard deviation*, and the *Z score*, all of which can be calculated as either sample statistics or population parameters.

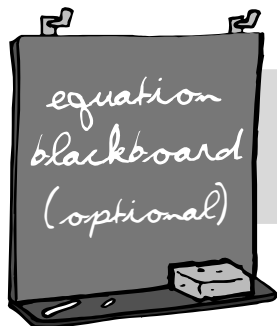
The Range

CONCEPT The difference between the largest and smallest data values in a set of values for a variable.

EXAMPLES In most everyday examples, the largest and smallest values are presented and the number that represents their difference is not shown: daily high and low temperatures, stock market 52-week high and low closing prices, best and worst times for timed sporting events.

INTERPRETATION The range is the number that represents the largest possible difference between any two values in a set of data values for a variable.


WORKED-OUT PROBLEM For the get-ready times data first presented on page 38, the range is 23 minutes ($52 - 29$). For the restaurant meal study, the ranges have already been calculated in the Microsoft Excel results shown on page 44. For the city meal cost data, the range is \$49, and the range for the suburban meal cost data is \$32. You can conclude that meal costs in the city show much more variation than suburban meal costs.



For a set of data values, the range is equal to:

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

interested
in
math?



The Variance and the Standard Deviation

CONCEPT Two related numbers that each individually measures how a set of data values for a variable fluctuate around the mean of that variable. The numbers are related because one of them, the standard deviation, is the positive square root of the other (the variance).

EXAMPLE The variance among SAT scores for incoming freshmen at a college, the standard deviation in the age of the workers in a company. The variance and standard deviation always appear as “variance” and “standard deviation” and should be accompanied by the mean.

INTERPRETATION The variance and standard deviation help you to know how a set of data values distributes around its mean. For almost all sets of data values, the majority of the values lie within an interval of plus and minus one standard deviation above and below the mean. Therefore, determining the mean and the standard deviation usually helps you define the range in which the majority of the data values occur.

The simplest measure of variation might take the difference between each value and the mean and sum these differences. However, by the properties of arithmetic and the definition of mean, the result of such calculations would be zero for *every* set of data values—not very helpful in comparing one set to another!

Instead, statisticians developed a method in which the difference between each data value and the mean is squared and the squared numbers are summed. This sum of squares (or SS) is then divided by either one less than the number of data values, for sample data, or the number of data values, for population data, to produce the variance. By definition, the standard

deviation is then the positive square root of the variance. Because calculating the variance includes squaring the difference between each value and the mean, a step that always produces a non-negative number, the variance itself can never be negative.

WORKED-OUT PROBLEM 1 You seek to calculate the variance and standard deviation for the get-ready times first presented on page 38. As first steps, you calculate the difference between each of the 10 individual times and the mean (39.6 minutes), square those differences, and sum the squares. (Table 3.1 shows these first steps.)

TABLE 3.1

First Steps Toward Calculating the Variance and Standard Deviation for the Get-Ready Times Data

Day	Time	Difference: Time Minus Mean (39.6)	Square of Difference
1	39	-0.6	0.36
2	29	-10.6	112.36
3	43	3.4	11.56
4	52	12.4	153.76
5	39	-0.6	0.36
6	44	4.4	19.36
7	40	0.4	0.16
8	31	-8.6	73.96
9	44	4.4	19.36
10	35	-4.6	<u>21.16</u>
Sum of squares:			412.40

Because these data are a sample of get-ready times, the sum of squares, 412.40, is divided by one less than the number of data values, 9, to get 45.82, the sample variance. In turn, the square root of 45.82 (6.77, after rounding) is the sample standard deviation. You can then reasonably conclude that most get-ready times are between 32.83 ($39.6 - 6.77$) minutes and 46.37 ($39.6 + 6.77$) minutes, a statement that inspection of the data values confirms.

WORKED-OUT PROBLEM 2 You seek to determine the standard deviation for the restaurant meal study. These values have already been calculated in the Microsoft Excel results shown on page 44. For city meal costs, the standard deviation is \$9.65, and you determine that the majority of meals will

cost between \$30.09 and \$49.39 (the mean $\$39.74 \pm \9.65). For suburban meal costs, the standard deviation is \$7.72, and you determine that the majority of those meals will cost between \$26.02 and \$41.46 (the mean $\$33.74 \pm \7.72).



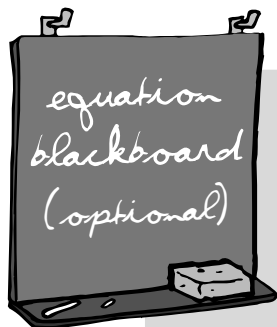
calculator keys

Mean, Median, Standard Deviation, Variance

To calculate descriptive statistics for a set of data values previously entered as the values of a variable, press [2nd] [STAT] [►] [►] (to display the Math menu) and select the appropriate statistic and press [ENTER]. Then enter the name of the variable and press [ENTER] to calculate the statistic.

For example, to calculate the mean for data entered as the values as variable L1, you would select 3:mean(, press [ENTER], press [2nd] [1] (to type the variable name L1), and then press [ENTER]. The value of the mean will appear on a new line, and your display will be similar to this:

```
mean(L1
39.6
```



interested
in
math?

Using symbols first introduced earlier in this chapter, you can express the sample variance and the sample standard deviation as:

$$\text{Sample variance} = s^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}$$

$$\text{Sample standard deviation} = S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}}$$

To calculate the variance and standard deviation for population data, change the divisor from one less than the number of

data values in the sample (the sample size) to the number of data values in the population, a value known as the **population size** and represented by an italicized uppercase N .

$$\text{Population variance} = \sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

$$\text{Population standard deviation} = \sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

Note that the lowercase Greek letter sigma represents the population standard deviation, replacing the uppercase italicized S . This follows a convention that symbols for population parameters are always Greek letters and explains the appearance of the lowercase Greek letter mu, μ , that represents the *population* mean, instead of the sample mean, \bar{X} .

Standard (Z) Scores

CONCEPT The number that is the difference between a data value for a variable and the mean of the variable, divided by the standard deviation, is its Z score.

EXAMPLE The Z score for a particular incoming freshman's SAT score, the Z score for the day 4 get-ready time.

INTERPRETATION Z scores help you determine whether a data value is an extreme value, or *outlier*—that is, far from the mean. As a general rule, a Z score that is less than -3 or greater than $+3$ indicates that the data value it represents is an extreme value.

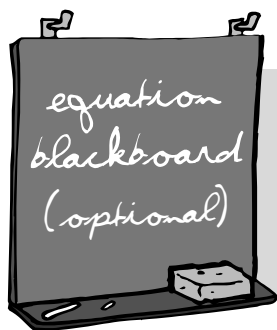
WORKED-OUT PROBLEM You want to determine whether any of the time values from the set of 10 get-ready times (see page 38) could be considered outliers. You calculate Z scores for each of those times and compare (see Table 3.2). From the table results you learn that the greatest positive Z score was 1.83 (for the day 4 value) and the greatest negative Z score was -1.27 (for the day 8 value). Because no Z score is less than -3 or greater than $+3$, you conclude that none of the get-ready times can be considered extreme.

TABLE 3.2

Table of Z Score Calculations for the Get-Ready Times Sample

Day	Time	Time Minus Mean	Z Score
1	39	-0.6	-0.09
2	29	-10.6	-1.57

Day	Time	Time Minus Mean	Z Score
3	43	3.4	0.50
4	52	12.4	1.83
5	39	-0.6	-0.09
6	44	4.4	0.65
7	40	0.4	0.06
8	31	-8.6	-1.27
9	44	4.4	0.65
10	35	-4.6	-0.68



Using symbols presented earlier in this chapter, you can express the Z score as follows:

$$Z \text{ score} = Z = \frac{X - \bar{X}}{S}$$

Shape of Distributions

A third important property that describes a set of numerical data is **shape**. Shape describes the pattern of the distribution of data values through the range of the data values. The shape may be either *symmetrical*, *left-skewed*, or *right-skewed*. Later in this book, you will learn that determining shape often has a second purpose—some statistical methods are invalid if the set of data values are too badly skewed.

Symmetrical Shape

CONCEPT A set of data values in which the mean equals the median value.

EXAMPLE Scores on a standardized exam, actual amount of soft drink in a one-liter bottle.

Left-Skewed Shape

CONCEPT A set of data values in which the mean is less than the median value. Also known as negative skew.

EXAMPLE Scores on an exam in which most students score between 80 and 100, whereas a few students score between 10 and 79.

Right-Skewed Shape

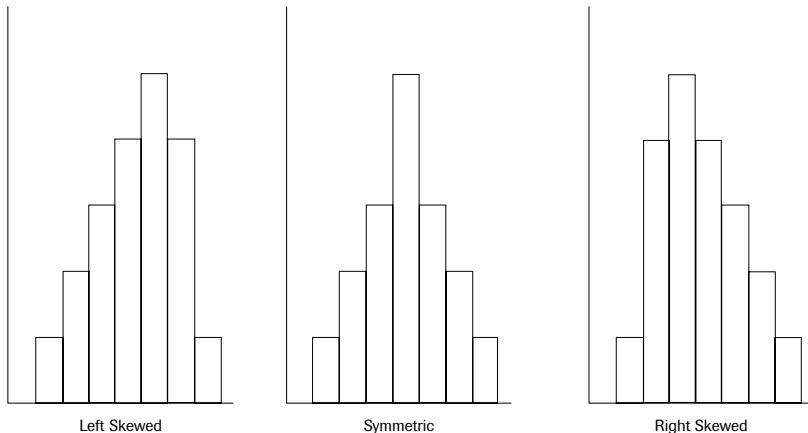
CONCEPT A set of data values in which the mean is greater than the median value. Also known as positive skew.

EXAMPLE Prices of new homes, annual family income.

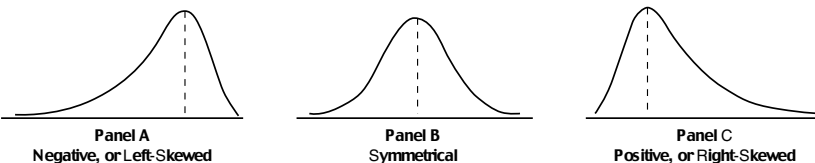
INTERPRETATION Right or positive skewness occurs when the data set contains some extremely high data values (that increase the mean). Negative skewness occurs when there are some extremely low values that decrease the mean. The set of data values for a variable are symmetrical when low and high values balance each other out.

When identifying shape, you should avoid the common pitfall of thinking that the side of the histogram in which the most data values cluster closely together is the skew “direction.” For example, the first histogram below shows that clustering appears toward the right of the histogram, but the pattern is properly labeled *left-skewed*. To see the shape more clearly, statisticians create **area-under-the-curve** or **distribution graphs**, in which a plotted, curved line represents the tops of all the bars. The second set of illustrations contains the distribution graphs that are equivalent to the histograms. If you remember that in such graphs the tail points to the skewness, you will never inadvertently confuse the direction of the skew.

Histograms of different distributions of data values



Distribution graphs equivalent to the histograms



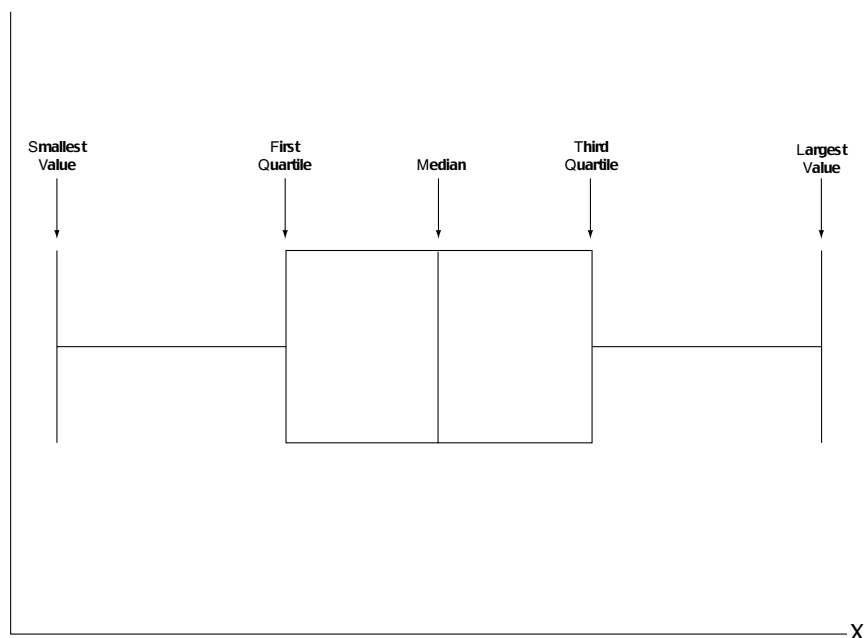
WORKED-OUT PROBLEM You want to identify the shape of the chemical viscosity data first presented on page 25. You examine the histogram and dot scale diagrams of these data (see pages 26 and 27) and determine that the distributions appear to be approximately symmetric because the low and high values approximately balance each other out.

A skewness statistic can also be calculated, a topic beyond the scope of this book. In an earlier worked-out problem, the skewness for the get-ready times data has been calculated as 0.086 by Microsoft Excel. Because a skewness statistic of zero means a perfectly symmetrical shape, you conclude that the distribution of get-ready times around the mean is also approximately symmetric.

The Box-and-Whisker Plot

CONCEPT For a set of data values for a variable, the five numbers that correspond to the smallest value, the first quartile Q_1 , the median, the third quartile Q_3 , and the largest value.

INTERPRETATION The five-number summary concisely summarizes the shape of a set of data values for a variable. This method determines the degree of symmetry (or skewness) based on the distances that separate the five numbers. To compare these distances effectively, you can create a **box-and-whisker plot** as shown below in which the five numbers are plotted as vertical lines, interconnected so as, with some imagination, to form a “box” from which a pair of cat whiskers sprout.



A box-and-whisker plot shows a **symmetric shape** for a set of data values if both of the following relationships are present in the plot:

- The distance from the line that represents the smallest value to the line that represents the median equals the distance from the line that represents the median to the line that represents the largest value.
- The distance from the line that represents the smallest value to the line that represents the first quartile equals the distance from the line that represents the third quartile to the line that represents the largest value.

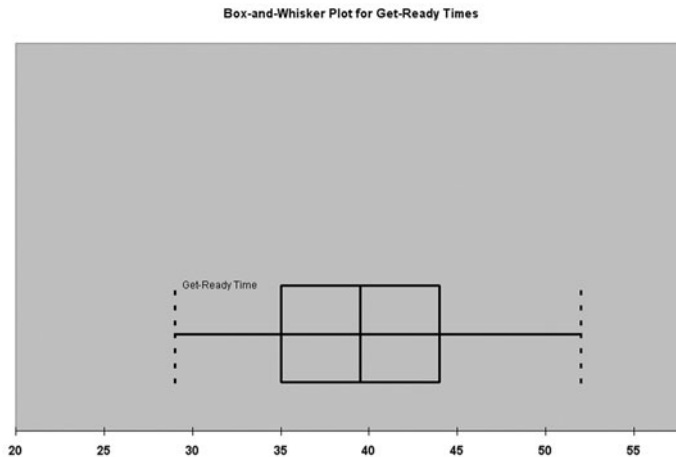
A box-and-whisker plot shows a **right-skewed shape** for a set of data values if the following relationships are both present in the plot:

- The distance from the line that represents the median to the line that represents the largest value is greater than the distance from the line that represents the smallest value to the line that represents the median.
- The distance from the line that represents the third quartile to the line that represents the largest value is greater than the distance from the line that represents the smallest value to the line that represents the first quartile.

A box-and-whisker plot shows a **left-skewed shape** for a set of data values if the following relationships are both present in the plot:

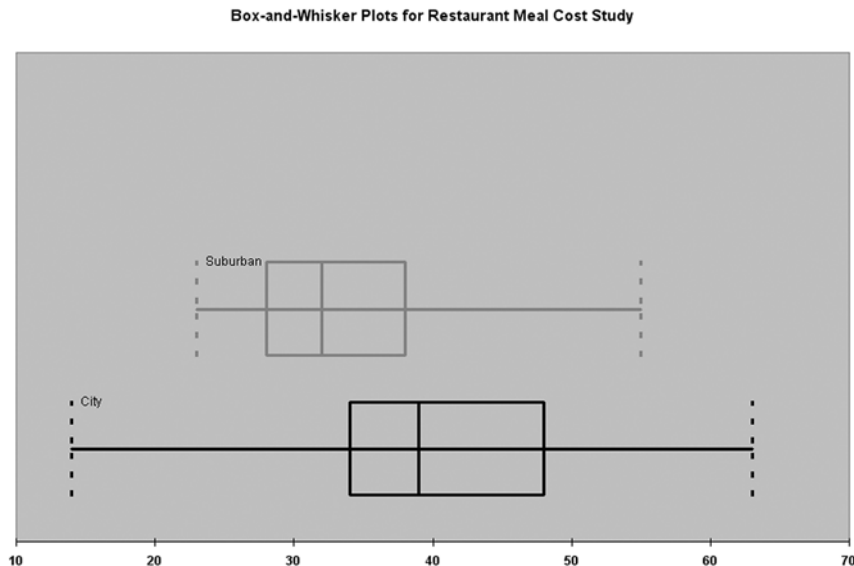
- The distance from the line that represents the smallest value to the line that represents the median is greater than the distance from the line that represents the median to the line that represents the largest value.
- The distance from the line that represents the smallest value to the line that represents the first quartile is greater than the distance from the line that represents the third quartile to the line that represents the largest value.

WORKED-OUT PROBLEM 1 The figure below represents the Microsoft Excel box-and-whisker plot of the times to get ready in the morning:



The box-and-whisker plot seems to indicate an approximately symmetric distribution of the time to get ready. The line that represents the median in the middle of the box is approximately equidistant between the ends of the box, and the length of the whiskers does not appear to be very different.

WORKED-OUT PROBLEM 2 You seek to better understand the shape of the restaurant meal cost study data used in an earlier worked-out problem. You produce box-and-whisker plots for the meal cost of both the city and suburban groups.



In examining the box-and-whisker plot for the city meal costs, you discover the following:

- The distance from the line that represents the smallest value (\$14) to the line that represents the median (\$39) is approximately the same as the distance from the line that represents the median to the line that represents the highest value (\$63).
- The distance from the line that represents the smallest value to the line that represents the first quartile (\$34) is more than the distance from the line that represents the third quartile (\$48) to the line that represents the highest value line.

You conclude that the city group of restaurant meal costs is slightly left-skewed.

In examining the box-and-whisker plot for the suburban meal costs, you discover the following:

- The distance from the line that represents the smallest value (\$23) to the line that represents the median (\$32) is much less than the distance

from the line that represents the median to the line that represents the highest value (\$55).

- The distance from the line that represents the smallest value to the line that represents the first quartile (\$28) is much less than the distance from the line that represents the third quartile (\$38) to the highest value line.

You conclude that the suburban group of restaurant meal costs is right-skewed.

In comparing the city and suburban meal cost, you conclude that the city cost is higher than the suburban cost, because the first quartile, median, and third quartile are substantially higher for the city restaurants (as is the maximum cost).



calculator keys

Box-and-Whisker Plots

To display a box-and-whisker plot for a set of data values previously entered as the values of a variable, press [2nd] [Y=] to display the Stat Plot menu and select 1:Plot1 and press [ENTER]. In the Plot1 screen, select On and press [ENTER], select the fifth type choice (a thumbnail box-and-whisker plot), and press [ENTER] and enter the variable name as the Xlist value. (Keep Freq as 1.) Press [GRAPH]. If you do not see your plot, press [ZOOM] and select 9:ZoomStat and press [ENTER] to re-center your graph on the plot. If you are plotting values in variable L1, your screen will look like this just before pressing [GRAPH]:

```

Plot1 Plot2 Plot3
On Off
Type: [L1] [L2] [L3] [L4] [L5] [L6] [L7] [L8] [L9] [L10]
Xlist:L1
Freq:1

```

Important Equations

Mean: (3.1) $\bar{X} = \frac{\sum X_i}{n}$

Median: (3.2) Median = $\frac{n+1}{2}$ th ranked value

First quartile Q_1 : (3.3) $Q_1 = \frac{n+1}{4}$ th ranked value

Third quartile Q_3 : (3.4) $Q_3 = \frac{3(n+1)}{4}$ th ordered value

Range: (3.5) Range = Largest value – Smallest value

Sample variance: (3.6) $s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$

Sample standard deviation: (3.7) $S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$

Population variance: (3.8) $\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$

Population standard deviation: (3.9) $\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$

Z score: (3.10) $Z = \frac{X - \bar{X}}{S}$

One-Minute Summary

The properties of central tendency, variation, and shape allow you to describe a set of data values for a numerical variable.

Numerical Descriptive Measures

- Central tendency
 - Mean
 - Median
 - Mode

- Variation
 - Range
 - Variance
 - Standard deviation
 - Z scores
- Shape
 - Five-number summary
 - Box-and-whisker plot

Test Yourself

1. Which of the following statistics are measures of central tendency?
 - (a) median
 - (b) range
 - (c) standard deviation
 - (d) all of these
 - (e) none of these
2. Which of the following statistics is *not* a measure of central tendency?
 - (a) mean
 - (b) median
 - (c) mode
 - (d) range
3. Which of the following statements about the median is not true?
 - (a) It is less affected by extreme values than the mean.
 - (b) It is a measure of central tendency.
 - (c) It is equal to the range.
 - (d) It is equal to the mode in bell-shaped “normal” distributions.
4. Which of the following statements about the mean is *not* true?
 - (a) It is more affected by extreme values than the median.
 - (b) It is a measure of central tendency.
 - (c) It is equal to the median in skewed distributions.
 - (d) It is equal to the median in symmetric distributions.
5. Which of the following measures of variability is dependent on every value in a set of data?
 - (a) range
 - (b) standard deviation
 - (c) each of these
 - (d) neither of these

6. Which of the following statistics *cannot* be determined from a box and whisker plot?
 - (a) standard deviation
 - (b) median
 - (c) range
 - (d) the first quartile
7. In a symmetric distribution:
 - (a) the median equals the mean
 - (b) the mean is less than the median
 - (c) the mean is greater than the median
 - (d) the median is less than the mode
8. The shape of a distribution is given by the:
 - (a) mean
 - (b) first quartile
 - (c) skewness
 - (d) variance
9. In a five-number summary, the following is not included:
 - (a) median
 - (b) third quartile
 - (c) mean
 - (d) minimum (smallest) value
10. In a right-skewed distribution:
 - (a) the median equals the mean
 - (b) the mean is less than the median
 - (c) the mean is greater than the median
 - (d) the median equals the mode
11. In a box-and-whisker plot, the box portion represents the data between the first and third quartile values.
 - (a) True
 - (b) False
12. The line drawn within the box of the box-and-whisker plot represents the mean.
 - (a) True
 - (b) False
13. The _____ is found as the middle value in a set of values placed in order from lowest to highest for an odd-sized sample of numerical data.
14. The standard deviation is a measure of _____.
15. If all the values in a data set are the same, the standard deviation will be _____.

16. A distribution that is negative-skewed is also called _____-skewed.
17. If each half of a distribution is a mirror image of the other half of the distribution, the distribution is called _____.
18. The median is a measure of _____.
- 19, 20, 21. The three characteristics that describe a set of numerical data are _____, _____, and _____.

For Questions 22 through 30, the number of days absent by a sample of 9 students during a semester was as follows:

9 1 1 10 7 11 5 8 2

22. The mean is equal to _____.
23. The median is equal to _____.
24. The mode is equal to _____.
25. The first quartile is equal to _____.
26. The third quartile is equal to _____.
27. The range is equal to _____.
28. The variance is approximately equal to _____.
29. The standard deviation is approximately equal to _____.
30. The data are:
 - (a) right-skewed
 - (b) left-skewed
 - (c) symmetrical

Answers to Test Yourself Questions

1. a
2. d
3. c
4. c
5. b
6. a
7. a
8. c
9. c
10. c
11. a
12. b
13. median

14. variation
15. 0
16. left
17. symmetric
18. central tendency
19. central tendency
20. variation
21. shape
22. 6
23. 7
24. 1
25. 1.5
26. 9.5
27. 10
28. 15.25
29. 3.91
30. b

References

1. Berenson, M. L., D. M. Levine, and T. C. Krehbiel. *Basic Business Statistics: Concepts and Applications, Ninth Edition*. Upper Saddle River, NJ: Prentice Hall, 2004.
2. Gitlow, H. S., and D. M. Levine. *Six Sigma for Green Belts and Champions*. Upper Saddle River, NJ: Financial Times - Prentice Hall, 2005.
3. Levine, D. M., T. C. Krehbiel, and M. L. Berenson. *Business Statistics: A First Course, Third Edition*. Upper Saddle River, NJ: Prentice Hall, 2003.
4. Levine, D. M., D. Stephan, T. C. Krehbiel, and M. L. Berenson. *Statistics for Managers Using Microsoft Excel, Fourth Edition*. Upper Saddle River, NJ: Prentice Hall, 2005.
5. Levine, D. M., P. P. Ramsey, and R. K. Smidt. *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Upper Saddle River, NJ: Prentice Hall, 2001.
6. Microsoft Excel 2002. Redmond, WA: Microsoft Corporation, 2001.
7. Sincich, T., D. M. Levine, and D. Stephan. *Practical Statistics by Example Using Microsoft Excel and Minitab, Second Edition*. Upper Saddle River, NJ: Prentice Hall, 2002.



Probability

4.1 Getting Started with Probability

4.2 Some Rules of Probability

4.3 Assigning Probabilities

One-Minute Summary

Test Yourself

You cannot properly learn the methods of inferential statistics without first knowing the basics of probability. If you are unfamiliar with probability, be sure to read this chapter closely before proceeding to Chapter 5. If you are already familiar with probability, you may want to skim this chapter and review the definitions of the probability concepts used in subsequent chapters.

4.1 Getting Started with Probability

Five basic concepts form the basis for understanding probability. You need to learn these concepts before you can learn about the role that probability and probability distributions play in supporting inferential statistics.

Event

CONCEPT An outcome of an experiment or survey.

EXAMPLES Rolling a die and turning up six dots, an individual who votes for the incumbent candidate in an election.

INTERPRETATION Recall from Chapter 1 (see page 6) that performing experiments or conducting surveys are two important types of data sources. When discussing probability, many statisticians use the word *experiment* broadly to include surveys, so you can use the shorter definition “an outcome of an experiment” if you understand this broader usage of *experiment*. Likewise, as you read this chapter and encounter the word *experiment*, you should use the broader meaning.

Elementary Event

CONCEPT An outcome that satisfies only one criterion.

EXAMPLES A red card from a regular deck of cards, a voter who selected the Republican candidate.

INTERPRETATION Elementary events are distinguished from *joint events*, which meet two or more criteria such as a card that is a red ace or a voter who selected the Republican candidate for president and the Democrat candidate for U.S. senator.

Random Variable

CONCEPT A variable whose numerical values represent the events of an experiment.

EXAMPLES Throwing a die, asking voters for their preferred candidate.

INTERPRETATION You use the phrase **random variable** to discuss a variable that has no data values until an experimental trial is performed or a survey question is asked and answered. This usage allows you to distinguish a random variable from the use of variable in the previous chapter in which the data values were already known before various descriptive methods were applied.

Random variables are either **discrete**, in which the possible numerical values are a set of integers; or **continuous**, in which the possible values are any number within a specific range.

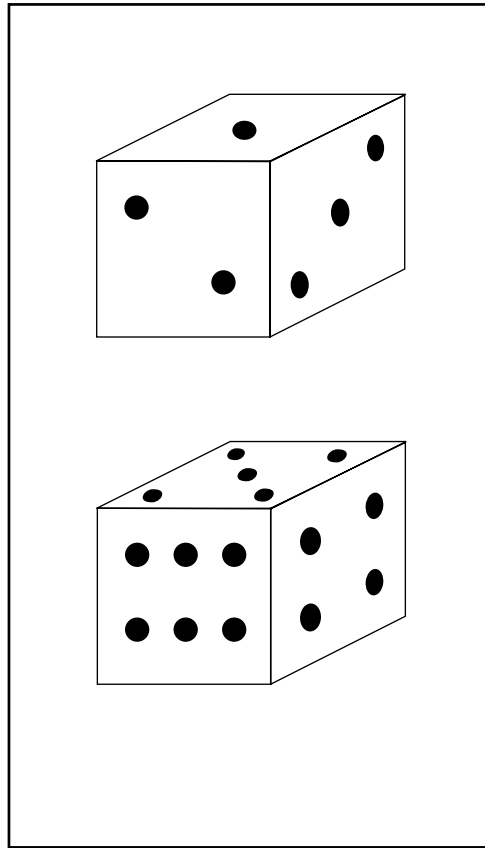
Probability

CONCEPT A number that represents the likelihood that a particular event will occur for a random variable.

EXAMPLES Odds of winning a random drawing, chance of rolling a seven when rolling two dice, likelihood of an incumbent winning reelection, percent chance of rain in a forecast.

INTERPRETATION Probability determines the likelihood that a random variable will be assigned a specific value. Probability considers things that may occur in the future, and its forward-looking nature provides a bridge to inferential statistics.

Probabilities can be developed for an elementary event of a random variable or any group of joint events. For example, when rolling a standard six-sided die (see illustration below), there are six possible elementary events that correspond to the six faces of the die that contain either one, two, three, four, five, or six dots. “Rolling a die and turning up an even number of dots” would be an example of an event formed from three elementary events (rolling a two, four, or six).



*important
point*

Probabilities are formally stated as decimal numbers in the range of 0 to 1. A probability of 0 indicates an event that never occurs (such an event is known as a **null event**). A probability of 1 indicates a **certain event**, an event that must occur. For example, when you roll a die, getting seven dots is a null event, because it can never happen, and getting six or fewer dots is a certain event, because you will always end up with a face that has six or fewer dots.

Probabilities can also be stated informally as the “percentage chance of (something)” or as quoted odds, such as a “50-50 chance.”

Collectively Exhaustive Events

CONCEPT A set of events that includes all the possible events.

EXAMPLES Head and tails in the toss of a coin, male and female, all six faces of a die.

INTERPRETATION When you have a set of collectively exhaustive events, one of them must occur. The coin must land on either heads or tails; the person must be male or female; the die must end with one numbered face up. The sum of the individual probabilities associated with a set of collectively exhaustive events is always 1.

4.2 Some Rules of Probability

A number of rules govern the calculation of the probabilities of elementary and joint events.

RULE 1 The probability of an event must be between 0 and 1. The smallest possible probability value is 0. You cannot have a negative probability. The largest possible probability value is 1.0. You cannot have a probability greater than 1.0.

EXAMPLE In the case of the die, the event of obtaining a face of seven has a probability of 0, because such an event cannot occur. The event obtaining a face with fewer than seven dots has a probability of 1.0, because it is certain that one of the elementary events of one, two, three, four, five, or six dots must occur.

RULE 2 The event that A does not occur is called A **complement** or simply **not A** , and is given the symbol A' . If $P(A)$ represents the probability of event A occurring, then $1 - P(A)$ represents the probability of event A not occurring.

EXAMPLE In the case of the die, the complement of obtaining the face that contains three dots is *not* obtaining the face that contains three dots. Because the probability of obtaining the face containing three dots is $1/6$, the probability of not obtaining the face that contains three dots is $(1 - 1/6) = 5/6$ or 0.833.

RULE 3 If two events A and B are **mutually exclusive**, the probability of both events A and B occurring is 0.

EXAMPLE On a single roll of a die, the face of the die turned up cannot have both three dots *and* have four dots, because such elementary events are

mutually exclusive. Either three dots can occur or four dots can occur, but not both.

RULE 4 If two events A and B are mutually exclusive, the probability of either event A or event B occurring is the sum of their separate probabilities.

EXAMPLE The probability of rolling a die and obtaining either a two or a three is $1/3$ or 0.333 , because this probability is the sum of the probability of rolling a two ($1/6$) and the probability of rolling a three ($1/6$).

INTERPRETATION You can extend this addition rule for mutually exclusive events to situations in which there are more than two events. In the case of rolling a die, the probability of turning up an even face (two, four, or six dots) is 0.50 , the sum of $1/6$ and $1/6$ and $1/6$ ($3/6$, or 0.50).

RULE 5 If events in a set are mutually exclusive and collectively exhaustive, the sum of their probabilities must add up to 1.0 .

EXAMPLE The events of a turning up a face with an even number of dots and turning up a face with an odd number of dots are mutually exclusive and collectively exhaustive. They are mutually exclusive, because even and odd cannot occur simultaneously on a single roll of a die. They are also collectively exhaustive, because either even or odd must occur on a particular roll. Therefore, for a single die, the probability of turning up a face with an even or odd face is the sum of the probability of turning up an even face plus the probability of turning up an odd face or 1.0 , as follows:

$$\begin{aligned} P(\text{even or odd face}) &= P(\text{even face}) + P(\text{odd face}) \\ &= \frac{3}{6} + \frac{3}{6} \\ &= \frac{6}{6} = 1 \end{aligned}$$

RULE 6 If two events A and B are *not* mutually exclusive, the probability of either event A or event B occurring is the sum of their separate probabilities minus the probability of their simultaneous occurrence (called **joint probability**).

EXAMPLE For rolling a single die, turning up a face with an even number of dots is not mutually exclusive to turning up a face with fewer than five dots, because both events include these (two) elementary events: turning up the face with two dots and turning up the face with four dots. To determine the probability of these two events, you add the probability of having a face with an even number of dots ($3/6$) to the probability of having a face with fewer than five dots ($4/6$) and then subtract the joint probability of simultaneously having a face with an even number of dots and having a face with fewer than five dots ($2/6$). You can express this as follows:

$$\begin{aligned}
 &P(\text{even face or face with fewer than five dots}) = \\
 &P(\text{even face}) + P(\text{face with fewer than five dots}) - \\
 &P(\text{even face and face with fewer than five dots}) \\
 &= \frac{3}{6} + \frac{4}{6} - \frac{2}{6} \\
 &= \frac{5}{6} = 0.833
 \end{aligned}$$

INTERPRETATION This rule requires that the joint probability be *subtracted*, because that probability has already been included twice (in the first event and in the second event). Because the joint probability has been “double counted,” it must be subtracted to provide the correct result.

RULE 7 If two events *A* and *B* are **independent**, the probability of both events *A* and *B* occurring is equal to the product of their respective probabilities. Two events are independent if the occurrence of one event in no way affects the probability of the second event.

EXAMPLE When rolling a die, each roll of the die is an independent event, because no roll can affect another (although gamblers who play dice games sometimes would like to think otherwise). Therefore, to determine the probability that two consecutive rolls both turn up the face with five dots, you would multiply the probability of turning up that face on roll one (1/6) by the probability of turning up that face on roll two (also 1/6). You can express this as follows:

$$\begin{aligned}
 &P(\text{face with five dots on roll one and face with five dots on roll two}) = \\
 &P(\text{face with five dots on roll one}) \times P(\text{face with five dots on roll two}) \\
 &= \frac{1}{6} \times \frac{1}{6} \\
 &= \frac{1}{36} = 0.028
 \end{aligned}$$

RULE 8 If two events *A* and *B* are *not* independent, the probability of both events *A* and *B* occurring is the product of the probability of event *A* times the conditional probability of event *B* occurring, given that event *A* has occurred.

EXAMPLE During the taping of a television game show, contestants are randomly selected from the audience watching the show. After a particular person has been chosen, he or she does not return to the audience and cannot be chosen later, therefore making this a case in which a conditional probability occurs.

If the audience were comprised of 30 women and 20 men (50 people), what would be the probability that the first two contestants chosen are male? The probability that the first contestant is male is simply 20/50 or 0.40. However,

the probability that the second contestant is male is *not* 20/50, because when the second selection is made, the eligible audience has now only 19 males and 49 people. Therefore, the probability that the second selection is male is 19/49 or 0.388, rounded. This means that the probability that the first two contestants are male is 0.155 as follows:

$$\begin{aligned} P(\text{male selection first and male selection second}) &= \\ P(\text{male selection first}) \times P(\text{male selection second}) &= \\ &= \frac{20}{50} \times \frac{19}{49} \\ &= \frac{380}{2,450} = 0.155 \end{aligned}$$

4.3 Assigning Probabilities

There are three distinct approaches for assigning probabilities to the events of a random variable: the *classical approach*, the *empirical approach*, and the *subjective approach*.

Classical Approach

CONCEPT Assigning probabilities based on prior knowledge of the process involved.

EXAMPLE Assigning the probability of rolling a die and turning up the face with three dots.

INTERPRETATION Classical probability often assumes that all elementary events are equally likely to occur. When this is true, the probability that a particular event will occur is defined by the number of ways the event can occur divided by the total number of elementary events. For example, if you roll a die, the probability of obtaining the face with three dots is 1/6 because there are six elementary events associated with rolling a die. This allows you to expect that 1,000 out of 6,000 rolls of a die would turn up the face with three dots.

Empirical Approach

CONCEPT Assigning probabilities based on frequencies obtained from empirically observed data.

EXAMPLE Probabilities determined by polling or market surveys.

INTERPRETATION The empirical approach does not use theoretical reasoning or assumed knowledge of a process to assign probabilities. Similar to the classical approach when all elementary events are equally likely, the empirical probability can be calculated by dividing the number of ways A can occur by the total number of elementary events. For example, if a poll of 500 registered voters reveals that 275 are likely to vote in the next election, you can assign the empirical probability of 0.55 (275 divided by 500).

Subjective Approach

CONCEPT Assign probabilities based on expert opinions or other subjective means such as “gut” feelings or hunches.

EXAMPLE Commentators stating odds about a political candidate’s chance of winning.

INTERPRETATION In this approach, you use your own intuition to judge the likeliest outcomes. You use the subjective approach when either the number of elementary events or actual data are not available for the calculation of relative frequencies. Because of the subjectivity, different individuals might assign different probabilities to the same event.

One-Minute Summary

Foundation Concepts

- Rules of probability
- Assigning probabilities

Test Yourself

1. If two events are collectively exhaustive, what is the probability that one or the other occurs?
 - (a) 0
 - (b) 0.50
 - (c) 1.00
 - (d) Cannot be determined from the information given
2. If two events are collectively exhaustive, what is the probability that both occur at the same time?
 - (a) 0
 - (b) 0.50
 - (c) 1.00
 - (d) Cannot be determined from the information given

3. If two events are mutually exclusive, what is the probability that both occur at the same time?
 - (a) 0
 - (b) 0.50
 - (c) 1.00
 - (d) Cannot be determined from the information given
4. If the outcome of event A is not affected by event B , then events A and B are said to be:
 - (a) mutually exclusive
 - (b) statistically independent
 - (c) collectively exhaustive
 - (d) None of the above

(Use the following problem description when answering questions 5 through 9)

A survey is taken among customers of a fast-food restaurant to determine preference for hamburger or chicken. Of 200 respondents selected, 125 were male and 75 were female. 120 preferred hamburger and 80 preferred chicken. Of the males, 85 preferred hamburger.

5. The probability that a randomly selected individual is a male is:
 - (a) $125/200$
 - (b) $75/200$
 - (c) $120/200$
 - (d) $200/200$
6. The probability that a randomly selected individual prefers hamburger or chicken is:
 - (a) $0/200$
 - (b) $125/200$
 - (c) $75/200$
 - (d) $200/200$
7. Suppose that two individuals are randomly selected. The probability that both prefer hamburger is:
 - (a) $(120/200)(120/200)$
 - (b) $(120/200)$
 - (c) $(120/200)(119/199)$
 - (d) $(85/200)$
8. The probability that a randomly selected individual prefers hamburger is:
 - (a) $0/200$
 - (b) $120/200$
 - (c) $75/200$
 - (d) $200/200$

9. The probability that a randomly selected individual prefers hamburger or is a male is:
 - (a) 0/200
 - (b) 125/200
 - (c) 160/200
 - (d) 200/200
10. The smallest possible value for a probability is _____.

Answers to Test Yourself Questions

1. c
2. d
3. a
4. b
5. a
6. d
7. c
8. b
9. c
10. 0

References

1. Berenson, M. L., D. M. Levine, and T. C. Krehbiel. *Basic Business Statistics: Concepts and Applications, Ninth Edition*. Upper Saddle River, NJ: Prentice Hall, 2004.
2. Gitlow, H. S., and D. M. Levine. *Six Sigma for Green Belts and Champions*. Upper Saddle River, NJ: Financial Times - Prentice Hall, 2005.
3. Levine, D. M., T. C. Krehbiel, and M. L. Berenson. *Business Statistics: A First Course, Third Edition*. Upper Saddle River, NJ: Prentice Hall, 2003.
4. Levine, D. M., D. Stephan, T. C. Krehbiel, and M. L. Berenson. *Statistics for Managers Using Microsoft Excel, Fourth Edition*. Upper Saddle River, NJ: Prentice Hall, 2005.
5. Levine, D. M., P. P. Ramsey, and R. K. Smidt. *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Upper Saddle River, NJ: Prentice Hall, 2001.

6. Microsoft Excel 2002. Redmond, WA: Microsoft Corporation, 2001.
7. Sincich, T., D. M. Levine, and D. Stephan. *Practical Statistics by Example Using Microsoft Excel and Minitab, Second Edition*. Upper Saddle River, NJ: Prentice Hall, 2002.

This page intentionally left blank



Probability Distributions

5.1 Probability Distributions for Discrete Variables

5.2 The Binomial and Poisson Probability Distributions

5.3 Continuous Probability Distributions and the Normal Distribution

5.4 The Normal Probability Plot

Important Equations

One-Minute Summary

Test Yourself

In Chapter 4, you learned to use the rules of probability to make inferences concerning a particular outcome. In practice, the probability of many events is unknown, so probability models are developed to estimate the probabilities of occurrence for different values.

5.1 Probability Distributions for Discrete Variables

Probability distributions for a random variable summarize or model the probabilities associated with the events for that random variable. Such distributions take different forms depending on whether the random variable is *discrete* or *continuous*.

This section reviews probability distributions for discrete random variables and statistics related to such distributions.

Discrete Probability Distribution

CONCEPT A listing of all possible distinct (elementary) events and their probabilities of occurring for a random variable.

EXAMPLE See the Worked-out Problem below.

INTERPRETATION In a probability distribution for a discrete variable, the sum of the probabilities of all the events always equals 1, an indirect way of saying that the (elementary) events listed are always **collectively exhaustive** (that is, that they have used up all possible occurrences). Although you can use a table of outcomes (see Worked-out Problem) to develop a probability distribution, probability distributions for certain types of random variables can also be calculated by using a formula that mathematically models the distribution.

WORKED-OUT PROBLEM You seek to determine the probability of getting 0, 1, 2, or 3 heads when you toss a fair coin three times in a row. Because getting 0, 1, 2, or 3 heads represent all possible distinct outcomes, you can form a table of all possible outcomes (eight) of tossing a fair coin three times as follows.

Outcome	First Toss	Second Toss	Third Toss
1	Head	Head	Head
2	Head	Head	Tail
3	Head	Tail	Head
4	Head	Tail	Tail
5	Tail	Head	Head
6	Tail	Head	Tail
7	Tail	Tail	Head
8	Tail	Tail	Tail

From this table of all eight possible outcomes, you can form this summary table shown in Table 5.1.

Table 5.1

Probability Distribution for Tossing a Fair Coin Three Times

Number of Heads	Number of Outcomes With That Number of Heads	Probability
0	1	$1/8 = 0.125$
1	3	$3/8 = 0.375$
2	3	$3/8 = 0.375$
3	1	$1/8 = 0.125$

From this probability distribution, you determine that the chance of rolling three heads in a row is 0.125 and that the sum of the probabilities is 1.0, as it should be for a distribution for a discrete variable.

Another way to obtain these probabilities is to extend Rule 8 on page 66, the multiplication rule, to three events (or tosses). To get the probability of three heads, which is equal to $1/8$ or 0.125 using Rule 8, you have:

$$P(H_1 \text{ and } H_2 \text{ and } H_3) = P(H_1) \times P(H_2) \times P(H_3)$$

Because each toss has a probability of heads of 0.5 :

$$P(H_1 \text{ and } H_2 \text{ and } H_3) = (0.5)(0.5)(0.5)$$

$$P(H_1 \text{ and } H_2 \text{ and } H_3) = 0.125$$

The Expected Value of a Random Variable

CONCEPT The sum of the products formed by multiplying each possible event in a discrete probability distribution by its corresponding probability.

INTERPRETATION The expected value tells you the value of the random variable that you could expect in the “long run,” after many experimental trials. The expected value of a random variable is sometimes known as the average value of a random variable.

WORKED-OUT PROBLEM If there are three tosses of a coin (Table 5.1), you can calculate the expected value of the number of heads as is done in Table 5.2.

Table 5.2

Computing the Expected Value of a Probability Distribution

Number of Heads	Probability	(Number of Heads) \times (Probability)
0	0.125	$(0) \times (0.125) = 0$
1	0.375	$(1) \times (0.375) = 0.375$
2	0.375	$(2) \times (0.375) = 0.75$
3	0.125	$(3) \times (0.125) = 0.375$
		Total = 1.50

Average value = Sum of [each value \times the probability of each value]

$$\mu = (0)(0.125) + (1)(0.375) + (2)(0.375) + (3)(0.125)$$

$$= 0 + 0.375 + 0.750 + 0.375 = 1.50$$

Notice that in this example, the average or expected value of the number of heads is 1.5 , a value for the number of heads that is impossible. The average of 1.5 heads tells you that, in the long run, if you toss three fair coins many times, the average number of heads you can expect is 1.5 .

Standard Deviation of a Random Variable (σ)

CONCEPT The measure of variation around the expected value of a random variable, calculated by first summing the products formed by multiplying the squared difference between each value and the expected value by its corresponding probability and then taking the square root of that sum.

EXAMPLE If there are three tosses of a coin (Table 5.1), you can calculate the standard deviation of the number of heads as is done in Table 5.3.

Table 5.3

Computing the Standard Deviation of a Probability Distribution

Number of Heads	Probability	(Number of Heads – (Average Number of Heads)) ² × (Probability)
0	0.125	$(0 - 1.5)^2 \times (0.125) = 2.25 \times (0.125) = 0.28125$
1	0.375	$(1 - 1.5)^2 \times (0.375) = 0.25 \times (0.375) = 0.09375$
2	0.375	$(2 - 1.5)^2 \times (0.375) = 0.25 \times (0.375) = 0.09375$
3	0.125	$(3 - 1.5)^2 \times (0.125) = 2.25 \times (0.125) = 0.28125$
		Total = 0.75

σ = Square root of [Sum of (Squared differences between a value and the expected value) × (Probability of the value)]

$$\begin{aligned}\sigma &= \sqrt{(0-1.5)^2(0.125) + (1-1.5)^2(0.375) + (2-1.5)^2(0.375) + (3-1.5)^2(0.125)} \\ &= \sqrt{2.25(0.125) + 0.25(0.375) + 0.25(0.375) + 2.25(0.125)} \\ &= \sqrt{0.75}\end{aligned}$$

and

$$\sigma = \sqrt{0.75} = 0.866$$

INTERPRETATION In financial analysis, you can use the standard deviation of a random variable about investment evaluations to assess the degree of risk of an investment, as the next Worked-out Problem illustrates.

WORKED-OUT PROBLEM Suppose that you are deciding between two alternative investments. Investment A is a mutual fund whose portfolio consists of a combination of stocks that make up the Dow Jones Industrial Average. Investment B consists of shares of a growth stock. You estimate the returns (per \$1,000 investment) for each investment alternative under three economic condition events (recession, stable economy, and expanding economy), and also provide your subjective probability of the occurrence of each economic condition as follows.

Estimated Return for Two Investments Under Three Economic Conditions

Probability	Economic Event	Investment	
		Dow Jones Fund (A)	Growth Stock (B)
0.2	Recession	-\$100	-\$200
0.5	Stable economy	+ 100	+ 50
0.3	Expanding economy	+ 250	+ 350

The mean return (expected value) for the two investments is as follows:

Mean = Sum of [Each value \times The probability of each value]

Mean for the Dow Jones fund = $(-100)(0.2) + (100)(0.5) + (250)(0.3) = \105

Mean for the growth stock = $(-200)(0.2) + (50)(0.5) + (350)(0.3) = \90

You can calculate the standard deviation for the two investments as done in Tables 5.4 and 5.5.

Table 5.4

Computing the Standard Deviation for Dow Jones Fund (A)

Probability	Economic Event	Dow Jones Fund (A)	(Return – Average return) ² \times Probability
0.2	Recession	-\$100	$(-100 - 105)^2 \times (0.2) =$ $(42,025) \times (0.2) = 8,405$
0.5	Stable economy	+ 100	$(100 - 105)^2 \times (0.5) = (25) \times$ $(0.5) = 12.5$
0.3	Expanding economy	+ 250	$(250 - 105)^2 \times (0.3) = (21,025)$ $\times (0.3) = 6,307.5$
			Total: 14,725

Table 5.5

Computing the Standard Deviation for Growth Stock (B)

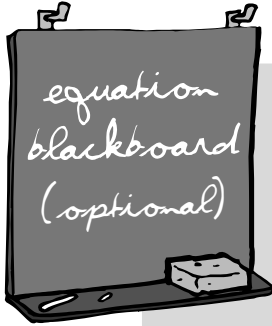
Probability	Economic Event	Growth Stock (B)	(Return – Average return) ² \times Probability
0.2	Recession	-\$200	$(-200 - 90)^2 \times (0.2) = (84,100)$ $\times (0.2) = 16,820$
0.5	Stable economy	+ 50	$(50 - 90)^2 \times (0.5) = (1,600) \times$ $(0.5) = 800$
0.3	Expanding economy	+ 350	$(350 - 90)^2 \times (0.3) = (67,600)$ $\times (0.3) = 20,280$
			Total: 37,900

σ = Square root of [Sum of (Squared differences between a value and the mean) \times (Probability of the value)]

$$\begin{aligned}\sigma_A &= \sqrt{(-100-105)^2(0.2) + (100-105)^2(0.5) + (250-105)^2(0.3)} \\ &= \sqrt{14,725} = \$121.35\end{aligned}$$

$$\begin{aligned}\sigma_B &= \sqrt{(-200-90)^2(0.2) + (50-90)^2(0.5) + (350-90)^2(0.3)} \\ &= \sqrt{37,900} = \$194.68\end{aligned}$$

The Dow Jones fund has a higher mean return than the growth fund and also has a lower standard deviation, indicating less variation in the return under the different economic conditions. Having a higher mean return with less variation makes the Dow Jones fund a more desirable investment than the growth fund.



To write the equations for the mean and standard deviation for a discrete probability distribution, you need the following symbols:

- An uppercase italic X , X , that represents a random variable.
- An uppercase italic X with an italic lowercase i subscript, X_i , that represents the i th event associated with random variable X .
- An uppercase italic N , N , that represents the number of elementary events for the random variable X . (In Chapter 3, this symbol was called the population size.)
- The symbol $P(X_i)$, which represents the probability of the event X_i .
- The population mean, μ .
- The population standard deviation, σ .

Using these symbols creates these equations:

The mean of a probability distribution:

$$\mu = \sum_{i=1}^N X_i P(X_i)$$

The standard deviation of a probability distribution:

$$\sigma = \sqrt{\sum_{i=1}^N (X_i - \mu)^2 P(X_i)}$$

5.2 The Binomial and Poisson Probability Distributions

As noted in the previous section, some probability distributions for certain types of discrete random variables can be modeled using a mathematical formula. This section looks at two widely used distributions that can be used to estimate probabilities. The first probability distribution, the binomial, is used for random variables that have only two mutually exclusive events. The second probability distribution, the Poisson, is used when you are counting the number of outcomes that occur in a unit.

The Binomial Distribution

CONCEPT The probability distribution for a discrete random variable that meets these criteria:

- The random variable is for a sample that consists of a fixed number of experimental trials.
- The random variable has only two mutually exclusive and collectively exhaustive events, typically labeled as success and failure.
- The probability of an event being classified as a success, p , and the probability of an event being classified as a failure, $1 - p$, are both constant in all experimental trials.
- The event (success or failure) of any single experimental trial is independent of (not influenced by) the event of any other trial.

important point



EXAMPLE The coin tossing experiment described in the Worked-out Problem on page 74.

INTERPRETATION Using the binomial distribution allows you to avoid having to determine the probability distribution by using a table of outcomes and applying the multiplication rule, as was done in Section 4.2. This distribution also does not require that the probability of success is 0.5, thereby allowing you to use it in more situations than the method discussed in Section 4.2.

You typically determine binomial probabilities by either using the formula in the EQUATION BLACKBOARD on page 81, by using a table of binomial probabilities, or by using software functions such as those used to produce the Microsoft Excel spreadsheet on page 80. (When the probability of success is 0.5, you can, of course, still use the table and multiplication rule method as was done in Table 5.1. Note the results of that table—that the probability of zero heads is 0.125, the probability of one head is 0.375, the probability of two heads is 0.375, and the probability of three heads is 0.125—agree with the spreadsheet results shown on page 80.)

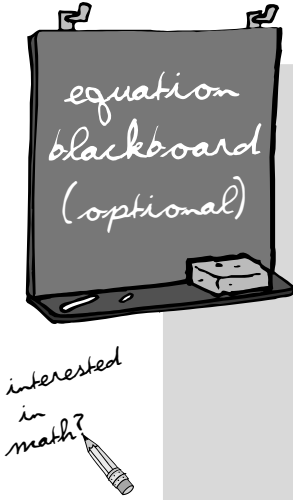
	A	B	C
1	Binomial Probabilities		
2			
3	Data		
4	Sample size	3	
5	Probability of success	0.5	
6			
7	Statistics		
8	Mean	1.5	
9	Variance	0.75	
10	Standard deviation	0.866025	
11			
12	Binomial Probabilities Table		
13		X	P(X)
14		0	0.125
15		1	0.375
16		2	0.375
17		3	0.125

Binomial distributions can be symmetrical or skewed. Whenever $p = 0.5$, the binomial distribution will be symmetrical regardless of how large or small the value of n . However, when $p \neq 0.5$, the distribution will be skewed. If $p < 0.5$, the distribution will be positive or right-skewed; if $p > 0.5$, the distribution will be negative or left-skewed. The closer p is to 0.5 and the larger the sample size, n , the more symmetrical the distribution will be.

For the binomial distribution, the number of experimental trials is equivalent to the term *sample size* introduced in Chapter 1. You calculate the mean and standard deviation for a random variable that can be modeled using the binomial distribution using the sample size, the probability of success, and the probability of failure as follows.

Binomial Distribution Characteristics

Mean	The sample size (n) times the probability of success or $n \times p$, remembering that the sample size is the number of experimental trials.
Variance	The product of these three: sample size, probability of success, and probability of failure ($1 - \text{Probability of success}$), or $n \times p \times (1 - p)$
Standard deviation	The square root of the variance, or $\sqrt{np(1-p)}$



For the equation for the binomial distribution, you take the symbols X (random variable), n (sample size), and p (probability of success) previously introduced and add these symbols:

- A lowercase italic X , x , which represents the number of successes in the sample.
- The symbol $P(X = x | n, p)$, which represents the probability of the value x , given sample size n and probability of success p .

You use these symbols to form two separate expressions. One expression represents the number of ways you can get a certain number of successes in a certain number of trials:

$$\frac{n!}{x!(n-x)!}$$

(The symbol $!$ means factorial, so that $n! = (n)(n-1)\dots(1)$ so that $3!$ is $3 \times 2 \times 1$, equals 6. $1!$ equals 1 and $0!$ is defined as being equal to 1.).

The second expression represents the probability of getting a certain number of successes in a certain number of trials *in a specific order*:

$$p^x \times (1-p)^{n-x}$$

Using these expressions forms the following equation:

$$P(X = x | n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

As an example, the calculations for determining the binomial probability of one head in three tosses of a fair coin (that is, for a problem in which $n = 3$, $p = 0.5$, and $x = 1$) are as follows:

$$\begin{aligned} P(X = 1 | n = 3, p = 0.5) &= \frac{3!}{1!(3-1)!} (0.5)^1 (1-0.5)^{3-1} \\ &= \frac{3!}{1!(2)!} (0.5)^1 (1-0.5)^2 \\ &= 3(0.5)(0.25) = 0.375 \end{aligned}$$

Using symbols previously introduced, you can write the equation for the mean and standard deviation of the binomial distribution:

(continues)

$$\mu = np$$

and

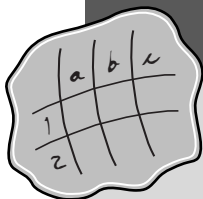
$$\sigma = \sqrt{np(1-p)}$$



calculator keys

Binomial Probabilities

Press [2nd] [VARS] (to display the Distr menu) and select either 0:binompdf or A:binomcdf and press [ENTER] to calculate an exact or cumulative probability. Enter the sample size, probability of success, and optionally, the number of successes, separated by commas, and press [ENTER]. If you do not enter a value for the number of successes, you will get a list of probabilities that you can view by using the cursor keys.



spreadsheet solution

Binomial Probabilities

Download and open the **Chapter 5 Binomial.xls** Excel file into which you can enter the sample size, probability of success, and number of successes to calculate an exact binomial probability.

WORKED-OUT PROBLEM Whether a would-be shopper stays and views a retail Web site for more than one minute is one of the measures of *success* of such sites. Suppose that the probability that the shopper does stay for more than one minute is 0.16. What is the probability that at least four (either four or five) of the next five shoppers will stay for more than one minute as well?

You need to sum the probabilities of four shoppers staying and five shoppers staying in order to determine the probabilities that at least four shoppers stay.

Microsoft Excel's and Texas Instruments statistical calculator's (partial) results for this study are:

	A	B	C
1	Binomial Probabilities		
2			
3	Data		
4	Sample size	5	
5	Probability of success	0.16	
6			
7	Statistics		
8	Mean	0.8	
9	Variance	0.672	
10	Standard deviation	0.8198	
11			
12	Binomial Probabilities Table		
13		X	P(X)
14		0	0.4182
15		1	0.3983
16		2	0.1517
17		3	0.0289
18		4	0.0028
19		5	0.0001

```
binompdf(5,0.16
(.4182119424 .3...
```

From the Microsoft Excel results:

$$P(X=4|n=5, p=0.16) = 0.0028$$

$$P(X=5|n=5, p=0.16) = 0.0001$$

Therefore, the probability of four or more shoppers staying is 0.0029 (which you compute by adding 0.0028 and 0.0001) or 0.29%.

The Poisson Distribution

CONCEPT The probability distribution for a discrete random variable that meets these criteria:

- You are counting the number of times a particular event occurs in a unit.
- The probability that an event occurs in a particular unit is the same for all other units.

important point



- The number of events that occur in a unit is independent of the number of events that occur in other units.
- As the unit gets smaller, the probability that two or more events will occur in that unit approaches zero.

EXAMPLES Number of computer network failures per day, number of surface defects per square yard of floor coverings, the number of fleas on the body of a dog.

INTERPRETATION To use the Poisson distribution, you define an **area of opportunity**, a continuous unit of area, time, or volume in which more than one occurrence of an event can occur. The Poisson distribution can model many random variables that count the number of defects per area of opportunity or count the number of times things are processed from a waiting line.

You determine Poisson probabilities by applying the formula in the EQUATION BLACKBOARD on pages 85–86, by using a table of Poisson values, or by using software functions that produce customized tables (see the figure on page 85). You can calculate the mean and standard deviation for a random variable that can be modeled using the Poisson distribution using the population mean as follows.

Poisson Distribution Characteristics

Mean	The population mean, λ .
Variance	The population mean, λ , that in the Poisson distribution is equal to the variance.
Standard deviation	The square root of the variance, or $\sqrt{\lambda}$.

WORKED-OUT PROBLEM You seek to determine the probabilities associated with the number of customers arriving at a large bank branch per one-minute interval during the lunch hour: Will zero customers arrive, one customer, two customers, and so on? You determine that you can use the Poisson distribution because of the following reasons:

- The random variable is a count per unit, customers per minute.
- You judge that the probability that a customer arrives during a one-minute interval is the same as the probability for all the other one-minute intervals.
- Each customer's arrival has no effect on (is independent of) all other arrivals.
- The probability that two or more customers will arrive in a given time period approaches zero as the time interval decreases from one minute.

Using historical data, you can determine the average number of arrivals of customers per minute during the lunch hour (three customers per minute).

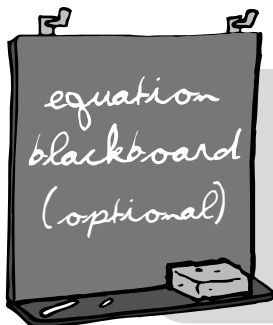
You use Microsoft Excel to generate these Poisson probabilities:

	A	B	C	D	E
1	Poisson Probabilities for Customer Arrivals				
2					
3	Data				
4	Average/Expected number of successes:				3
5					
6	Poisson Probabilities Table				
7		X	P(X)		
8		0	0.049787		
9		1	0.149361		
10		2	0.224042		
11		3	0.224042		
12		4	0.168031		
13		5	0.100819		
14		6	0.050409		
15		7	0.021604		
16		8	0.008102		
17		9	0.002701		
18		10	0.000810		
19		11	0.000221		
20		12	0.000055		
21		13	0.000013		
22		14	0.000003		
23		15	0.000001		

From the results, you note the following:

- The probability of zero arrivals is 0.049787.
- The probability of one arrival is 0.149361.
- The probability of two arrivals is 0.224042.

Therefore, the probability of two or fewer customer arrivals per minute at the bank during the lunch hour is 0.42319, the sum of the probabilities for zero, one, and two arrivals ($0.049787 + 0.149361 + 0.224042 = 0.42319$).




For the equation for the Poisson distribution, you take the symbols X (random variable), n (sample size), p (probability of success) previously introduced and add these symbols:

- A lowercase italic E , e , which represents the mathematical constant approximated by the value 2.71828.

(continues)

interested
in
math?



- A lowercase Greek symbol lambda, λ , which represents the average number of times that the event occurs per area of opportunity.
- A lowercase italic X , x , which represents the number of times the event occurs per area of opportunity.
- The symbol $P(X = x | \lambda)$, which represents the probability of x , given λ .

Using these symbols forms the following equation:

$$P(X = x | \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

As an example, the calculations for determining the Poisson probability of exactly 2 arrivals in the next minute given an average of three arrivals per minute is as follows:

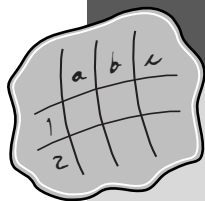
$$\begin{aligned} P(X = 2 | \lambda = 3) &= \frac{e^{-3}(3)^2}{2!} \\ &= \frac{(2.71828)^{-3}(3)^2}{2!} \\ &= \frac{(0.049787)(9)}{(2)} \\ &= 0.224042 \end{aligned}$$



calculator keys

Poisson Probabilities

Press [2nd] [VARS] (to display the Distr menu) and select either **B:poissonpdf** or **C:poissoncdf** to calculate an exact or cumulative Poisson probability. Enter the average number of successes and number of successes and press [ENTER].



spreadsheet solution

Poisson Probabilities

Download and open the **Chapter 5 Poisson.xls** Excel file into which you can enter the average/expected number of successes to produce a table of Poisson probabilities.

5.3 Continuous Probability Distributions and the Normal Distribution

Probability distributions for a continuous random variable differ from discrete distributions in several important ways:



- An event can take on any value within the range of the random variable and not just integers.
- The probability of any specific value is zero.
- Probabilities are expressed in terms of an area under a curve that represents the continuous distribution.

One continuous distribution, the **normal distribution**, dominates statistics, because it can model many different types of continuous random variables. Probabilities associated with such diverse things as physical characteristics such as height and weight, scores on standardized exams, and the dimension of industrial parts tend to follow a normal distribution. Under certain circumstances, the normal distribution also approximates various discrete probability distributions such as the binomial and Poisson and provides the basis for classical statistical inference discussed in Chapters 6 through 9. For these reasons, the normal distribution is the focus of this section.

Normal Distribution

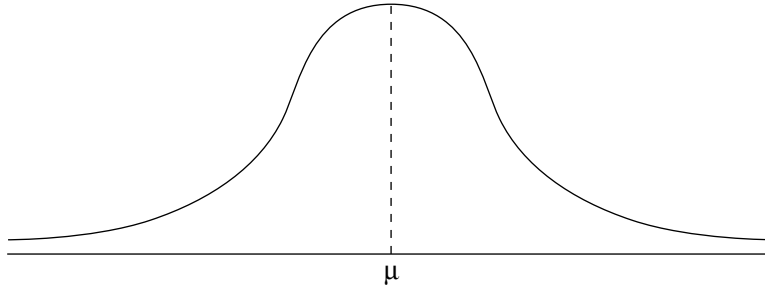
CONCEPT The probability distribution for a continuous random variable that meets these criteria:



- The graphed curve of the distribution is bell-shaped and symmetrical.
- The mean, median, and mode are all the same value.
- The population mean, μ , and the population standard deviation, σ , determine probabilities.
- The distribution extends from negative to positive infinity. (The distribution has an infinite range.)

Probabilities are always cumulative and expressed as inequalities, such as $P < X$ or $P \geq X$, where X is a value for the variable.

EXAMPLE The normal distribution appears as a bell-shaped curve as pictured below.



INTERPRETATION The importance of the normal distribution to statistics, already stated in the introduction to this section, cannot be overstated.

You determine normal probabilities by using a table of normal probabilities (such as Table C.1 in Appendix C) or by using software functions. (You do not use a formula to directly determine the probabilities, because the complexities of the formula rule out its everyday use.) Normal probability tables (including Table C.1) and some software functions use a standardized normal distribution that require you to convert an X value of a variable to its corresponding Z score (see Section 3.2). You perform this conversion by subtracting the population mean μ from the X value and dividing the resulting difference by the population standard deviation σ , expressed algebraically as follows:

$$Z = \frac{X - \mu}{\sigma}$$

important point 

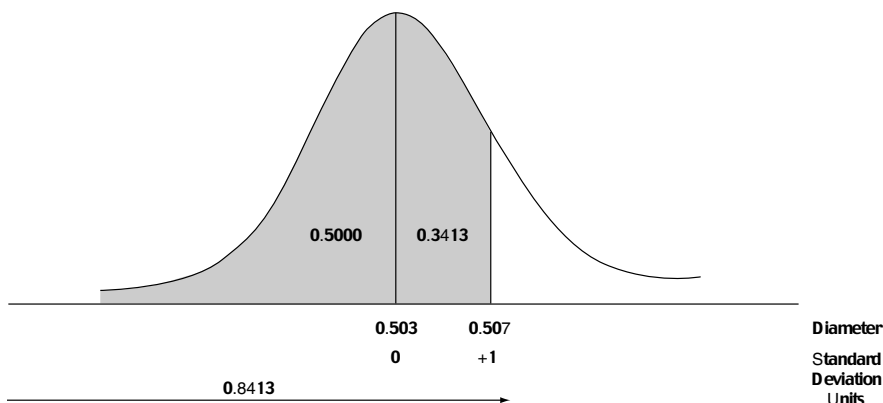
Note that when the mean is 0 and the standard deviation is 1, the X value and Z score will be the same and no conversion is necessary.

WORKED-OUT PROBLEM A certain machine uses ball bearings that must be between 0.49 inches (lower) and 0.51 inches (upper) in diameter. Past experience has indicated that the actual diameter of ball bearings used is approximately normally distributed with a mean $\mu = 0.503$ inches and a standard deviation $\sigma = 0.004$ inches. Suppose that you want to determine the probability that a single ball bearing used will have a diameter between 0.503 and 0.507 inches using Table C.1, the table of the probabilities of the cumulative standardized normal distribution.

To use Table C.1, you must first convert the diameters to their Z scores by subtracting the mean and the dividing by the standard deviation, as shown here:

$$Z \text{ (lower)} = \frac{0.503 - 0.503}{0.004} = 0 \qquad Z \text{ (upper)} = \frac{0.507 - 0.503}{0.004} = 1.0$$

Therefore, you need to determine the probability that corresponds to the area between 0 and +1 Z units (standard deviations). To do this, you take the cumulative probability associated with 0 Z units and subtract it from the probability associated with +1 Z units. Using Table C.1, you determine that these probabilities are 0.8413 and 0.5000, respectively. Therefore, the probability of obtaining a single ball bearing that is between 0.503 and 0.507 inches is 0.3413 ($0.8413 - 0.5000 = 0.3413$).



A Microsoft Excel worksheet that calculates various normal probabilities shows the same results:

	A	B	C	D	E
1	Normal Probabilities				
2					
3	Common Data				
4	Mean	0.503			
5	Standard Deviation	0.004			
6					
7	Probability for X <=		Probability for a Range		
8	X Value	0.503	From X Value	0.503	
9	Z Value	0	To X Value	0.507	
10	P(X<=0.503)	0.5	Z Value for 0.503	0	
11			Z Value for 0.507	1	
12	Probability for X >		P(X<=0.503)	0.5000	
13	X Value	0.507	P(X<=0.507)	0.8413	
14	Z Value	1	P(0.503<=X<=0.507)	0.3413	
15	P(X>0.507)	0.1587			
16					
17	Probability for X<0.503 or X>0.507				
18	P(X<0.503 or X>0.507)	0.6587			

Using Standard Deviation Units

Because of the equivalence between Z scores and standard deviation units, probabilities of the normal distribution are often expressed as ranges of

plus-or-minus standard deviation units. Such probabilities can be determined directly from Table C.1, the table of the probabilities of the cumulative standardized normal distribution.

For example, to determine the normal probability associated with the range plus-or-minus 3 standard deviations, you would use Table C.1 to look up the probabilities associated with $Z = -3.00$ and $Z = +3.00$:

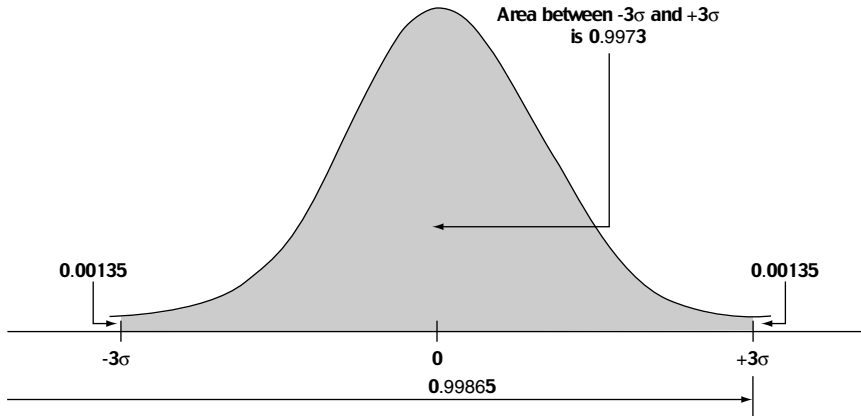


Table 5.6 represents the appropriate portion of Table C.1 for $Z = -3.00$. From this table excerpt, you can determine that the probability of a value less than $Z = -3$ units is 0.00135.

Table 5.6

Partial Table C.1 for Obtaining a Cumulative Area Below $-3 Z$ Units

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00103	0.00100

Source: Extracted from Table C.1

Table 5.7 represents the appropriate portion of Table C.1 for $Z = +3.00$. From this table excerpt, you can determine that the probability of a value less than $Z = +3$ units is 0.99865.

Table 5.7

Partial Table C.1 for Obtaining a Cumulative Area Below $+3 Z$ Units

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
+3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99897	0.99900

Source: Extracted from Table C.1

Therefore, the normal probability associated with the range plus-or-minus 3 standard deviations is 0.9973 ($0.99865 - 0.00135$). Stated another way, there is the probability of 0.0027 (2.7 out of a thousand chance) that a value will not be within the range of plus-or-minus 3 standard deviations. Table 5.8 summarizes probabilities for several different ranges of standard deviation units.

Table 5.8

Normal Probabilities for Selected Number of Standard Deviation Units

Standard Deviation Unit Ranges	Probability or Area Outside These Units	Probability or Area Within These Units
-1σ to $+1\sigma$	0.3174	0.6826
-2σ to $+2\sigma$	0.0455	0.9545
-3σ to $+3\sigma$	0.0027	0.9973
-6σ to $+6\sigma$	0.000000002	0.999999998

Finding the Z Value from the Area Under the Normal Curve

Each of the previous examples involved using the normal tables to find an area under the normal curve that corresponded to a specific Z value. There are many circumstances when you want to do the opposite of this and find the Z value that corresponds to a specific area. For example, you might want to find the Z value that corresponds to a cumulative area of 1%, 5%, 95%, or 99%. You might also want to find lower and upper Z values between which 95% of the area under the curve is contained.

To find the Z value that corresponds to a cumulative area, you locate the cumulative area in the body of the normal table, or the closest value to the cumulative area you seek, and then determine the Z value that corresponds to this cumulative area.

WORKED-OUT PROBLEM You want to find the Z values such that 95% of the normal curve is contained between a lower Z value and an upper Z value with 2.5% below the lower Z value, and 2.5% above the upper Z value. Using the figure at the top of p. 92, you determine that you need to find the Z value that corresponds to a cumulative area of 0.025 and the Z value that corresponds to a cumulative area of 0.975.

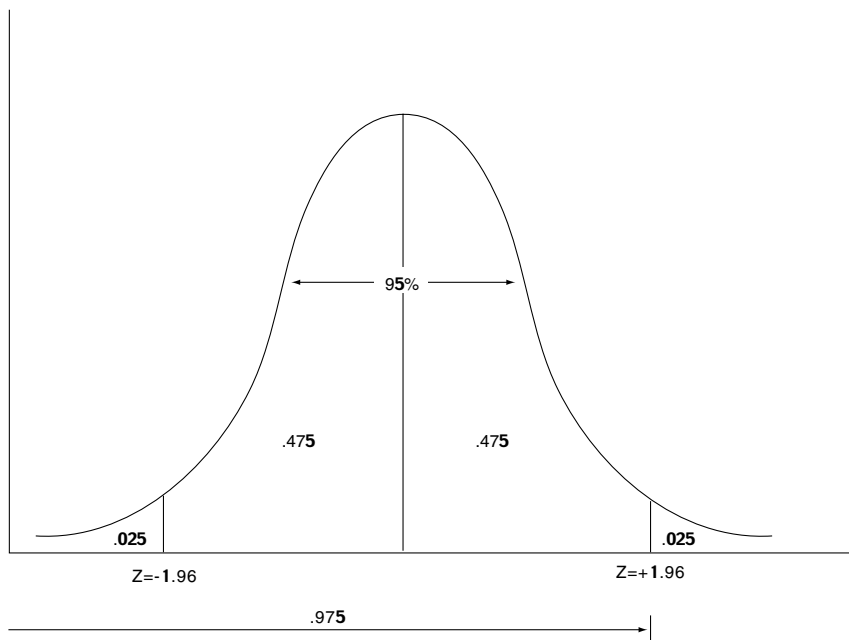


Table 5.9 contains a portion of Table C.1 that is needed to find the Z value that corresponds to a cumulative area of 0.025. Table 5.10 contains a portion of Table C.1 that is needed to find the Z value that corresponds to a cumulative area of 0.975.

Table 5.9

Partial Table C.1 for Finding Z Value That Corresponds to a Cumulative Area of 0.025

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.
.
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233

Table 5.10

Partial Table C.1 for Finding Z Value That Corresponds to a Cumulative Area of 0.975

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.
.
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

To find the Z value that corresponds to a cumulative area of 0.025, you look in the body of Table 5.9 until you see the value of 0.025. Then you determine the row and column that this value corresponds to. Locating the value of 0.025, you see that it is located in the -1.9 row and the $.06$ column. Thus the Z value that corresponds to a cumulative area of 0.025 is -1.96 .

To find the Z value that corresponds to a cumulative area of 0.975, you look in the body of Table 5.10 until you see the value of 0.975. Then you determine the corresponding row and column that this value belongs to. Locating the value of 0.975, you see that it is in the 1.9 row and the $.06$ column. Thus the Z value that corresponds to a cumulative area of 0.975 is 1.96 . Taking this result along with the Z value of -1.96 for a cumulative area of 0.025 means that 95% of all the values will be between $Z = -1.96$ and $Z = 1.96$.



calculator keys

Normal Probabilities

To calculate the cumulative normal probability for a specific X value:

Press [2nd] [VARS] (to display the Distr menu) and select 1:normalpdf and press [ENTER]. Enter the X value, the arithmetic mean, and the standard deviation, separated by commas, and press [ENTER].

To calculate the normal probability for a range:

Press [2nd] [VARS] (to display the Distr menu) and select 2:normalcdf and press [ENTER]. Enter the lower value, the upper value, the arithmetic mean, and the standard deviation, separated by commas, and press [ENTER].

To find a Z value from the area under the normal curve:

Press [2nd] [VARS] (to display the Distr menu) and select 3:invNorm(. Enter the area value and press [ENTER].



spreadsheet solution

Normal Probabilities

Download and open the **Chapter 5 Normal.xls** Excel file into which you can enter the mean, standard deviation, and X value(s) to determine the normal probability for several types of problems.

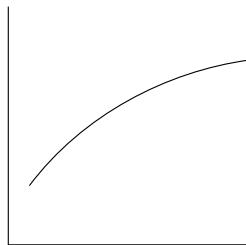
To find a Z value from the area under the normal curve, download and open the **Chapter 5 ZValue.xls** Excel file and enter the area value in the appropriate cell.

S.4 The Normal Probability Plot

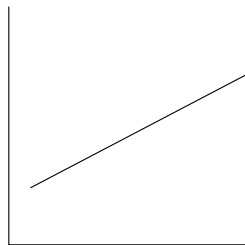
You need to establish that a set of data values follows a normal distribution in order to use many inferential statistical methods. One technique for showing that the data follow the normal distribution is the normal probability plot.

CONCEPT A graph that plots the relationship between ranked data values and the Z scores to which those values would correspond if the set of data values follows a normal distribution. If the data values follow a normal distribution, the graph will be linear (a straight line).

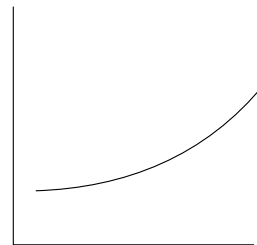
EXAMPLES



Left-Skewed



Normal

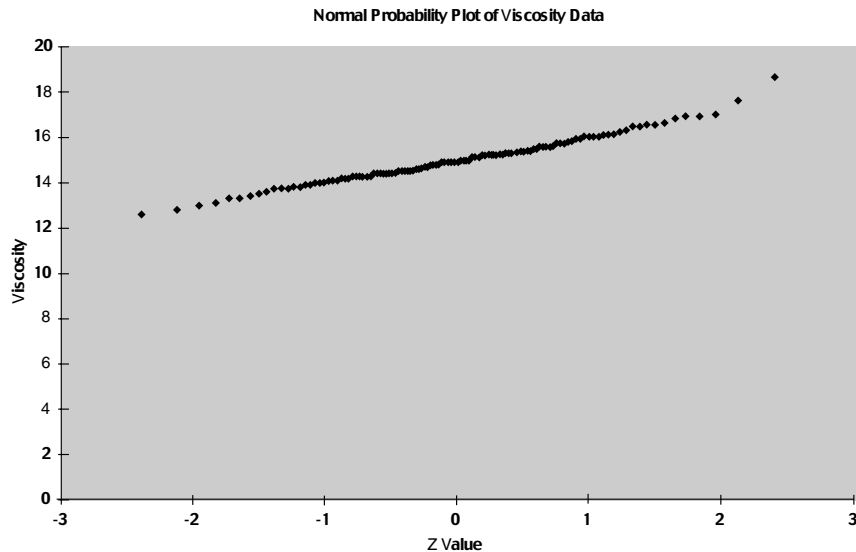


Right-Skewed

INTERPRETATION Normal probability plots are based on the idea that each ranked value will have a Z score greater than its immediate predecessor and that Z scores increase at a predictable rate in data that follow a normal distribution. The exact details to produce a normal probability plot can vary, but one common approach is called the **quantile–quantile plot**. In this method, each ranked value is transformed to a Z score and plotted along with the ranked values of the variable. If the data are normally distributed, a plot of the data in order from lowest to highest will follow a straight line. If the data are left-skewed, the curve will rise more rapidly at first, and then level off. If the data are right-skewed, the data will rise more slowly at first, and then rise

at a faster rate for higher values of the variable being plotted. These relationships are shown in the examples on page 94.

WORKED-OUT PROBLEM You seek to determine whether the viscosity measurements taken from 120 manufacturing batches (**Chemical**), first presented in Chapter 2, follows a normal distribution. You decide to use Microsoft Excel to produce the following normal probability plot:



Consistent with the results of the histogram in Section 2.3, the approximate straight line that the data follow in this normal probability plot appears to indicate that the viscosity data are approximately normally distributed.



calculator keys

Normal Probability Plots

To display a normal probability plot for a set of data values previously entered as the values of a variable, press [2nd] [Y=] to display the Stat Plot menu and select 1:Plot1 and press [ENTER]. On the Plot 1 screen, select On and press [ENTER], the sixth type choice (a thumbnail normal probability plot) and press [ENTER], and enter the name of the variable as the Data List. Press [GRAPH]. If you do not see your plot, press [ZOOM] and select 9:ZoomStat and press [ENTER] to re-center your graph on the plot.

Important Equations

The mean of a discrete probability distribution:

$$(5.1) \quad \mu = \sum_{i=1}^N X_i P(X_i)$$

The standard deviation of a discrete probability distribution:

$$(5.2) \quad \sigma = \sqrt{\sum_{i=1}^N (X_i - \mu)^2 P(X_i)}$$

The binomial distribution:

$$(5.3) \quad P(X = x) | n, p = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

The mean of the binomial distribution:

$$(5.4) \quad \mu = np$$

The standard deviation of the binomial distribution:

$$(5.5) \quad \sigma_x = \sqrt{np(1-p)}$$

The Poisson Distribution:

$$(5.6) \quad P(X = x | \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

The normal distribution:

$$(5.7) \quad Z = \frac{X - \mu}{\sigma}$$

One-Minute Summary

Probability Distributions

- Discrete probability distributions

Expected value

Variance σ^2 and standard deviation σ

Is there a fixed sample size n and is each observation classified into one of two categories?

- If yes, use the binomial distribution, subject to other conditions.
- If no, use the Poisson distribution, subject to other conditions.
- Continuous probability distributions

Normal distribution
Normal probability plot

Test Yourself

1. The expected value is most similar to the:
 - (a) mean
 - (b) median
 - (c) standard deviation
 - (d) variance
2. The largest number of possible successes in a binomial distribution is:
 - (a) 0
 - (b) 1
 - (c) n
 - (d) infinite
3. The smallest number of possible successes in a binomial distribution is:
 - (a) 0
 - (b) 1
 - (c) n
 - (d) infinite
4. Which of the following about the binomial distribution is not a true statement?
 - (a) The probability of success must be constant from trial to trial.
 - (b) Each outcome is independent of the other.
 - (c) Each outcome may be classified as either “success” or “failure.”
 - (d) The random variable of interest is continuous.
5. Whenever $p = 0.5$, the binomial distribution will:
 - (a) always be symmetric
 - (b) be symmetric only if n is large
 - (c) be right-skewed
 - (d) be left-skewed
6. What type of probability distribution will the consulting firm most likely employ to analyze the insurance claims in the following problem?

An insurance company has called a consulting firm to determine whether the company has an unusually high number of false insurance claims. It is known that the industry proportion for false claims is 6%. The consulting firm has decided to randomly and independently sample 50 of the company's insurance claims. They believe the number of these 50 that are false will yield the information the company desires.

- (a) Binomial distribution
 - (b) Poisson distribution
 - (c) Normal distribution
 - (d) None of the above
7. What type of probability distribution will most likely be used to analyze warranty repair needs on new cars in the following problem?

The service manager for a new automobile dealership reviewed dealership records of the past 20 sales of new cars to determine the number of warranty repairs he will be called on to perform in the next 30 days. Corporate reports indicate that the probability any one of their new cars needs a warranty repair in the first 30 days is 0.035. The manager assumes that calls for warranty repair are independent of one another and is interested in predicting the number of warranty repairs he will be called on to perform in the next 30 days for this batch of 20 new cars sold.

- (a) Binomial distribution
 - (b) Poisson distribution
 - (c) Normal distribution
 - (d) None of the above
8. The quality control manager of Marilyn's Cookies is inspecting a batch of chocolate chip cookies. When the production process is in control, the average number of chocolate chip parts per cookie is 9.0. The manager is interested in analyzing the probability that any particular cookie being inspected has fewer than 10.0 chip parts. What probability distribution should be used?
- (a) Binomial distribution
 - (b) Poisson distribution
 - (c) Normal distribution
 - (d) None of the above
9. The smallest number of possible successes in a Poisson distribution is:
- (a) 0
 - (b) 1
 - (c) n
 - (d) infinite
10. Based on past experience, your time spent on e-mails per day has a mean of 10 minutes and a standard deviation of 3 minutes. To compute the probability of spending at least 12 minutes on e-mails, you might use what probability distribution?
- (a) Binomial distribution
 - (b) Poisson distribution
 - (c) Normal distribution
 - (d) None of the above

11. A computer lab at a university has 100 personal computers. The probability that any one of them will require repair on a given day is 0.05. To find the probability that exactly 20 of the computers will require repair on a given day, you will use what probability distribution?
 - (a) Binomial distribution
 - (b) Poisson distribution
 - (c) Normal distribution
 - (d) None of the above
12. On the average, 1.8 customers per minute arrive at any one of the checkout counters of a grocery store. What probability distribution can be used to find out the probability that there will be no customers arriving at a checkout counter in the next minute?
 - (a) Binomial distribution
 - (b) Poisson distribution
 - (c) Normal distribution
 - (d) None of the above
13. A multiple-choice test has 30 questions. There are 4 choices for each question. A student who has not studied for the test decides to answer all questions randomly. What probability distribution can be used to figure out his chance of getting at least 20 questions right?
 - (a) Binomial distribution
 - (b) Poisson distribution
 - (c) Normal distribution
 - (d) None of the above
14. Which of the following about the normal distribution is/are true?
 - (a) Theoretically, the mean, median, and mode are the same.
 - (b) About 99.7% of the values fall within 3 standard deviations from the mean.
 - (c) It is defined by two characteristics μ and σ .
 - (d) All of the above are true.
15. Which of the following about the normal distribution is not true?
 - (a) Theoretically, the mean, median, and mode are the same.
 - (b) About two thirds of the observations fall within 1 standard deviation from the mean.
 - (c) It is a discrete probability distribution.
 - (d) Its parameters are the mean, μ , and standard deviation, σ .
16. The probability that Z is less than -1.0 is _____ the probability that Z is greater than $+1.0$.
 - (a) less than
 - (b) the same as
 - (c) greater than

17. The normal distribution is _____ in shape:
 - (a) right-skewed
 - (b) left-skewed
 - (c) symmetric
18. If a particular set of data is approximately normally distributed, you would find that approximately:
 - (a) 2 of every 3 observations would fall between 1 standard deviation around the mean
 - (b) 4 of every 5 observations would fall between 1.28 standard deviations around the mean
 - (c) 19 of every 20 observations would fall between 2 standard deviations around the mean
 - (d) All the above
19. Theoretically, the mean, median, and the mode are all equal for a normal distribution.
 - (a) True
 - (b) False
20. Another name for the mean of a probability distribution is its expected value.
 - (a) True
 - (b) False
21. The diameters of 100 randomly selected bolts follow a binomial distribution.
 - (a) True
 - (b) False
22. Suppose that a judge's decisions are upheld by an appeals court 90% of the time. In her next 10 decisions, the probability that 8 or more of her decisions are upheld by an appeals court is _____.
23. The number of power outages at a power plant has a Poisson distribution with a mean of 4 outages per year. The probability that there will be at least 3 power outages in a year is _____.
24. Given that X is a normally distributed random variable with a mean of 50 and a standard deviation of 2, the probability that X is between 47 and 54 is _____.
25. A company that sells annuities must base the annual payout on the probability distribution of the length of life of the participants in the plan. Suppose the lifetimes of the participants are approximately normally distributed with a mean of 72 years and a standard deviation of 5 years. What proportion of the plan recipients die before they reach the standard retirement age of 65?
26. The owner of a fish market determined that the average weight for salmon is 12.3 pounds with a standard deviation of 2 pounds.

Assuming the weights of salmon are normally distributed, the probability that a randomly selected salmon will weigh between 12 and 15 pounds is _____.

A venture capitalist firm that specializes in funding risky high-technology startup companies has determined that only 1 in 10 of its companies is a “success” that makes a substantive profit within 6 years. Given this historical record, what is the probability that:

27. The firm will have exactly one success in the next 3 startups it finances?

In the next 6 startups it finances, what is the probability that the firm will have:

28. Exactly 2 successes?

29. Less than 2 successes?

30. At least 2 successes?

A campus program enrolls undergraduate and graduate students. Of the students, 70% are undergraduates. If a random sample of 4 students is selected from the program to be interviewed about the introduction of a new fast-food outlet on the ground floor of the campus building, what is the probability that all 4 students selected are:

31. Undergraduate students?

32. Graduate students?

Answers to Test Yourself Questions

1. a
2. c
3. a
4. d
5. a
6. a
7. a
8. b
9. a
10. c
11. a
12. b
13. a
14. d
15. c

- 16. b
- 17. c
- 18. d
- 19. a
- 20. a
- 21. b
- 22. 0.9298
- 23. 0.7619
- 24. 0.9104
- 25. 0.0808
- 26. 0.4711
- 27. 0.2430
- 28. 0.0984
- 29. 0.8857
- 30. 0.1143
- 31. 0.2401
- 32. 0.0081

References

1. Berenson, M. L., D. M. Levine, and T. C. Krehbiel. *Basic Business Statistics: Concepts and Applications, Ninth Edition*. Upper Saddle River, NJ: Prentice Hall, 2004.
2. Gitlow, H. S., and D. M. Levine. *Six Sigma for Green Belts and Champions*. Upper Saddle River, NJ: Financial Times - Prentice Hall, 2005.
3. Levine, D. M., T. C. Krehbiel, and M. L. Berenson. *Business Statistics: A First Course, Third Edition*. Upper Saddle River, NJ: Prentice Hall, 2003.
4. Levine, D. M., D. Stephan, T. C. Krehbiel, and M. L. Berenson. *Statistics for Managers Using Microsoft Excel, Fourth Edition*. Upper Saddle River, NJ: Prentice Hall, 2005.
5. Levine, D. M., P. P. Ramsey, and R. K. Smidt. *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Upper Saddle River, NJ: Prentice Hall, 2001.
6. Microsoft Excel 2002. Redmond, WA: Microsoft Corporation, 2001.
7. Sincich, T., D. M. Levine, and D. Stephan. *Practical Statistics by Example Using Microsoft Excel and Minitab, Second Edition*. Upper Saddle River, NJ: Prentice Hall, 2002.



Sampling Distributions and Confidence Intervals

- 6.1 Sampling Distributions
 - 6.2 Basic Concepts of Confidence Intervals
 - 6.3 Confidence Interval Estimate for the Mean Using the t Distribution (σ Unknown)
 - 6.4 Confidence Interval Estimate for the Proportion
- Important Equations
- One-Minute Summary
- Test Yourself

You will recall from Section 1.1 that **inferential statistics** are defined as those in which conclusions about a large set of data, called the **population**, are made from a subset of the data, called the **sample**. Inferential statistical methods use the results of a sample statistic to draw conclusions about a population parameter by using just a single sample. Drawing conclusions about a whole thing based on looking at only a (possibly fairly) small part does not seem intuitively correct to many people and brings to mind the old joke about what would happen if a group of scientists examined different parts of the same large elephant in the dark. The “intuition” leads many to be dismissive of statistics and to wrongly claim that when you look at a sample, you “only” learn about that subset of data. (Even in the old joke, most, if not all of the scientists would probably agree that they were examining an animal and not, say, a rock or a plant.)

In this chapter, you will learn the following:

- The concept of a sampling distribution
- The concept of a confidence interval
- How to obtain confidence interval estimates for the mean and the proportion

These concepts, the first step toward learning about inferential statistics, explain and illustrate how a small part of the whole can allow you to make plausible inferences about the whole.

6.1 Sampling Distributions

Your knowledge of **sampling distributions**, combined with the probability and probability distribution concepts of the previous two chapters, provides you with the theoretical justifications that allow you to draw conclusions about an entire population based on a single sample.

Sampling Distribution

CONCEPT The distribution of a sample statistic, such as the mean, for all possible samples of a given size n .

EXAMPLES Sampling distribution of the mean, sampling distribution of the proportion.

INTERPRETATION Consider a population that includes 1,000 items. The sampling distribution of the mean for samples of 15 items consists of the mean of every single different sample of 15 items from the population. Imagine the distribution of all the means that could possibly occur: Some means would be smaller than most, some would be larger than most, and many would have similar values.

Calculating the means for all of the samples would be an involved and time-consuming task. Actually, you do not have to develop specific sampling distributions yourself, because statisticians have extensively studied sampling distributions for many different statistics, including the widely used distribution for the mean and the distribution for the proportion. These well-known sampling distributions are used extensively, starting in this chapter and continuing through Chapter 9, as a basis for inferential statistics.

Sampling Distribution of the Mean and the Central Limit Theorem

The mean is the most widely used measure in statistics. Recall from Section 3.1 that the mean is the number equal to the sum of the data values for a variable, divided by the number of data values that were summed, and that because the mean uses all the data values, it has one great weakness: an individual extreme value can distort the mean.

Through several insights—including the observation that the probability of getting such a distorted mean is relatively low, whereas the probability of

getting a mean similar to many other sample means is much greater—statisticians have developed the **central limit theorem**, which states that regardless of the shape of the distribution of the individual values in the population:

As the sample size gets *large enough*, the sampling distribution of the mean can be approximated by a normal distribution.



As a general rule, statisticians have found that for many population distributions, a sample size of at least 30 is “large enough.” However, you may be able to apply the central limit theorem for even smaller sample sizes if the distribution is known to be approximately bell-shaped. In the uncommon case in which the distribution is extremely skewed or has more than one mode, sample sizes larger than 30 may be needed in order to apply the theorem.

Figure 6.1 on page 106 contains the sampling distribution of the mean for three different populations. For each population, the sampling distribution of the sample mean is shown for all samples of $n = 2$, $n = 5$, and $n = 30$.

Panel A illustrates the sampling distribution of the mean selected from a population that is normally distributed. When the population is normally distributed, the sampling distribution of the mean will be normally distributed regardless of the sample size. If the sample size increases, the variability of the sample mean from sample to sample will decrease.

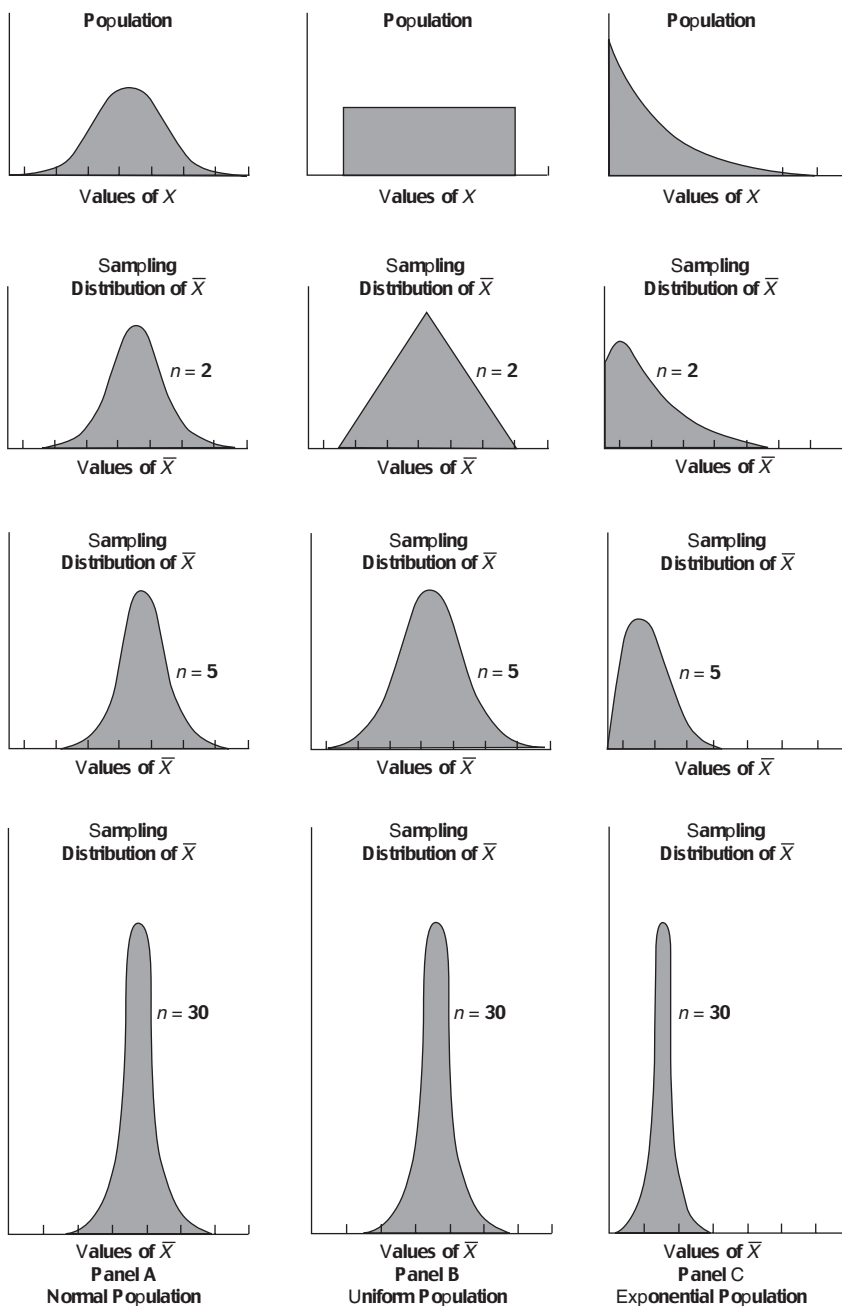
Panel B displays the sampling distribution of the mean from a population with a uniform (or rectangular) distribution. When samples of size $n = 2$ are selected, there is a **central limiting** effect already working in which there are more sample means in the center than there are individual values. For $n = 5$, the sampling distribution is bell-shaped and approximately normal. When $n = 30$, the sampling distribution appears to be very similar to a normal distribution. In general, the larger the sample size, the more closely the sampling distribution will follow a normal distribution. As with all cases, the mean of each sampling distribution is equal to the mean of the population, and the variability decreases as the sample size increases.

Panel C presents an exponential distribution. This population is heavily skewed to the right. When $n = 2$, the sampling distribution is still highly skewed to the right, but less so than the distribution of the population. For $n = 5$, the sampling distribution of the mean is only slightly skewed to the right. When $n = 30$, the sampling distribution appears to be approximately normally distributed. Again, the mean of each sampling distribution is equal to the mean of the population and the variability decreases as the sample size increases.

Observations, such as those just made from using this figure, allow you to state the following relationships between the normal distribution and the sampling distribution of the mean:

- For most population distributions, regardless of shape, the sampling distribution of the mean is approximately normally distributed, if samples of at least 30 observations are selected.

FIGURE 6.1



- If the population distribution is fairly symmetrical, the sampling distribution of the mean is approximately normal, if samples of at least 15 observations are selected.
- If the population is normally distributed, the sampling distribution of the mean is normally distributed, regardless of the sample size.

Sampling Distribution of the Proportion

Recall from Section 5.2 that a binomial distribution can be used to determine probabilities for categorical variables that have only two categories, traditionally labeled “success” and “failure.” As the sample size increases for such variables, the normal distribution can be used to approximate the sampling distribution of the number of successes or the proportion of successes. Specifically, as a general rule, the normal distribution can be used to approximate the binomial distribution when the average number of successes and the average number of failures are *each* at least 5. For most cases in which you are estimating the proportion, the sample size will be more than sufficient to meet the conditions for using the normal approximation.

What You Need to Know About Sampling Distributions



Sampling distributions are the key to making the statistical inferences that are discussed in the remainder of this chapter and continuing through Chapter 9. Remember the following aspects of sampling distributions as you read through those sections:

- Every sample statistic has a sampling distribution.
- A specific sample statistic is used to estimate its corresponding population characteristic.

6.2 Sampling Error and Confidence Intervals

Taking one sample and obtaining the results of a sample statistic, such as the mean, creates a **point estimate** of the population parameter. This single estimate will almost certainly not be the same if a different sample is selected. For example, the results generated from the 20 different samples of size 10 from chemical viscosity data for 120 batches with the population mean, μ , 14.978 and the population standard deviation, σ , 1.003, first presented in Chapter 2, are shown in Table 6.1.

Table 6.1*Results from 20 Samples of $n = 10$ Selected from a Population of $N = 120$*

Sample	Mean	Standard Deviation	Minimum	Median	Maximum	Range
1	14.69	0.858	13.4	14.60	16.1	2.7
2	15.31	0.590	14.5	15.30	16.3	1.8
3	14.65	0.824	13.3	14.65	16.0	2.7
4	14.91	1.049	13.7	14.85	16.9	3.2
5	14.78	0.847	13.7	14.65	16.5	2.8
6	14.63	1.145	12.6	14.35	16.6	4.0
7	14.61	1.034	12.8	14.90	16.4	3.6
8	15.04	1.422	13.6	14.70	18.6	5.0
9	15.34	1.055	13.7	15.65	16.8	3.1
10	15.37	0.572	14.3	15.30	16.2	1.9
11	15.23	0.864	14.2	15.10	16.9	2.7
12	15.16	0.749	14.4	14.95	16.9	2.5
13	15.12	0.840	14.0	15.15	16.6	2.6
14	14.86	0.696	14.2	14.75	16.5	2.3
15	15.68	0.750	14.4	15.45	17.0	2.6
16	15.13	0.699	14.0	15.15	16.3	2.3
17	14.47	0.715	13.3	14.60	15.3	2.0
18	15.25	0.985	13.8	15.15	16.8	3.0
19	14.72	0.888	13.3	14.50	16.0	2.7
20	15.40	0.968	14.3	15.15	17.6	3.3

From these results, note the following:

- The sample statistics differ from sample to sample. The sample means vary from 14.47 to 15.68, the sample standard deviations vary from 0.572 to 1.422, the sample medians vary from 14.35 to 15.65, and the sample ranges vary from 1.8 to 3.6.
- Some of the sample means are higher than the population mean of 14.978, and some of the sample means are lower than the population mean.
- Some of the sample standard deviations are higher than the population standard deviation of 1.003, and some of the sample standard deviations are lower than the population standard deviation.

- The variation in the sample range from sample to sample is much more than the variation in the sample standard deviation.

You should note that sample statistics almost always vary from sample to sample. This expected variation is called the **sampling error**.

Sampling Error

CONCEPT The variation that occurs due to selecting a single sample from the population.

EXAMPLE In polls, the plus-or-minus margin of the results; as in “42%, plus or minus 3%, said they were likely to vote for the incumbent.”

INTERPRETATION The size of the sampling error is primarily based on the variation in the population itself and on the size of the sample selected. Larger samples will have less sampling error, but will be more costly to obtain.

In practice, only one point estimate (that is, one sample) is used as the basis for estimating a population parameter. To account for the differences among the point estimates from each of the samples, statisticians have developed the concept of a **confidence interval estimate** which indicates the likelihood that a stated interval with a lower and upper limit properly estimates the parameter.

Confidence Interval Estimate

CONCEPT An estimate of a population parameter stated as a range between a lower and upper limit with a specific degree of certainty.

INTERPRETATION All that you need to know to develop a confidence interval estimate is the sample statistic used to estimate the population parameter and the sampling distribution for the sample statistic. This is always true regardless of the population parameter being estimated.

Because you are developing an interval using one sample and not precisely determining a value, there is no way that you can be 100% certain that your interval correctly estimates the population parameter as noted earlier and illustrated by the Worked-out Problem that appears on page 110. However, by setting the level of certainty to a value below 100%, you can use the interval estimate to obtain plausible inferences about the population with that given degree of certainty.

There is a trade-off between the level of confidence and the width of the interval. For a given sample size, if you want more confidence that your interval will be correct, you will have a wider interval and therefore a less-precise estimate.

Given this trade-off, what level of certainty should you use? Expressed as a percentage, the most common level of certainty used is 95%. If more confidence is needed, 99% is typically used; if less confidence is needed, 90% is typically used.

Because of this factor, the degree of certainty, or **confidence**, must always be stated when reporting an interval estimate. When you hear an “interval estimate with 95% confidence,” or simply, a “95% confidence interval estimate,” you can conclude that if all possible samples of the same size n were selected, 95% of them would include the population parameter somewhere within the interval and 5% would not.

WORKED-OUT PROBLEM You seek to develop 95% confidence interval estimates for the mean from 20 samples of size 10 for the chemical viscosity data for 120 batches first presented in Chapter 2. Unlike most real-life problems, the population mean, μ , 14.978, and the population standard deviation, σ , 1.003, are already known, so the confidence interval estimate for the mean developed from each sample can be compared to the actual value of the population mean.

Table 6.2

95% Confidence Interval Estimates from 20 Samples of $n = 10$ Selected from a Population of $N = 120$

	$\mu=14.978$	$\sigma=1.003$	95% Confidence	
Sample	Mean	Standard deviation	Lower Limit	Upper Limit
1	14.69	0.858	14.0683	15.3117
2	15.31	0.590	14.6883	15.9317
3	14.65	0.824	14.0283	15.2717
4	14.91	1.049	14.2883	15.5317
5	14.78	0.847	14.1583	15.4017
6	14.63	1.145	14.0083	15.2517
7	14.61	1.034	13.9883	15.2317
8	15.04	1.422	14.4183	15.6617
9	15.34	1.055	14.7183	15.9617
10	15.37	0.572	14.7483	15.9917
11	15.23	0.864	14.6083	15.8517
12	15.16	0.749	14.5383	15.7817
13	15.12	0.840	14.4983	15.7417
14	14.86	0.696	14.2383	15.4817
15	15.68	0.750	15.0583	16.3017

(continues)

	$\mu=14.978$	$\sigma=1.003$	95% Confidence	
Sample	Mean	Standard deviation	Lower Limit	Upper Limit
16	15.13	0.699	14.5083	15.7517
17	14.47	0.715	13.8483	15.0917
18	15.25	0.985	14.6283	15.8717
19	14.72	0.888	14.0983	15.3417
20	15.40	0.968	14.7783	16.0217

From the results, you can conclude the following:

- For sample 1, the sample mean is 14.69, the sample standard deviation is 0.858, and the interval estimate for the population mean is 14.0683 to 15.3117. This allows you to conclude with 95% certainty that the population mean is between 14.0683 and 15.3117. This is a correct estimate, because the population mean of 14.978 is included within this interval. Although their sample means and standard deviations differ, the confidence interval estimates for samples 2 through 14 and 16 through 20 lead to an interval estimate that includes the population mean value.
- For sample 15, the sample mean is 15.68, the sample standard deviation is 0.750, and the interval estimate for the population mean is 15.0583 to 16.3017 (highlighted in the results). This is an incorrect estimate, because the population mean of 14.978 is not included within this interval.

You note that these results are not surprising, because the percentage of correct results (19 out of 20) is 95%, just as statistical theory would claim. Of course, with other specific sets of 20 samples, the percentage of correct results might not be exactly 95%—it could be higher or lower—but in the long run, 95% of all samples used will result in a correct estimate.

6.3 Confidence Interval Estimate for the Mean Using the t Distribution (σ Unknown)

The most common confidence interval estimate involves estimating the mean of a population. In virtually all cases, the population mean is estimated from sample data in which only the sample mean and sample standard deviation—and not the population standard deviation—are available. To sidestep this complication, statisticians (see Reference 1) have developed the t distribution.

***t* Distribution**

CONCEPT The sampling distribution that allows you to develop a confidence interval estimate of the mean using the sample standard deviation.

INTERPRETATION The *t* distribution assumes that the variable being studied is normally distributed. In practice, however, as long as the sample size is large enough and the population is not very skewed, the *t* distribution can be used to estimate the population mean when the population standard deviation σ is unknown. You should be concerned about the validity of the confidence interval primarily when dealing with a small sample size and a skewed population distribution. The assumption of normality in the population can be assessed by evaluating the shape of the sample data using a histogram, box-and-whisker plot, or normal probability plot.

WORKED-OUT PROBLEM You seek to determine the average increase in tuition costs for both in-state and out-of-state students attending public universities during a one-year period for a sample of 67 universities. Table 6.3 contains the change in tuition costs for in-state students, and Table 6.4 contains the change in tuition for out-of-state students for this sample.

(Tuition)

Table 6.3

Change in Tuition for In-State Students for a Sample of 67 Universities

638	176	617	116	876	609	1,442	303	604	274
642	462	572	676	274	359	1,522	1,202	490	243
236	448	434	210	1,324	291	454	280	836	1,048
1,021	116	364	534	353	730	658	918	0	79
1,010	312	861	625	794	308	738	802	1,006	
312	262	1750	144	1,100	711	189	616	70	
1,001	354	1,121	394	220	303	792	642	494	

Table 6.4

Change in Tuition for Out-of-State Students for a Sample of 67 Universities

1,730	1,660	890	703	1,038	1,975	3,353	1,627	1,413	2,171
1,802	1,868	700	1,665	721	677	1,426	1,380	1,892	784
750	1,702	434	912	1,718	743	1,747	1,568	0	1,308
1,270	116	994	1,194	1,015	1,082	738	1,452	1,300	1,124
1,008	672	2,333	1,157	2,012	1,058	1,754	1,246	996	
690	2,694	2,380	518	1,600	711	1,489	747	843	
3,273	2,473	1,497	2,051	847	265	1,533	1,204	776	

important point

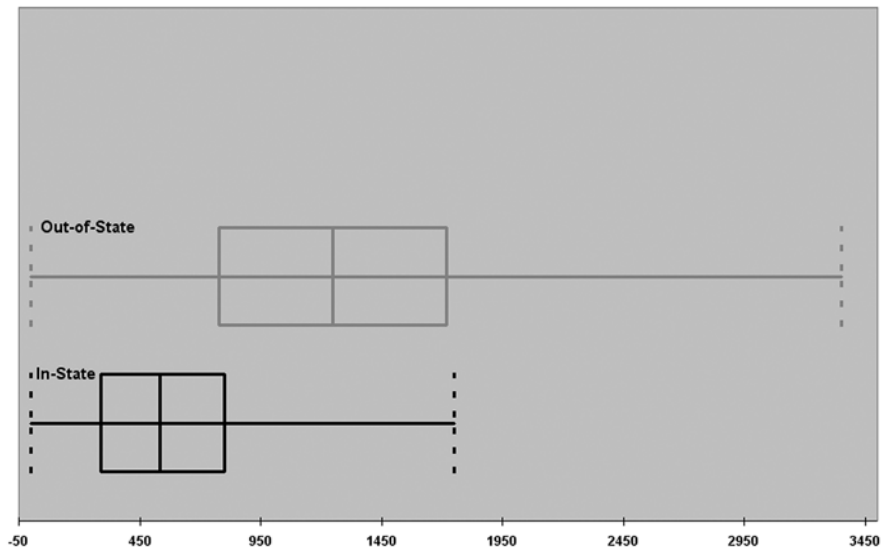


The confidence interval estimate of the population mean prepared in Microsoft Excel for the average change in tuition costs for in-state students and the average change in tuition costs for out-of-state students is as follows:

	A	B	C
1	Confidence Interval Estimate for the Mean		
2			
3	Data	In-State	Out-of-State
4	Sample Standard Deviation	378.8789	670.5598
5	Sample Mean	587.4925	1320.4328
6	Sample Size	67	67
7	Confidence Level	95%	95%
8			
9	Intermediate Calculations		
10	Standard Error of the Mean	46.2874	81.9219
11	Degrees of Freedom	66	66
12	t Value	1.9966	1.9966
13	Interval Half Width	92.4158	163.5624
14			
15	Confidence Interval	In-State	Out-of-State
16	Interval Lower Limit	495.0767	1156.8704
17	Interval Upper Limit	679.9084	1483.9952

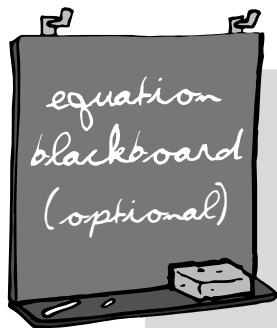
To evaluate the assumption of normality necessary to use these estimates, you develop box-and-whisker plots for the in-state and the out-of-state tuition increases (shown below).

Box-and-Whisker Plot for Tuition Study Data



Note that the box-and-whisker plots contain a long tail on the right, indicating right-skewness due to some large changes in tuition. However, you can also observe that for both of the box-and-whisker plots, the values in the box between the first and third quartiles are symmetric. Given the relatively large sample size, you can conclude that any departure from the normality assumption will not seriously affect the validity of the confidence interval estimate.

Based on these results, with 95% confidence you can conclude that the average change in tuition costs for in-state students is between \$495.08 and \$679.91 and between \$1,156.87 and \$1,484.00 for out-of-state students. You conclude that tuition costs have increased by different amounts for the two groups of students.



interested
in
math?

You take the symbols \bar{X} (sample mean), μ (population mean), S (sample standard deviation), and n (sample size), all introduced earlier, and include the new symbol t_{n-1} , which represents the critical value of the t distribution with $n - 1$ degrees of freedom for an area of $\alpha/2$ in the upper tail, to write the formula for the confidence interval for the mean in cases in which the population standard deviation, σ , is unknown. For the symbol t_{n-1} :

$n - 1$ is one less than the sample size

α is equivalent to 1 minus the confidence percentage. For 95% confidence, α is 0.05 ($1 - 0.95$), so the upper tail area is 0.025.

Using these symbols creates the following equation:

$$\bar{X} \pm t_{n-1} \frac{S}{\sqrt{n}}$$

or expressed as a range

$$\bar{X} - t_{n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1} \frac{S}{\sqrt{n}}$$

For the worked-out tuition costs problem of this section, $\bar{X} = 587.4925$, and $S = 378.8789$, and because the sample size is 67, there are 66 degrees of freedom. Given 95% confidence, α is 0.05, and the area in the upper tail of the t distribution is 0.025 ($0.05/2$). Using Table C.2, the critical value for the row

(continues)

with 66 degrees of freedom and the column with an area of 0.025 is 1.9966. Substituting these values yields the following result:

$$\begin{aligned}\bar{X} \pm t_{n-1} \frac{S}{\sqrt{n}} \\ &= 587.4925 \pm (1.9966) \frac{378.8789}{\sqrt{67}} \\ &= 587.4925 \pm 92.4158 \\ 495.08 &\leq \mu \leq 679.91\end{aligned}$$

The interval is estimated to be between \$495.08 and \$679.91 with 95% confidence.



calculator keys

Confidence Interval Estimate for the Mean When σ Is Unknown

Press [STAT] [◀] (to display the Tests menu) and select 8:TInterval and press [ENTER] to display the TInterval screen. In this screen, select Stats as the Inpt type and press [ENTER]. Enter values for the sample mean \bar{X} , the sample standard deviation S_x , and the sample size, n . Also enter the confidence level (C-Level) as the decimal fraction equivalent to a percentage—for example, .95 for 95% (see the first illustration below). Select Calculate and press [ENTER]. The lower and upper limits of interval estimate will appear as an ordered pair of values enclosed in parentheses as shown in second illustration.

```
TInterval
Inpt:Data Stats
x̄:587.4925
Sx:378.8789
n:67
C-Level:.95
Calculate
```

```
TInterval
(495.08,679.91)
x̄=587.4925
Sx=378.8789
n=67
```

(continues)



spreadsheet solution

Confidence Interval Estimate for the Mean When σ is Unknown

Download and open the **Chapter 6 SigmaUnknown.xls** Excel file into which you can enter the values for the sample standard deviation, the sample mean, the sample size, and the confidence level as a percentage.

6.4 Confidence Interval Estimation for the Proportion

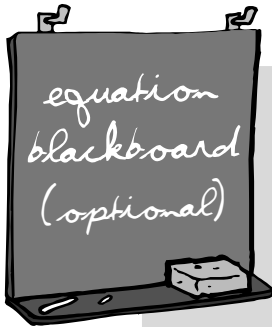
A confidence interval estimate for a categorical variable can be developed to estimate the proportion of successes in a given category. Instead of using the sample mean to estimate the population mean, you use the sample proportion of successes, equal to the number of successes divided by the sample size, to estimate the population proportion. The sample statistic p follows a binomial distribution which can be approximated by the normal distribution for most studies.

For a given sample size, confidence intervals for proportions are wider than those for numerical variables. With continuous variables, the measurement on each respondent contributes more information than for a categorical variable. In other words, a categorical variable with only two possible values is a very crude measure compared with a continuous variable, so each observation contributes only a little information about the parameter being estimated.

WORKED-OUT PROBLEM You want to estimate the proportion of newspapers that are printed in which a nonconforming attribute is present (such as excessive rub-off, improper page setup, missing pages, or duplicate pages). You select a random sample of 200 newspapers and discover that 35 contain some type of nonconformance.

Based on the 95% confidence interval estimate prepared in Microsoft Excel for the percentage of nonconforming newspapers (see the following figure), you estimate that between 12.2% and 22.8% of the newspapers printed have some type of nonconformance.

	A	B
1	Confidence Interval Estimate for the Proportion	
2		
3	Data	
4	Sample Size	200
5	Number of Successes	35
6	Confidence Level	95%
7		
8	Intermediate Calculations	
9	Sample Proportion	0.1750
10	Z Value	-1.9600
11	Standard Error of the Proportion	0.0269
12	Interval Half Width	0.0527
13		
14	Confidence Interval	
15	Interval Lower Limit	0.1223
16	Interval Upper Limit	0.2277



You take the symbols p (sample proportion of success), n (sample size), and Z (Z score), previously introduced, and the symbol π for the population proportion, to assemble the equation for the confidence interval estimate for the proportion:

$$p \pm Z \sqrt{\frac{p(1-p)}{n}}$$

or expressed as a range

$$p - Z \sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + Z \sqrt{\frac{p(1-p)}{n}}$$

For the Worked-out Problem, $n = 200$ and $p = 35/200 = 0.175$. For a 95% level of confidence, the lower tail area of 0.025 provides a Z value from the normal distribution of -1.96 , and the upper tail area of 0.025 provides a Z value from the normal distribution of $+1.96$. Substituting these numbers yields the following result:

(continues)

interested
in
math?

$$\begin{aligned}
 p \pm z \sqrt{\frac{p(1-p)}{n}} \\
 &= 0.175 \pm (1.96) \sqrt{\frac{(0.175)(0.825)}{200}} \\
 &= 0.175 \pm (1.96)((0.0269) \\
 &= 0.175 \pm 0.053 \\
 0.122 \leq \pi \leq 0.228
 \end{aligned}$$

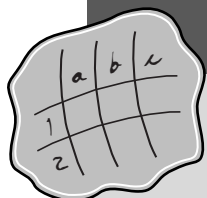
The proportion of nonconforming newspapers is estimated to be between 12.2% and 22.8%.



calculator keys

Confidence Interval Estimate for the Proportion

Press [STAT] [◀] (to display the Tests menu) and select A:1-PropZInt to display the 1-PropZInt screen. In this screen, enter values for the number of successes x , the sample size n , and the confidence level (C-Level) as the decimal fraction equivalent to a percentage (for example, .95 for 95%). Select Calculate and press [ENTER]. The lower and upper limits of interval estimate will appear as an ordered pair of values enclosed in parentheses.



spreadsheet solution

Confidence Interval Estimate for the Proportion

Download and open the **Chapter 6 Proportion.xls** Excel file into which you can enter the values for the sample size, the number of successes, and the confidence level as a percentage.

Important Equations

Confidence interval for the mean with σ unknown:

$$\bar{X} \pm t_{n-1} \frac{S}{\sqrt{n}}$$

(6.1) or

$$\bar{X} - t_{n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1} \frac{S}{\sqrt{n}}$$

Confidence interval estimate for the proportion:

$$p \pm Z \sqrt{\frac{p(1-p)}{n}}$$

(6.2) or

$$p - Z \sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + Z \sqrt{\frac{p(1-p)}{n}}$$

One-Minute Summary

For what type of variable are you developing a confidence interval estimate?

- If it is a numerical variable, use the confidence interval estimate for the mean.
- If it is a categorical variable, use the confidence interval estimate for the proportion.

Test Yourself

1. The sampling distribution of the mean can be approximated by the normal distribution:
 - (a) as the number of samples gets “large enough”
 - (b) as the sample size (number of observations in each sample) gets large enough
 - (c) as the size of the population standard deviation increases
 - (d) as the size of the sample standard deviation decreases
2. The sampling distribution of the mean requires _____ sample size to reach a normal distribution if the population is skewed than if the population is symmetrical.
 - (a) the same
 - (b) a smaller
 - (c) a larger
 - (d) The two distributions cannot be compared

3. Which of the following is true regarding the sampling distribution of the mean for a large sample size?
 - (a) It has the same shape and mean as the population.
 - (b) It has a normal distribution with the same mean as the population.
 - (c) It has a normal distribution with a different mean from the population.
4. For sample of $n = 30$, the sampling distribution of the mean will be approximately normally distributed:
 - (a) regardless of the shape of the population
 - (b) only if the shape of the population is symmetrical
 - (c) if the standard deviation of the mean is known
 - (d) only if the population is normally distributed
5. For sample of $n = 1$, the sampling distribution of the mean will be normally distributed:
 - (a) regardless of the shape of the population
 - (b) only if the shape of the population is symmetrical
 - (c) if the standard deviation of the mean is known
 - (d) only if the population is normally distributed
6. A 99% confidence interval estimate can be interpreted to mean that:
 - (a) If all possible samples are taken and confidence interval estimates are developed, 99% of them would include the true population mean somewhere within their interval
 - (b) You have 99% confidence that you have selected a sample whose interval does include the population mean
 - (c) both a and b are true
 - (d) neither a nor b is true
7. Which of the following statements is false?
 - (a) There is a different critical value for each level of alpha.
 - (b) Alpha is the proportion in the tails of the distribution that is outside the confidence interval.
 - (c) You can construct a 100% confidence interval estimate of μ .
 - (d) In practice, the population mean is the unknown quantity that is to be estimated.
8. Sampling distributions describe the distribution of:
 - (a) parameters
 - (b) statistics
 - (c) both parameters and statistics
 - (d) neither parameters nor statistics

9. In the construction of confidence intervals, if all other quantities are unchanged, an increase in the sample size will lead to _____ interval.
- (a) a narrower
 - (b) a wider
 - (c) a less significant
 - (d) the same
10. As an aid to the establishment of personnel requirements, the manager of a bank wants to estimate the mean number of people who arrive at the bank during the two-hour lunch period from 12 noon to 2 p.m. The director randomly selects 64 different two-hour lunch periods from 12 noon to 2 p.m. and determines the number of people who arrive for each. For this sample, $\bar{X} = 49.8$ and $S^2 = 25$. Which of the following assumptions is necessary in order for a confidence interval to be valid?
- (a) The population sampled from has an approximate normal distribution.
 - (b) The population sampled from has an approximate t distribution.
 - (c) The mean of the sample equals the mean of the population.
 - (d) None of these assumptions are necessary.
11. A university dean is interested in determining the proportion of students who are planning to attend graduate school. Rather than examine the records for all students, the dean randomly selects 200 students and finds that 118 of them are planning to attend graduate school. The 95% confidence interval for p is 0.59 ± 0.07 . Interpret this interval.
- (a) You are 95% confident that the true proportion of all students planning to attend graduate school is between 0.52 and 0.66.
 - (b) There is 95% chance of selecting a sample that finds that between 52% and 66% of the students are planning to attend graduate school.
 - (c) You are 95% confident that between 52% and 66% of the sampled students are planning to attend graduate school.
 - (d) You are 95% confident that 59% of the students are planning to attend graduate school.
12. Other things being equal, as the confidence level for a confidence interval increases, the width of the interval increases.
- (a) True
 - (b) False
13. In estimating the population mean with the population standard deviation unknown, if the sample size is 12, there will be _____ degrees of freedom.

14. As the sample size increases, the effect of an extreme value on the sample mean becomes smaller.
 - (a) True
 - (b) False
15. A sampling distribution is defined as the probability distribution of possible sample sizes that can be observed from a given population.
 - (a) True
 - (b) False
16. The t distribution is used to construct confidence intervals for the population mean when the population standard deviation is unknown.
 - (a) True
 - (b) False

Answers to Test Yourself Questions

1. b
2. c
3. b
4. a
5. d
6. c
7. c
8. b
9. a
10. d
11. a
12. a
13. 11
14. a
15. b
16. a

References

1. Berenson, M. L., D. M. Levine, and T. C. Krehbiel. *Basic Business Statistics: Concepts and Applications, Ninth Edition*. Upper Saddle River, NJ: Prentice Hall, 2004.
2. Cochran, W. G. *Sampling Techniques, Third Edition*. New York: Wiley, 1977.

3. Gitlow, H. S., and D. M. Levine. *Six Sigma for Green Belts and Champions*. Upper Saddle River, NJ: Financial Times - Prentice Hall, 2005.
4. Levine, D. M., T. C. Krehbiel, and M. L. Berenson. *Business Statistics: A First Course, Third Edition*. Upper Saddle River, NJ: Prentice Hall, 2003.
5. Levine, D. M., D. Stephan, T. C. Krehbiel, and M. L. Berenson. *Statistics for Managers Using Microsoft Excel, Fourth Edition*. Upper Saddle River, NJ: Prentice Hall, 2005.
6. Levine, D. M., P. P. Ramsey, and R. K. Smidt, *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Upper Saddle River, NJ: Prentice Hall, 2001.
7. Microsoft Excel 2002. Redmond, WA: Microsoft Corporation, 2001.
8. Sincich, T., D. M. Levine, and D. Stephan, *Practical Statistics by Example Using Microsoft Excel and Minitab, Second Edition*. Upper Saddle River, NJ: Prentice Hall, 2002.

This page intentionally left blank



Fundamentals of Hypothesis Testing

7.1 The Null and Alternative Hypotheses

7.2 Hypothesis Testing Issues

7.3 Decision-Making Risks

7.4 Performing Hypothesis Testing

7.5 Types of Hypothesis Tests

Important Equations

One-Minute Summary

Test Yourself

Science progresses by first stating tentative explanations, or **hypotheses**, about natural phenomena and then by proving (or disproving) those explanations through investigation and testing. This **scientific method** has been adapted by statisticians to an inferential method called **hypothesis testing** that seeks to evaluate a claim made about the value of a population parameter by using a sample statistic. In this chapter, you will learn the basic concepts and principles of hypothesis testing and the statistical assumptions necessary for performing hypothesis testing.

7.1 The Null and Alternative Hypotheses

Unlike the broader hypothesis testing of science, statistical hypothesis testing always involves evaluating a claim made about the value of a population parameter. This claim is stated as a pair of statements: the *null hypothesis* and the *alternative hypothesis*.

Null Hypothesis

CONCEPT The statement that a population parameter is equal to a specific value, or that the population parameters from two or more groups are equal.

EXAMPLES “The population mean time to answer customer complaints was 4 minutes in 2003,” “the average height for women is the same as the average height for men,” “the proportion of food orders filled correctly for drive-through customers is the same as the proportion of food orders filled correctly for sit-down customers.”

INTERPRETATION The null hypothesis *always* expresses an equality, either between a population parameter and a specific value or between two or more population parameters, and is always paired with another statement, the **alternative hypothesis**. You use the symbol H_0 to identify the null hypothesis and write a null hypothesis using an equal sign and the symbol for the population parameter, as in $H_0: \mu = 4$ or $H_0: \mu_1 = \mu_2$ or $H_0: \pi_1 = \pi_2$. (Recall that in statistics the symbol π represents the population proportion and not the ratio of the circumference to the diameter of a circle, as the symbol represents in geometry.)



A null hypothesis is considered true until evidence indicates otherwise. If you can conclude that the null hypothesis is false, then **the alternative hypothesis must be true**.

Alternative Hypothesis

CONCEPT The statement paired with a null hypothesis that is mutually exclusive to the null hypothesis.

EXAMPLES “The population mean for the time to answer customer complaints was not 4 minutes in 2003” (which would be paired with the first null hypothesis example above); “the average height for women is not the same as the average height for men” (paired with the second null hypothesis); “the proportion of food orders filled correctly for drive-through customers is not the same as the proportion of food orders filled correctly for sit-down customers” (paired with the third null hypothesis example).

INTERPRETATION The alternative hypothesis is typically the idea you are studying about your data. The alternative hypothesis *always* expresses an inequality, either between a population parameter and a specific value or between two or more population parameters and is always paired with the null hypothesis. You use the symbol H_1 to identify the alternative hypothesis and write an alternative hypothesis using either a not-equal sign or a less than or greater than sign, along with the symbol for the population parameter, as in $H_1: \mu \neq 2$ or $H_1: \mu_1 \neq \mu_2$ or $H_0: \pi_1 \neq \pi_2$.



The alternative hypothesis represents the conclusion reached by rejecting the null hypothesis. You reject the null hypothesis if evidence from the sample

statistic indicates that the null hypothesis is unlikely to be true. However, if you cannot reject the null hypothesis, you *cannot claim to have proven the null hypothesis*. Failure to reject the null hypothesis means (only) that you have failed to prove the alternative hypothesis.

7.2 Hypothesis Testing Issues

In hypothesis testing, you use the sample statistic to estimate the population parameter named in the null hypothesis. For example, to evaluate the null hypothesis “the population mean time to answer customer complaints was 4 minutes in 2003,” you would use the sample mean time to estimate the population mean time. As Chapter 6 establishes, a sample statistic is unlikely to be identical to its corresponding population parameter, and in that chapter you learned to apply sampling distributions to develop an interval estimate for the parameter based on the statistic.

If the sample statistic is not the same as the population parameter, as it almost never is, the issue of whether to reject the null hypothesis involves deciding how dissimilar the sample statistic is to its corresponding population parameter. (In the case of two groups, the issue can be expressed, under certain conditions, as deciding how dissimilar the sample statistics of each group are to each other.)

Without a rigorous procedure that includes a clear operational definition of dissimilarity, you would find it hard to decide on a consistent basis whether a null hypothesis is false and, therefore, whether to reject or not reject the null hypothesis. Statistical hypothesis-testing methods provide such definitions and enable you to restate the decision-making process as the probability of obtaining a given sample statistic, if the null hypothesis were true through the use of a test statistic and a risk factor.

important point



Test Statistic

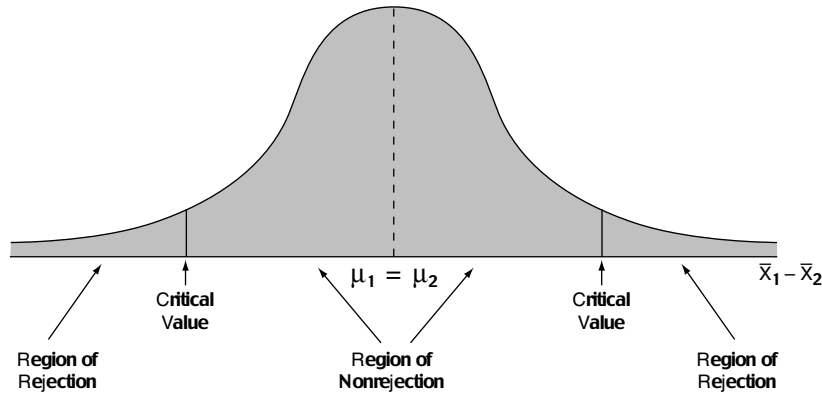
CONCEPT The value based on the sample statistic and the sampling distribution for the sample statistic.

EXAMPLES If you are testing whether the mean of a population was equal to a specific value, the sample statistic is the sample mean. The test statistic is based on the difference between the sample mean and the value of the population mean stated in the null hypothesis. This test statistic follows a statistical distribution called the t distribution that is discussed in Sections 8.2 and 8.3.

If you are testing whether the mean of population one is equal to the mean of population two, the sample statistic is the difference between the mean in sample one and the mean in sample two. The test statistic is based on the

difference between the mean in sample one and the mean in sample two. This test statistic also follows the t distribution.

INTERPRETATION The sampling distribution of the test statistic is divided into two regions, a **region of rejection** (also known as the **critical region**) and a **region of nonrejection**. If the test statistic falls into the region of nonrejection, the null hypothesis is not rejected.



The region of rejection contains the values of the test statistic that are unlikely to occur if the null hypothesis is true. If the null hypothesis is false, these values are likely to occur. Therefore, if you observe a value of the test statistic that falls into the rejection region, the null hypothesis is rejected, because that value is unlikely if the null hypothesis is true.

To make a decision concerning the null hypothesis, you first determine the **critical value** of the test statistic that separates the nonrejection region from the rejection region. You determine the critical value by using the appropriate sampling distribution and deciding on the risk you are willing to take of rejecting the null hypothesis when it is true.

Practical Significance Versus Statistical Significance

Another issue in hypothesis testing concerns the distinction between a statistically significant difference and a practically significant difference. Given a *large enough* sample size, it is always possible to detect a statistically significant difference. This is because no two things in nature are exactly equal. So, with a large enough sample size, you can always detect the natural difference between two populations. You need to be aware of the real-world practical implications of the statistical significance.

important point

7.3 Decision-Making Risks

In hypothesis testing, you always face the possibility that either you will wrongly reject the null hypotheses or wrongly not reject the null hypothesis. These possibilities are labeled *type I* and *type II errors*, respectively.

Type I Error

CONCEPT The error that occurs if the null hypothesis H_0 is rejected when in fact it is true and should not be rejected.

INTERPRETATION The risk, or probability, of a type I error occurring is identified by the Greek lowercase alpha, α . Alpha is also known as the **level of significance** of the statistical test. Traditionally, you control the probability of a type I error by deciding the risk level α you are willing to tolerate of rejecting the null hypothesis when it is true. Because you specify the level of significance before performing the hypothesis test, the risk of committing a type I error, α , is directly under your control. The most common α values are 0.01, 0.05, and 0.10, and researchers traditionally select a value of 0.05 or smaller.

When you specify the value for α , you determine the rejection region, and using the appropriate sampling distribution, the critical value or values that divide the rejection and nonrejection regions are determined.

Type II Error

CONCEPT The error that occurs if the null hypothesis H_0 is not rejected when in fact it is false and should be rejected.

INTERPRETATION The risk, or probability, of a type II error occurring is identified by the Greek lowercase beta, β . The probability of a type II error depends on the size of the difference between the value of the population parameter stated in the null hypothesis and the actual population value. Unlike the type I error, the type II error is not directly established by you. Because large differences are easier to find, as the difference between the value of the population parameter stated in the null hypothesis and its corresponding population parameter increases, the probability of a type II error decreases. Therefore, if the difference between the value of the population parameter stated in the null hypothesis and the corresponding parameter is small, the probability of a type II error will be large.

The arithmetic complement of beta, $1 - \beta$, is known as the **power of the test** and represents the probability of rejecting the null hypothesis when it is false and should be rejected.

Risk Trade-Off

The types of errors and their associated risks are summarized in Table 7.1. The probabilities of the two types of errors have an inverse relationship. When you decrease α , you always increase β and when you decrease β , you always increase α .

Table 7.1
Risks and Decisions in Hypothesis Testing

		Actual Situation	
		H_0 True	H_0 False
Statistical Decision	Do not reject H_0	Correct decision Confidence = $1 - \alpha$	Type II error $P(\text{Type II error}) = \beta$
	Reject H_0	Type I error $P(\text{Type I error}) = \alpha$	Correct decision Power = $1 - \beta$

One way in which you can lower β without affecting the value of α is to increase the size of the sample. Larger sample sizes generally permit you to detect even very small differences between the hypothesized and actual values of the population parameter. For a given level of α , increasing the sample size will decrease β and therefore increase the power of the test to detect that the null hypothesis H_0 is false.

In establishing a value for α , you need to consider the negative consequences of a type I error. If these consequences are substantial, you can set $\alpha = 0.01$ instead of 0.05 and tolerate the greater β that results. If the negative consequences of a type II error most concern you, you can select a larger value for α (for example, 0.05 rather than 0.01) and benefit from the lower β that results.

7.4 Performing Hypothesis Testing

When you perform a hypothesis test, you should follow the steps of hypothesis testing in this order:

1. State the null hypothesis, H_0 , and the alternative hypothesis, H_1 .
2. Evaluate the risks of making type I and II errors, and choose the level of significance, α , and the sample size as appropriate.
3. Determine the appropriate test statistic and sampling distribution to use and identify the critical values that divide the rejection and nonrejection regions.

4. Collect the data, calculate the appropriate test statistic, and determine whether the test statistic has fallen into the rejection or the nonrejection region.
5. Make the proper statistical inference. Reject the null hypothesis if the test statistic falls into the rejection region. Do not reject the null hypothesis if the test statistic falls into the nonrejection region.

The p -Value Approach to Hypothesis Testing

Most modern statistical software, including the functions found in Microsoft Excel and statistical calculators, can compute a probability value known as the p -value that you can use as a second way of drawing the proper statistical inference during hypothesis testing.

p -Value

CONCEPT The probability of obtaining a test statistic equal to or more extreme than the result obtained from the sample data, given that the null hypothesis H_0 is true.

INTERPRETATION The p -value is the smallest level at which H_0 can be rejected for a given set of data. You can consider the p -value the actual risk of having a type I error for a given set of data. Using p -values, you reject the null hypothesis if the p -value is less than α and do not reject the null hypothesis if the p -value is greater than or equal to α . Many people confuse this rule, mistakenly believing that a high p -value is grounds for rejection. You can avoid this confusion by informally remembering that “if the p -value is low, then H_0 must go.”

important point



In practice, most researchers today use p -values for several reasons, including efficiency of the presentation of results. The p -value is also known as the **observed level of significance**. When using p -values, you can restate the steps of hypothesis testing as follows:

1. State the null hypothesis, H_0 , and the alternative hypothesis, H_1 .
2. Evaluate the risks of making type I and II errors, and choose the level of significance, α , and the sample size as appropriate.
3. Collect the data and calculate the sample value of the appropriate test statistic.
4. Calculate the p -value based on the test statistic and compare the p -value to α .
5. Make the proper statistical inference. Reject the null hypothesis if the p -value is less than α . Do not reject the null hypothesis if the p -value is greater than or equal to α .

7.5 Types of Hypothesis Tests

Your choice of which statistical test to use when performing hypothesis testing is influenced by the following factors:

- Number of groups of data: one, two, or multiple (more than two)
- Relationship stated in alternative hypothesis H_1 : not equal to or inequality (less than, greater than)
- Type of variable (population parameter): numerical (mean) or categorical (proportion)

Number of Groups

One group of hypothesis tests, more formally known as **one-sample tests**, are of limited practical use, because if you are interested in examining the value of a population parameter, you can usually use one of the confidence interval estimate methods of Chapter 6. **Two-sample tests**, examining the differences between two groups, have been the focus of this chapter and can be found in the Worked-out Problems of Sections 8.1 through 8.3. Tests for more than two groups are discussed in Chapter 9.

Relationship Stated in Alternative Hypothesis H_1

Alternative hypotheses can be stated either using the not-equal sign, as in, $H_1: \mu_1 \neq \mu_2$; or by using an inequality, such as $H_1: \mu_1 > \mu_2$. Alternative hypotheses that use the not-equal sign require you to perform what statisticians call a **two-tail test**; alternative hypotheses that consist of an inequality require that you perform what statisticians call a **one-tail test**.

One-tail and two-tail test procedures are very similar and differ mainly in the way they use critical values to determine the region of rejection. Throughout this book, two-tail hypothesis tests are featured. One-tail tests are not further described in this book, although the Worked-out Problem 2b on page 155 illustrates one possible use for such tests.

Type of Variable

The type of variable, numerical or categorical, also influences the choice of hypothesis test used. For a numerical variable, the test will examine the population mean or the differences among the means, if two or more groups are used. For a categorical variable, the test will examine the population proportion or the differences among the population proportions if two or more groups are used. Tests involving two groups for each type of variable can be found in the Worked-out Problems of Sections 8.1 through 8.3. Tests involving more than two groups for each type of variable are featured in Chapter 9.

One-Minute Summary

Hypotheses

- Null hypothesis
- Alternative hypothesis

Types of errors

- Type I error
- Type II error

Hypothesis testing approach

- Test statistic
- p -value

Hypothesis test relationship

- One-tail test
- Two-tail test

Test Yourself

1. A type II error is committed when:
 - (a) you reject a null hypothesis that is true
 - (b) you don't reject a null hypothesis that is true
 - (c) you reject a null hypothesis that is false
 - (d) you don't reject a null hypothesis that is false
2. A type I error is committed when:
 - (a) you reject a null hypothesis that is true
 - (b) you don't reject a null hypothesis that is true
 - (c) you reject a null hypothesis that is false
 - (d) you don't reject a null hypothesis that is false
3. Which of the following is an appropriate null hypothesis?
 - (a) The difference between the means of two populations is equal to 0.
 - (b) The difference between the means of two populations is not equal to 0.
 - (c) The difference between the means of two populations is less than 0.
 - (d) The difference between the means of two populations is greater than 0.

4. Which of the following is *not* an appropriate alternative hypothesis?
 - (a) The difference between the means of two populations is equal to 0.
 - (b) The difference between the means of two populations is not equal to 0.
 - (c) The difference between the means of two populations is less than 0.
 - (d) The difference between the means of two populations is greater than 0.
5. The power of a test is the probability of:
 - (a) rejecting a null hypothesis that is true
 - (b) not rejecting a null hypothesis that is true
 - (c) rejecting a null hypothesis that is false
 - (d) not rejecting a null hypothesis that is false
6. If the p -value is less than α in a two-tail test:
 - (a) the null hypothesis should not be rejected
 - (b) the null hypothesis should be rejected
 - (c) a one-tail test should be used
 - (d) no conclusion should be reached
7. A test of hypothesis has a type I error probability (α) of 0.01. Therefore:
 - (a) if the null hypothesis is true, you don't reject it 1% of the time
 - (b) if the null hypothesis is true, you reject it 1% of the time
 - (c) if the null hypothesis is false, you don't reject it 1% of the time
 - (d) if the null hypothesis is false, you reject it 1% of the time
8. Which of the following statements is *not* true about the level of significance in a hypothesis test?
 - (a) The larger the level of significance, the more likely you are to reject the null hypothesis.
 - (b) The level of significance is the maximum risk you are willing to accept in making a type I error.
 - (c) The significance level is also called the α level.
 - (d) The significance level is another name for type II error.
9. If you reject the null hypothesis when it is false, then you have committed:
 - (a) a type II error
 - (b) a type I error
 - (c) no error
 - (d) a type I and type II error
10. The probability of a type ____ error is also called "the level of significance."
11. The probability of a type I error is represented by the symbol _____.

12. The value that separates a rejection region from a non-rejection region is called the _____.

The following are True or False questions:

13. For a given level of significance, if the sample size is increased, the power of the test will increase.
14. For a given level of significance, if the sample size is increased, the probability of committing a type I error will increase.
15. The statement of the null hypothesis always contains an equality.
16. The larger the p -value, the more likely you are to reject the null hypothesis.
17. The statement of the alternative hypothesis always contains an equality.

Answers to Test Yourself Questions

1. d
2. a
3. a
4. a
5. c
6. b
7. b
8. d
9. c
10. I
11. α
12. critical value
13. True
14. False
15. True
16. False
17. False

References

1. Berenson, M. L., D. M. Levine, and T. C. Krehbiel. *Basic Business Statistics: Concepts and Applications, Ninth Edition*. Upper Saddle River, NJ: Prentice Hall, 2004.
2. Cochran, W. G. *Sampling Techniques, Third Edition*. New York: Wiley, 1977.

3. Gitlow, H. S., and D. M. Levine. *Six Sigma for Green Belts and Champions*. Upper Saddle River, NJ: Financial Times - Prentice Hall, 2005.
4. Levine, D. M., T. C. Krehbiel, and M. L. Berenson. *Business Statistics: A First Course, Third Edition*. Upper Saddle River, NJ: Prentice Hall, 2003.
5. Levine, D. M., D. Stephan, T. C. Krehbiel, and M. L. Berenson. *Statistics for Managers Using Microsoft Excel, Fourth Edition*. Upper Saddle River, NJ: Prentice Hall, 2005.
6. Levine, D. M., P. P. Ramsey, and R. K. Smidt, *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Upper Saddle River, NJ: Prentice Hall, 2001.
7. Microsoft Excel 2002. Redmond, WA: Microsoft Corporation, 2001.
8. Sincich, T., D. M. Levine, and D. Stephan, *Practical Statistics by Example Using Microsoft Excel and Minitab, Second Edition*. Upper Saddle River, NJ: Prentice Hall, 2002.



Hypothesis Testing: Z and t Tests

8.1 Testing for the Difference Between Two Proportions

8.2 Testing for the Difference Between the Means of Two Independent Groups

8.3 Paired t Test

Important Equations

One-Minute Summary

Test Yourself

In Chapter 7, you learned about the fundamentals of hypothesis testing. In this chapter, you will learn about tests that involve two groups, more formally known as two-sample tests. The following tests are discussed:

- The hypothesis test which examines the differences between the proportions of two groups
- The hypothesis test which examines the differences between the means of two groups

You will also learn how to evaluate the statistical assumptions behind these tests—what to do if the assumptions do not hold—as well as learn how to choose the appropriate test for any two-group set of data.

8.1 Testing for the Difference Between Two Proportions

Often you want to analyze differences between two groups in the proportion of items that are in a particular category. The sample statistics needed to analyze these differences are the proportion of occurrences in group 1 and the

proportion of occurrences in group 2. With a sufficient sample size in each group, the sampling distribution of the difference between the two proportions approximately follows a normal distribution (see Section 5.3).

WORKED-OUT PROBLEM 1 The following two-way table summarizes some findings of a manufacturing plant study that investigated factors involved in producing good and bad silicon wafers. This table presents the joint responses of whether a specific wafer was “good” or “bad” and whether that wafer had particles present on it.

Counts of Particles Found Cross-Classified by Wafer Condition

		Wafer Condition		Total
		Good	Bad	
Particles Present	Yes	14	36	50
	No	320	80	400
	Total	334	116	450

Of 334 wafers that were classified as good, 320 had no particles found on the dye that produced the wafer. Of 116 wafers that were classified as bad, 80 had no particles found on the dye that produced the wafer. You seek to determine whether the proportion of wafers with no particles is the same for good and bad wafers using a level of significance of $\alpha = 0.05$.

For these data,

the proportion of good wafers without particles is $\frac{320}{334} = 0.9581$

and

the proportion of bad wafers without particles is $\frac{80}{116} = 0.6897$

Because the number of good and bad wafers that have no particles (320 and 80) is large, as is the number of good and bad wafers that have particles (14 and 36), the sampling distribution for the difference between the two proportions is approximately normally distributed. The null and alternative hypotheses are as follows:

$H_0: \pi_1 = \pi_2$ (no difference between the proportions for the “good” and “bad” groups)

$H_1: \pi_1 \neq \pi_2$ (a difference between the proportions for the two groups)

Microsoft Excel and statistical calculator results for the manufacturing plant study are as follows:

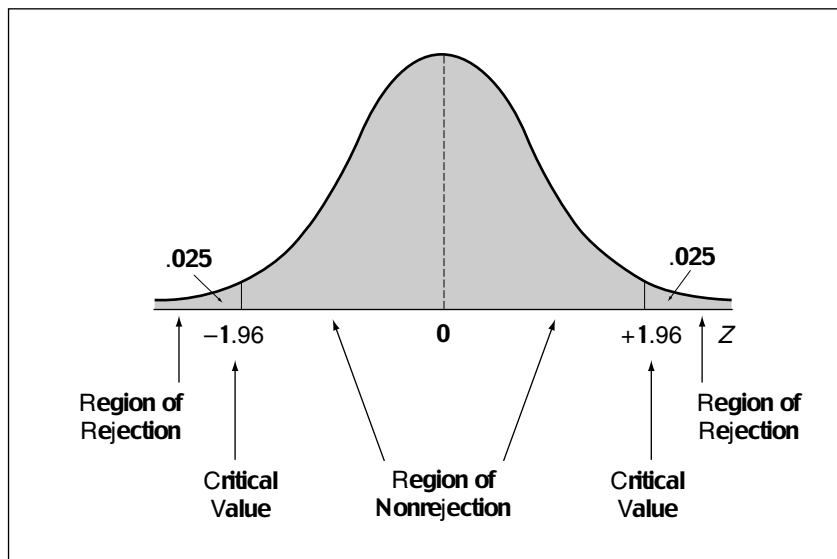
	A	B
1	Z Test for the Difference in Two Proportions	
2		
3	Data	
4	Hypothesized Difference	0
5	Level of Significance	0.05
6	Group 1	
7	Number of Successes	320
8	Sample Size	334
9	Group 2	
10	Number of Successes	80
11	Sample Size	116
12		
13	Intermediate Calculations	
14	Group 1 Proportion	0.9581
15	Group 2 Proportion	0.6897
16	Difference in Two Proportions	0.2684
17	Average Proportion	0.8889
18	Z Test Statistic	7.9254
19		
20	Two-Tail Test	
21	Lower Critical Value	-1.9600
22	Upper Critical Value	1.9600
23	p-Value	0.0000
24	Reject the null hypothesis	

```

2-PropZTest
P1≠P2
z=7.92542153
P=2.296343E-15
P1=.9580838323
P2=.6896551724
P=.8888888889

```

You decide to use the critical value approach. With a level of significance of 0.05, the lower tail area is 0.025, and the upper tail area is 0.025. Using the cumulative normal distribution table (Table C.1), the lower critical value of 0.025 corresponds to a Z value of -1.96, and an upper critical value of 0.025 (cumulative area of 0.975) corresponds to Z value of +1.96, as shown in the following diagram:



Given these regions, you will reject H_0 if $Z < -1.96$ or if $Z > +1.96$; otherwise you will not reject H_0 . Calculations done in a Microsoft Excel worksheet (see example on page 139) determine that the Z test statistic is 7.93. Because $Z = 7.93$ is greater than the upper critical value of $+1.96$, you reject the null hypothesis. You report that there is evidence of a difference in the proportion of good and bad wafers that have no particles.

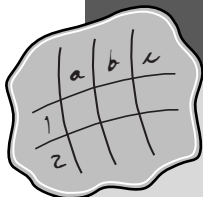
WORKED-OUT PROBLEM 2 You decide to use the p -value approach to hypothesis testing. Calculations done in a Microsoft Excel worksheet (see the figure on page 139) determine that the p -value is 0.0000. This means that the probability of obtaining a Z value greater than 7.93 is virtually zero (0.0000). Because the p -value is *less than* the level of significance $\alpha = 0.05$, you reject the null hypothesis. You report that there is evidence of a difference in the proportion of good and bad wafers that have no particles. Clearly, the good wafers are much more likely to have no particles than the bad wafers.



calculator keys

Z Test for the Differences in Two Proportions

Press [STAT] [◀] (to display the Tests menu), select 6:2-PropZTest, and press [ENTER] to display the 2-PropZTest screen. In this screen, enter values for the number of successes and sample size for group 1, the number of successes and sample size for group 2, and the first alternative hypothesis choice. Select Calculate and press [ENTER]. (You cannot set the level of significance that is preset to $\alpha = 0.05$.) If the p -value is a very small value, it may appear as number in exponential notation such as 5.7009126E-7. You should consider such values to be equivalent to zero.



spreadsheet solution

Z Test for the Differences in Two Proportions

Download and open the **Chapter 8 Z Two Proportions.xls** Excel file into which you can enter the values for the hypothesized difference, level of significance, and number of successes and sample size for each group. Already entered into the worksheet as an example are the data of Worked-out Problem 2.

WORKED-OUT PROBLEM 3 You seek to analyze the results of a famous health-study experiment that investigated the effectiveness of aspirin in the reduction of the incidence of heart attacks, using a level of significance of $\alpha = 0.05$. In this experiment, 22,071 male U.S. physicians were randomly assigned to either a group that was given one 325mg buffered aspirin tablet every other day or a group that was given a placebo (a pill that contained no active ingredients). Of 11,037 physicians taking aspirin, 104 suffered heart attacks during the 5-year period of the study. An additional 11,034 physicians were assigned to a group that took a placebo every other day. The results of this study were as follows.

		Study Group		Totals
		Aspirin	Placebo	
Results	Heart attack	104	189	293
	No heart attack	10,933	10,845	21,778
	Totals	11,037	11,034	22,071

The null and alternative hypotheses are as follows:

$H_0: \pi_1 = \pi_2$ (no difference in the proportion of heart attacks between the group that was given aspirin and the group that was given the placebo)

$H_1: \pi_1 \neq \pi_2$ (a difference in the proportion of heart attacks between the two groups)

Microsoft Excel and statistical calculator results for the health study experiment are as follows:

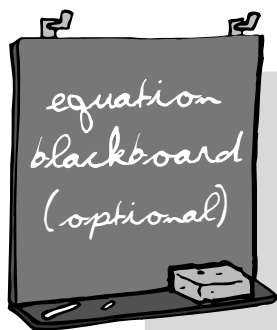
	A	B
1	Z Test for the Difference in Two Proportions	
2		
3	Data	
4	Hypothesized Difference	0
5	Level of Significance	0.05
6	Group 1	
7	Number of Successes	104
8	Sample Size	11037
9	Group 2	
10	Number of Successes	189
11	Sample Size	11034
12		
13	Intermediate Calculations	
14	Group 1 Proportion	0.0094
15	Group 2 Proportion	0.0171
16	Difference in Two Proportions	-0.0077
17	Average Proportion	0.0133
18	Z Test Statistic	-5.0014
19		
20	Two-Tail Test	
21	Lower Critical Value	-1.9600
22	Upper Critical Value	1.9600
23	p-Value	0.0000
24	Reject the null hypothesis	

```

2-PropZTest
P1≠P2
z=-5.001388204
P=5.7009126E-7
P1=.0094228504
P2=.0171288744
ΔP=.0132753387

```

Calculations determine that the p -value is 0.0000 and the value of the test statistic is $Z = 5.00$. This means that the chance of obtaining a Z value greater than 5.00 is virtually zero. Because the p -value = 0.0000 is *less than* the level of significance (0.05), you reject the null hypothesis and accept the alternative hypothesis that there is a difference in the proportion of heart attacks between the two groups. (Alternatively, using the critical value approach, because $Z = 5.00$ is greater than the upper critical value of +1.96 at the 0.05 level of significance, the null hypothesis is rejected.) You conclude that there is evidence of a difference in the proportion of doctors that have had heart attacks between those who took the aspirin and those who did not take the aspirin. The study group who took the aspirin had a significantly lower proportion of heart attacks over the study period.



The Worked-out Problems use the Z test for the difference between two proportions. You need the subscripted symbols for the number of successes, X , the sample sizes, n , sample proportions, p , and population proportions, π , as well as the symbol for the pooled estimate of the population proportion, \bar{p} , to express the Z test statistic calculation as an equation.

To write the Z test statistic equation, you first define the symbols for equations for the pooled estimate of the population proportion and the sample proportions for the two groups:

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} \quad p_1 = \frac{X_1}{n_1} \quad p_2 = \frac{X_2}{n_2}$$

You next use the just defined \bar{p} , p_1 , and p_2 along with the symbols for the sample sizes and population proportion to form the equation for the Z test for the difference between two proportions:

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

As an example, the calculations for determining the Z test statistic for the manufacturing plant study Worked-out Problem, (with $\alpha = 0.05$, reject H_0 if $Z < -1.96$ or if $Z > +1.96$; otherwise do not reject H_0) are as follows:

$$p_1 = \frac{X_1}{n_1} = \frac{320}{334} = 0.9581 \quad p_2 = \frac{X_2}{n_2} = \frac{80}{116} = 0.6897$$

and

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{320 + 80}{334 + 116} = \frac{400}{450} = 0.8889$$

so that

$$\begin{aligned} Z &= \frac{(0.9581 - 0.6897) - (0)}{\sqrt{0.8889(1 - 0.8889)\left(\frac{1}{334} + \frac{1}{116}\right)}} \\ &= \frac{0.2684}{\sqrt{0.8889(1 - 0.8889)(0.0116)}} \\ &= \frac{0.2684}{\sqrt{(0.09876)(0.0116)}} \\ &= \frac{0.2684}{\sqrt{0.0011456}} \\ &= \frac{0.2684}{0.03385} = +7.93 \end{aligned}$$

Because $Z = 7.93$ is greater than the upper critical value of $+1.96$, reject the null hypothesis.

8.2 Testing for the Difference Between the Means of Two Independent Groups

Many studies compare a numerical population parameter of two populations or groups. Statisticians distinguish between using two **independent** groups, the meaning of groups as previously defined in this chapter, and two **related** groups, in which the observations are matched according to a relevant characteristic or repeated measurements of the same items are taken. For studies involving two *independent* groups, the most common test of hypothesis used is the *pooled-variance t test*.

Pooled-Variance t Test

CONCEPT The hypothesis test for the difference between the population means of two independent groups that requires that the sample variances of each group be combined (“pooled”) into one estimate of the variance common in the two groups.

INTERPRETATION For this test, the test statistic is based on the difference in the sample means of the two groups, and the sampling distribution for the difference in the two sample means approximately follows the t distribution.

In a pooled variance t test, the null hypothesis of no difference in the means of two independent populations is:

$$H_0: \mu_1 = \mu_2 \text{ (The two population means are equal)}$$

and the alternative hypothesis is:

$$H_1: \mu_1 \neq \mu_2 \text{ (The two population means are not equal)}$$

WORKED-OUT PROBLEM 1 You seek to determine at a level of significance of $\alpha = 0.05$ whether the average surface hardness of steel intaglio printing plates prepared using a new treatment differs from the average hardness of plates that are untreated. You review the following results of an experiment in which 40 steel plates, 20 treated and 20 untreated, were tested for surface hardness.

Surface Hardness of 20 Untreated Steel Plates and 20 Treated Steel Plates

Untreated		Treated	
164.368	177.135	158.239	150.226
159.018	163.903	138.216	155.620
153.871	167.802	168.006	151.233
165.096	160.818	149.654	158.653
157.184	167.433	145.456	151.204
154.496	163.538	168.178	150.869
160.920	164.525	154.321	161.657
164.917	171.230	162.763	157.016
169.091	174.964	161.020	156.67
175.276	166.311	167.706	147.920

(Intaglio)

Microsoft Excel results for the printing plates study are as follows:

	A	B	C
1	t-Test: Two-Sample Assuming Equal Variances		
2			
3		<i>Untreated</i>	<i>Treated</i>
4	Mean	165.0948	155.73135
5	Variance	41.6934168	62.41409971
6	Observations	20	20
7	Pooled Variance	52.05375826	
8	Hypothesized Mean Difference	0	
9	df	38	
10	t Stat	4.104023608	
11	P(T<=t) one-tail	0.000103572	
12	t Critical one-tail	1.685954461	
13	P(T<=t) two-tail	0.000207144	
14	t Critical two-tail	2.024394147	

Calculations determine that the t statistic (labeled as “t Stat” in the Microsoft Excel results) is 4.10 and the p -value [labeled as $P(T \leq t)$ two-tail] is 0.0002. Because the p -value is 0.0002, which is less than $\alpha = 0.05$, the null hypothesis is rejected. This means that the chance of obtaining a t value greater than 4.10 is virtually zero (0.0002). You conclude that the mean surface hardness is higher for the untreated steel plates (sample average of 165.09) than for the treated steel plates (sample average of 155.70).



calculator keys

Pooled Variance t Test for the Differences in Two Means

Press [STAT] [◀] (to display the Tests menu), select 4:2-SampTTest, and press [ENTER] to display the 2-SampTTest screen. Proceed as appropriate:

When using sample data:

Select **Data** as the **Inpt** type and press [ENTER]. Enter the names of the list variables for each sample and make sure Freq1 and Freq2 are both set to 1. Select the first alternative hypothesis and press [ENTER]. Select **Yes** (Pooled) and press [ENTER]. Press [▼], select **Calculate**, and press [ENTER].

When using sample statistics:

Select **Stats** as the **Inpt** type and press [ENTER]. Enter the sample mean, sample standard deviation, and sample size of group 1, followed by those statistics for group 2. Press [▼], select the first alternative hypothesis, and press [ENTER]. Press [▼], select **Yes** (Pooled) and press [ENTER]. Press [▼], select **Calculate**, and press [ENTER].

As with the Z test for the differences in two proportions, you cannot set the level of significance (preset to $\alpha = 0.05$.) and if the p -value is a very small value, it may appear as a number in exponential notation, such as 2.296343E-15, that you should consider to be equivalent to zero.



spreadsheet solution

Pooled Variance t Test for the Differences in Two Arithmetic Means

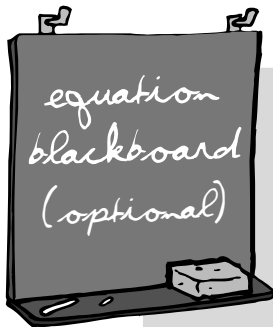
When using sample data:

Select **Tools** → **Data Analysis**, and in the Data Analysis dialog box select **t-Test: Two-Sample Assuming Equal Variances** and click **OK**. In the t-Test dialog box, enter the group 1 cell range as the Variable 1 Range and enter the group 2 cell range as the Variable 2 Range. Enter 0 as the Hypothesized Mean Difference and, if appropriate, check **Labels**. Enter 0.05 as the Alpha value, select the **New Worksheet Ply** option, and click **OK**. Results appear on a new worksheet. (See Appendix D.3 for more information about the Data Analysis feature.)

When using sample statistics:

Download and open the **Chapter 8 Pooled T.xls**

Excel file into which you can enter the values for the hypothesized difference, the level of significance, and the sample size, sample mean, and sample standard deviation for each group. Already entered into the worksheet as an example are the data of Worked-out Problem 1.



The Worked-out Problems use the pooled variance t test for the difference between the population means of two independent groups. You need the subscripted symbols for the sample means, \bar{X}_1 and \bar{X}_2 , the sample size, n , and population means, μ_1 and μ_2 , as well as the symbol for the pooled estimate of the variance, S_p^2 , to express the t test statistic calculation as an equation.

To write the t test statistic equation, you first define the symbols for the equation for the pooled estimate of the population variance:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

(continues)

interested
in
math?

You next use the just defined S_p^2 and the symbols for the sample means, the population means, and the sample sizes to form the equation for the pooled-variance t test for the difference between two means:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

The calculated test statistic t follows a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

As an example, the calculations for determining the t test statistic for the printing plates Worked-out Problem, with $\alpha = 0.05$, are as follows:

$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} \\ &= \frac{19(41.6934) + 19(62.4141)}{19 + 19} = 52.05375 \end{aligned}$$

Using 52.05375 as the value for S_p^2 in the original equation

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

produces the following:

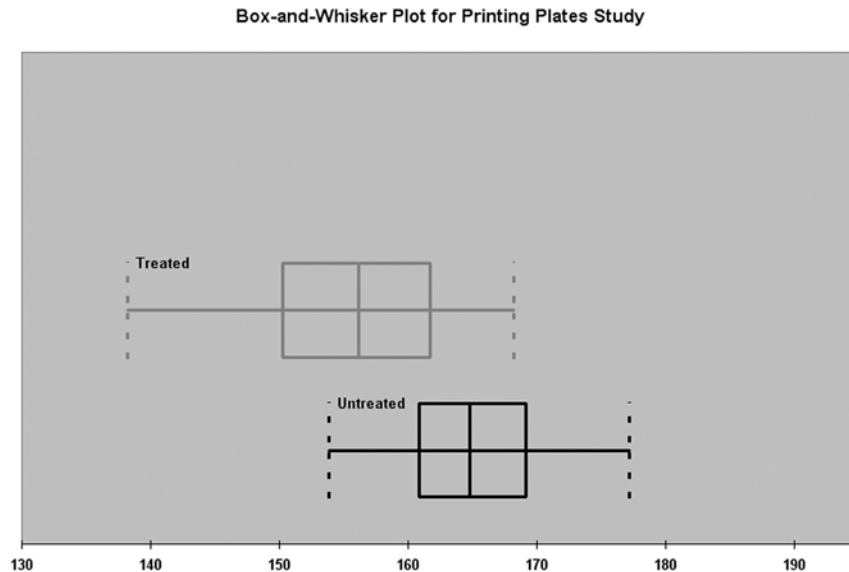
$$\begin{aligned} t &= \frac{(165.0948 - 155.73135) - 0}{\sqrt{52.05375 \left(\frac{1}{20} + \frac{1}{20} \right)}} \\ t &= \frac{165.0948 - 155.73135}{\sqrt{52.05375(0.10)}} \\ &= \frac{9.36345}{\sqrt{5.205375}} = +4.104 \end{aligned}$$

Using the $\alpha = 0.05$ level of significance, with $20 + 20 - 2 = 38$ degrees of freedom, the critical value of t is 2.0244 (0.025 in the upper tail of the t distribution). Because $t = +4.104 > 2.0244$, you reject H_0 .

Pooled-Variance t Test Assumptions

In testing for the difference between the means, you assume that the populations from which the two independent samples are drawn are normally distributed with equal variances. For situations in which the two populations have equal variances, the pooled-variance t test is not sensitive to moderate departures from this assumption, provided that the sample sizes are large. In such situations, the pooled-variance t test can be used without serious effect on its power.

To check the assumption of normality in each of the two groups, you can prepare a side-by-side box-and-whisker plot similar to the following plot:



In this box-and-whisker plot, the medians in each of the two groups (the center lines) seem to be equally spaced away from the ends of the box (the quartiles). The lower tail in the box-and-whisker plot for the treated group is longer than the upper tail. Comparing these box-and-whisker plots to the box-and-whisker plot for a normal distribution discussed on page 53, you can conclude that there appears to be only moderate departure from normality. Therefore, the assumption of normality needed for the t test is not seriously violated.

If the data in each group cannot be assumed to be from normally distributed populations, two choices exist. A **nonparametric** procedure, such as the **Wilcoxon rank sum test** (see References 1, 4, 5, and 8), can be used that does not depend on the assumption of normality for the two populations, or a transformation (see Reference 1) on each of the outcomes can be made and the pooled-variance t test can then be used.



The pooled-variance t test also assumes that the population variances are equal. If this assumption cannot be made, the pooled-variance t test is inappropriate. Instead, the **separate-variance t test** (see References 1 and 2) is used. Although the computations for the separate-variance t test are cumbersome, Microsoft Excel can be used to perform them.

WORKED-OUT PROBLEM 1 You seek to determine at a level of significance of $\alpha = 0.05$ whether the average payment made by online customers of a Web site differs according to two methods of payment. You obtained the following statistics based on a random sample of 50 transactions.

	Method 1	Method 2
Sample size	22	28
Sample mean	30.37	23.17
Sample standard deviation	12.006	7.098

Microsoft Excel and statistical calculator results for this study prepared are as follows:

	A	B
1	t Test for Differences in Two Means	
2		
3	Data	
4	Hypothesized Difference	0
5	Level of Significance	0.05
6	Group 1 Sample	
7	Sample Size	22
8	Sample Mean	30.37
9	Sample Standard Deviation	12.006
10	Group 2 Sample	
11	Sample Size	28
12	Sample Mean	23.17
13	Sample Standard Deviation	7.098
14		
15	Intermediate Calculations	
16	Population 1 Sample Degrees of Freedom	21
17	Population 2 Sample Degrees of Freedom	27
18	Total Degrees of Freedom	48
19	Pooled Variance	91.4027
20	Difference in Sample Means	7.2000
21	t Test Statistic	2.6434
22		
23	Two-Tail Test	
24	Lower Critical Value	-2.0106
25	Upper Critical Value	2.0106
26	p-Value	0.0111
27	Reject the null hypothesis	

```

2-SampTTest
μ1≠μ2
t=2.643372782
p=.0110544755
df=48
x1=30.37
x2=23.17

```

Calculations determine that the t statistic is 2.64 and the p -value (labeled as p by the TI-83) is 0.0111. Because the p -value is 0.0111 is less than $\alpha = 0.05$, you reject the null hypothesis. (Using the critical value approach, $t = 2.64 > 2.01$.) This means that the chance of obtaining a t value greater than 2.64 is very small (0.0111). You conclude that the mean payment amount is higher

for method 1 (sample average of \$30.37) than for method 2 (sample average of \$23.17).

8.3 The Paired t Test

The tests for the pooled variance t test and the difference between two population proportions presented earlier in this chapter were based on differences between two *independent* groups. Often, data are obtained from observations in which the groups are **related**.

Two approaches that involve related data between groups are possible. In one approach, you **pair**, or match, the items or individuals under study according to some other variable. For example, in testing whether a new drug treatment lowers blood pressure, a sample of patients could be *paired* according to their blood pressure at the beginning of the study. For example, assuming there were two patients in the study who had a diastolic blood pressure of 140, one would be randomly assigned to the new drug, and the other would be assigned to either a placebo or an old drug. Assigning patients in this manner allows the experimenter to factor out the initial effect of the patient's blood pressure so the analysis can focus on the difference between the drugs.

In the second approach, **repeated measurements** are obtained from the same set of items or individuals. This approach is based on the theory that the same items or individuals will behave alike if treated alike. The objective of the analysis is to show that any differences between two measurements of the same items or individuals are due to different treatment conditions. For example, when performing an experiment on the effect of a diet, each person can be used as his or her own control so that repeated measurements on the same people are obtained. One measurement is taken just prior to starting the diet and the other measurement is taken at the end of a specified time period.

In either approach, the variable of interest is the *difference between the values* of the paired items or individuals rather than the *values* themselves, stated algebraically as follows:

$$\text{Difference } (D) = \text{Related value in sample 1} - \text{Related value in sample 2}$$

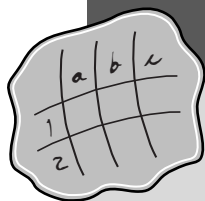
With related groups and a numeric variable of interest, the null hypothesis becomes that there is no difference in the population means of the two related groups, and the alternative hypothesis becomes that there is a difference in the population means of the two related groups. Using the symbol μ_D to represent the difference between the population means, the null and alternative hypotheses can be expressed as follows:

$$H_0: \mu_D = 0$$

and

$$H_1: \mu_D \neq 0$$

To decide whether to reject the null hypothesis, you use a paired t test.



spreadsheet solution

Paired t Test

Select Tools → Data Analysis, and in the Data Analysis dialog box select **t-Test: Paired Two Sample for Means** and click OK. In the t-Test dialog box, enter the group 1 sample data cell range as the Variable 1 Range and enter the group 2 sample data cell range as the Variable 2 Range. Enter 0 as the Hypothesized Mean Difference and, if appropriate, check **Labels**. Enter 0.05 as the Alpha value, select the **New Worksheet Ply** option and click OK. Results appear on a new worksheet. (See Appendix D.3 for more information about the Data Analysis feature.)

WORKED-OUT PROBLEM You seek to determine, using a level of significance of $\alpha = 0.05$, whether there is a difference between the Doppler echocardiography measurements of 23 patients as taken by two different observers. Study data for this problem is as follows.

Patient	Observer A	Observer B	Difference (D)
1	4.8	5.8	-1.0
2	5.6	6.1	-0.5
3	6.0	7.7	-1.7
4	6.4	7.8	-1.4
5	6.5	7.6	-1.1
6	6.6	8.1	-1.5
7	6.8	8.0	-1.2
8	7.0	8.1	-1.1
9	7.0	6.6	0.4
10	7.2	8.1	-0.9
11	7.4	9.5	-2.1

(continues)

Patient	Observer A	Observer B	Difference (D)
12	7.6	9.6	-2.0
13	7.7	8.5	-0.8
14	7.7	9.5	-1.8
15	8.2	9.1	-0.9
16	8.2	10.0	-1.8
17	8.3	9.1	-0.8
18	8.5	10.8	-2.3
19	9.3	11.5	-2.2
20	10.2	11.5	-1.3
21	10.4	11.2	-0.8
22	10.6	11.5	-0.9
23	11.4	12.0	-0.6

(Cardiac)

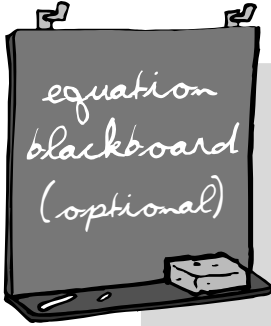
Source: M. L. R. Guerra Ernst and W. R. Schucany, "Scatterplots for Unordered Pairs," *The American Statistician*, 1996, 50, 260-265.

Because the two observers measure the same set of patients, the two groups of measurements are related and the differences between the two groups are tested.

Microsoft Excel results for this study are as follows:

	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3		Observer A	Observer B
4	Mean	7.8000	9.0304
5	Variance	2.8027	3.2213
6	Observations	23	23
7	Pearson Correlation	0.9346	
8	Hypothesized Mean Difference	0	
9	df	22	
10	t Stat	-9.2421	
11	P(T<=t) one-tail	0.0000	
12	t Critical one-tail	1.7171	
13	P(T<=t) two-tail	0.0000	
14	t Critical two-tail	2.0739	

Calculations done in this worksheet determine that the t statistic (labeled as t Stat in the Microsoft Excel results) is -9.24 and the p -value [labeled as $P(T \leq t)$ two-tail] is 0.0000 . Because the p -value is 0.0000 is less than $\alpha = 0.05$, you reject the null hypothesis. This means that the chance of obtaining a t value less than -9.24 is virtually zero. You conclude that there is a difference between the observers.



interested
in
math?

The Worked-out Problem uses the equation for the paired t test. You need the symbols for the sample size, n , the difference between the population arithmetic means, μ_D , the sample standard deviation, S_D , the subscripted symbol for the differences in the paired values, D_i , all previously introduced, and the symbol for the average difference, \bar{D} , to express the t test statistic calculation as an equation.

To write the t test statistic equation, you first define the symbols for the equation for the average difference, \bar{D} :

$$\bar{D} = \frac{\sum_{i=1}^n D_i}{n}$$

You next use the just defined \bar{D} , the symbols for the sample size, and the differences in the paired values to form the equation for the sample standard deviation, S_D :

$$S_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n - 1}}$$

Finally, you assemble the two just-defined symbols and the remaining symbols to form the equation for the paired t test for the difference between two related means:

$$t = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}}$$

The test statistic t follows a t distribution with $n - 1$ degrees of freedom.

As an example, the calculations for determining the t test statistic for the two observers Worked-out Problem, with $\alpha = 0.05$, are as follows:

$$\begin{aligned}\bar{D} &= \frac{\sum_{i=1}^n D_i}{n} \\ &= \frac{-28.3}{23} = -1.23\end{aligned}$$

This makes $S_D = 0.638$ (calculation not shown). Substituting these values produces the following:

(continues)

$$t = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}} = \frac{-1.23 - 0}{\frac{0.638}{\sqrt{23}}} = -9.24$$

Using the $\alpha = 0.05$ level of significance, with $23 - 1 = 22$ degrees of freedom, the critical value of t is -2.0739 (0.025 in the lower tail of the t distribution). Because $t = -9.24 < -2.0739$, you reject the null hypothesis H_0 .

WORKED-OUT PROBLEM 2a You seek to determine, using a level of significance of $\alpha = 0.05$, whether there are differences in monthly sales between the new package design and the old package design of a laundry stain remover. The new package was test marketed over a period of one month in a sample of supermarkets in a particular city. A random sample of ten pairs of supermarkets is matched according to weekly sales volume and a set of demographic characteristics. The data of the test-marketing study is as follows.

Monthly Sales of Laundry Stain Remover

Pair	New Package	Old Package	Difference
1	458	437	21
2	519	488	31
3	394	409	-15
4	632	587	45
5	768	753	15
6	348	400	-52
7	572	508	64
8	704	695	9
9	527	496	31
10	584	513	71

(Supermarket)

Because the ten pairs of supermarkets were matched, you use the paired t test, which provides more statistical power than if samples were selected from two independent groups. Test results for this study prepared using Microsoft Excel are as follows:

	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3		New Package	Old Package
4	Mean	550.6	528.6
5	Variance	17188.2667	13785.1556
6	Observations	10	10
7	Pearson Correlation	0.9631	
8	Hypothesized Mean Difference	0	
9	df	9	
10	t Stat	1.9116	
11	P(T<=t) one-tail	0.0441	
12	t Critical one-tail	1.8331	
13	P(T<=t) two-tail	0.0882	
14	t Critical two-tail	2.2622	

Calculations done in this worksheet determine that the t statistic (labeled as t Stat in the Microsoft Excel results) is 1.91 and the p -value [labeled as $P(T \leq t)$ two-tail] is 0.0882. This means that the chance of obtaining a t value greater than 1.91 or less than -1.91 is 0.0882 or 8.82%. Because the p -value is 0.0882 is greater than $\alpha = 0.05$, you do not reject the null hypothesis. (Using the critical value approach, $t = 1.91 < 2.262$.) You can conclude that there is insufficient evidence of a difference between the new and old package design.

WORKED-OUT PROBLEM 2b Before the test marketing, you may have wanted to determine whether the new package design produced *more* sales than the old package design and not just a difference in sales. Such a situation would be a good application of a one-tail test, in which the alternative hypothesis would be $H_1: \mu_D > 0$.

In such a case, you would use the one-tail p -value (or one-tail critical value). For the test market sales data, the one-tail p -value is 0.0441 (and the one-tail critical value is 1.8331). Because the p -value is less than $\alpha = 0.05$ (or because using the critical value approach, $t = 1.911$ is greater than 1.8331), you reject the null hypothesis and conclude that the average sales from the new package design were higher than the average sales for the old package design—a different conclusion than you made from the two-tail test.

Important Equations

Z test for the difference between two proportions:

$$(8.1) \quad Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Pooled-variance t test for the difference between the population arithmetic means of two independent groups:

$$(8.2) \quad t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Paired t test for the difference between the means of two related groups:

$$(8.3) \quad t = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}}$$

One-Minute Summary

For tests for the differences between two groups, first determine if your data are categorical or numerical.

- If your data are categorical, use the Z test for the difference between two proportions.
- If your data are numerical, determine if you have independent or related groups:
 - If you have independent groups, use the pooled variance t test for the difference between two means.
 - If you have related groups, use the paired t test.

Test Yourself

1. The t test for the difference between the means of two independent populations assumes that the respective:
 - (a) Sample sizes are equal
 - (b) Sample medians are equal
 - (c) Populations are approximately normal
 - (d) All of the above
2. In testing for differences between the means of two related populations, the null hypothesis is:
 - (a) $H_0: \mu_D = 2$
 - (b) $H_0: \mu_D = 0$
 - (c) $H_0: \mu_D < 0$
 - (d) $H_0: \mu_D > 0$

3. A researcher is curious about the effect of sleep on students' test performances. He chooses 100 students and gives each two exams. One is given after 4 hours' sleep and one after 8 hours' sleep. The statistical test the researcher should use is the:
 - (a) Z test for the difference between two proportions
 - (b) Pooled-variance t test
 - (c) Paired t test
4. A statistics professor wanted to test whether the grades on a statistics test were the same for her morning class and her afternoon class. For this situation, the professor should use the:
 - (a) Z test for the difference between two proportions
 - (b) Pooled-variance t test
 - (c) Paired t test

The following are True or False questions:

5. The sample size in each independent sample must be the same in order to test for differences between the means of two independent populations.
6. In testing a hypothesis about the difference between two proportions, the p -value is computed to be 0.043. The null hypothesis should be rejected if the chosen level of significance is 0.05.
7. In testing a hypothesis about the difference between two proportions, the p -value is computed to be 0.034. The null hypothesis should be rejected if the chosen level of significance is 0.01.
8. In testing a hypothesis about the difference between two proportions, the Z test statistic is computed to be 2.04. The null hypothesis should be rejected if the chosen level of significance is 0.01 and a two-tail test is used.
9. The sample size in each independent sample must be the same in order to test for differences between the proportions of two independent populations.
10. When you are sampling the same individuals and taking a measurement before treatment and after treatment, you should use the paired t test.

Answers to Test Yourself Questions

1. c
2. b
3. c
4. b
5. False

6. True
7. False
8. False
9. False
10. True

References

1. Berenson, M. L., D. M. Levine, and T. C. Krehbiel. *Basic Business Statistics: Concepts and Applications, Ninth Edition*. Upper Saddle River, NJ: Prentice Hall, 2004.
2. Cochran, W. G. *Sampling Techniques, Third Edition*. New York: Wiley, 1977.
3. Gitlow, H. S., and D. M. Levine. *Six Sigma for Green Belts and Champions*. Upper Saddle River, NJ: Financial Times - Prentice Hall, 2005.
4. Levine, D. M., T. C. Krehbiel, and M. L. Berenson. *Business Statistics: A First Course, Third Edition*. Upper Saddle River, NJ: Prentice Hall, 2003.
5. Levine, D. M., D. Stephan, T. C. Krehbiel, and M. L. Berenson. *Statistics for Managers Using Microsoft Excel, Fourth Edition*. Upper Saddle River, NJ: Prentice Hall, 2005.
6. Levine, D. M., P. P. Ramsey, and R. K. Smidt, *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Upper Saddle River, NJ: Prentice Hall, 2001.
7. Microsoft Excel 2002. Redmond, WA: Microsoft Corporation, 2001.
8. Sincich, T., D. M. Levine, and D. Stephan, *Practical Statistics by Example Using Microsoft Excel and Minitab, Second Edition*. Upper Saddle River, NJ: Prentice Hall, 2002.



Hypothesis Testing: Chi-Square Tests and the One-Way Analysis of Variance (ANOVA)

9.1 Chi-Square Test for Two-Way Cross-Classification
Tables

9.2 One-Way Analysis of Variance (ANOVA): Testing
for the Differences Among the Means of More
Than Two Groups

Important Equations

One-Minute Summary

Test Yourself

In Chapter 8, you learned about two specific two-sample hypothesis tests that are used for analyzing the data of two groups. Recall from Chapter 7, hypothesis testing can also be done for multiple groups—that is, more than two groups of data. In this chapter, you will learn about the following tests that you can use when you have *multiple* groups:

- The chi-square (χ^2) test for categorical variables that determine whether there is a difference in the population proportions between two or more groups
- The one-way analysis of variance (ANOVA) for numerical variables that determine whether there is a difference in the means among more than two groups

9.1 Chi-Square Test for Two-Way Tables

CONCEPT The hypothesis tests for the difference in the proportion of successes in two or more groups or a relationship between two categorical variables in a two-way cross-classification table.

INTERPRETATION Recall from Chapter 2 that a two-way cross-classification table presents the count of joint responses to two categorical variables. The categories of one variable form the rows of the table, and the categories of the other variable form the columns. The chi-square test determines whether there is a relationship between the row variable and the column variable. By the way, when there are only two rows and two columns (the simplest case of a two-way table), the test is equivalent to the Z test for the difference between two proportions discussed in Section 8.1.

The null and alternative hypotheses for the two-way cross-classification table are as follows:

H_0 (There is no relationship between the row variable and the column variable.)

H_1 (There is a relationship between the row variable and the column variable.)

When there are two rows and two columns in the cross-classification table, the null and alternative hypotheses are (those of the Z test for the difference between two proportions) as follows:

$H_0: \pi_1 = \pi_2$ (no difference between the two proportions)

$H_1: \pi_1 \neq \pi_2$ (a difference between the two proportions)

The chi-square test is based on a comparison of the actual count (or frequency) in each cell, the intersection of a row and column, with the frequency that would be expected to occur if the null hypothesis were true. The expected frequency for each cell is obtained by multiplying the row total of that cell by the column total of that cell and dividing by the total sample size:

$$\text{expected frequency} = \frac{(\text{row total})(\text{column total})}{\text{sample size}}$$

The test statistic compares the actual frequency in each cell to the expected frequency that would occur if the null hypothesis were true. Because some differences are positive and some are negative, each difference is squared; then each squared difference is divided by the expected frequency. The results for all cells are then added to produce a statistic that follows the chi-square distribution.

For the special case of a two-way table of two rows and two columns, the expected frequency of each cell must be at least 5. If the expected frequency is less than 5, other procedures such as Fisher's exact test (see References 2 and 3) can be used. For all other cross-classification tables having more than two rows or more than two columns, all expected frequencies should be greater than 1.0.

important point



WORKED-OUT PROBLEM 1a You seek to determine, with a level of significance $\alpha = 0.05$, whether the results of the manufacturing plant study first presented in Chapter 2 show that there is a difference between the propor-

tions of good and bad wafers that have particles (Particles present = Yes). The study data are as follows.

Counts of Particles Found Cross-Classified by Wafer Condition

		Wafer Condition		Total
		Good	Bad	
Particles Present	Yes	14	36	50
	No	320	80	400
	Total	334	116	450

Row 1's total means that there are 50 wafers that have particles present. Column 1's total means that there are 334 good wafers. The expected frequency for good wafers with particles present would be 37.11—the total number of wafers with particles present (50) multiplied by the total number of good wafers (334) and divided by the total, or sample size (450).

$$\begin{aligned}\text{expected frequency} &= \frac{(50)(334)}{450} \\ &= 37.11\end{aligned}$$

The expected frequencies for all four cells are as follows.

		Wafer Condition		Total
		Good	Bad	
Particles Present	Yes	37.11	12.89	50
	No	296.89	103.11	400
	Total	334	116	450

Microsoft Excel and statistical calculator results for this study are shown on page 162.

Because the p -value for this chi-square test, 0.0000 (2.273737E-15 is an extremely small number that you interpret as zero), is less than the level of significance α of 0.05, you reject the null hypothesis. You conclude that there is a relationship between having a particle on the wafer and the condition of the wafer. In other words, there is a significant difference between good and bad wafers in the proportion of wafers that have particles. Clearly, the percentage of good wafers that have particles is less than the percentage of bad wafers that have particles.

WORKED-OUT PROBLEM 1b You decide to use the critical value approach for the manufacturing plant study. The computed chi-square statistic (see results on page 162) is 62.81. The number of degrees of freedom for the chi-square test equals the number of rows minus 1 multiplied by the number of columns minus 1:

$$\text{Degrees of freedom} = (\text{Number of rows} - 1)(\text{Number of columns} - 1)$$

Using the table of the chi-square distribution (Table C.3), with $\alpha = 0.05$ and the degrees of freedom $= (2 - 1)(2 - 1) = 1$, the critical value of chi-square is equal to 3.841. Because $62.81 > 3.841$, you reject the null hypothesis.

	A	B	C	D
1	Chi-Square Test			
2				
3	Observed Frequencies			
4		Wafer Condition		
5	Particles Present	Good	Bad	Total
6	Yes	14	36	50
7	No	320	80	400
8	Total	334	116	450
9				
10	Expected Frequencies			
11		Wafer Condition		
12	Particles Present	Good	Bad	Total
13	Yes	37.11111	12.88889	50
14	No	296.8889	103.1111	400
15	Total	334	116	450
16				
17	Data			
18	Level of Significance	0.05		
19	Number of Rows	2		
20	Number of Columns	2		
21	Degrees of Freedom	1		
22				
23	Results			
24	Critical Value	3.8415		
25	Chi-Square Test Statistic	62.8123		
26	p-Value	0.0000		
27	Reject the null hypothesis			
28				
29	Expected frequency assumption			
30	is met.			

```

χ²-Test
χ²=62.81230642
P=2.273737E-15
df=1

```



calculator keys

Chi-Square Tests

To enter the table of observed frequencies:

Press [2nd] [x^{-1}] [◀] and press [ENTER] to display the MATRIX[A] screen. Enter the number of rows, press [▶], enter the number of columns, and press [▶] again. A table of the entered size appears. Into that table, enter the observed frequencies of each cell. (Press [ENTER] to advance to the next cell.) Press [2nd] [MODE] after you have finished entering all of the cell values.

To perform a chi-square test:

Press [STAT] [◀] (to display the Tests menu) and select C: χ^2 -Test and press [ENTER] to display the χ^2 -Test screen. Verify that Observed is (matrix) A and that Expected is (matrix) B. Select Calculate and press [ENTER].



spreadsheet solution

Chi-Square Tests

Download and open the **Chapter 9 Chi Square.xls** Excel file. Select, as appropriate, either the **ChiSquare 2 X 2**, **ChiSquare 2 X 3**, or **ChiSquare 3 X 4** worksheet and follow the instructions displayed in the worksheet to fill in the observed frequencies table.

WORKED-OUT PROBLEM 2a Fast-food chains are evaluated each year on many variables and the results are summarized in *QSR Magazine*. One important variable is the accuracy of the order. You seek to determine, with a level of significance of $\alpha = 0.05$, whether there is a difference in the proportions of food orders filled correctly at Burger King, Wendy's, and McDonald's. You use the following data that report the results of placing an order consisting of a main item, a side item, and a drink.

		Fast-Food Chain		
		Burger King	Wendy's	McDonald's
Order Filled Correctly	Yes	440	430	422
	No	60	70	78
Total		500	500	500

The null and alternative hypotheses are as follows:

$H_0: \pi_1 = \pi_2 = \pi_3$ (no difference in the proportion of correct orders among Burger King, Wendy's, and McDonald's)

$H_1: \pi_1 \neq \pi_2 \neq \pi_3$ (a difference in the proportion of correct orders among Burger King, Wendy's, and McDonald's)

Microsoft Excel and statistical calculator results for this study are on page 164.

Because the p -value for this chi-square test, 0.2562, is greater than the level of significance α of 0.05, you cannot reject the null hypothesis. There is insufficient evidence of a difference in the proportion of correct orders filled among Burger King, Wendy's, and McDonald's.

WORKED-OUT PROBLEM 2b Using the critical value approach for the same problem, the computed chi-square statistic (see results on page 164) is 2.72. At the 0.05 level of significance with the 2 degrees of freedom $[(2 - 1)(3 - 1) = 2]$, the chi-square critical value from Table C.3 is 5.991. Because the computed test statistic is less than 5.991, you cannot reject the null hypothesis.

	A	B	C	D	E
1	Chi-Square Test				
2					
3	Observed Frequencies				
4		Column variable			
5	Row variable	Burger King	Wendy's	McDonald's	Total
6	Yes	440	430	422	1292
7	No	60	70	78	208
8	Total	500	500	500	1500
9					
10	Expected Frequencies				
11		Column variable			
12	Row variable	Burger King	Wendy's	McDonald's	Total
13	Yes	430.6666667	430.666667	430.6666667	1292
14	No	69.33333333	69.3333333	69.33333333	208
15	Total	500	500	500	1500
16					
17	Data				
18	Level of Significance	0.05			
19	Number of Rows	2			
20	Number of Columns	3			
21	Degrees of Freedom	2			
22					
23	Results				
24	Critical Value	5.9915			
25	Chi-Square Test Statistic	2.7239			
26	p-Value	0.2562			
27	Do not reject the null hypothesis				
28					
29	Expected frequency assumption				
30	is met.				

χ^2 -Test
 $\chi^2=2.723862824$
 $P=.2561655376$
 $df=2$

WORKED-OUT PROBLEM 3a You seek to determine, with a level of significance of $\alpha = 0.05$, whether there was a relationship between numbers selected for the Vietnam War era military draft lottery system and the time of the year a man was born. The following shows how many low (1–122), medium (123–244), and high (245–366) numbers were drawn for birth dates in each quarter of the year.

		Quarter of Year				Total
		Jan.–Mar.	Apr.–June	July–Sept.	Oct.–Dec.	
Number Set	Low	21	28	35	38	122
	Medium	34	22	29	37	122
	High	36	41	28	17	122
	Total	91	91	92	92	366

The null and alternative hypotheses are:

H_0 (no relationship between the number selected and the time of the year a man was born)

H_1 (a relationship between the number selected and the time of the year a man was born)

Microsoft Excel and statistical calculator results for this study are as follows:

	A	B	C	D	E	F
1	Chi-Square Test					
2						
3	Observed Frequencies					
4	Column variable					
5	Row variable	Jan.-Mar.	Apr.-Jun.	Jul.-Sep.	Oct.-Dec.	Total
6	Low	21	28	35	38	122
7	Medium	34	22	29	37	122
8	High	36	41	28	17	122
9	Total	91	91	92	92	366
10	Expected Frequencies					
11	Column variable					
12	Row variable	Jan.-Mar.	Apr.-Jun.	Jul.-Sep.	Oct.-Dec.	Total
13	Low	30.333333	30.333333	30.66667	30.66667	122
14	Medium	30.333333	30.333333	30.66667	30.66667	122
15	High	30.333333	30.333333	30.66667	30.66667	122
16	Total	91	91	92	92	366
17						
18	Data					
19	Level of Significance	0.05				
20	Number of Rows	3				
21	Number of Columns	4				
22	Degrees of Freedom	6				
23						
24	Results					
25	Critical Value	12.5916				
26	Chi-Square Test Statistic	20.6804				
27	p-Value	0.0021				
28	Reject the null hypothesis					
29						
30						
31	Expected frequency assumption					
32	is met.					

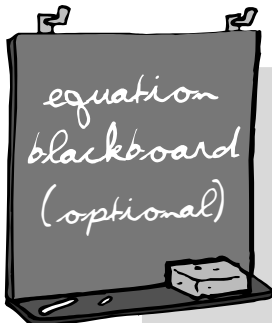
```

x2-Test
x2=20.68036312
P=.0020935834
df=6

```

Because the p -value for this chi-square test, 0.0021, is less than the level of significance α of 0.05, you reject the null hypothesis. There is evidence of a relationship between the number selected and the time of the year in which the man was born. It appears that men who were born between January and June were more likely than expected to have high numbers, whereas men born between July and December were more likely than expected to have low numbers.

WORKED-OUT PROBLEM 3b Using the critical value approach for the same problem, the computed chi-square statistic (see results above) is 20.68. At the 0.05 level of significance with the 6 degrees of freedom $[(3 - 1)(4 - 1) = 6]$, the chi-square critical value from Table C.3 is 12.592. Because the computed test statistic is greater than 12.592, you reject the null hypothesis.




You need the subscripted symbols for the observed cell frequencies, f_o , and the expected cell frequencies, f_e , to write the equation for the chi-square test for a two way cross-classification table:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

(continues)

interested
in
math?



For the manufacturing plant study (Worked-out Problem 1), this would evaluate as follows.

f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
14	37.11	-23.11	534.0721	14.3916
320	296.89	23.11	534.0721	1.7989
36	12.89	23.11	534.0721	41.4331
80	103.11	-23.11	534.0721	5.1796
				<hr/> 62.8032

Using the level of significance $\alpha = 0.05$, with $(2 - 1)(2 - 1) = 1$ degree of freedom from Table C.3, the critical value is 3.841. Because $62.80 > 3.841$, you reject the null hypothesis.

9.2 One-Way Analysis of Variance (ANOVA): Testing for the Differences Among the Means of More Than Two Groups

Many applications involve experiments in which differences in more than two groups need to be tested. Evaluating differences between groups is often viewed as a **one-factor experiment** (also known as a **completely randomized design**) in which the variable defining the groups is called the **factor of interest**. A factor of interest can have several *numerical levels* such as baking temperature (e.g., 300°, 350°, 400°, 450°) for an industrial process study, or a factor can have several *categorical levels* such as brand preference (Brand A, Brand B, Brand C) for a marketing study.

One-Way ANOVA

CONCEPT The hypothesis test that simultaneously compares the differences among the population means of more than two groups for a one-factor experiment.

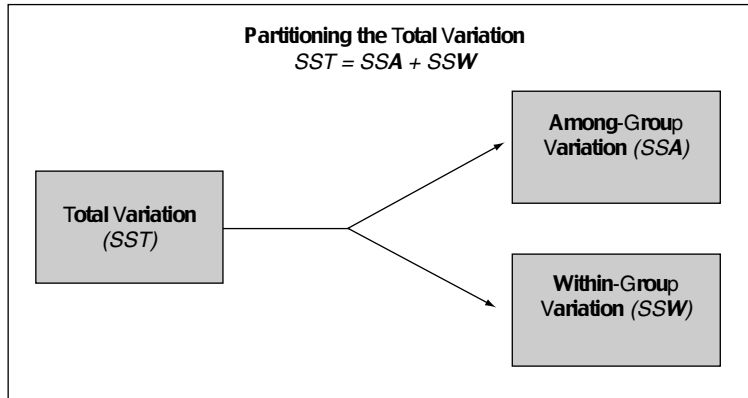
INTERPRETATION Unlike the t test, which compares differences in two means, the analysis of variance simultaneously compares the differences among the means of more than two groups. Although ANOVA is an acronym for Analysis Of VAriance, the term is misleading, because the objective is to analyze differences among the group means, not the variances. The null hypothesis of this test is:

H_0 : (All the population means are equal.)

and the alternative hypothesis is:

H_1 : (Not all the population means are equal.)

In ANOVA, the total variation in the measurements is subdivided into variation that is due to differences *among* the groups and variation that is due to variation *within* the groups (see the figure below). Within group variation is called **experimental error**, and the group variation that represents variation due to the factor of interest is called the **treatment effect**.



The **sum of squares total (SST)** is the total variation that represents the sum of the squared differences between each individual value and the mean of all the values that is based on all the values in all the groups combined:

$$SST = \text{Sum of } (\text{Each value} - \text{Mean of all values})^2$$

The **sum of squares among groups (SSA)** is the among-group variation that represents the sum of the squared differences between the sample mean of each group and the mean of all the values, weighted by the sample size in each group:

$$SSA = \text{sum of } [(\text{Sample size in each group})(\text{Group mean} - \text{Mean of all values})^2]$$

The **sum of squares within groups (SSW)** is the within-group variation that measures the difference between each value and the mean of its own group and sums the squares of these differences over all groups:

$$SSW = \text{Sum of } [(\text{Each value in the group} - \text{Group mean})^2]$$

The one-way analysis of variance is the simplest type of experimental design, because it evaluates only one factor of interest. More complicated experimental designs examine at least two factors of interest simultaneously. For more information on these designs, see References 1, 4 through 6, 8, and 9.

The Three Variances of ANOVA

ANOVA derives its name from the fact that the comparison of the means of the groups is achieved by comparing variances. In Section 3.2 on page 48, the variance was computed as a sum of squared differences around the mean divided by the sample size minus 1. This sample size minus 1 represents the actual number of values that are free to vary once the mean is known and is called the **degrees of freedom**.

In the analysis of variance, there are three different variances: the variance among groups, the variance within groups, and the total variance. These variances are referred to in the analysis-of-variance terminology as **mean squares**. The **mean square among groups (MSA)** is equal to the sum of squares among groups (SSA) divided by the number of groups minus 1. The **mean square within groups (MSW)** is equal to the sum of squares within groups (SSW) divided by the sample size minus the number of groups. The **mean square total (MST)** is equal to the sum of squares total (SST) divided by the sample size minus 1.

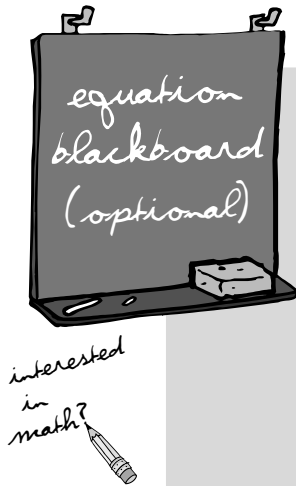
To test the null hypothesis:

$$H_0: \text{All the population means are equal.}$$

against the alternative:

$$H_1: \text{Not all the population means are equal.}$$

You calculate the test statistic F , which follows the F distribution (see Table C.4), as the ratio of two of the variances, MSA to MSW.



To form the equations for the three mean squares and the test statistic F , you assemble these symbols:

- $\bar{\bar{X}}$, pronounced as “X double bar,” which represents the overall or grand mean of all the values
- A subscripted X bar, \bar{X}_j , which represents the mean of a group
- A double-subscripted uppercase italic X, X_{ij} , which represents individual values of the independent variable X
- A subscripted lowercase italic n, n_j , which represents the sample size of a group
- A lowercase italic n, n , which represents the total sample size (sum of the sample sizes of each group)
- A lowercase italic c, c , which represents the number of groups

(continues)

First form the equation for the grand mean as follows:

$$\bar{\bar{X}} = \frac{\sum_{j=1}^c \sum_{i=1}^{n_j} X_{ij}}{n} = \text{grand mean}$$

\bar{X}_j = sample mean of group j

X_{ij} = i th value in group j

n_j = number of values in group j

n = total number of values in all groups combined

(that is, $n = n_1 + n_2 + \dots + n_c$)

c = number of groups of the factor of interest

With $\bar{\bar{X}}$ defined, next form the equations that define the sum of squares total, SST , the sum of squares among groups, SSA , and the sum of squares within groups, SSW :

$$SST = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{\bar{X}})^2$$

$$SSA = \sum_{j=1}^c n_j (\bar{X}_j - \bar{\bar{X}})^2$$

$$SSW = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

Next, using these definitions, form the equations for the mean squares:

$$MSA = \frac{SSA}{\text{number of groups} - 1}$$

$$MSW = \frac{SSW}{\text{sample size} - \text{number of groups}}$$

$$MST = \frac{SST}{\text{sample size} - 1}$$

Finally, using the definitions of MSA and MSW , form the equation for the test statistic F :

$$F = \frac{MSA}{MSW}$$

ANOVA Summary Table

The results of an analysis of variance are usually displayed in an ANOVA **summary table** (shown below). The entries in this table include the sources of variation (among-group, within-group, and total), the degrees of freedom, the sums of squares, the mean squares (or variances), and the F test statistic. In addition, when you use software such as Microsoft Excel, the p -value is included in the ANOVA table.

Analysis of Variance Summary Table

Source	Degrees of Freedom	Sum of Squares	Mean Square (Variance)	F
Among groups	Number of groups - 1	SSA	$MSA = \frac{SSA}{\text{number of groups} - 1}$	$F = \frac{MSA}{MSW}$
Within groups	Sample size - number of groups	SSW	$MSW = \frac{SSW}{\text{sample size} - \text{number of groups}}$	
Total	Sample size - 1	SST		

important point



After performing a one-way ANOVA and finding a significant difference among groups, you do not know which groups are significantly different. All that is known is that there is sufficient evidence to state that the population means are not all the same. To determine exactly which groups differ, all possible pairs of groups would need to be compared. Many statistical procedures for making these comparisons have been developed (see References 1 through 6 and 8 through 10).

WORKED-OUT PROBLEM 1a You seek to determine, with a level of significance $\alpha = 0.05$, whether differences exist among three sets of mathematics learning materials (labeled A, B, and C). You devise an experiment that randomly assigns 24 students to one of the three sets of materials. At the end of a school year, all 24 students are given the same standardized mathematics test that is scored on a 0 to 100 scale, the results of which are as follows.

A	B	C
87	58	81
80	63	62
74	64	70
82	75	64
74	70	70
81	73	72
97	80	92
71	62	63

(Math)

Microsoft Excel and statistical calculator (two screens) results for this study are as follows:

	A	B	C	D	E	F	G
1	Anova: Single Factor for Learning Materials Study						
2							
3	SUMMARY						
4	Groups	Count	Sum	Average	Variance		
5	A	8	646	80.75	70.2143		
6	B	8	545	68.125	56.9821		
7	C	8	574	71.75	104.7857		
8							
9							
10	ANOVA						
11	Source of Variation	SS	df	MS	F	P-value	F crit
12	Between Groups	676.0833	2	338.0417	4.3716	0.0259	3.4668
13	Within Groups	1623.8750	21	77.3274			
14							
15	Total	2299.9583	23				

```

One-way ANOVA
F=4.37156493
p=.0258667931
Factor
df=2
SS=676.083333
↓ MS=338.041667

```

```

One-way ANOVA
↑ MS=338.041667
Error
df=21
SS=1623.875
MS=77.327381
SxP=8.79359886

```

Because the p -value for this test, 0.0259, is less than the level of significance $\alpha = 0.05$, you reject the null hypothesis. You can conclude that the mean scores are not the same for all the sets of mathematics materials. From the output provided, you see that the mean for materials A is 80.75, for materials B it is 68.125, for materials C it is 71.75. It appears that the mean score is higher for materials A than for materials B and C.

WORKED-OUT PROBLEM 1b Using the critical value approach for the same problem, the computed F statistic is 4.37 (see results above). To determine the critical value of F , you refer to the table of the F statistic (Table C.4). This table requires these degrees of freedom:

- The numerator degrees of freedom, equal to the number of groups minus 1
- The denominator degrees of freedom, equal to the sample size minus the number of groups

With three groups and a sample size of 24, the numerator degrees of freedom are 2 ($3 - 1 = 2$), and the denominator degrees of freedom are 21 ($24 - 3 = 21$). With the level of significance $\alpha = 0.05$, the critical value of F from Table C.4 is 3.47 (also shown in the Microsoft Excel worksheet). Because the decision rule is to reject H_0 if $F > \text{critical value of } F$, and $F = 4.37 > 3.47$, you reject the null hypothesis.



calculator keys

One-Way ANOVA

To perform a one-way ANOVA on group data previously entered as the values of a list variable (see Chapter 1), press [STAT] [◀] (to display the Tests menu), select F:ANOVA, and press [ENTER] to display the ANOVA function. Enter the list variable names, separated by commas, that contain the data values for each group, and press [ENTER].



spreadsheet solution

One-Way ANOVA

Select Tools → Data Analysis, and in the Data Analysis dialog box select ANOVA: Single Factor and click OK. In the ANOVA dialog box, enter the cell range that contains the sample data for *all* groups as the Input Range and select either the Columns or Rows option, depending upon whether each group has been placed in its own column or row. (Data that are available for this book by download are arranged by columns.) If appropriate, check Labels in First Row. Then enter 0.05 as the Alpha value, select the New Worksheet Ply option, and click OK. Results appear on a new worksheet. (See Appendix D.3 for more information about the Data Analysis feature.)

WORKED-OUT PROBLEM 2a You seek to determine, with a level of significance $\alpha = 0.05$, whether differences exist among the four plants that fill boxes of a particular brand of cereal. You select samples of 20 cereal boxes from each of the four plants. The weights of these cereal boxes (in grams) are as follows.

Plant 1		Plant 2		Plant 3		Plant 4	
361.43	364.78	370.26	360.27	367.53	390.12	361.95	369.36
368.91	376.75	357.19	362.54	388.36	335.27	381.95	363.11
365.78	353.37	360.64	352.22	359.33	366.37	383.90	400.18
389.70	372.73	398.68	347.28	367.60	371.49	358.07	358.61
390.96	363.91	380.86	350.43	358.06	358.01	382.40	370.87

Plant 1		Plant 2		Plant 3		Plant 4	
372.62	375.68	334.95	376.50	369.93	373.18	386.20	380.56
390.69	380.98	359.26	369.27	355.84	377.40	373.47	376.21
364.93	354.61	389.56	377.36	382.08	396.30	381.16	380.97
387.13	378.03	371.38	368.50	381.45	354.82	379.41	365.78
360.77	374.24	373.06	363.86	356.20	383.78	382.01	395.55

(Boxfills)

Microsoft Excel and statistical calculator (two screens) results for this study are as follows:

	A	B	C	D	E	F	G
1	Anova: Single Factor for Cereal Filling Study						
2							
3	SUMMARY						
4	Groups	Count	Sum	Average	Variance		
5	Plant 1	20	7448	372.4	132.1037		
6	Plant 2	20	7324.07	366.2035	218.1177		
7	Plant 3	20	7393.12	369.656	222.0002		
8	Plant 4	20	7531.72	376.586	131.1284		
9							
10							
11	ANOVA						
12	Source of Variation	SS	df	MS	F	P-value	F crit
13	Between Groups	1155.9485	3	385.3162	2.1913	0.0959	2.7249
14	Within Groups	13363.6500	76	175.8375			
15							
16	Total	14519.5985	79				

```

One-way ANOVA
F=2.191319698
p=.0959375598
Factor
df=3
SS=1155.94853
↓ MS=385.316178

```

```

One-way ANOVA
↑ MS=385.316178
Error
df=76
SS=13363.65
MS=175.8375
SxP=13.2603733

```

Because the p -value for this test, 0.0959, is greater than the level of significance $\alpha = 0.05$, you cannot reject the null hypothesis. You conclude that there is insufficient evidence of a difference in the mean cereal weights among the four plants.

WORKED-OUT PROBLEM 3b Using the critical value approach for the same problem, the computed F statistic is 2.19. At the level of significance $\alpha = 0.05$, with 3 degrees of freedom in the numerator ($4 - 1$) and 76 degrees of freedom in the denominator ($80 - 4$), the critical value of F from Table C.4 is 2.725. Because the computed F test statistic 2.19 is less than 2.725, you do not reject the null hypothesis.

One-Way ANOVA Assumptions

There are three major assumptions you must make to use the one-way ANOVA F test: randomness and independence, normality, and homogeneity of variance.

The first assumption, **randomness and independence**, always must be met, because the validity of any experiment depends on random sampling and/or randomizing the assignment of items or subjects to groups. Departures from this assumption can seriously affect inferences from the analysis of variance. These problems are discussed more thoroughly in Reference 8.

The second assumption, **normality**, states that the values in each group are drawn from normally distributed populations. The one-way ANOVA F test is not very sensitive to departures from this assumption of normality. As long as the distributions are not very skewed, the level of significance of the ANOVA F test is usually not greatly affected by lack of normality, particularly for large samples. When only the normality assumption is seriously violated, nonparametric alternatives to the one-way ANOVA F test are available (see References 1–3, 5, 6, and 8 through 10).

The third assumption, **equality of variances**, states that the variance within each population should be equal for all populations. Although the one-way ANOVA F test is relatively robust or insensitive with respect to the assumption of equal group variances, large departures from this assumption may seriously affect the level of significance and the power of the test. Therefore, various procedures have been developed to test the assumption of homogeneity of variance (see References 1, 4, and 5).

One way to evaluate the assumptions is to plot a side-by-side box-and-whisker plot of the groups to study their central tendency, variation, and shape.

Important Equations

Chi-square test for a two-way cross-classification table:

$$(9.1) \quad \chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

ANOVA calculations:

$$(9.2) \quad SST = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{\bar{X}})^2$$

$$(9.3) \quad SSA = \sum_{j=1}^c n_j (\bar{X}_j - \bar{\bar{X}})^2$$

$$(9.4) \quad SSW = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

$$(9.5) \quad MSA = \frac{SSA}{\text{number of groups} - 1}$$

$$(9.6) \quad MSW = \frac{SSW}{\text{sample size} - \text{number of groups}}$$

$$(9.7) \quad MST = \frac{SST}{\text{sample size} - 1}$$

$$(9.8) \quad F = \frac{MSA}{MSW}$$

One-Minute Summary

Tests for the differences among more than two groups:

- If your data are categorical, use chi-square (χ^2) tests (can also use for two groups).
- If your data are numerical and if you have one factor, use the one-way ANOVA.

Test Yourself

1. In a one-way ANOVA, if the F test statistic is greater than the critical F value, you:
 - (a) reject H_0 because there is evidence all the means differ
 - (b) reject H_0 because there is evidence at least one of the means differs from the others
 - (c) do not reject H_0 because there is no evidence of a difference in the means
 - (d) do not reject H_0 because one mean is different from the others
2. In a one-way ANOVA, if the p -value is greater than the level of significance, you:
 - (a) reject H_0 because there is evidence all the means differ
 - (b) reject H_0 because there is evidence at least one of the means differs from the others.
 - (c) do not reject H_0 because there is no evidence of a difference in the means
 - (d) do not reject H_0 because one mean is different from the others

3. The F test statistic in a one-way ANOVA is:
 - (a) MSW/MSA
 - (b) SSW/SSA
 - (c) MSA/MSW
 - (d) SSA/SSW
4. In a one-way ANOVA, the null hypothesis is always:
 - (a) all the population means are different
 - (b) some of the population means are different
 - (c) some of the population means are the same
 - (d) all of the population means are the same
5. A car rental company wants to select a computer software package for its reservation system. Three software packages (A, B, and C) are commercially available. The car rental company will choose the package that has the lowest average number of renters for whom a car is not available at the time of pickup. An experiment is set up in which each package is used to make reservations for five randomly selected weeks. How should the data be analyzed?
 - (a) Chi-square test for differences in proportions
 - (b) One-way ANOVA F test
 - (c) t test for the differences in means
 - (d) t test for the mean difference

The following should be used to answer Questions 6 through 9.

For fast-food restaurants, the drive-through window is an increasing source of revenue. The chain that offers that fastest service is considered most likely to attract additional customers. In a study of 20 drive-through times (from menu board to departure) at 5 fast-food chains, the following ANOVA table was developed.

Source	DF	Sum of Squares	Mean Squares	F
Among groups (chains)		6,536	1,634.0	12.51
Within groups (chains)	95		130.6	
Total	99	18,943		

6. Referring to the table above, the among groups degrees of freedom is:
 - (a) 3
 - (b) 4
 - (c) 12
 - (d) 16

7. Referring to the table on page 176, the within groups sum of squares is:
 - (a) 12,407
 - (b) 95
 - (c) 130.6
 - (d) 4
8. Referring to the table on page 176, the within groups mean squares is:
 - (a) 12,407
 - (b) 95
 - (c) 130.6
 - (d) 4
9. Referring to the table on page 176, at the 0.05 level of significance, you:
 - (a) do not reject the null hypothesis and conclude that there is no difference in the average
 - (b) do not reject the null hypothesis and conclude that there is a difference in the average drive-up time between the fast-food chains
 - (c) reject the null hypothesis and conclude that there is a difference in the average drive-up time between the fast-food chains
 - (d) reject the null hypothesis and conclude that there is no difference in the average drive-up time between the fast-food chains
10. When testing for independence in a contingency table with three rows and four columns, there are _____ degrees of freedom.
 - (a) 5
 - (b) 6
 - (c) 7
 - (d) 12
11. In testing a hypothesis using the chi-square test, the theoretical frequencies are based on the:
 - (a) null hypothesis
 - (b) alternative hypothesis
 - (c) normal distribution
 - (d) t distribution
12. An agronomist is studying three different varieties of tomato to determine whether there is a difference in the proportion of seeds that germinate. Random samples of 100 seeds of each of three varieties are subjected to the same starting conditions. How should the data be analyzed?
 - (a) Chi-square test for differences in proportions
 - (b) One-way ANOVA F test
 - (c) t test for the differences in means
 - (d) t test for the mean difference

The following are True or False Questions:

13. A test for the difference between two proportions can be performed using the chi-square distribution.
14. The analysis-of-variance (ANOVA) tests hypotheses about the difference between population proportions.
15. The one-way analysis-of-variance (ANOVA) tests hypotheses about the difference between population means.

Answers to Test Yourself Questions

1. b
2. c
3. c
4. d
5. b
6. b
7. a
8. c
9. c
10. b
11. a
12. a
13. True
14. False
15. True

References

1. Berenson, M. L., D. M. Levine, and T. C. Krehbiel. *Basic Business Statistics: Concepts and Applications, Ninth Edition*. Upper Saddle River, NJ: Prentice Hall, 2004.
2. Conover, W. J. *Practical Nonparametric Statistics, Third Edition*. New York: Wiley, 2000.
3. Daniel, W. *Applied Nonparametric Statistics, Second Edition*. Boston: Houghton Mifflin, 1990.
4. Gitlow, H. S., and D. M. Levine. *Six Sigma for Green Belts and Champions*. Upper Saddle River, NJ: Financial Times - Prentice Hall, 2005.

5. Levine, D. M., D. Stephan, T. C. Krehbiel, and M. L. Berenson. *Statistics for Managers Using Microsoft Excel, Fourth Edition*. Upper Saddle River, NJ: Prentice Hall, 2005.
6. Levine, D. M., P. P. Ramsey, and R. K. Smidt. *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Upper Saddle River, NJ: Prentice Hall, 2001.
7. Microsoft Excel 2002. Redmond, WA: Microsoft Corporation, 2001.
8. Montgomery, D. C. *Design and Analysis of Experiments, Fifth Edition*. New York: John Wiley, 2001.
9. Neter, J., M. H. Kutner, C. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models, Fourth Edition*. Homewood, IL: Richard D. Irwin, 1996.
10. Sincich, T., D. M. Levine, and D. Stephan. *Practical Statistics by Example Using Microsoft Excel and Minitab, Second Edition*. Upper Saddle River, NJ: Prentice Hall, 2002.

This page intentionally left blank



Regression Analysis

10.1 Basics of Regression Analysis

10.2 Determining the Simple Linear Regression Equation

10.3 Measures of Variation

10.4 Regression Assumptions

10.5 Residual Analysis

10.6 Inferences About the Slope

10.7 Common Mistakes Using Regression Analysis

Important Equations

One-Minute Summary

Test Yourself

Chapter 7 compared the inferential methods of hypothesis testing to the parts of the scientific method that state tentative hypotheses about natural phenomena and then attempt to study those hypotheses through investigation and testing.

The goal of scientific investigation and testing is to develop statements, laws, and theorems that explain or predict natural phenomena. One famous statement, developed by Albert Einstein and known by the equation $E = mc^2$, relates to the amount of energy, E , that a quantity of matter, m , contains.

Likewise, when doing statistical analysis, you may want to explore how the values of one variable might influence another variable. For example, managers of a growing chain of retail stores may wonder if larger-sized stores generate greater sales. Descriptive and inferential statistical methods known as **regression analysis** allow you to explore such educated guesses and allow you to explain or predict mathematical relationships among variables. In this chapter, you will learn the basic concepts and principles of regression analysis and the statistical assumptions necessary for performing regression analysis as well as how to do the following:

- Predict the value of a variable of interest
- Understand the meaning of the Y intercept and the slope
- Make inferences about the significance of the slope

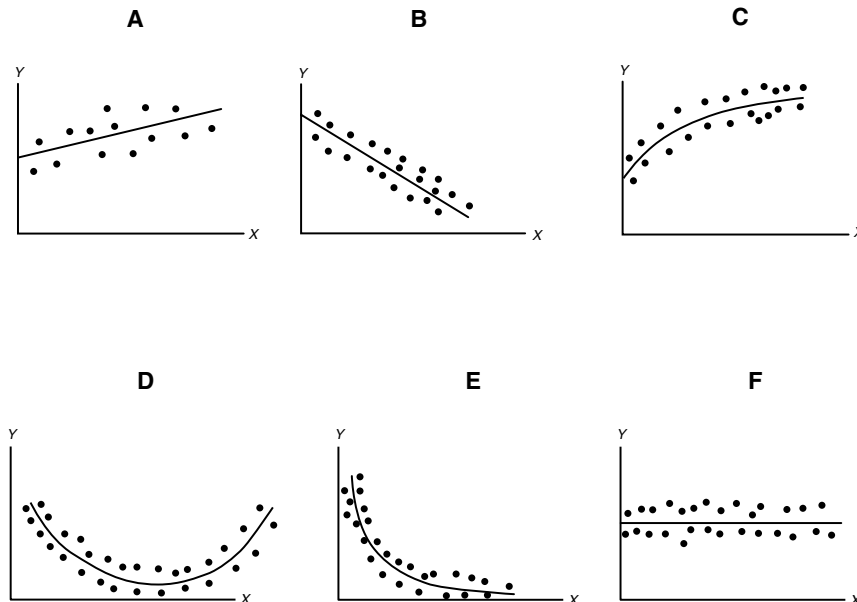
10.1 Basics of Regression Analysis

In regression analysis, you seek to develop a model that can be used to predict the values of a **dependent** or **response variable** based on the values of one or more **explanatory** or **independent variables**. In this chapter, you will learn about *simple linear regression*, in which you use a single explanatory numerical variable (such as size of store) to predict a numerical dependent variable (such as store sales). Other types of regression such as *multiple regression*, in which several explanatory variables are used to predict a numerical dependent variable (see References 2 through 4, 6, and 7) and *logistic regression*, in which the dependent variable is categorical (see Reference 2), are not further described in this book.

Simple Linear Regression

CONCEPT A statistical technique that uses a straight-line relationship to predict a numerical dependent variable Y from a *single* numerical independent variable X .

INTERPRETATION Simple *linear* regression attempts to discover whether the values of the dependent Y (such as store sales) and the independent X variable (such as the size of the store), when graphed on a scatter plot (see Section 2.2), would suggest a straight-line relationship of the values. The figure below shows the different types of patterns that you could discover when plotting the values of the X and Y variables.



The patterns shown on page 182 can be described as follows:

- Panel A, positive straight-line or linear relationship between X and Y .
- Panel B, negative straight-line or linear relationship between X and Y .
- Panel C, a positive curvilinear relationship between X and Y . The values of Y are increasing as X increases, but this increase tapers off beyond certain values of X .
- Panel D, a U-shaped relationship between X and Y . As X increases, at first Y decreases. However, as X continues to increase, Y not only stops decreasing but actually increases above its minimum value.
- Panel E, an exponential relationship between X and Y . In this case, Y decreases very rapidly as X first increases, but then decreases much less rapidly as X increases further.
- Panel F, values that have very little or no relationship between X and Y . High and low values of Y appear at each value of X .

Scatter plots only informally help you identify the relationship between the dependent variable Y and the independent variable X in a simple regression. To specify the numeric relationship between the variables, you need to develop an equation that best represents the relationship.

10.2 Determining the Simple Linear Regression Equation

After you determine that a straight-line relationship exists between a dependent variable Y and the independent variable X , you need to determine which straight line to use to represent the relationship. Two values define any straight line: the Y *intercept* and the *slope*.

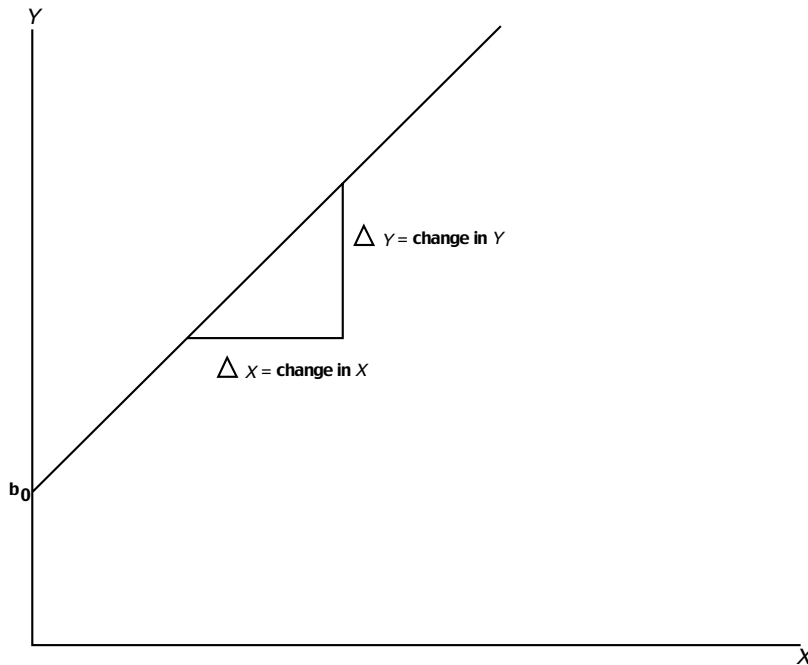
Y intercept

CONCEPT The value of Y when $X = 0$, represented by the symbol b_0 .

Slope

CONCEPT The change in Y per unit change in X represented by the symbol b_1 . Positive slope means Y increases as X increases. Negative slope means Y decreases as X increases.

INTERPRETATION The Y intercept and the slope are known as the **regression coefficients**. The symbol b_0 is used for the Y intercept, and the symbol b_1 is used for the slope. Multiplying a specific X value by the slope and then adding the Y intercept generates the corresponding Y value. The equation



$Y = b_0 + b_1X$ is used to express this relationship for the entire line. (Some sources use the symbol a for the Y intercept and b for the slope to form the equation $Y = a + bX$.)

Least-Squares Method

CONCEPT The simple linear regression method that seeks to minimize the sum of the squared differences between the actual values of the dependent variable Y and the predicted values of Y .

INTERPRETATION For plotted sets of X and Y values, there are many possible straight lines, each with its own values of b_0 and b_1 , that might seem to fit the data. The least-squares method finds the values for the Y intercept and the slope that makes the sum of the squared differences between the actual values of the dependent variable Y and the predicted values of Y as small as possible.

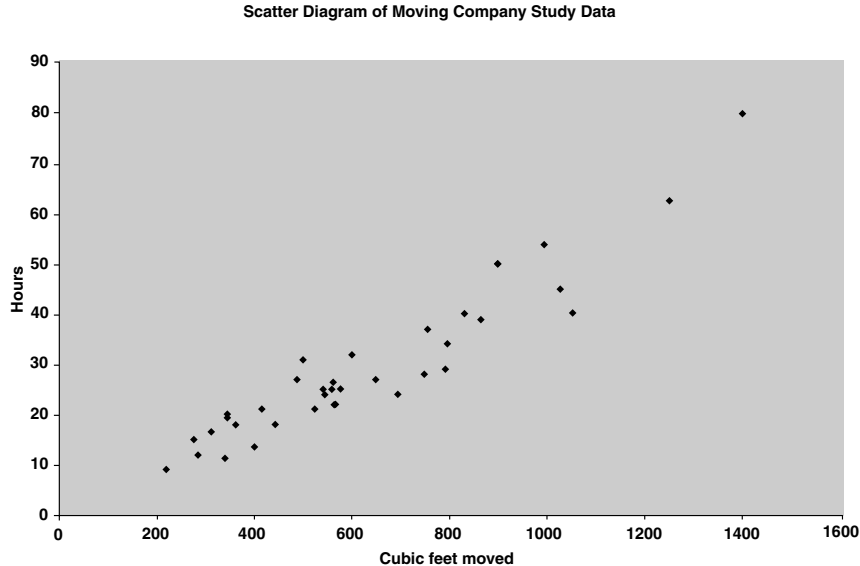
Calculating the Y intercept and the slope using the least-squares method is tedious and can be subject to rounding errors if you use a simple four-function calculator. You will get more accurate results faster if you use regression software routines to perform the calculations.

WORKED-OUT PROBLEM 1 You seek to assist a moving company owner to develop a more accurate method of predicting the labor hours needed for a moving job by using the volume of goods (in cubic feet) that is being

moved. The manager has collected the following data for 36 moves and has eliminated the travel-time portion of the time needed for the move.

Hours	Cubic Feet Moved	
24	545	
13.5	400	
26.25	562	
25	540	
9	220	
20	344	
22	569	
11.25	340	
50	900	
12	285	
38.75	865	
40	831	
19.5	344	
18	360	
28	750	
27	650	
21	415	
15	275	
25	557	
45	1,028	
29	793	
21	523	
22	564	
16.5	312	
37	757	
32	600	
34	796	
25	577	
31	500	
24	695	
40	1,054	
27	486	
18	442	
62.5	1,249	
53.75	995	
79.5	1,397	(Moving)

The scatter diagram for these data (shown below) indicates an increasing relationship between cubic feet moved (X) and labor hours (Y). As the cubic footage moved increases, labor hours increase approximately as a straight line.



Microsoft Excel regression results for this study are as follows:

	A	B	C	D	E	F	G
1	Regression Analysis for Moving Company Study						
2							
3	Regression Statistics						
4	Multiple R	0.9430					
5	R Square	0.8892					
6	Adjusted R Square	0.8860					
7	Standard Error	5.0314					
8	Observations	36					
9							
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	1	6910.7189	6910.7189	272.9864	0.0000	
13	Residual	34	860.7186	25.3153			
14	Total	35	7771.4375				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	-2.3697	2.07326	-1.14296	0.26104	-6.58303	1.84371
18	Cubic Feet Moved	0.0501	0.00303	16.52230	0.00000	0.04392	0.05624

The results show that $b_1 = 0.05$ and $b_0 = -2.37$. Thus, the equation for the best straight line for these data is this:

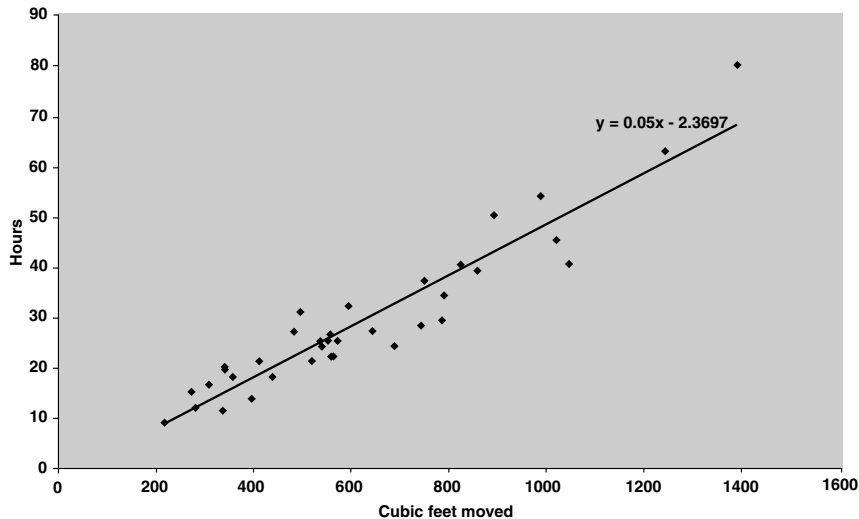
$$\text{Predicted value of labor hours} = -2.37 + 0.05 \times \text{Cubic feet moved}$$

The slope b_1 was computed as +0.05. This means that for each increase of 1 unit in X , the average value of Y is estimated to increase by 0.05 units. In

other words, for each increase of 1 cubic foot to be moved, the fitted model predicts that the expected labor hours are estimated to increase by 0.05 hours.

The Y intercept b_0 was computed to be -2.37 . The Y intercept represents the average value of Y when X equals 0. Because the cubic feet moved cannot be 0, the Y intercept has no practical interpretation. The sample linear regression line for these data, along with the actual values, is shown below.

Scatter Diagram of Moving Company Study Data

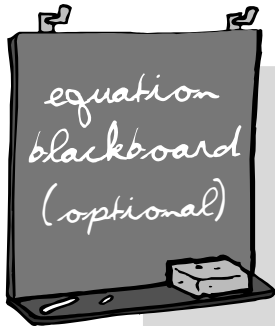


Regression Model Prediction

Once developed, you can use a regression model for predicting values of a dependent variable Y from the independent variable X . However, you are restricted to the **relevant range** of the independent variable in making predictions. This range is all the values from the smallest to the largest X used to develop the regression model. You should not extrapolate beyond the range of X values. For example, when you use the model developed in Worked-out Problem 1, predictions of labor hours should be made *only* for moves whose cubic footage is between 220 and 1,397.

important point

WORKED-OUT PROBLEM 2 Using the regression model developed in Worked-out Problem 1, you want to predict the average labor hours for a moving job that consists of 800 cubic feet. You predict that the average labor hours for a move would be 37.69 ($-2.3697 + 0.0501 \times 800$).



You use the symbols for Y intercept, b_0 , and the slope, b_1 , the sample size, n , and these symbols:

- The subscripted \hat{Y} , \hat{Y}_i , for predicted Y values
- The subscripted italic capital X for the independent X values
- The subscripted italic capital Y for the dependent Y values
- \bar{X} for the mean or average of the X values
- \bar{Y} for the mean or average of the Y values

To write the equation for a simple linear regression model:

$$\hat{Y}_i = b_0 + b_1 X_i$$

You use this equation and these summations:

- $\sum_{i=1}^n X_i$, the sum of the X values
- $\sum_{i=1}^n Y_i$, the sum of the Y values
- $\sum_{i=1}^n X_i^2$, the sum of the squared X values
- $\sum_{i=1}^n X_i Y_i$, the sum of the cross product of X and Y

to define the equation of the slope, b_1 , as:

$$b_1 = \frac{SSXY}{SSX}, \text{ in which}$$

$$SSXY = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}$$

and

$$SSX = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}$$

These equations, in turn, allow you to define the Y intercept as:

$$b_0 = \bar{Y} - b_1 \bar{X}$$

(continues)

For the moving company problem, these sums and the sum of the squared Y values ($\sum_{i=1}^n Y_i^2$) used for calculating the sum of squares total (SST) on page 192 are as follows:

Move	Hours (Y)	Cubic Feet Moved (X)	X^2	Y^2	XY
1	24	545	297,025	576	13,080
2	13.5	400	160,000	182.25	5,400
3	26.25	562	315,844	689.0625	14,752.5
4	25	540	291,600	625	13,500
5	9	220	48,400	81	1,980
6	20	344	118,336	400	6,880
7	22	569	323,761	484	12,518
8	11.25	340	115,600	126.5625	3,825
9	50	900	810,000	2,500	45,000
10	12	285	81,225	144	3,420
11	38.75	865	748,225	1,501.5625	33,518.75
12	40	831	690,561	1,600	33,240
13	19.5	344	118,336	380.25	6,708
14	18	360	129,600	324	6,480
15	28	750	562,500	784	21,000
16	27	650	422,500	729	17,550
17	21	415	172,225	441	8,715
18	15	275	75,625	225	4,125
19	25	557	310,249	625	13,925
20	45	1,028	1,056,784	2,025	46,260
21	29	793	628,849	841	22,997
22	21	523	273,529	441	10,983
23	22	564	318,096	484	12,408
24	16.5	312	97,344	272.25	5,148
25	37	757	573,049	1,369	28,009
26	32	600	360,000	1,024	19,200

(continues)

Move	Hours (Y)	Cubic Feet Moved (X)	X ²	Y ²	XY
27	34	796	633,616	1,156	27,064
28	25	577	332,929	625	14,425
29	31	500	250,000	961	15,500
30	24	695	483,025	576	16,680
31	40	1,054	1,110,916	1,600	42,160
32	27	486	236,196	729	13,122
33	18	442	195,364	324	7,956
34	62.5	1249	1,560,001	3,906.25	78,062.5
35	53.75	995	990,025	2,889.0625	53,481.25
36	79.5	1,397	1,951,609	6,320.25	111,061.5
Sums:	1,042.5	22,520	16,842,944	37,960.50	790,134.50

Using these sums, you can compute the values of the slope b_1 :

$$\begin{aligned}
 SSXY &= \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right)}{n} \\
 SSXY &= 790,134.5 - \frac{(22,520)(1,042.5)}{36} \\
 &= 790,134.5 - 652,141.66 \\
 &= 137,992.84
 \end{aligned}$$

$$\begin{aligned}
 SSX &= \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i \right)^2}{n} \\
 &= 16,842,944 - \frac{(22,520)^2}{36} \\
 &= 16,842,944 - 14,087,511.11 \\
 &= 2,755,432.889
 \end{aligned}$$

$$\text{Because } b_1 = \frac{SSXY}{SSX}$$

$$\begin{aligned}
 b_1 &= \frac{137,992.84}{2,755,432.889} \\
 &= 0.05008
 \end{aligned}$$

(continues)

With the value for slope b_1 , you can calculate the Y intercept as follows:

First calculate the average Y (\bar{Y}) and the average X (\bar{X}) values:

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} = \frac{1,042.5}{36} = 28.9583$$

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{22,520}{36} = 625.5555$$

Then use the results in the following equation:

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$$\begin{aligned} b_0 &= 28.9583 - (0.05008)(625.5555) \\ &= -2.3695 \end{aligned}$$

10.3 Measures of Variation

After a regression model has been fit to a set of data, three measures of variation determine how much of the variation in the dependent variable Y can be explained by variation in the independent variable X.

Regression Sum of Squares (SSR)

CONCEPT The variation that is due to the relationship between X and Y.

INTERPRETATION The regression sum of squares (SSR) is equal to the sum of the squared differences between the Y values that are predicted from the regression equation and the average value of Y:

$$SSR = \text{Sum (Predicted Y value} - \text{Average Y value)}^2$$

Error Sum of Squares (SSE)

CONCEPT The variation that is due to factors other than the relationship between X and Y.

INTERPRETATION The error sum of squares (SSE) is equal to the sum of the squared differences between each observed Y value and the predicted value of Y:

$$SSE = \text{Sum (Observed Y value} - \text{predicted Y value)}^2$$

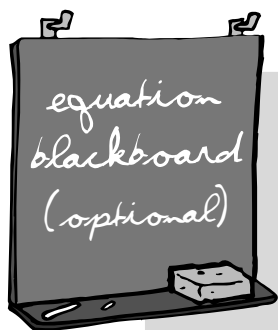
Total Sum of Squares (SST)

CONCEPT The measure of variation of the Y_i values around their mean.

INTERPRETATION The total sum of squares (SST) is equal to the sum of the squared differences between each observed Y value and the average value of Y :

$$SST = \text{Sum (Observed } Y \text{ value} - \text{Average } Y \text{ value)}^2$$

The total sum of squares is also equal to the sum of the regression sum of squares and the error sum of squares. For the Worked-out Problem of the previous section, the SSR is 6,910.7188867, the SSE (called residual) is 860.7186332, and the SST is 7,771.4375. (Note that 7,771.4375 is the sum of 6,910.7188867 and 860.7186332.)



You use symbols introduced earlier in this chapter to write the equations for the three measures of variation used in regression.

The equation for total sum of squares (SST) can be expressed in either of two ways:

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2 \text{ which is equivalent to } \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i \right)^2}{n}$$

or as:

$$SST = SSR + SSE$$

The equation for the regression sum of squares (SSR) is:

$$= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

which is equivalent to

$$= b_0 \sum_{i=1}^n Y_i + b_1 \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n Y_i \right)^2}{n}$$

The equation for the error sum of squares (SSE) is as follows:

SSE = unexplained variation or error sum of squares

$$= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \text{ which is equivalent to}$$

$$= \sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n X_i Y_i$$

(continues)

For the moving company problem on page 189:

$$SST = \text{total sum of squares} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n}$$

$$= 37,960.5 - \frac{(1,042.5)^2}{36}$$

$$= 37,960.5 - 30,189.0625$$

$$= 7,771.4375$$

SSR = regression sum of squares

$$= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$$= b_0 \sum_{i=1}^n Y_i + b_1 \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n}$$

$$= (-2.3695)(1,042.5) + (0.05008)(790,134.5) - \frac{(1,042.5)^2}{36}$$

SSE = error sum of squares

$$= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

$$= \sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n X_i Y_i$$

$$= 37,960.5 - (-2.3695)(1,042.5) - (0.05008)(790,134.5)$$

$$= 860.768$$

Calculated as $SSR + SSE$, the total sum of squares SST is 7,771.439, slightly different from the results from the first equation because of rounding errors.

The Coefficient of Determination

CONCEPT The ratio of the regression sum of squares to the total sum of squares, represented by the symbol r^2 .

INTERPRETATION By themselves, SSR , SSE , and SST provide little that can be directly interpreted. The ratio of the regression sum of squares (SSR) to the total sum of squares (SST) measures the proportion of variation in Y that is explained by the independent variable X in the regression model. The ratio can be expressed as follows:

$$r^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}} = \frac{SSR}{SST}$$

For the moving company problem, the $SSR = 6,910.7188867$ and the $SST = 7,771.4375$ (see regression results on page 186). Therefore:

$$r^2 = \frac{6,910.719}{7,771.4375} = 0.8892$$

This value means that 89% of the variation in labor hours can be explained by the variability in the cubic footage to be moved. This shows a strong positive linear relationship between two variables, because the use of a regression model has reduced the variability in predicting labor hours by 89%. Only 11% of the sample variability in labor hours can be explained by factors other than what is accounted for by the linear regression model that uses only cubic footage.

The Coefficient of Correlation

CONCEPT The measure of the strength of the linear relationship between two variables, represented by the symbol r .

INTERPRETATION The values of this coefficient vary from -1 , which indicates perfect negative correlation, to $+1$, which indicates perfect positive correlation. The sign of the correlation coefficient r is the same as the sign of the slope. If the slope is positive, r is positive. If the slope is negative, r is negative. The coefficient of correlation (r) is the square root of the coefficient of determination r^2 .

For the moving company problem, the coefficient of correlation, r , is 0.943 , the square root of 0.8892 (r^2). (Microsoft Excel labels the coefficient of correlation as “multiple r .”) Because the coefficient is very close to $+1.0$, you can say that the relationship between cubic footage moved and labor hours is very strong. You can plausibly conclude that the increased volume that had to be moved is associated with increased labor hours.



In general, you must remember that just because two variables are strongly correlated, you cannot always conclude that there is a cause-and-effect relationship between the variables.

Standard Error of the Estimate

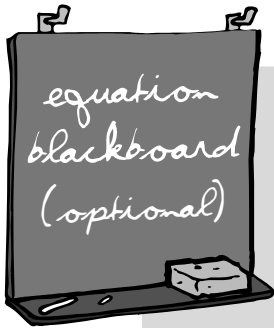
CONCEPT The standard deviation around the fitted line of regression that measures the variability of the actual Y values from the predicted Y , represented by the symbol S_{YX} .

INTERPRETATION Although the least-squares method results in the line that fits the data with the minimum amount of variation, unless the

coefficient of determination $r^2 = 1.0$, the regression equation is not a perfect predictor.

The variability around the line of regression was illustrated in the figure on page 187, which showed the scatter diagram and the line of regression for the moving company data. You can see from that figure that there are values above the line of regression as well as values below the line of regression. For the moving company problem, the standard error of the estimate (labeled as Standard Error in the figure on page 186) is equal to 5.03 hours.

Just as the standard deviation measures variability around the mean, the standard error of the estimate measures variability around the fitted line of regression. As you will see in Section 10.6, the standard error of the estimate can be used to determine whether a statistically significant relationship exists between the two variables.



interested
in
math?

You use symbols introduced earlier in this chapter to write the equation for the standard error of the estimate:

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}}$$

For the moving company problem, with SSE equal to 860.7186:

$$S_{YX} = \sqrt{\frac{860.7186}{36-2}}$$

$$S_{YX} = 5.0314$$

10.4 Regression Assumptions

The assumptions necessary for regression are similar to those of hypothesis testing. These assumptions are as follows:

- Normality of the variation around the line of regression
- Equality of variation in the Y values for all values of X
- Independence of the variation around the line of regression

The first assumption, **normality**, requires that the variation around the line of regression be normally distributed at each value of X. Like the *t* test and the ANOVA *F* test, regression analysis is fairly insensitive to departures from

important
point

the normality assumption. As long as the distribution of the variation around the line of regression at each level of X is not extremely different from a normal distribution, inferences about the line of regression and the regression coefficients will not be seriously affected.

The second assumption, **equality of variation**, requires that the variation around the line of regression be constant for all values of X . This means that the variation is the same when X is a low value as when X is a high value. The equality of variation assumption is important for using the least-squares method of determining the regression coefficients. If there are serious departures from this assumption, other methods (see References 2 and 6) can be used.

The third assumption, **independence of the variation around the line of regression**, requires that the variation around the regression line be independent for each value of X . This assumption is particularly important when data are collected over a period of time. In such situations, the variation around the line for a specific time period is often correlated with the variation of the previous time period.

10.5 Residual Analysis

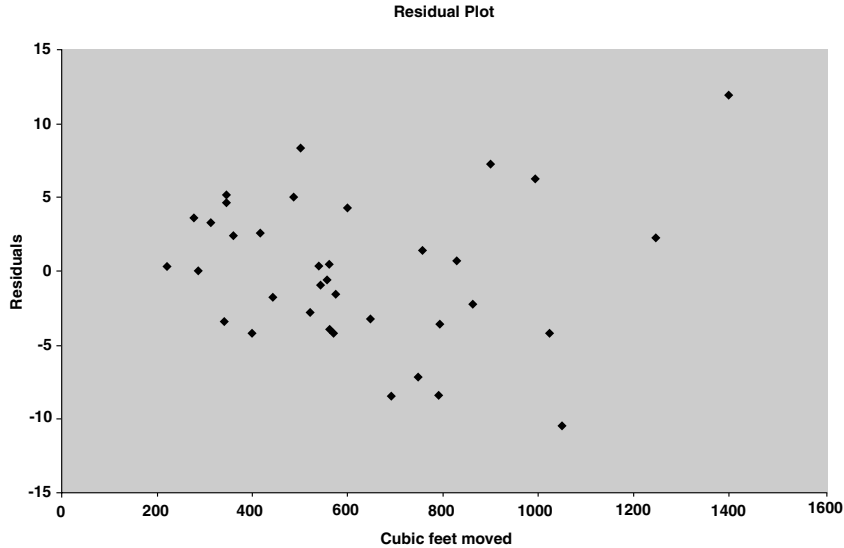
The graphical method, **residual analysis**, allows you to evaluate whether the regression model that has been fitted to the data is an appropriate model *and* determine whether there are violations of the assumptions of the regression model.

Residual

CONCEPT The difference between the observed and predicted values of the dependent variable Y for a given value of X .

INTERPRETATION To evaluate the aptness of the fitted model, you plot the residuals on the vertical axis against the corresponding X values of the independent variable on the horizontal axis. If the fitted model is appropriate for the data, there will be no apparent pattern in this plot. However, if the fitted model is not appropriate, there will be a clear relationship between the X values and the residuals.

A residual plot for the moving company problem fitted line of regression appears on page 197. In this figure, the cubic feet are plotted on the horizontal X -axis and the residuals are plotted on the vertical Y -axis. You see that although there is widespread scatter in the residual plot, there is no apparent pattern or relationship between the residuals and X . The residuals appear to be evenly spread above and below 0 for the differing values of X . This result enables you to conclude that the fitted straight-line model is appropriate for the moving company data.



Evaluating the Assumptions

Different techniques, all involving the residuals, allow you to evaluate the regression assumptions.

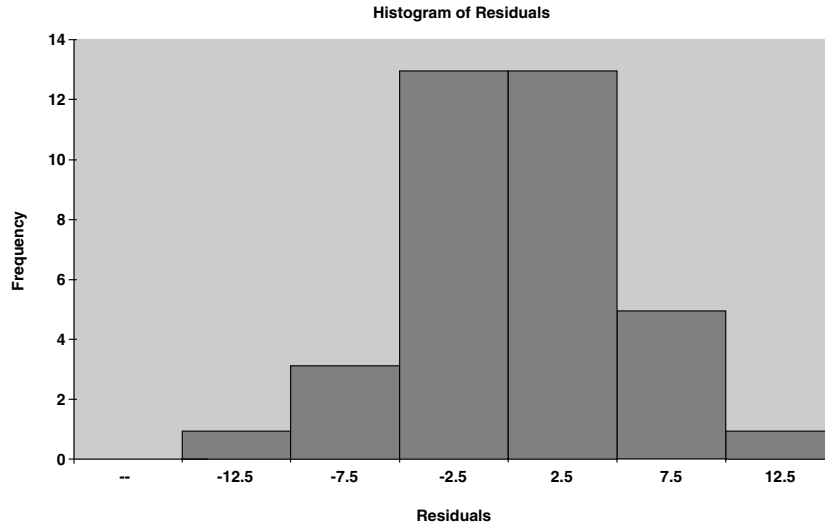
For equality of variation, you use the same plot to evaluate the aptness of the fitted model. For the moving company problem residual plot shown above, there do not appear to be major differences in the variability of the residuals for different X values. You can conclude that for this fitted model, there is no apparent violation in the assumption of equal variation at each level of X .

For the normality of the variation around the line of regression, you plot the residuals in a histogram (see Section 2.2), box-and-whisker plot (see Section 3.3), or a normal probability plot (see Section 5.4). From the histogram shown on page 198 for the moving company problem, you can see that the data appear to be approximately normally distributed, with most of the residuals concentrated in the center of the distribution.

For the independence of the variation around the line of regression, you plot the residuals in the order or sequence in which the observed data was obtained, looking for a relationship between consecutive residuals. If you can see such a relationship, the assumption of independence is violated.

10.6 Inferences About the Slope

You can make inferences about the linear relationship between the variables in a population based on your sample results after using residual analysis to



show that the assumptions of a least-squares regression model have not been seriously violated and that the straight-line model is appropriate.

***t* Test for the Slope**

You can determine the existence of a significant relationship between the X and Y variables by testing whether β_1 (the population slope) is equal to 0. If this hypothesis is rejected, you conclude that there is evidence of a linear relationship. The null and alternative hypotheses are as follows:

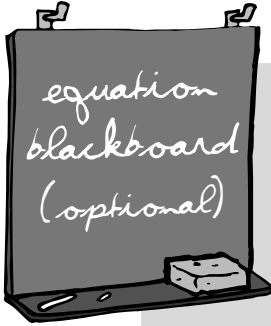
$$H_0: \beta_1 = 0 \text{ (There is no linear relationship.)}$$

$$H_1: \beta_1 \neq 0 \text{ (There is a linear relationship.)}$$

The test statistic follows the t distribution with the degrees of freedom equal to the sample size minus 2. The test statistic is equal to the sample slope divided by the standard error of the slope:

$$t = \frac{\text{sample slope}}{\text{standard error of the slope}}$$

For the moving company problem, the critical value of t for a level of significance of $\alpha = 0.05$ is 2.0322, the value of t is 16.52, and the p -value is 0.0000. (Microsoft Excel labels the t statistic “ t Stat” on page 186.) Using the p -value approach, you reject H_0 because the p -value of 0.00000 is less than $\alpha = 0.05$. Using the critical value approach, you reject H_0 because $t = 16.52 > 2.0322$. You can conclude that there is a significant linear relationship between labor hours and the cubic footage moved.



interested
in
math?

You assemble symbols introduced earlier and the symbol for the standard error of the slope, S_{b_1} , to form the equation for the t statistic used in testing a hypothesis for a population slope β_1 .

You begin by forming the equation for the standard error of the slope, S_{b_1} as follows:

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}}$$

In turn, you use the standard error of the slope S_{b_1} to define t :

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

The test statistic t follows a t distribution with $n - 2$ degrees of freedom.

For the moving company problem, to test whether there is a significant relationship between the cubic footage and the labor hours at the level of significance $\alpha = 0.05$, refer to the calculation of SSX on page 190 and the standard error of the estimate on page 195.

$$\begin{aligned} S_{b_1} &= \frac{S_{YX}}{\sqrt{SSX}} \\ &= \frac{5.0314}{\sqrt{2,755,432.889}} \\ &= 0.00303 \end{aligned}$$

Therefore, to test the existence of a linear relationship at the 0.05 level of significance, with

$$b_1 = +0.05008 \quad n = 36 \quad S_{b_1} = 0.00303$$

$$\begin{aligned} t &= \frac{b_1 - \beta_1}{S_{b_1}} \\ &= \frac{0.05008 - 0}{0.00303} = 16.52 \end{aligned}$$

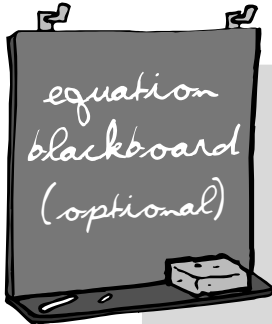
Confidence Interval Estimate of the Slope (β_1)

You can also test the existence of a linear relationship between the variables by calculating a confidence interval estimate of β_1 and seeing whether the hypothesized value ($\beta_1 = 0$) is included in the interval.

You calculate the confidence interval estimate of the slope β_1 by multiplying the t statistic by the standard error of the slope and then adding and subtracting this product to the sample slope.

For the moving company problem, the Microsoft Excel regression results on page 186 include the calculated lower and upper limits of the confidence interval estimate for the slope of cubic footage and labor hours. With 95% confidence, the lower limit is 0.0439 and the upper limit is 0.0562.

Because these values are above 0, you conclude that there is a significant linear relationship between labor hours and cubic footage moved. The confidence interval indicates that for each increase of 1 cubic foot moved, average labor hours are estimated to increase by at least 0.0439 hours but less than 0.0562 hours. Had the interval included 0, you would have concluded that no relationship exists between the variables.



interested
in
math?

You assemble symbols introduced earlier to form the equation for the confidence interval estimate of the slope β_1 :

$$b_1 \pm t_{n-2} S_{b_1}$$

For the moving company problem, b_1 has already been calculated on page 190, and the standard error of the slope, S_{b_1} , has already been calculated on page 199.

$$b_1 = +0.05008 \quad n = 36 \quad S_{b_1} = 0.00303$$

Thus, using 95% confidence, with degrees of freedom = $36 - 2 = 34$:

$$\begin{aligned} & b_1 \pm t_{n-2} S_{b_1} \\ & = +0.05008 \pm (2.0322)(0.00303) \\ & = +0.05008 \pm 0.0061 \\ & +0.0439 \leq \beta_1 \leq +0.0562 \end{aligned}$$

10.7 Common Mistakes Using Regression Analysis

Some of the common mistakes that people make when using regression analysis are as follows:

important point



- Lacking an awareness of the assumptions of least-squares regression
- Knowing how to evaluate the assumptions of least-squares regression
- Knowing what the alternatives to least-squares regression are if a particular assumption is violated
- Using a regression model without knowledge of the subject matter
- Predicting Y outside the relevant range of X

Most software regression analysis routines do not double-check for these mistakes. You must always use regression analysis wisely and always double-check that others who provide you with regression results have avoided these mistakes as well.

For example, the following four sets of data illustrate some of the mistakes that you can make during regression analysis.

Data Set A		Data Set B		Data Set C		Data Set D	
X_i	Y_i	X_i	Y_i	X_i	Y_i	X_i	Y_i
10	8.04	10	9.14	10	7.46	8	6.58
14	9.96	14	8.10	14	8.84	8	5.76
5	5.68	5	4.74	5	5.73	8	7.71
8	6.95	8	8.14	8	6.77	8	8.84
9	8.81	9	8.77	9	7.11	8	8.47
12	10.84	12	9.13	12	8.15	8	7.04
4	4.26	4	3.10	4	5.39	8	5.25
7	4.82	7	7.26	7	6.42	19	12.50
11	8.33	11	9.26	11	7.81	8	5.56
13	7.58	13	8.74	13	12.74	8	7.91
6	7.24	6	6.13	6	6.08	8	6.89

Source: F. J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, Vol. 27 (1973), 17–21.

Anscombe (Reference 1) showed that for the four data sets, the regression results are identical:

predicted value of $Y = 3.0 + 0.5X_i$

standard error of the estimate = 1.237

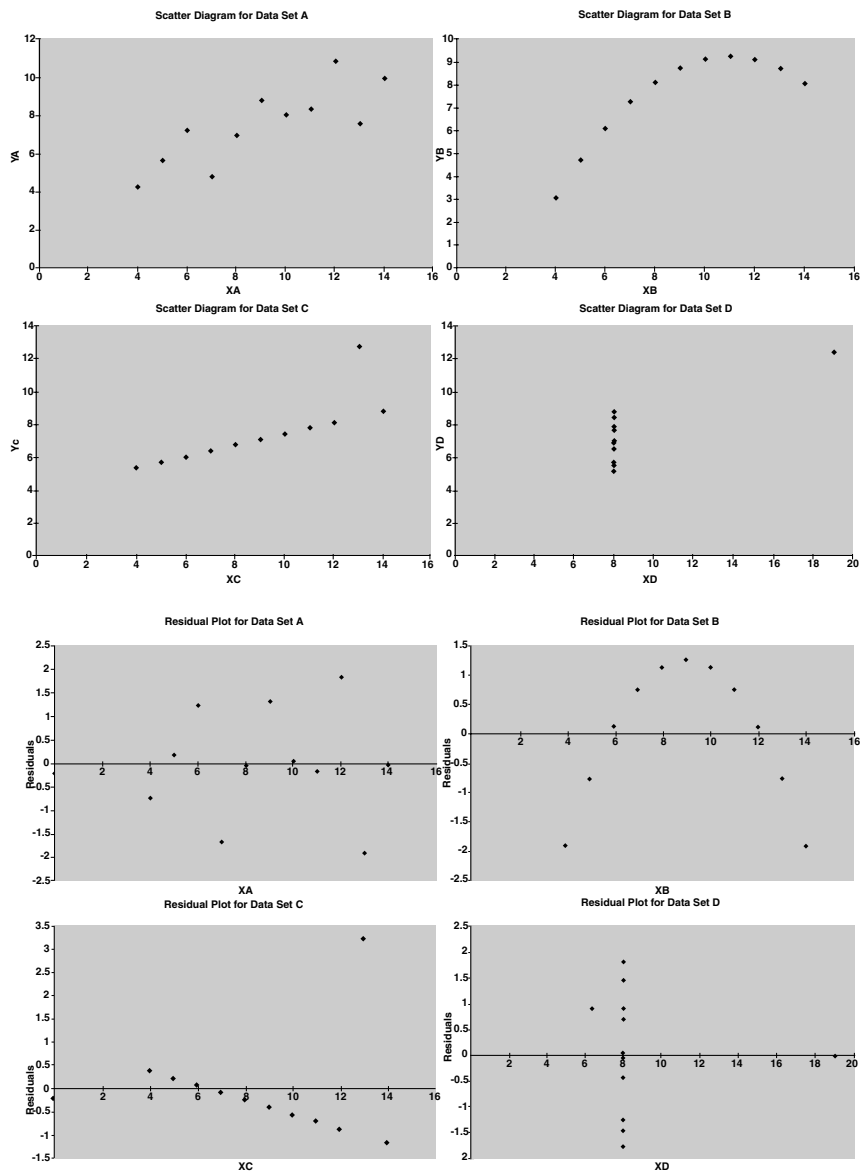
$$r^2 = .667$$

SSR = regression sum of squares = 27.51

SSE = error sum of squares = 13.76

SST = total sum of squares = 41.27

However, the four data sets are actually quite different as scatter diagrams and residual plots for the four sets reveal.



From the scatter diagrams and the residual plots, you see how different the data sets are. The only data set that seems to follow an approximate straight line is data set A. The residual plot for data set A does not show any obvious patterns or outlying residuals. This is certainly not the case for data sets B, C, and D. The scatter plot for data set B shows that a curvilinear regression model should be considered. The residual plot reinforces this conclusion for B. The scatter diagram and the residual plot for data set C clearly depict what is an extreme value. Similarly, the scatter diagram for data set D represents the unusual situation in which the fitted model is heavily dependent on the outcome of a single response ($X = 19$ and $Y = 12.50$). Any regression model fit for these data should be evaluated cautiously, because its regression coefficients are heavily dependent on a single observation.

To avoid the common mistakes of regression analysis, you can use the following process:

- Always start with a scatter plot to observe the possible relationship between X and Y .
- Check the assumptions of regression after the regression model has been fit, before using the results of the model.
- Plot the residuals versus the independent variable. This will enable you to determine whether the model fit to the data is an appropriate one and will allow you to check visually for violations of the equal variation assumption.
- Use a histogram, box-and-whisker plot, or normal probability plot of the residuals to graphically evaluate whether the normality assumption has been seriously violated.
- If the evaluation of the residuals indicates violations in the assumptions, use alternative methods to least-squares regression or alternative least-squares models (see References 2 and 6), depending on what the evaluation has indicated.
- If the evaluation of the residuals does not indicate violations in the assumptions, then you can undertake the inferential aspects of the regression analysis. A test for the significance of the slope and a confidence interval estimate of the slope can be carried out.

Important Equations

Regression equation:

$$(10.1) \quad \hat{Y}_i = b_0 + b_1 X_i$$

Slope:

$$(10.2) \quad b_1 = \frac{SSXY}{SSX}$$

$$(10.3) \quad SSXY = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}$$

and

$$SSX = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}$$

Y intercept:

$$(10.4) \quad b_0 = \bar{Y} - b_1 \bar{X}$$

Total sum of squares:

$$(10.5) \quad SST = \sum_{i=1}^n (Y_i - \bar{Y})^2 \text{ which is equivalent to } \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n}$$

$$(10.6) \quad SST = SSR + SSE$$

Regression sum of squares:

SSR = explained variation or regression sum of squares

$$= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

(10.7) which is equivalent to

$$= b_0 \sum_{i=1}^n Y_i + b_1 \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n}$$

Error sum of squares:

SSE = unexplained variation or error sum of squares

$$(10.8) \quad = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \text{ which is equivalent to}$$

$$= \sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n X_i Y_i$$

Coefficient of determination:

$$(10.9) \quad r^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}} = \frac{SSR}{SST}$$

Coefficient of correlation:

(10.10) $r = \sqrt{r^2}$ If b_1 is positive, r is positive. If b_1 is negative, r is negative.

Standard error of the estimate:

$$(10.11) \quad S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}}$$

t test for the slope:

$$(10.12) \quad t = \frac{b_1 - \beta_1}{S_{b_1}}$$

One-Minute Summary

Simple Linear Regression

- Least-squares method
- Measures of variation
- Residual analysis
 - t test for the significance of the slope
 - Confidence interval estimate of the slope

Test Yourself

1. The Y intercept (b_0) represents the:
 - (a) predicted value of Y when $X = 0$
 - (b) change in estimated average Y per unit change in X
 - (c) predicted value of Y
 - (d) variation around the regression line
2. The slope (b_1) represents:
 - (a) predicted value of Y when $X = 0$
 - (b) change in Y per unit change in X
 - (c) predicted value of Y
 - (d) variation around the regression line
3. The standard error of the estimate is a measure of:
 - (a) total variation of the Y variable
 - (b) the variation around the regression line
 - (c) explained variation
 - (d) the variation of the X variable

4. The coefficient of determination (r^2) tells you:
 - (a) that the coefficient of correlation (r) is larger than 1
 - (b) whether the slope has any significance
 - (c) whether the regression sum of squares is greater than the total sum of squares
 - (d) the proportion of total variation that is explained
5. In performing a regression analysis involving two numerical variables, you assume:
 - (a) the variances of X and Y are equal
 - (b) the variation around the line of regression is the same for each X value
 - (c) that X and Y are independent
 - (d) All of the above
6. Which of the following assumptions concerning the distribution of the variation around the line of regression (the residuals) is correct?
 - (a) The distribution is normal.
 - (b) All of the variations are positive.
 - (c) The variation increases as X increases.
 - (d) Each variation is dependent on the previous variation.
7. The residuals represent:
 - (a) the difference between the actual Y values and the mean of Y
 - (b) the difference between the actual Y values and the predicted Y values
 - (c) the square root of the slope
 - (d) the predicted value of Y when $X = 0$
8. If the coefficient of determination (r^2) = 1.00, then:
 - (a) the Y intercept must equal 0
 - (b) the regression sum of squares (SSR) equals the error sum of squares (SSE)
 - (c) the error sum of squares (SSE) equals 0
 - (d) the regression sum of squares (SSR) equals 0
9. If the coefficient of correlation (r) = -1.00, then:
 - (a) all of the data points must fall exactly on a straight line with a slope that equals 1.00
 - (b) all of the data points must fall exactly on a straight line with a negative slope
 - (c) all of the data points must fall exactly on a straight line with a positive slope
 - (d) all of the data points must fall exactly on a horizontal straight line with a zero slope

10. Assuming a straight line (linear) relationship between X and Y , if the coefficient of correlation (r) equals -0.30 :
 - (a) there is no correlation
 - (b) the slope is negative
 - (c) variable X is larger than variable Y
 - (d) the variance of X is negative
11. The strength of the linear relationship between two numeric variables is measured by the:
 - (a) predicted value of Y
 - (b) coefficient of determination
 - (c) total sum of squares
 - (d) Y intercept
12. In a simple linear regression problem, the coefficient of correlation and the slope:
 - (a) may have opposite signs
 - (b) must have the same sign
 - (c) must have opposite signs
 - (d) are equal

The following are True or False Questions:

13. The regression sum of squares (SSR) can never be greater than the total sum of squares (SST).
14. The coefficient of determination represents the ratio of SSR to SST .
15. Regression analysis is used for prediction, while correlation analysis is used to measure the strength of the association between two numeric variables.
16. The value of r is always positive.
17. When the coefficient of correlation $r = -1$, a perfect relationship exists between X and Y .
18. If there is no apparent pattern in the residual plot, the regression model fit is appropriate for the data.
19. If the range of the X variable is between 100 and 300, you should not make a prediction for $X = 400$.
20. If the p -value for a t test for the slope is 0.021, the results are significant at the 0.01 level of significance.

Answers to Test Yourself Questions

1. a
2. b

3. b
4. d
5. b
6. a
7. b
8. c
9. b
10. b
11. b
12. b
13. True
14. True
15. True
16. False
17. True
18. True
19. True
20. False

References

1. Anscombe, F. J. "Graphs in Statistical Analysis." *American Statistician* 27 (1973): 17–21.
2. Berenson, M. L., D. M. Levine, and T. C. Krehbiel. *Basic Business Statistics: Concepts and Applications, Ninth Edition*. Upper Saddle River, NJ: Prentice Hall, 2004.
3. Levine, D. M., D. Stephan, T. C. Krehbiel, and M. L. Berenson. *Statistics for Managers Using Microsoft Excel, Fourth Edition*. Upper Saddle River, NJ: Prentice Hall, 2005.
4. Levine, D. M., P. P. Ramsey, and R. K. Smidt. *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Upper Saddle River, NJ: Prentice Hall, 2001.
5. Microsoft Excel 2002. Redmond, WA: Microsoft Corporation, 2001.
6. Neter, J., M. H. Kutner, C. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models, Fourth Edition*. Homewood, IL: Richard D. Irwin, 1996.
7. Sincich, T., D. M. Levine, and D. Stephan, *Practical Statistics by Example Using Microsoft Excel and Minitab, Second Edition*. Upper Saddle River, NJ: Prentice Hall, 2002.



Quality and Six Sigma Management Applications of Statistics

11.1 Total Quality Management

11.2 Six Sigma Management

11.3 Control Charts: The p Chart

11.4 The Parable of the Red Bead Experiment:
Understanding Process Variability

11.5 Variables Control Charts for the Mean and Range
Important Equations

One-Minute Summary

Test Yourself

In recent times, improving quality and productivity have become essential goals for all organizations. However, monitoring and measuring such improvements can be problematic if subjective judgments about quality are made. A set of techniques and management practices known as **statistical process control** helps by relating quality to measurable sources of variation.

11.1 Total Quality Management

During the past 20 years, the renewed interest in quality and productivity in the United States followed as a reaction to perceived improvements of Japanese industry that had begun as early as 1950. Individuals such as W. Edwards Deming, Joseph Juran, and Kaoru Ishikawa developed an approach that focuses on continuous improvement of products and services through an increased emphasis on statistics, process improvement, and optimization of the total system. This approach, widely known as **total quality management (TQM)**, is characterized by these themes:

- The primary focus is on process improvement.
- Most of the variation in a process is due to the system and not the individual.

*important
point*



- Teamwork is an integral part of a quality management organization.
- Customer satisfaction is a primary organizational goal.
- Organizational transformation must occur in order to implement quality management.
- Fear must be removed from organizations.
- Higher quality costs less, not more, but requires an investment in training.

As this approach became more familiar, the federal government of the United States began efforts to encourage increased quality in American business, starting, for example, the annual competition for the Malcolm Baldrige Award, given to companies making the greatest strides in improving quality and customer satisfaction with their products and services. W. Edwards Deming became a more prominent consultant and widely discussed his “14 points for management.”

1. Create constancy of purpose for improvement of product and service.
2. Adopt the new philosophy.
3. Cease dependence on inspection to achieve quality.
4. End the practice of awarding business on the basis of price tag alone. Instead, minimize total cost by working with a single supplier.
5. Improve constantly and forever every process for planning, production, and service.
6. Institute training on the job.
7. Adopt and institute leadership.
8. Drive out fear.
9. Break down barriers between staff areas.
10. Eliminate slogans, exhortations, and targets for the workforce.
11. Eliminate numerical quotas for the workforce and numerical goals for management.
12. Remove barriers that rob people of pride of workmanship. Eliminate the annual rating or merit system.
13. Institute a vigorous program of education and self-improvement for everyone.
14. Put everyone in the company to work to accomplish the transformation.

Although Deming's points were thought-provoking, some criticized his approach for lacking a formal, objective accountability. Many managers of large-scale organizations, used to seeing economic analyses of policy changes, needed a more prescriptive approach.

11.2 Six Sigma Management

One methodology, inspired by earlier TQM efforts, that attempts to apply quality improvement with increased accountability is the Six Sigma approach, originally conceived by Motorola in the mid-1980s. Refined and enhanced over the years, and famously applied to other large firms such as General Electric, Six Sigma was developed as a way to cut costs while improving efficiency. As with earlier total quality management approaches, Six Sigma relies on statistical process control methods to find and eliminate defects and reduce product variation.

Six Sigma

CONCEPT The quality management approach that is designed to create processes that result in no more than 3.4 defects per million.

INTERPRETATION Six Sigma considers the variation of a process. Recall from Chapter 3 that the lowercase Greek letter sigma (σ) represents the population standard deviation, and recall from Chapter 5 that the range -6σ to $+6\sigma$ in a normal distribution includes virtually all (specifically, 0.99999998) of the probability or area under the curve. The Six Sigma approach assumes that the process may shift as much as 1.5 standard deviations over the long term. Six standard deviations minus a 1.5 standard deviation shift produces a 4.5 standard deviation goal. The area under the normal curve outside 4.5 standard deviations is approximately 3.4 out of a million (0.0000034).

The Six Sigma DMAIC Model

Unlike other quality management approaches, Six Sigma seeks to help managers achieve measurable, bottom-line results in a relatively short three to six-month period of time. This has enabled Six Sigma to obtain strong support from top management of many companies (see References 6 and 7).

To guide managers in their task of affecting short-term results, Six Sigma uses a five-step process known as the **DMAIC model**, for the names of steps in the process: Define, Measure, Analyze, Improve, and Control. This model can be summarized as follows:

- **Define**—The problem to be solved needs to be defined along with the costs, benefits of the project, and the impact on the customer.
- **Measure**—Operational definitions for each critical-to-quality (CTQ) characteristic must be developed. In addition, the measurement procedure must be verified so that it is consistent over repeated measurements.
- **Analyze**—The root causes of why defects can occur need to be determined along with the variables in the process that cause these defects to occur. Data are collected to determine the underlying value for each

important point



process variable often using control charts (to be discussed in Sections 11.3 through 11.5).

- **Improve**—The importance of each process variable on the Critical-To-Quality (CTQ) characteristic are studied using designed experiments. The objective is to determine the best level for each variable that can be maintained in the long term.
- **Control**—Maintain the gains that have been made with a revised process in the long term by avoiding potential problems that can occur when a process is changed.

Implementation of the Six Sigma approach requires intensive training in the DMAIC model as well as a data-oriented approach to analysis that uses designed experiments and various statistical methods, such as the control chart methods discussed in the remainder of the chapter.

11.3 Control Charts

Control charts monitor variation in a characteristic of a product or service by focusing on the variation in a process over time. Control charts aid in quality improvement by letting you assess the stability and capability of a process.

Control charts are divided into two types called *attribute control charts* and *variables control charts*. **Attribute control charts**, such as the p chart discussed later in this section, are used to evaluate categorical data. If you wanted to study the proportion of newspaper ads that have errors or the proportion of trains that are late, you would use attribute control charts.

Variables control charts are used for continuous data. If you wanted to study the waiting time at a bank or the weight of packages of candy, you would use variables control charts. **Variables control charts** contain more information than attribute charts and are generally used in pairs, such as the range chart and the mean chart.

The principal focus of the control chart is the attempt to separate *special or assignable causes of variation* from *chance or common causes of variation*.

Special or Assignable Causes of Variation

CONCEPT Variation that represents large fluctuations or patterns in the data that are not inherent to a process.

EXAMPLE If during your process of getting ready to go to work or school there is a leak in a toilet that needs immediate attention, your time to get ready will certainly be affected. This is special cause variation, because it is not a cause of variation that can be expected to occur every day, and therefore it is not part of your everyday process of getting ready (at least you hope it is not!).

INTERPRETATION Special cause variation is the variation that is not always present in every process. It is variation that occurs for special reasons that usually can be explained.

Chance or Common Causes of Variation

CONCEPT Variation that represents the inherent variability that exists in a process over time. These consist of the numerous small causes of variability that operate randomly or by chance.

EXAMPLE Your process of getting ready to go to work or school has common cause variation, because there are small variations in how long it takes you to perform the activities, such as making breakfast and getting dressed, that are part of your get-ready process from day to day.

INTERPRETATION Common cause variation is the variation that is always present in every process. Typically, this variation can be reduced only by changing the process itself.



Distinguishing between these two causes of variation is crucial, because only special causes of variation are not considered part of a process and therefore are correctable, or exploitable, without changing the system. Common causes of variation occur randomly or by chance and can be reduced only by changing the system.

Control charts allow you to monitor the process and determine the presence of special causes. Control charts help prevent two types of errors. The first type of error involves the belief that an observed value represents special cause variation when in fact the error is due to the common cause variation of the system. An example of this type of error occurs if you were to single out someone for disciplinary action based on having more errors than anyone else when in fact the variation in errors was just due to common cause variation in the system. Treating common causes of variation as special cause variation can result in overadjustment of a process that results in an accompanying increase in variation. The second type of error involves treating special cause variation as if it is common cause variation and not taking immediate corrective action when it is necessary. An example of this type of error occurs if you did *not* single someone out for disciplinary action based on having more errors than anyone else when in fact the large number of errors made by the person could be explained and subsequently corrected. Although these errors can still occur when a control chart is used, they are far less likely.

Control Limits

The most typical form of control chart sets **control limits** that are within ± 3 standard deviations of the average of the process located at the **center line**. Depending on the control chart being used, the process average could be the

average proportion, the average of the means, or the average of the ranges. The value that is +3 standard deviations above the process average is called the **upper control limit (UCL)**; the value that is -3 standard deviations below the process average is called the **lower control limit (LCL)**. Should the value that is -3 standard deviations be less than 0, the lower control limit is set to 0.

After these control limits are set, you evaluate the control chart from the perspective of discerning any pattern that might exist in the values over time and determining whether any points fall outside the control limits.



The simplest rule for detecting the presence of a special cause is one or more points falling beyond the ± 3 standard deviation limits of the chart. The chart can be made more sensitive and effective in detecting out-of-control points if other signals and patterns that are unlikely to occur by chance alone are considered.

Two other simple rules enable you to detect a shift in the average level of a process:

- Eight or more *consecutive points* lie above the center line, or eight or more *consecutive points* lie below the center line.
- Eight or more *consecutive points* move upward in value, or eight or more *consecutive points* move downward in value.

The *p* Chart

CONCEPT The control chart used to study a process that involves the proportion of items with a characteristic of interest, such as the number of newspaper ads with errors. Sample sizes in a *p* chart may remain constant or may vary.

INTERPRETATION In the *p* chart, the process average is the average proportion of nonconformances. The average proportion is computed from:

$$\text{average proportion} = \frac{\text{total number of nonconformances}}{\text{total number in all samples}}$$

To calculate the control limits, the average sample size first needs to be calculated:

$$\text{average sample size} = \frac{\text{total number in all samples}}{\text{number of groups}}$$

The control limits are:

$$\text{Upper control limit (UCL)} =$$

$$\text{Average proportion} + 3 \sqrt{\frac{(\text{average proportion})(1 - \text{average proportion})}{\text{average sample size}}}$$

Lower control limit (LCL) =

$$\text{Average proportion} - 3 \sqrt{\frac{(\text{average proportion})(1 - \text{average proportion})}{\text{average sample size}}}$$

To use a p chart, the following three statements must be true:

- There are only two possible outcomes for an event. An item must be found to be either conforming or nonconforming.
- The probability, p , of a nonconforming item is constant.
- Successive items are independent.

WORKED-OUT PROBLEM You are part of a team in an advertising production department of a newspaper that is trying to reduce the number and dollar amount of the advertising errors. You collect data that tracks the number of ads with errors on a daily basis, excluding Sundays (which is considered to be substantially different from the other days). Data relating to the number of ads with errors in the last month are shown in the following table.

Day	Number of Ads with Errors	Number of Ads
1	4	228
2	6	273
3	5	239
4	3	197
5	6	259
6	7	203
7	8	289
8	14	241
9	9	263
10	5	199
11	6	275
12	4	212
13	3	207
14	5	245
15	7	266
16	2	197
17	4	228

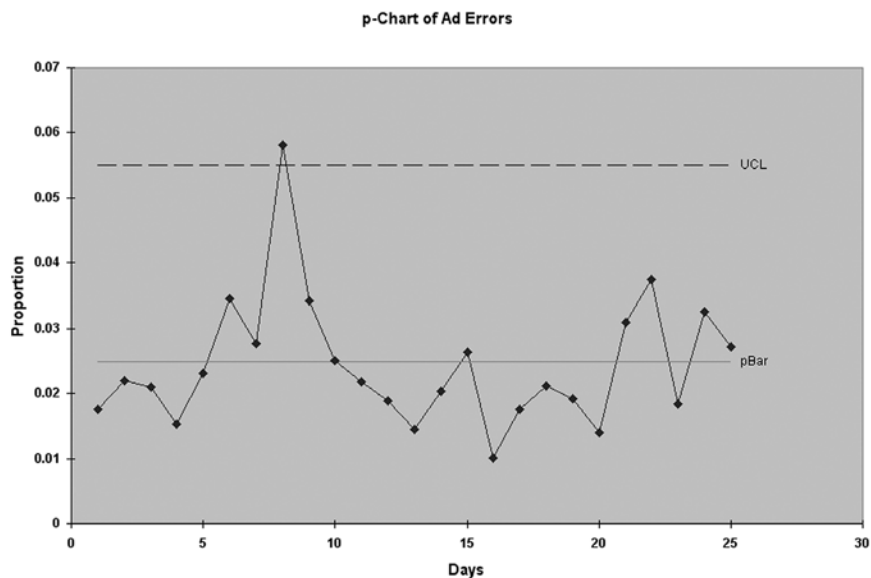
(continues)

Day	Number of Ads with Errors	Number of Ads
18	5	236
19	4	208
20	3	214
21	8	258
22	10	267
23	4	217
24	9	277
25	7	258

(Aderrors)

These data are appropriate for a p chart, because each ad is classified as with errors or without errors, the probability of an ad with an error is assumed to be constant from day to day, and each ad is considered independent of the other ads.

A p chart prepared in Microsoft Excel for the newspaper ads data is shown here:



For these data, the total number of nonconformances is 148, the total number in all samples is 5,956, and the number of groups is 25. Using these values, the average proportion is 0.0248, and the average sample size is 238.24, as shown:

$$\begin{aligned}\text{average proportion} &= \frac{\text{total number of nonconformances}}{\text{total number in all samples}} \\ &= \frac{148}{5,956} = 0.0248\end{aligned}$$

$$\begin{aligned}\text{average sample size} &= \frac{\text{total number in all samples}}{\text{number of groups}} \\ &= \frac{5,956}{25} = 238.24\end{aligned}$$

Using the average proportion and average sample size values, the UCL is 0.0551 and the LCL is 0, as shown:

Upper control limit (UCL) =

$$0.0248 + 3 \sqrt{\frac{(\text{average proportion})(1 - \text{average proportion})}{\text{average sample size}}}$$

$$0.0248 + 3 \sqrt{\frac{(0.0248)(1 - 0.0248)}{238.24}}$$

$$UCL = 0.0248 + 0.0303 = 0.0551$$

Lower control limit (LCL) =

$$0.0248 - 3 \sqrt{\frac{(\text{average proportion})(1 - \text{average proportion})}{\text{average sample size}}}$$

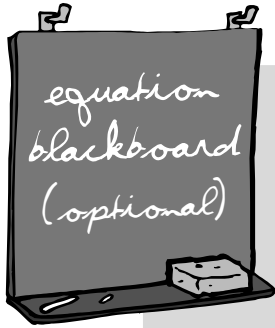
$$0.0248 - 3 \sqrt{\frac{(0.0248)(1 - 0.0248)}{238.24}}$$

$$LCL = 0.0248 - 0.0303 = -0.0054$$

Because the calculated value is less than 0, the LCL is set at 0.

Using the rules for determining out-of-control points, you observe that point 8 is above the upper control limit. None of the other rules seems to be violated. There are no instances when eight consecutive points move upward or downward, nor are there eight consecutive points on one side of the center line.

Upon further investigation, you learn that point 8 corresponds to the day there was an employee from another work area assigned to the processing of the ads, because several employees were out ill. Your group brainstorms ways of avoiding such a problem in the future and recommends that a team of people from other work areas receive training on the work done by this area.



interested
in
math?

You use these symbols to write the equations for the lower and upper control limits of a p chart:

- A subscripted uppercase italic X , X_i , for the number of nonconforming items in a group
- A subscripted lowercase italic n , n_i , for the sample or subgroup size for a group
- A lowercase italic k , k , for the number of groups taken
- A lowercase italic n bar, \bar{n} , for the average group size
- A subscripted lowercase italic p , p_i , for the proportion of nonconforming items for a group
- A lowercase italic p bar, \bar{p} , for the average proportion of nonconforming items

You first use some of the symbols to define \bar{p} and \bar{n} as follows:

$$\bar{p} = \frac{\sum_{i=1}^k X_i}{\sum_{i=1}^k n_i} = \frac{\text{total number of nonconformances}}{\text{total sample size}}$$

$$\text{and } \bar{n} = \frac{\sum_{i=1}^k n_i}{k} = \frac{\text{total sample size}}{\text{number of groups}}$$

You then use these just-defined symbols to write the equations for the control limits as follows:

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}$$

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}$$

Although not necessary for the equations for the control limits, you can use some of the symbols to define the proportion of nonconforming items for group i as follows:

(continues)

$$p_i = \frac{X_i}{n_i}$$

For the advertising errors data, the number of ads with errors is 148, the total sample size is 5,956, and there are 25 groups. Thus:

$$\bar{p} = \frac{148}{5,956} = 0.0248 \quad \text{and}$$

$$\bar{n} = \frac{\sum_{i=1}^k n_i}{k} = \frac{\text{total sample size}}{\text{number of groups}} = \frac{5,956}{25} = 238.24$$

so that:

$$0.0248 + 3 \sqrt{\frac{(0.0248)(1 - 0.0248)}{238.24}}$$

$$UCL = 0.0248 + 0.0303 = 0.0551$$

$$0.0248 - 3 \sqrt{\frac{(0.0248)(1 - 0.0248)}{238.24}}$$

$$LCL = 0.0248 - 0.0303 = -0.0054$$

Therefore, because LCL is less than 0, it is set at 0.

11.4 The Parable of the Red Bead Experiment: Understanding Process Variability

Imagine that you have been selected to appear on a new reality television series about job performance excellence. Over several days, you are assigned different tasks and your results are compared with your competitors.

The current task involves visiting the W.E. Beads Company and helping to select groups of 50 white beads for sale from a pool of 4,000 beads. You are told that W.E. Beads regularly tries to produce and sell only white beads, but that an occasional red bead gets produced in error. Unknown to you, the producer of the series has arranged that the pool of 4,000 beads contains 800 red beads to see how you and the other participants will react to this special challenge.

To select the groups of 50 beads, you and your competitors will be sharing a special scoop that can extract exactly 50 beads in one motion. You are told to hold the scoop at exactly an angle of 41 degrees to the vertical and that you will have three turns, simulating three days of production. At the end of each “day,” two judges will independently count the number of red beads you select

with the scoop and report their findings to a chief judge who may give out an award for exceptional job performance. To make things more fair, after a group of 50 beads has been extracted, they will be returned to the pool so that every participant will always be selecting from the same pool of 4,000.

At the end of the three days, the judges, plus two famous business executives, will meet in a management council to discuss which worker deserves a promotion to the next task and which worker should be sent home from the competition.

The results of the competition are as follows.

Contestant	Day 1	Day 2	Day 3	All 3 Days
You	9 (18%)	11 (22%)	6 (12%)	26 (17.33%)
A	12 (24%)	12 (24%)	8 (16%)	32 (21.33%)
B	13 (26%)	6 (12%)	12 (24%)	31 (20.67%)
C	7 (14%)	9 (18%)	8 (16%)	24 (16.0%)
All four workers	41	38	34	113
Average (\bar{X})	10.25	9.5	8.5	9.42
Proportion	20.5%	19%	17%	18.83%

From the preceding above, you observe several phenomena. On each day, some of the workers were above the average and some below the average. On day 1, C did best; but on day 2, B (who had the worst record on day 1) was best; and on day 3, you were the best. You are hopeful that your great job performance on day 3 will attract notice; if the decisions are solely based on job performance, however, whom would you promote and whom would you fire?

Deming's Red Bead Experiment

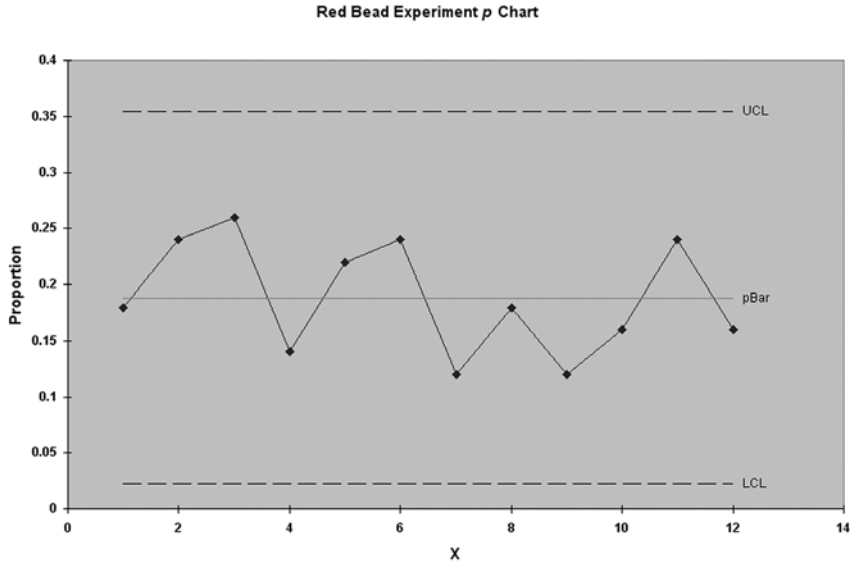
The description of the reality series is very similar to a famous demonstration that has become known as the **red bead experiment** that the statistician W. Edwards Deming performed during many lectures. In both the experiment and the imagined reality series, the workers have very little control over their production, even though common management practice might imply otherwise, and there are way too many managers officiating. Among the points about the experiment that Deming would make during his lectures are these:

- Variation is an inherent part of any process.
- Workers work within a system over which they have little control. It is the system that primarily determines their performance.
- Only management can change the system.
- Some workers will always be above the average, and some workers will always be below the average.

important point



How then can you explain all the variation? A p chart of the data puts the numbers into perspective and reveals that all of the values are within the control limits, and there are no patterns in the results (see below). The differences between you and the other participants merely represent the common cause variation expected in a stable process.



11.5 Variables Control Charts for the Mean and Range

Variables control charts can be used to monitor a process for a numerical variable such as bank waiting time. Because numerical variables provide more information than the proportion of nonconforming items, these charts are more sensitive in detecting special cause variation than the p chart. Variables charts are typically used in pairs. One chart monitors the variation in a process, while the other monitors the process average. The chart that monitors variability must be examined first, because if it indicates the presence of out-of-control conditions, the interpretation of the chart for the average will be misleading.

One of the most commonly employed pair of charts is the \bar{X} chart used in conjunction with the R chart. The group range, R , is plotted on the R chart, which monitors process variability. The group average, \bar{X} , is plotted on the \bar{X} chart, which monitors the central tendency of the process.

WORKED-OUT PROBLEM You want to study waiting times of customers for teller service at a bank during the peak 12 noon to 1 p.m. lunch hour. You select a group of four customers (one at each 15-minute interval during

the hour) and measure the time in minutes from the point each customer enters the line to when he or she begins to be served. The results over a 4-week period are as follows.

Day	Time in Minutes			
1	7.2	8.4	7.9	4.9
2	5.6	8.7	3.3	4.2
3	5.5	7.3	3.2	6.0
4	4.4	8.0	5.4	7.4
5	9.7	4.6	4.8	5.8
6	8.3	8.9	9.1	6.2
7	4.7	6.6	5.3	5.8
8	8.8	5.5	8.4	6.9
9	5.7	4.7	4.1	4.6
10	3.7	4.0	3.0	5.2
11	2.6	3.9	5.2	4.8
12	4.6	2.7	6.3	3.4
13	4.9	6.2	7.8	8.7
14	7.1	6.3	8.2	5.5
15	7.1	5.8	6.9	7.0
16	6.7	6.9	7.0	9.4
17	5.5	6.3	3.2	4.9
18	4.9	5.1	3.2	7.6
19	7.2	8.0	4.1	5.9
20	6.1	3.4	7.2	5.9

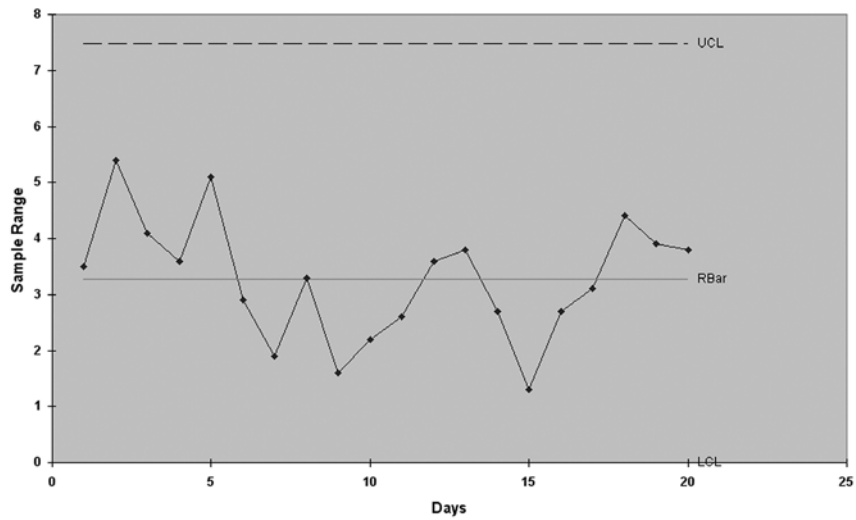
(Banktime)

\bar{R} and \bar{X} charts prepared in Microsoft Excel for these data are shown on page 223:

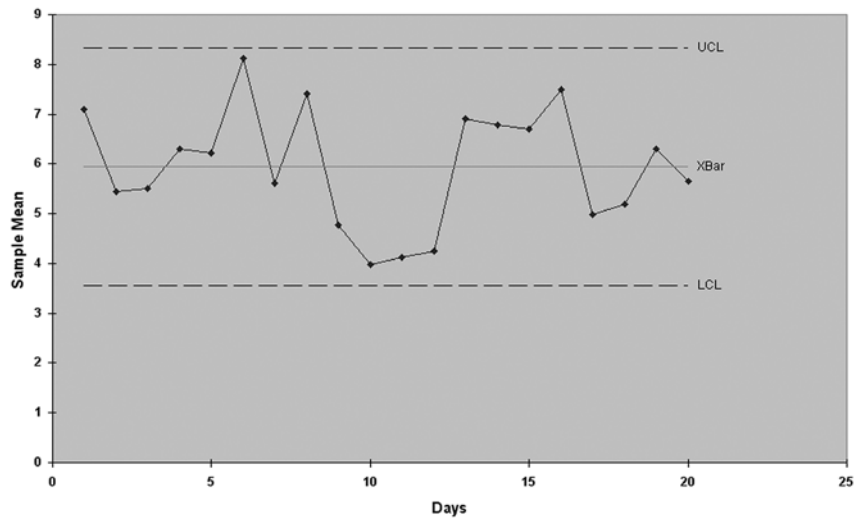
Reviewing the \bar{R} chart, you note that none of the points are outside of the control limits, and there are no other signals indicating a lack of control. This suggests that there are no special causes of variation present.

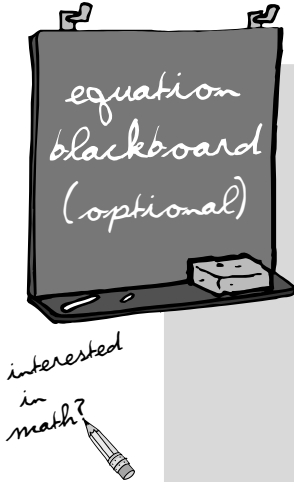
Reviewing the \bar{X} chart, you note that none of the points are outside of the control limits, and there are no other signals indicating a lack of control. This also suggests that there are no special causes of variation present. If management wants to reduce the variation in the waiting times or lower the average waiting time, you conclude that changes in the process need to be made.

R Chart of Bank Waiting Time



XBar Chart of Bank Waiting Time





Equations for the Lower and Upper Control Limits for the Range

You use the following symbols to write the equations for the lower and upper control limits for the range:

- A subscripted \bar{X} bar, \bar{X}_i , for the sample mean of n observations at time i
- A subscripted uppercase italic R , R_i , for the range of n observations at time i
- A lowercase italic k , k , for the number of groups

You use these symbols to first define \bar{R} as follows:

$$\bar{R} = \frac{\sum_{i=1}^k R_i}{k} = \frac{\text{sum of all the ranges}}{\text{number of groups}}$$

You then use these newly defined symbols to write the equations for the control limits:

$$LCL = \bar{R} - 3\bar{R} \frac{d_3}{d_2}$$

$$UCL = \bar{R} + 3\bar{R} \frac{d_3}{d_2}$$

in which the symbols d_3 and d_2 represent control chart factors obtained from Table C.5.

By using the D_3 factor that is equal to $1 - 3(d_3/d_2)$ and the D_4 factor, equal to $1 + 3(d_3/d_2)$, for which values for different subgroup sizes are listed in Table C.5, the equations can be simplified as follows:

$$LCL = D_3 \bar{R}$$

$$UCL = D_4 \bar{R}$$

For the data of the table of bank waiting times, the sum of the ranges is 65.5 and the number of groups is 20. Therefore:

$$\begin{aligned} \bar{R} &= \frac{\text{sum of all the ranges}}{\text{number of groups}} \\ &= \frac{65.5}{20} = 3.275 \end{aligned}$$

(continues)

For a subgroup size = 4, $D_3 = 0$ and $D_4 = 2.282$

$$LCL = (0)(3.275) = 0$$

$$UCL = (2.282)(3.275) = 7.4736$$

Equations for the Lower and Upper Control Limits for the Mean

You use the following symbols to write the equations for the lower and upper control limits for the mean:

- A subscripted \bar{X} , \bar{X}_i , for the sample mean of n observations at time i
- A subscripted uppercase italic R , R_i , for the range of n observations at time i
- A lowercase italic k , k , for the number of groups

You use these symbols to first define $\bar{\bar{X}}$ and $\bar{\bar{R}}$ as follows:

$$\bar{\bar{X}} = \frac{\sum_{i=1}^k \bar{X}_i}{k} = \frac{\text{sum of the sample means}}{\text{number of groups}}$$

and

$$\bar{\bar{R}} = \frac{\sum_{i=1}^k R_i}{k} = \frac{\text{sum of all the ranges}}{\text{number of groups}}$$

You then use these newly defined symbols to write the equations for the control limits:

$$LCL = \bar{\bar{X}} - 3 \frac{\bar{\bar{R}}}{d_2 \sqrt{n}}$$

$$UCL = \bar{\bar{X}} + 3 \frac{\bar{\bar{R}}}{d_2 \sqrt{n}}$$

in which the lowercase italic subscripted D , d_2 , represents a control chart factor obtained from Table C.5. By using the A_2 factor that is equal to $3/(d_2 \sqrt{n})$, and for which values for different subgroup sizes are listed in Table C.5, the equations can be simplified as follows:

$$LCL = \bar{\bar{X}} - A_2 \bar{\bar{R}}$$

$$UCL = \bar{\bar{X}} + A_2 \bar{\bar{R}}$$

(continues)

For the bank waiting time data, the sum of the ranges is 65.5, the sum of the sample means is 118.825, and the number of groups is 20. Therefore:

$$\begin{aligned}\bar{R} &= \frac{\text{sum of all the ranges}}{\text{number of groups}} \\ &= \frac{65.5}{20} = 3.275 \\ \bar{\bar{X}} &= \frac{\text{sum of the sample means}}{\text{number of groups}} \\ &= \frac{118.825}{20} = 5.94125\end{aligned}$$

For a group size = 4, $A_2 = 0.729$

$$LCL = 5.94125 - (0.729)(3.275) = 3.553775$$

$$UCL = 5.94125 + (0.729)(3.275) = 8.328725$$

Important Equations

Lower and upper control limits for the p chart:

$$(11.1) \quad LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}$$

$$(11.2) \quad UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}$$

Lower and upper control limits for the range:

$$(11.3) \quad LCL = \bar{R} - 3\bar{R}\frac{d_3}{d_2}$$

$$(11.4) \quad UCL = \bar{R} + 3\bar{R}\frac{d_3}{d_2}$$

$$(11.5) \quad LCL = D_3\bar{R}$$

$$(11.6) \quad UCL = D_4\bar{R}$$

Lower and upper control limits for the mean:

$$(11.7) \quad LCL = \bar{\bar{X}} - 3\frac{\bar{R}}{d_2\sqrt{n}}$$

$$(11.8) \quad UCL = \bar{\bar{X}} + 3\frac{\bar{R}}{d_2\sqrt{n}}$$

$$(11.9) \quad LCL = \bar{\bar{X}} - A_2 \bar{R}$$

$$(11.10) \quad UCL = \bar{\bar{X}} + A_2 \bar{R}$$

One-Minute Summary

Quality management approaches

- Total quality management (TQM)
- Six Sigma DMAIC model

Process control techniques

- If a categorical variable, use attribute control charts such as p charts.
- If a continuous numerical variable, use variables control charts such as R and \bar{X} charts.

Test Yourself

1. The control chart:
 - (a) focuses on the time dimension of a system
 - (b) captures the natural variability in the system
 - (c) can be used for categorical or numerical variables
 - (d) All of the above
2. Variation signaled by individual fluctuations or patterns in the data is called:
 - (a) special causes of variation
 - (b) common causes of variation
 - (c) Six Sigma
 - (d) the red bead experiment
3. Variation due to the inherent variability in a system of operation is called:
 - (a) special causes of variation
 - (b) common causes of variation
 - (c) Six Sigma
 - (d) the red bead experiment
4. Which of the following is *not* one of Deming's 14 points?
 - (a) Believe in mass inspection.
 - (b) Create constancy of purpose for improvement of product or service.
 - (c) Adopt and institute leadership.
 - (d) Drive out fear.

5. The principal focus of the control chart is the attempt to separate special or assignable causes of variation from common causes of variation. What cause of variation can be reduced only by changing the system?
 - (a) Special or assignable causes
 - (b) Common causes
 - (c) Total causes
 - (d) None of the above
6. After the control limits are set for a control chart, you attempt to:
 - (a) discern patterns that might exist in values over time
 - (b) determine whether any points fall outside the control limits
 - (c) Both of the above
 - (d) None of the above
7. Which of the following situations suggests a process that appears to be operating in a state of statistical control?
 - (a) A control chart with a series of consecutive points that are above the center line and a series of consecutive points that are below the center line
 - (b) A control chart in which no points fall outside either the upper control limit or the lower control limit and no patterns are present
 - (c) A control chart in which several points fall outside the upper control limit
 - (d) All of the above
8. Which of the following situations suggests a process that appears to be operating out of statistical control?
 - (a) A control chart with a series of eight consecutive points that are above the center line
 - (b) A control chart in which points fall outside the lower control limit
 - (c) A control chart in which points fall outside the upper control limit
 - (d) All of the above
9. A process is said to be out of control if:
 - (a) a point falls above the upper or below the lower control limits
 - (b) eight or more consecutive points are above the center line
 - (c) Either (a) or (b)
 - (d) Neither (a) or (b)
10. One of the morals of the red bead experiment is:
 - (a) variation is part of the process
 - (b) only management can change the system
 - (c) it is the system that primarily determines performance
 - (d) All of the above
11. The cause of variation that can be reduced only by changing the system is _____ cause variation.

12. _____ causes of variation are correctable without modifying the system.

The following are True or False questions:

13. The control limits are based on the standard deviation of the process.
14. The purpose of a control chart is to eliminate common cause variation.
15. Special causes of variation are signaled by individual fluctuations or patterns in the data.
16. Common causes of variation represent variation due to the inherent variability in the system.
17. Common causes of variation are correctable without modifying the system.
18. Changes in the system to reduce common cause variation are the responsibility of management.
19. The p chart is a control chart used for monitoring the proportion of items that have a certain characteristic.
20. It is not possible for the \bar{X} chart to be out of control when the R chart is in control.

Answers to Test Yourself Questions

1. d
2. a
3. b
4. a
5. b
6. c
7. b
8. d
9. c
10. d
11. common
12. special
13. True
14. False
15. True
16. True

17. False
18. True
19. True
20. False

References

1. Berenson, M. L., D. M. Levine, and T. C. Krehbiel. *Basic Business Statistics: Concepts and Applications, Ninth Edition*. Upper Saddle River, NJ: Prentice Hall, 2004.
2. Deming, W. E. *Out of the Crisis*. Cambridge, MA: MIT Center for Advanced Engineering Study, 1986.
3. Deming, W. E. *The New Economics for Business, Industry, and Government*. Cambridge, MA: MIT Center for Advanced Engineering Study, 1993.
4. Friedman, T. L. *The Lexus and the Olive Tree: Understanding Globalization*. New York: Farrar, Straus and Giroux, 1999.
5. Halberstam, D. *The Reckoning*. New York: Morrow, 1986.
6. Gitlow, H. S., and D. M. Levine. *Six Sigma for Green Belts and Champions*. Upper Saddle River, NJ: Financial Times - Prentice Hall, 2005.
7. Gitlow, H. G., A. Oppenheim, R. Oppenheim, and D. M. Levine. *Quality Management, Third Edition*. New York: McGraw-Hill-Irwin, 2005.
8. Levine, D. M., T. C. Krehbiel, and M. L. Berenson. *Business Statistics: A First Course, Third Edition*. Upper Saddle River, NJ: Prentice Hall, 2003.
9. Levine, D. M., D. Stephan, T. C. Krehbiel, and M. L. Berenson. *Statistics for Managers Using Microsoft Excel, Fourth Edition*. Upper Saddle River, NJ: Prentice Hall, 2005.
10. Levine, D. M., P. C. Ramsey, and R. K. Smidt. *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Upper Saddle River, NJ: Prentice Hall, 2001.
11. Montgomery, D. C. *Introduction to Statistical Quality Control, Fourth Edition*. New York: John Wiley, 2000.
12. Walton, M. *The Deming Management Method*. New York: Perigee Books, Putnam Publishing Group, 1986.



TI Statistical Calculator Settings and Microsoft Excel Settings

a.1 TI Statistical Calculator Settings

“Ready State” Assumptions

CALCULATOR KEYS procedures in this book always assume that you are beginning from the main screen and are not in the middle of some calculator activity. You should always be at the main screen, with the calculator in a “ready” state, before entering any of the keystrokes of a CALCULATOR KEYS section. (Most of the time, pressing [CLEAR] will place your calculator in a “ready” state.)

Menu Selections

When CALCULATOR KEYS sections require you to make choices from an onscreen menu list, the instructions in this book will tell you to “select **n**:Choice and press [ENTER].” To do this, you should use the down (or up) cursor key to highlight the choice and then press the [ENTER] key. You can also make selections by pressing the key that corresponds with the **n** value (and *not* pressing [ENTER]); you can use this alternative method if that is your preference.

Statistical Function Entries by Menus

This book always uses menus to enter the information to perform statistical functions. This means that many CALCULATOR KEYS procedures begin with the instruction “Press [STAT] [◀]” that will display the Stat Tests menu. If you are an advanced calculator user, you may be familiar with an alternative way of choosing a statistical function that involves typing command lines on the main screen, and you can use that method if you prefer.

Primary Key Legend Convention

Keystroke instructions in this book, unlike such instructions that appear in certain Texas Instruments manuals, always name keys by their primary legend, the label that is physically on the key, and not by their second function name that is printed on the face of the calculator in tiny yellow type above a key. For example, to display the List variables screen, this book would say “press [2nd] [STAT]” and not “press [2nd] [LIST]” as some TI materials would.

Mode Settings

The instructions in this book were written for a calculator set to “normal” numeric notation and floating-decimal format. **To set your calculator to these settings (or to verify them):**

- Press [MODE] and select **Normal**, and press [ENTER].
- Press [▼], select **Float**, and press [ENTER].
- Press [2nd][MODE] to return to the main screen.

Calculator Clearing and Reset

If, at any time, you want to clear all data from the memory of the calculator and reset all settings to the factory default, press [2nd] [+] to display the Memory screen. Select **7:Reset**, and from the RAM screen press [1] [2] to clear memory, or press [2] [2] to reset the calculator.

Data Storage

Some models in the TI-83 and TI-84 families include Flash memory, a computer data-link, or USB cable. If your calculator contains Flash memory, you may be able to store a set of values in Flash memory for later use. If your calculator has a data-link or USB cable, and if you have installed the appropriate Texas Instruments software on your personal computer, you will be able to upload and download variable data to and from your calculator.

Storing data in either of these ways can facilitate your use of your calculator with more complicated statistical methods discussed in the later chapters of this book. If you want to use either feature, consult your documentation or visit the Texas Instruments Web site at <http://education.ti.com>.

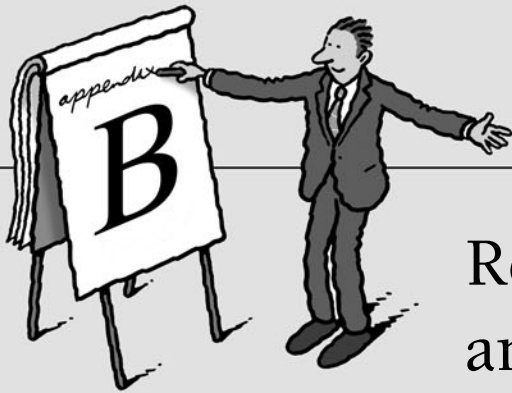
a.2 Microsoft Excel Settings

This book assumes no special Microsoft Excel settings other than the inclusion of the Data Analysis add-in that is used in several SPREADSHEET SOLUTION sections. To verify that your copy of Microsoft Excel has this add-in already installed:

- Open Microsoft Excel.
- Select **Tools → Add-Ins**.
- In the Add-Ins dialog box that appears, select the **Analysis ToolPak** and **Analysis ToolPak – VBA** check boxes from the Add-Ins Available list and click the OK button.
- Exit Microsoft Excel (to save the selections).

If the Analysis ToolPak choice does not appear in the Add-Ins Available list, you will need to rerun the Microsoft Excel (or Office) setup program using your original Microsoft Office/Excel CD-ROM or DVD to install this component.

This page intentionally left blank



Review of Arithmetic and Algebra

The authors understand and realize that there are wide differences in the mathematical background of readers of this book. Some of you may have taken various courses in calculus and matrix algebra, whereas others may not have taken any mathematics courses in a long period of time. Because the emphasis of this book is on statistical concepts and the interpretation of Microsoft Excel and statistical calculator output, no prerequisite beyond elementary algebra is needed. To assess your arithmetic and algebraic skills, you may want to answer the following questions and then read the review that follows.

Assessment Quiz

Part 1

Fill in the correct answer.

1. $\frac{1}{2} = \frac{\quad}{3}$

2. $(0.4)^2 =$

3. $1 + \frac{2}{3} =$

4. $\left(\frac{1}{3}\right)^{(4)} =$

5. $\frac{1}{5} =$ (in decimals)

6. $1 - (-0.3) =$

7. $4 \times 0.2 \times (-8) =$

8. $\left(\frac{1}{4} \times \frac{2}{3}\right) =$

9. $\left(\frac{1}{100}\right) + \left(\frac{1}{200}\right) =$

10. $\sqrt{16} =$

Part 2

Select the correct answer.

1. If $a = bc$, then $c =$
 - (a) ab
 - (b) b/a
 - (c) a/b
 - (d) None of the above
2. If $x + y = z$, then y
 - (a) z/x
 - (b) $z + x$
 - (c) $z - x$
 - (d) None of the above
3. $(x^3)(x^2) =$
 - (a) x^5
 - (b) x^6
 - (c) x^1
 - (d) None of the above
4. $x^0 =$
 - (a) x
 - (b) 1
 - (c) 0
 - (d) None of the above

5. $x(y - z) =$
(a) $xy - xz$
(b) $xy - z$
(c) $(y - z)/x$
(d) None of the above
6. $(x + y)/z =$
(a) $(x/z) + y$
(b) $(x/z) + (y/z)$
(c) $x + (y/z)$
(d) None of the above
7. $x/(y + z) =$
(a) $(x/y) + (1/z)$
(b) $(x/y) + (x/z)$
(c) $(y + z)/x$
(d) None of the above
8. If $x = 10$, $y = 5$, $z = 2$, and $w = 20$, then $(xy - z^2)/w =$
(a) 5
(b) 2.3
(c) 46
(d) None of the above
9. $(8x^4)/(4x^2) =$
(a) $2x^2$
(b) 2
(c) $2x$
(d) None of the above
10. $\sqrt{\frac{X}{Y}} =$
(a) \sqrt{Y}/\sqrt{X}
(b) $\sqrt{1}/\sqrt{XY}$
(c) \sqrt{X}/\sqrt{Y}
(d) None of the above

The answers to both parts of the quiz appear at the end of this appendix.

Symbols

Each of the four basic arithmetic operations—addition, subtraction, multiplication, and division—is indicated by a symbol:

+ add

\times or \cdot multiply

– subtract

\div or / divide

In addition to these operations, the following symbols are used to indicate equality or inequality:

= equals

\neq not equal

\cong approximately equal to

> greater than

< less than

\geq greater than or equal to

\leq less than or equal to

Addition

Addition refers to the summation or accumulation of a set of numbers. In adding numbers, there are two basic laws: the *commutative law* and the *associative law*.

The **commutative law** of addition states that the order in which numbers are added is irrelevant. This can be seen in the following two examples:

$$1 + 2 = 3 \qquad 2 + 1 = 3$$

$$x + y = z \qquad y + x = z$$

In each example, which number was listed first and which number was listed second did not matter.

The **associative law** of addition states that in adding several numbers, any subgrouping of the numbers can be added first, last, or in the middle. You can see this in the following examples:

$$2 + 3 + 6 + 7 + 4 + 1 = 23$$

$$(5) + (6 + 7) + 4 + 1 = 23$$

$$5 + 13 + 5 = 23$$

$$5 + 6 + 7 + 4 + 1 = 23$$

In each of these examples, the order in which the numbers have been added has no effect on the results.

Subtraction

The process of subtraction is the opposite or inverse of addition. The operation of subtracting 1 from 2 (i.e., $2 - 1$) means that one unit is to be taken away from two units, leaving a remainder of one unit. In contrast to addition, the commutative and associative laws do not hold for subtraction. Therefore, as indicated in the following examples:

$$\begin{array}{lll} 8 - 4 = 4 & \text{but} & 4 - 8 = -4 \\ 3 - 6 = -3 & \text{but} & 6 - 3 = 3 \\ 8 - 3 - 2 = 3 & \text{but} & 3 - 2 - 8 = -7 \\ 9 - 4 - 2 = 3 & \text{but} & 2 - 4 - 9 = -11 \end{array}$$

When subtracting negative numbers, remember that that same result occurs when subtracting a negative number as when adding a positive number. Thus:

$$\begin{array}{ll} 4 - (-3) = +7 & 4 + 3 = 7 \\ 8 - (-10) = +18 & 8 + 10 = 18 \end{array}$$

Multiplication

The operation of multiplication is a shortcut method of addition when the same number is to be added several times. For example, if 7 is to be added 3 times ($7 + 7 + 7$), you could multiply 7 times 3 to obtain the product of 21.

In multiplication as in addition, the commutative laws and associative are in operation so that:

$$\begin{array}{l} a \times b = b \times a \\ 4 \times 5 = 5 \times 4 = 20 \\ (2 \times 5) \times 6 = 10 \times 6 = 60 \end{array}$$

A third law of multiplication, the *distributive law*, applies to the multiplication of one number by the sum of several numbers. Here:

$$\begin{array}{l} a(b + c) = ab + ac \\ 2(3 + 4) = 2(7) = 2(3) + 2(4) = 14 \end{array}$$

The resulting product is the same regardless of whether b and c are summed and multiplied by a , or a is multiplied by b and by c and the two products are added together.

You also need to remember that when multiplying negative numbers, a negative number multiplied by a negative number equals a positive number. Thus:

$$\begin{array}{l} (-a) \times (-b) = ab \\ (-5) \times (-4) = +20 \end{array}$$

Division

Just as subtraction is the opposite of addition, division is the opposite or inverse of multiplication. Division can be viewed as a shortcut to subtraction. When 20 is divided by 4, you are actually determining the number of times that 4 can be subtracted from 20. In general, however, the number of times one number can be divided by another may not be an exact integer value, because there could be a remainder. For example, if 21 is divided by 4, the answer is 5 with a remainder of 1, or $5 \frac{1}{4}$.

As in the case of subtraction, neither the commutative nor associative law of addition and multiplication holds for division.

$$a \div b \neq b \div a$$

$$9 \div 3 \neq 3 \div 9$$

$$6 \div (3 \div 2) = 4$$

$$(6 \div 3) \div 2 = 1$$

The distributive law will hold only when the numbers to be added are contained in the numerator, not the denominator. Thus:

$$\frac{a+b}{c} = \frac{a}{c} + \frac{b}{c} \quad \text{but} \quad \frac{a}{b+c} \neq \frac{a}{b} + \frac{a}{c}$$

For example:

$$\frac{6+9}{3} = \frac{6}{3} + \frac{9}{3} = 2 + 3 = 5$$

$$\frac{1}{2+3} = \frac{1}{5} \quad \text{but} \quad \frac{1}{2+3} \neq \frac{1}{2} + \frac{1}{3}$$

The last important property of division states that if the numerator and the denominator are both multiplied or divided by the same number, the resulting quotient will not be affected. Therefore:

$$\frac{80}{40} = 2$$

then

$$\frac{5(80)}{5(40)} = \frac{400}{200} = 2$$

and

$$\frac{80 \div 5}{40 \div 5} = \frac{16}{8} = 2$$

Fractions

A fraction is a number that consists of a combination of whole numbers and/or parts of whole numbers. For instance, the fraction $\frac{1}{3}$ consists of only one portion of a number, whereas the fraction $\frac{7}{6}$ consists of the whole number 1 plus the fraction $\frac{1}{6}$. Each of the operations of addition, subtraction, multiplication, and division can be used with fractions. When adding and subtracting fractions, you must obtain the lowest common denominator for each fraction prior to adding or subtracting them. Thus, in adding $\frac{1}{3} + \frac{1}{5}$, the lowest common denominator is 15, so:

$$\frac{5}{15} + \frac{3}{15} = \frac{8}{15}$$

In subtracting $\frac{1}{4} - \frac{1}{6}$, the same principles applies, so that the lowest common denominator is 12, producing a result of:

$$\frac{3}{12} - \frac{2}{12} = \frac{1}{12}$$

Multiplying and dividing fractions do not have the lowest common denominator requirement associated with adding and subtracting fractions. Thus, if a/b is multiplied by c/d , the result is $\frac{ac}{bd}$.

The resulting numerator, ac , is the product of the numerators a and c , whereas the denominator, bd , is the product of the two denominators b and d . The resulting fraction can sometimes be reduced to a lower term by dividing the numerator and denominator by a common factor. For example, taking:

$$\frac{2}{3} \times \frac{6}{7} = \frac{12}{21}$$

and dividing the numerator and denominator by 3 produces the result $\frac{4}{7}$.

Division of fractions can be thought of as the inverse of multiplication, so the divisor can be inverted and multiplied by the original fraction. Thus:

$$\frac{9}{5} \div \frac{1}{4} = \frac{9}{5} \times \frac{4}{1} = \frac{36}{5}$$

The division of a fraction can also be thought of as a way of converting the fraction to a decimal number. For example, the fraction $\frac{2}{5}$ can be converted to a decimal number by dividing its numerator, 2, by its denominator, 5, to produce the decimal number 0.40.

Exponents and Square Roots

Exponentiation (raising a number to a power) provides a shortcut in writing numerous multiplications. For example, $2 \times 2 \times 2 \times 2 \times 2$ can be written as $2^5 = 32$. The 5 represents the exponent (or power) of the number 2, telling you that 2 is to be multiplied by itself five times.

Several rules can be applied for multiplying or dividing numbers that contain exponents.

Rule 1: $x^a \cdot x^b = x^{(a+b)}$

If two numbers involving a power of the same number are multiplied, the product is the same number raised to the sum of the powers.

$$4^2 \cdot 4^3 = (4 \cdot 4)(4 \cdot 4 \cdot 4 \cdot 4) = 4^5$$

Rule 2: $(x^a)^b = x^{ab}$

If you take the power of a number that is already taken to a power, the result will be a number that is raised to the product of the two powers. For example,

$$(4^2)^3 = (4^2)(4^2)(4^2) = 4^6$$

Rule 3: $\frac{x^a}{x^b} = x^{(a-b)}$

If a number raised to a power is divided by the same number raised to a power, the quotient will be the number raised to the difference of the powers. Thus:

$$\frac{3^5}{3^3} = \frac{3 \cdot 3 \cdot 3 \cdot 3 \cdot 3}{3 \cdot 3 \cdot 3} = 3^2$$

If the denominator has a higher power than the numerator, the resulting quotient will be a negative power. Thus:

$$\frac{3^3}{3^5} = \frac{3 \cdot 3 \cdot 3}{3 \cdot 3 \cdot 3 \cdot 3 \cdot 3} = \frac{1}{3^2} = 3^{-2} = \frac{1}{9}$$

If the difference between the powers of the numerator and denominator is 1, the result will be the number itself. In other words, $x^1 = x$. For example:

$$\frac{3^3}{3^2} = \frac{3 \cdot 3 \cdot 3}{3 \cdot 3} = 3^1 = 3$$

If, however, there is no difference in the power of the numbers in the numerator and denominator, the result will be 1. Thus:

$$\frac{x^a}{x^a} = x^{a-a} = x^0 = 1$$

Therefore, any number raised to the 0 power equals 1. For example:

$$\frac{3^3}{3^3} = \frac{3 \cdot 3 \cdot 3}{3 \cdot 3 \cdot 3} = 3^0 = 1$$

The square root, represented by the symbol $\sqrt{\quad}$, is a special power of number, the $1/2$ power. It indicates the value that when multiplied by itself, will produce the original number.

Equations

In statistics, many formulas are expressed as equations where one unknown value is a function of another value. Thus, it is important that you know how to manipulate equations into various forms. The rules of addition, subtraction, multiplication, and division can be used to work with equations. For example, the equation:

$$x - 2 = 5$$

can be solved for x by adding 2 to each side of the equation. This results in:

$$x - 2 + 2 = 5 + 2. \text{ Therefore } x = 7.$$

If $x + y = z$, you could solve for x by subtracting y from both sides of the equation so that

$$x + y - y = z - y \text{ Therefore } x = z - y.$$

If the product of two variables is equal to a third variable, such as:

$$x \cdot y = z$$

you can solve for x by dividing both sides of the equation by y . Thus:

$$\begin{aligned} \frac{x \cdot y}{y} &= \frac{z}{y} \\ x &= \frac{z}{y} \end{aligned}$$

Conversely, if $\frac{x}{y} = z$, you can solve for x by multiplying both sides of the equation by y :

$$\begin{aligned} \frac{xy}{y} &= zy \\ x &= zy \end{aligned}$$

In summary, the various operations of addition, subtraction, multiplication, and division can be applied to equations as long as the same operation is performed on each side of the equation, thereby maintaining the equality.

Answers to Quiz

Part 1

1. $\frac{3}{2}$
2. 0.16
3. $\frac{5}{3}$
4. $\frac{1}{81}$
5. 0.20
6. 1.30
7. -6.4
8. $+\frac{1}{6}$
9. $\frac{3}{200}$
10. 4

Part 2

1. c
2. c
3. a
4. b
5. a
6. b
7. d
8. b
9. a
10. c



Statistical Tables

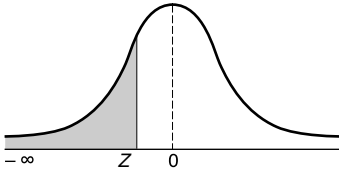
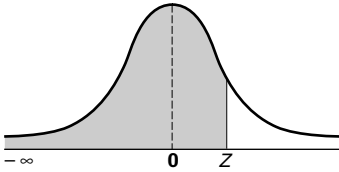


Table C.1

*The Cumulative Standardized Normal Distribution*Entry represents area under the cumulative standardized normal distribution from $-\infty$ to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003
-3.8	0.00007	0.00007	0.00007	0.00006	0.00006	0.00006	0.00006	0.00005	0.00005	0.00005
-3.7	0.00011	0.00010	0.00010	0.00010	0.00009	0.00009	0.00008	0.00008	0.00008	0.00008
-3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
-3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017
-3.4	0.00034	0.00032	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
-3.3	0.00048	0.00047	0.00045	0.00043	0.00042	0.00040	0.00039	0.00038	0.00036	0.00035
-3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
-3.1	0.00097	0.00094	0.00090	0.00087	0.00084	0.00082	0.00079	0.00076	0.00074	0.00071
-3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00103	0.00100
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2388	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2482	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

(continues)



Entry represents area under the standardized normal distribution from $-\infty$ to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7518	0.7549
0.7	0.7580	0.7612	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99897	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997
4.0	0.99996832									
4.5	0.99999660									
5.0	0.99999971									
5.5	0.99999998									
6.0	0.99999999									

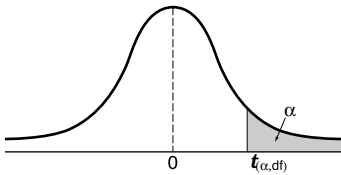


Table C.2
Critical Values of t

Degrees of Freedom	Upper-Tail Areas					
	0.25	0.10	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0322
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784

Degrees of Freedom	Upper-Tail Areas					
	0.25	0.10	0.05	0.025	0.01	0.005
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5177	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500
31	0.6825	1.3095	1.6955	2.0395	2.4528	2.7740
32	0.6822	1.3086	1.6939	2.0369	2.4487	2.7385
33	0.6820	1.3077	1.6924	2.0345	2.4448	2.7333
34	0.6818	1.3070	1.6909	2.0322	2.4411	2.7284
35	0.6816	1.3062	1.6896	2.0301	2.4377	2.7238
36	0.6814	1.3055	1.6883	2.0281	2.4345	2.7195
37	0.6812	1.3049	1.6871	2.0262	2.4314	2.7154
38	0.6810	1.3042	1.6860	2.0244	2.4286	2.7116
39	0.6808	1.3036	1.6849	2.0227	2.4258	2.7079
40	0.6807	1.3031	1.6839	2.0211	2.4233	2.7045
41	0.6805	1.3025	1.6829	2.0195	2.4208	2.7012
42	0.6804	1.3020	1.6820	2.0181	2.4185	2.6981
43	0.6802	1.3016	1.6811	2.0167	2.4163	2.6951
44	0.6801	1.3011	1.6802	2.0154	2.4141	2.6923
45	0.6800	1.3006	1.6794	2.0141	2.4121	2.6896
46	0.6799	1.3022	1.6787	2.0129	2.4102	2.6870

(continues)

Degrees of Freedom	Upper-Tail Areas					
	0.25	0.10	0.05	0.025	0.01	0.005
47	0.6797	1.2998	1.6779	2.0117	2.4083	2.6846
48	0.6796	1.2994	1.6772	2.0106	2.4066	2.6822
49	0.6795	1.2991	1.6766	2.0096	2.4049	2.6800
50	0.6794	1.2987	1.6759	2.0086	2.4033	2.6778
51	0.6793	1.2984	1.6753	2.0076	2.4017	2.6757
52	0.6792	1.2980	1.6747	2.0066	2.4002	2.6737
53	0.6791	1.2977	1.6741	2.0057	2.3988	2.6718
54	0.6791	1.2974	1.6736	2.0049	2.3974	2.6700
55	0.6790	1.2971	1.6730	2.0040	2.3961	2.6682
56	0.6789	1.2969	1.6725	2.0032	2.3948	2.6665
57	0.6788	1.2966	1.6720	2.0025	2.3936	2.6649
58	0.6787	1.2963	1.6716	2.0017	2.3924	2.6633
59	0.6787	1.2961	1.6711	2.0010	2.3912	2.6618
60	0.6786	1.2958	1.6706	2.0003	2.3901	2.6603
61	0.6785	1.2956	1.6702	1.9996	2.3890	2.6589
62	0.6785	1.2954	1.6698	1.9990	2.3880	2.6575
63	0.6784	1.2951	1.6694	1.9983	2.3870	2.6561
64	0.6783	1.2949	1.6690	1.9977	2.3860	2.6549
65	0.6783	1.2947	1.6686	1.9971	2.3851	2.6536
66	0.6782	1.2945	1.6683	1.9966	2.3842	2.6524
67	0.6782	1.2943	1.6679	1.9960	2.3833	2.6512
68	0.6781	1.2941	1.6676	1.9955	2.3824	2.6501
69	0.6781	1.2939	1.6672	1.9949	2.3816	2.6490
70	0.6780	1.2938	1.6669	1.9944	2.3808	2.6479
71	0.6780	1.2936	1.6666	1.9939	2.3800	2.6469
72	0.6779	1.2934	1.6663	1.9935	2.3793	2.6459
73	0.6779	1.2933	1.6660	1.9930	2.3785	2.6449
74	0.6778	1.2931	1.6657	1.9925	2.3778	2.6439

Degrees of Freedom	Upper-Tail Areas					
	0.25	0.10	0.05	0.025	0.01	0.005
75	0.6778	1.2929	1.6654	1.9921	2.3771	2.6430
76	0.6777	1.2928	1.6652	1.9917	2.3764	2.6421
77	0.6777	1.2926	1.6649	1.9913	2.3758	2.6412
78	0.6776	1.2925	1.6646	1.9908	2.3751	2.6403
79	0.6776	1.2924	1.6644	1.9905	2.3745	2.6395
80	0.6776	1.2922	1.6641	1.9901	2.3739	2.6387
81	0.6775	1.2921	1.6639	1.9897	2.3733	2.6379
82	0.6775	1.2920	1.6636	1.9893	2.3727	2.6371
83	0.6775	1.2918	1.6634	1.9890	2.3721	2.6364
84	0.6774	1.2917	1.6632	1.9886	2.3716	2.6356
85	0.6774	1.2916	1.6630	1.9883	2.3710	2.6349
86	0.6774	1.2915	1.6628	1.9879	2.3705	2.6342
87	0.6773	1.2914	1.6626	1.9876	2.3700	2.6335
88	0.6773	1.2912	1.6624	1.9873	2.3695	2.6329
89	0.6773	1.2911	1.6622	1.9870	2.3690	2.6322
90	0.6772	1.2910	1.6620	1.9867	2.3685	2.6316
91	0.6772	1.2909	1.6618	1.9864	2.3680	2.6309
92	0.6772	1.2908	1.6616	1.9861	2.3676	2.6303
93	0.6771	1.2907	1.6614	1.9858	2.3671	2.6297
94	0.6771	1.2906	1.6612	1.9855	2.3667	2.6291
95	0.6771	1.2905	1.6611	1.9853	2.3662	2.6286
96	0.6771	1.2904	1.6609	1.9850	2.3658	2.6280
97	0.6770	1.2903	1.6607	1.9847	2.3654	2.6275
98	0.6770	1.2902	1.6606	1.9845	2.3650	2.6269
99	0.6770	1.2902	1.6604	1.9842	2.3646	2.6264
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259
110	0.6767	1.2893	1.6588	1.9818	2.3607	2.6213
120	0.6765	1.2886	1.6577	1.9799	2.3578	2.6174
∞	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758

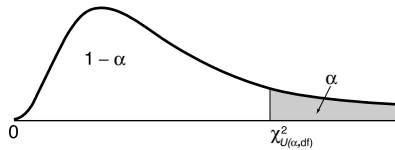


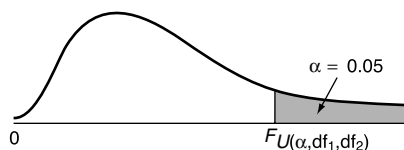
Table C.3
Critical Values of χ^2

For a particular number of degrees of freedom, entry represents the critical value of χ^2 corresponding to a specified upper-tail area (α).

Degrees of Freedom	Upper-Tail Areas (α)											
	0.995	0.99	0.975	0.95	0.90	0.75	0.25	0.10	0.05	0.025	0.01	0.005
1			0.001	0.004	0.016	0.102	1.323	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.575	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	6.626	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.455	7.841	10.645	12.592	14.449	16.812	18.458
7	0.989	1.239	1.690	2.167	2.833	4.255	9.037	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	5.071	10.219	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	5.899	11.389	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	6.737	12.549	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	7.584	13.701	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	8.438	14.845	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	9.299	15.984	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	10.165	17.117	21.064	23.685	26.119	29.141	31.319

Degrees of Freedom	Upper-Tail Areas (α)											
	0.995	0.99	0.975	0.95	0.90	0.75	0.25	0.10	0.05	0.025	0.01	0.005
15	4.601	5.229	6.262	7.261	8.547	11.037	18.245	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	11.912	19.369	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	12.792	20.489	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	13.675	21.605	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	14.562	22.718	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	15.452	23.828	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	16.344	24.935	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	17.240	26.039	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	18.137	27.141	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	19.037	28.241	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	19.939	29.339	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	20.843	30.435	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	21.749	31.528	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	22.657	32.620	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	23.567	33.711	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	24.478	34.800	40.256	43.773	46.979	50.892	53.672

For larger values of degrees of freedom (df) the expression $Z = \sqrt{2x^2 - \sqrt{2(df) - 1}}$ may be used and the resulting upper-tail area can be obtained from the table of the cumulative standardized normal distribution (Table C.1).

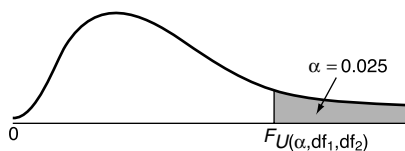
**Table C.4***Critical Values of F*

For a particular combination of numerator and denominator degrees of freedom, entry represents the critical values of F corresponding to a specified upper-tail area (α).

Denominator, df_2	Numerator, df_1								
	1	2	3	4	5	6	7	8	9
1	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

Numerator, df_1									
10	12	15	20	24	30	40	60	120	∞
241.90	243.90	245.90	248.00	249.10	250.10	251.10	252.20	253.30	254.30
19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
2.30	2.23	2.15	2.07	2.03	1.98	1.91	1.89	1.84	1.78
2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

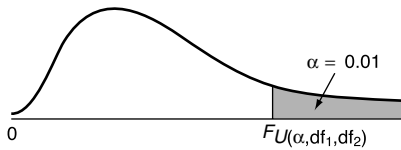
(continues)



Denominator, df_2	Numerator, df_1								
	1	2	3	4	5	6	7	8	9
1	647.80	799.50	864.20	899.60	921.80	937.10	948.20	956.70	963.30
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.39	39.39
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11

Numerator, df_1									
10	12	15	20	24	30	40	60	120	∞
968.60	976.70	984.90	993.10	997.20	1,001.00	1,006.00	1,010.00	1,014.00	1,018.00
39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

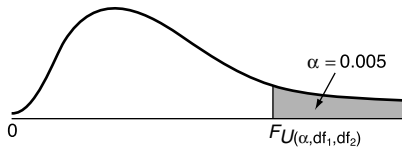
(continues)



Denominator, df_2	Numerator, df_1								
	1	2	3	4	5	6	7	8	9
1	4,052.00	4,999.50	5,403.00	5,625.00	5,764.00	5,859.00	5,928.00	5,982.00	6,022.00
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

Numerator, df_1									
10	12	15	20	24	30	40	60	120	∞
6,056.00	6,106.00	6,157.00	6,209.00	6,235.00	6,261.00	6,287.00	6,313.00	6,339.00	6,366.00
99.40	99.42	99.43	94.45	99.46	99.47	99.47	99.48	99.49	99.50
27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.81	2.75
3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

(continues)



Denominator, df_2	Numerator, df_1								
	1	2	3	4	5	6	7	8	9
1	16,211.00	20,000.00	21,615.00	22,500.00	23,056.00	23,437.00	23,715.00	23,925.00	24,091.00
2	198.50	199.00	199.20	199.20	199.30	199.30	199.40	199.40	199.40
3	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88
4	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14
5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77
6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39
7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51
8	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34
9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54
10	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97
11	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54
12	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20
13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94
14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72
15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54
16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38
17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25
18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14
19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.02	3.88
22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75
24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69
25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64
26	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60
27	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56
28	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52
29	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48
30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45
40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22
60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01
120	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81
∞	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62

Numerator, df_1									
10	12	15	20	24	30	40	60	120	∞
24,224.00	24,426.00	24,630.00	24,836.00	24,910.00	25,044.00	25,148.00	25,253.00	25,359.00	25,465.00
199.40	199.40	199.40	199.40	199.50	199.50	199.50	199.50	199.50	199.50
43.69	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.83
20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32
13.62	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.11
10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88
8.38	8.18	7.97	7.75	7.65	7.53	7.42	7.31	7.19	7.08
7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95
6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.19
5.85	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	1.61
5.42	5.24	5.05	4.86	4.75	4.65	4.55	4.44	4.34	4.23
5.09	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.90
4.82	4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.65
4.60	4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.41
4.42	4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.26
4.27	4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	3.11
4.14	3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	2.98
4.03	3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.89	2.87
3.93	3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.78
3.85	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69
3.77	3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	2.61
3.70	3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	2.55
3.64	3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.48
3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.43
3.54	3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.38
3.49	3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.33
3.45	3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.29
3.41	3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.25
3.38	3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.21
3.34	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.18
3.12	2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	1.93
2.90	2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.69
2.71	2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	1.43
2.52	2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.00

Table C.5
Control Chart Factors

Number of Observations in Sample	d_2	d_3	D_3	D_4	A_2
2	1.128	0.853	0	3.267	1.880
3	1.693	0.888	0	2.575	1.023
4	2.059	0.880	0	2.282	0.729
5	2.326	0.864	0	2.114	0.577
6	2.534	0.848	0	2.004	0.483
7	2.704	0.833	0.076	1.924	0.419
8	2.847	0.820	0.136	1.864	0.373
9	2.970	0.808	0.184	1.816	0.337
10	3.078	0.797	0.223	1.777	0.308
11	3.173	0.787	0.256	1.744	0.285
12	3.258	0.778	0.283	1.717	0.266
13	3.336	0.770	0.307	1.693	0.249
14	3.407	0.763	0.328	1.672	0.235
15	3.472	0.756	0.347	1.653	0.223
16	3.532	0.750	0.363	1.637	0.212
17	3.588	0.744	0.378	1.622	0.203
18	3.640	0.739	0.391	1.609	0.194
19	3.689	0.733	0.404	1.596	0.187
20	3.735	0.729	0.415	1.585	0.180
21	3.778	0.724	0.425	1.575	0.173
22	3.819	0.720	0.435	1.565	0.167
23	3.858	0.716	0.443	1.557	0.162
24	3.895	0.712	0.452	1.548	0.157
25	3.931	0.708	0.459	1.541	0.153

Source: Reprinted from ASTM-STP 15D by kind permission of the American Society for Testing and Materials.



Using Microsoft Excel Wizards

Wizards are a series of linked dialog boxes that guide you, step by step, through the task of generating a Microsoft Office object. You make selections and enter information about the object as you step through a series of dialog boxes, clicking **Next** buttons. Clicking a **Finish** button, typically done in the last dialog box, generates the object. At any point, you can cancel the operation of the wizard by clicking a **Cancel** button or move to a previous dialog box by clicking **Back**.

d.1 Using the Chart Wizard

You use the Microsoft Excel Chart Wizard to generate a wide variety of charts. You begin this wizard by selecting **Insert → Chart** from the Excel menu bar and making these entries in the Chart Wizard dialog boxes:

Step 1 dialog box. Choose the chart type.

Step 2 dialog box. Enter the workbook locations of the source data for the values to be plotted and the source data for chart labeling information (if any).

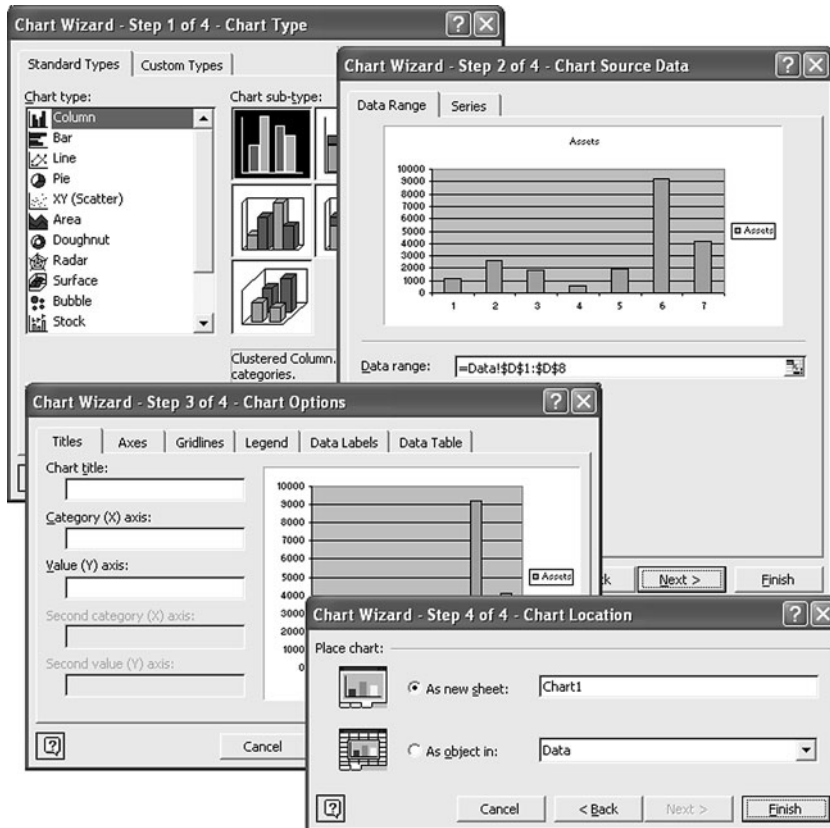
Step 3 dialog box. Specify the formatting and labeling options for the chart.

Step 4 dialog box. Choose the workbook location of the chart. You will always create a better-scaled chart if you choose the **As new sheet** and not the **As object in (a worksheet)** option.

The Chart Wizard dialog boxes for Microsoft Excel 2003 are shown in Figure D.1.

FIGURE D.1

Chart Wizard dialog boxes



Choosing the Best Chart Options

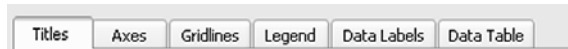
The default settings in the Step 3 dialog box create imperfectly designed charts. When you use the Chart Wizard to generate your own charts, consider selecting the tabs of this dialog box (see Figure D.2) and making the changes listed below to your chart. (Not every tab displays for every chart type; only three tabs that are appropriate for the chart type you selected in the Step 2 dialog box will display.) The suggested changes are as follows:

- Select the **Titles** tab and enter a title and axis labels, if appropriate.

- Select the **Axes** tab and then select both the (X) axis and (Y) axis check boxes. Also select the **Automatic** option button under the (X) axis check box.
- Select the **Gridlines** tab and deselect (uncheck) all the choices under the (X) axis heading and under the (Y) axis heading.
- Select the **Legend** tab and deselect (uncheck) the **Show legend** check box.
- Select the **Data Labels** tab, and in the Data labels group select the **None** option button.

FIGURE D.2

Chart Wizard
Step 3 dialog
box tabs



If you overlook making any suggested change while you are using the wizard, you can always return to these tabs by right-clicking the chart and selecting **Chart Options** from the shortcut menu that appears.

d.2 Using the PivotTable Wizard

You use the PivotTable Wizard to generate **PivotTables**, summary tables that update themselves automatically as the data on which they are based changes. PivotTables have many applications involving database querying, retrieval, and **drill-down**, the ability to display the underlying raw data that is being summarized by the table. When reading this book, you can use the wizard to generate one-way and two-way frequency distribution tables for categorical data (see Chapter 2).

To use the PivotTable Wizard, you first select **Data → PivotTable and PivotChart Report (Data → PivotTable Report** if you are using Microsoft Excel 97). Then, you enter information about the design of the table as you step through the dialog boxes by clicking a **Next** button. You click the **Finish** button in the last dialog box to create the table. At any point, you can cancel the operation of the wizard by clicking **Cancel**, or move to a previous dialog box by clicking **Back**. In the Step 3 dialog box, you must click additional buttons to display supplemental dialog boxes that add two additional, unnumbered steps to the process of creating a PivotTable.

The PivotTable Wizards of the various versions of Microsoft Excel differ slightly. For Microsoft Excel 2003, the three-step wizard (see Figure D.3) requires you to do the following:

Step 1: Select the source for the data for the PivotTable and the type of report to be produced in the Step 1 dialog box. In this text you will

always select **Microsoft Excel list or database** as the source and **PivotTable** as the report type. (You do not select a report type in Microsoft Excel 97; PivotTable is assumed.)

- Step 2:** Enter the cell range of the data that will be summarized in the PivotTable. The first row of this cell range should contain column headings that the wizard will later use as the variable name(s).
- Step 3:** Choose the location of the PivotTable. In this text, you will always select the **New worksheet** option. Click **Layout** to display the supplemental Layout dialog box.

In the Layout dialog box, you drag name labels from a list of variables that appears on the right side (obscured in Figure D.4) into a template that contains page, row, column, and data areas. When you drag a label into the Data area, the label changes to **Count of variable** to indicate that the PivotTable will automatically tally the variable. When you have finished dragging labels, click **OK** to return to the main Step 3 dialog box and then click **Options** to display the supplemental PivotTable Options dialog box.

In the PivotTable Options dialog box, enter a self-descriptive table name in the **Name** box and usually enter **0** in the **For empty cells**, **show** box and leave all other settings as is. Then click **OK** to return to the main Step 3 dialog box and click **Finish** to produce the PivotTable.

FIGURE D.3

PivotTable
Wizard

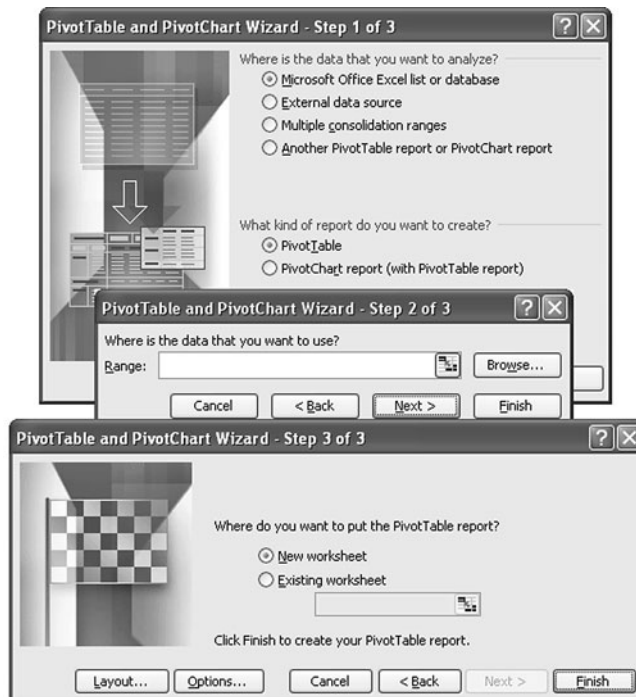
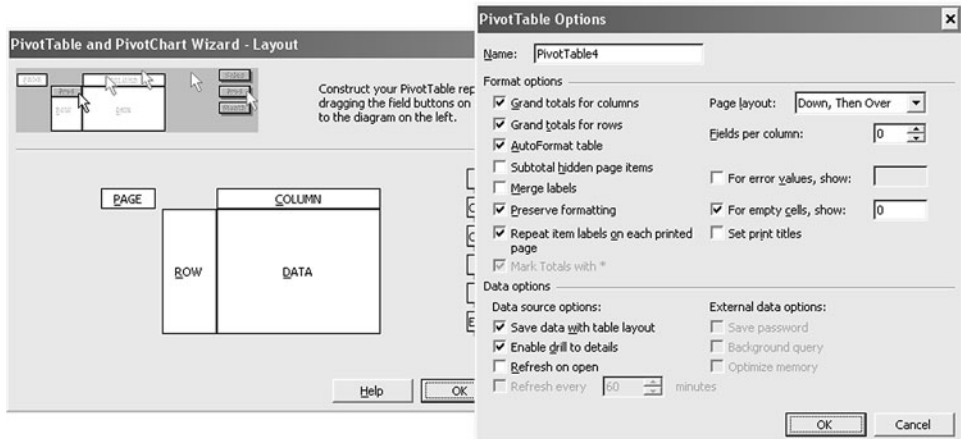


FIGURE D.4

PivotTable Layout (slightly obscured) and Options dialog boxes

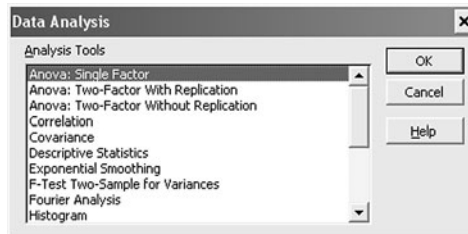


d.3 Using the Data Analysis Tools

The Data Analysis Tools are a set of statistical procedures included with Microsoft Excel. To use the Data Analysis tools, first verify that they are properly installed (see Section A.2). Then select **Tools** → **Data Analysis** to display the Data Analysis dialog box (see Figure D.5).

FIGURE D.5

Data Analysis dialog box



In the Analysis Tools list, select a procedure and click the OK button. For most procedures, a second dialog box appears in which you make entries and selections.

d.4 Simple Linear Regression

Open a worksheet in which you have placed data for the variables for the regression analysis in separate columns. Select **Tools** → **Data Analysis**. Select **Regression** from the Data Analysis list and click OK. In the procedure's dialog box, enter the cell range of the Y variable data as the **Input Y Range**, enter the cell range of the X variable data as the **Input X Range**, select **Labels**

if the variable columns include a column heading in their first rows, select **Confidence Level**, and Click **OK**. Results appear on a separate worksheet.

If you use the Chart Wizard to generate a scatter diagram for this analysis, select the **XY (Scatter)** from the **Standard Types Chart type** box and leave the first **Chart subtype** selected in the Step 1 dialog box.



Glossary

Alternative hypothesis (H_1)—The opposite of the null hypothesis (H_0).

Analysis of variance (ANOVA)—A statistical method that tests the significance of different factors on a variable of interest.

Arithmetic mean—The balance point in a set of data that is calculated by summing the observed numerical values in a set of data and then dividing by the number of values involved.

Bar chart—A chart containing rectangles (“bars”) in which the length of each bar represents the count, amount, or percentage of responses of one category.

Binomial distribution—A distribution that finds the probability of a given number of successes for a given probability of success and sample size.

Box-and-whisker plot—A graphical representation of the five-number summary that consists of the smallest value, the first quartile (or 25th percentile), the median, the third quartile (or 75th percentile), and the largest value.

Categorical variable—The values of these variables are selected from an established list of categories.

Cell—Intersection of a row and a column in a two-way cross-classification table

Chi-square (χ^2) distribution—Distribution used to test relationships in two-way cross-classification tables.

Coefficient of correlation—Measures the strength of the linear relationship between two variables.

Coefficient of determination—Measures the proportion of variation in Y that is explained by the independent variable X in the regression model.

Collectively exhaustive events—One in a set of events must occur.

Common causes of variation—Represent the inherent variability that exists in the system.

Completely randomized design—An experimental design in which there is only a single factor.

Confidence interval estimate—An estimate of the population parameter given by an interval with a lower and upper limit.

Continuous numerical variables—Values of these variables are measurements.

Control chart—A tool for distinguishing between the common and special causes of variation.

Critical value—Divides the nonrejection region from the rejection region.

Degrees of freedom—The actual number of values that are free to vary after the mean is known.

Dependent variable—The variable to be predicted in a regression analysis.

Descriptive statistics—The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

Discrete numeric variables—The values are counts of things.

Dot scale diagram—A chart in which each response is represented as a point above a number line that includes the range of all values.

Error sum of squares (SSE)—Consists of variation that is due to factors other than the relationship between X and Y .

Event—Each possible type of occurrence.

Expected frequency—Frequency expected in a particular cell if the null hypothesis is true.

Expected value—The mean of a probability distribution.

Experiments—A process that uses controlled conditions to study the effect on the variable of interest of varying the value(s) of another variable or variables.

Explanatory variable—The variable used to predict the dependent or response variable in a regression analysis.

F distribution—A distribution used for testing the ratio of two variances.

First quartile (Q_1)—The value such that 25.0% of the observations are smaller and 75.0% are larger.

Five-number summary—Consists of smallest value, Q_1 , median, Q_3 , and largest value.

Frame—The list of all items in the population from which samples will be selected.

Frequency distribution—A table of grouped numerical data in which the names of each group are listed in the first column and the percentages of each group of numerical data are listed in the second column.

Histogram—A special bar chart for grouped numerical data in which the frequencies or percentages of each group of numerical data are represented as individual bars.

Hypothesis testing—Methods used to make inferences about the hypothesized values of population parameters using sample statistics.

Independent events—Events in which the occurrence of one event in no way affects the probability of the second event.

Independent variable—The variable used to predict the dependent or response variable in a regression analysis.

Inferential statistics—The branch of statistics that analyzes sample data to draw conclusions about a population.

Level of significance—Probability of committing a type I error.

Mean—The balance point in a set of data that is calculated by summing the observed numerical values in a set of data and then dividing by the number of values involved.

Mean squares—The variances in an analysis-of-variance table.

Median—The middle value in a set of data that has been ordered from the lowest to highest value.

Mode—The value in a set of data that appears most frequently.

Mutually exclusive events—events are mutually exclusive if both events *cannot* occur at the same time.

Normal distribution—The normal distribution is defined by its mean (μ) and standard deviation (σ) and is bell-shaped.

Normal probability plot—A graphical device for helping to evaluate whether a set of data follows a normal distribution.

Null hypothesis—A statement about a parameter equal to a specific value, or the statement that there is no difference between the parameters for two or more populations.

Numerical variables—The values of these variables involve a counted or measured value.

Observed frequency—Actual tally in a particular cell of a cross-classification table.

***p*-chart**—Used to study a process that involves the proportion of items with a characteristic of interest.

***p*-value**—The probability of getting a test statistic equal to or more extreme than the result obtained from the sample data, given that the null hypothesis H_0 is true.

Paired samples—Items are matched according to some characteristic and the differences between the matched values are analyzed.

Parameter—A numerical measure that describes a characteristic of a population.

Pareto diagram—A special type of bar chart in which the count, amount, or percentage of responses of each category are presented in descending order left to right, along with a superimposed plotted line that represents a running cumulative percentage.

Percentage distribution—A table of grouped numerical data in which the names of each group are listed in the first column and the percentages of each group of numerical data are listed in the second column.

Pie chart—A chart in which wedge-shaped areas (“pie slices”) represent the count, amount, or percentage of each category and the circle (the “pie”) itself represents the total.

Placebo—A substance that has no medical effect.

Poisson distribution—A distribution to find the probability of the number of occurrences in an area of opportunity.

Population—All the members of a group about which you want to draw a conclusion.

Power of a statistical test—The probability of rejecting the null hypothesis when it is false and should be rejected.

Probability—The numeric value representing the chance, likelihood, or possibility a particular event will occur.

Probability distribution for a discrete random variable—A listing of all possible distinct outcomes and their probabilities of occurring.

Probability sampling—A sampling process that takes into consideration the chance of occurrence of each item being selected.

Published sources—Data available in print or in electronic form, including data found on Internet Web sites.

Range—The difference between the *largest* and *smallest* values in a set of data.

Region of rejection—Consists of the values of the test statistic that are unlikely to occur if the null hypothesis is true.

Regression sum of squares (SSR)—Consists of variation that is due to the relationship between *X* and *Y*.

Residual—The difference between the observed and predicted values of the dependent variable for a given value of *X*.

Response variable—The variable to be predicted in a regression analysis.

Sample—The part of the population selected for analysis.

Sampling—The process by which members of a population are selected for a sample.

Sampling distribution—The distribution of a sample statistic (such as the arithmetic mean) for all possible samples of a given size *n*.

Sampling error—Variation of the sample statistic from sample to sample.

Sampling with replacement—A sampling method in which each selected item is returned to the frame from which it was selected so that it has the same probability of being selected again.

Sampling without replacement—A sampling method in which each selected item is not returned to the frame from which it was selected. Using this technique, an item can be selected no more than one time.

Scatter plot—A chart that plots the values of two variables for each response. In a scatter plot, the *X*-axis (the horizontal axis) always represents units of one variable, and the *Y*-axis (the vertical axis) always represents units of the second variable.

Simple linear regression—A statistical technique that uses a *single* numerical independent variable *X* to predict the numerical dependent variable *Y*.

Simple random sampling—The probability sampling process in which every individual or item from a population has the same chance of selection as every other individual or item.

Six Sigma management —A method for breaking processes into a series of steps in order to eliminate defects and produce near perfect results.

Skewness—A skewed distribution is not symmetric. There are extreme values either in the lower portion of the distribution or in the upper portion of the distribution.

Slope—The change in Y per unit change in X .

Special causes of variation—Represent large fluctuations or patterns in the data that are not inherent to a process.

Standard deviation—Measure of variation around the mean of a set of data.

Standard error of the estimate—The standard deviation around the line of regression.

Statistic—A numerical measure that describes a characteristic of a sample.

Statistics—The branch of mathematics that consists of methods of processing and analyzing data to better support rational decision-making processes.

Sum of squares among groups (SSA)—The sum of the squared differences between the sample mean of each group and the mean of all the values, weighted by the sample size in each group.

Sum of squares total (SST)—Represents the sum of the squared differences between each individual value and the mean of all the values.

Sum of squares within groups (SSW)—Measures the difference between each value and the mean of its own group and sums the squares of these differences over all groups.

Summary table—A two-column table in which the names of the categories are listed in the first column, and the count, amount, or percentage of responses are listed in a second column.

Survey—A process that uses questionnaires or similar means to gather values for the responses from a set of participants.

Symmetry—Distribution in which each half of a distribution is a mirror image of the other half of the distribution.

t distribution—A distribution used to estimate the mean of a population and to test hypotheses about means.

Test statistic—The statistic used to determine whether to reject the null hypothesis.

Third quartile (Q_3)—The value such that 75.0% of the observations are smaller and 25.0% are larger.

Time-series plot—A chart in which each point represents a response at a specific time. In a time series plot, the X -axis (the horizontal axis) always represents units of time, and the Y -axis (the vertical axis) always represents units of the numerical responses.

Two-way cross-classification table—A table that presents the count or percentage of joint responses to two categorical variables (a mutually exclusive pairing, or cross-classifying, of categories from each variable). The categories of one variable form the rows of the table, and the categories of the other variable form the columns.

Type I error—Occurs if the null hypothesis H_0 is rejected when in fact it is true and should not be rejected. The probability of a type I error occurring is α .

Type II error—Occurs if the null hypothesis H_1 is not rejected when in fact it is false and should be rejected. The probability of a type II error occurring is β .

Variable—A characteristic of an item or an individual that will be analyzed using statistics.

Variance—The square of the standard deviation.

Variation—The amount of dispersion, or “spread,” in the data.

Y intercept—The value of Y when $X = 0$.

Z score—The difference between the value and the mean, divided by the standard deviation.

This page intentionally left blank



Index

A

α , 129

Alternative hypothesis, 126, 269

Analysis of variance (ANOVA)
see One-Way Analysis of
Variance

ANOVA summary table, 170

Arithmetic mean (see mean)

Arithmetic and algebra review,
235–244

B

β , 129

Bar chart, 18–19, 269

Binomial distribution, 79–82, 269

Box-and-whisker plot, 52–55,
269

C

Calculator keys,

Binomial distribution, 82

Box-and-whisker plot, 55

Chi-square tests, 162

Confidence interval estimate for
the mean (σ unknown), 115

Confidence interval estimate
for the proportion, 118

Entering data, 9

Mean, 48

Median, 48

Normal probabilities, 93

Normal probability plots, 95

One-Way Analysis of Variance
(ANOVA), 172

Poisson probabilities, 86

Pooled-variance t test for the
differences in two means,
145

Standard deviation, 48

Variance, 48

Z test for the differences in
two proportions, 140
Calculator settings, 231–233
Categorical variable, 3–4, 269
Cell, 160, 270
Central limit theorem, 105
Certain event, 63
Chart Wizard, 263–265
Chi-square distribution, 270
Chi-square distribution tables,
252–253
Chi-square test, 159–166
Classical approach to probability,
67
Coefficient of correlation, 194,
270
Coefficient of determination,
193–194, 270
Collectively exhaustive events,
64, 270
Common causes of variation,
213, 270
Complement, 64
Completely randomized design
(see One-Way Analysis of
Variance)
Confidence interval estimate,
111, 270
for the mean (σ unknown),
111–115
for the proportion, 116–118
for the slope, 200
Continuous values, 3, 270
Control chart factor tables, 262
Control chart, 212, 270
 p -chart, 214–219
Range (R) chart, 221–226
 \bar{X} chart, 221–226
Control limits, 213–214
Critical value, 130, 270

D

Data Analysis Tool, 267

Degrees of freedom, 114, 147,
154, 161, 168, 270
Dependent variable, 182, 270
Descriptive statistics, 4, 270
Discrete values, 3, 270
Discrete probability distribution,
73–75
DMAIC model, 211–212
Dot scale diagram, 27, 270
Double-blind study, 6

E

Elementary event, 62
Empirical approach to
probability, 67–68
Equation Blackboard,
Binomial distribution, 81–82
Chi-square tests, 165–166
Confidence interval estimate
for the mean (σ unknown),
114–115
Confidence interval estimate
for the proportion, 117–118
Confidence interval estimate
for the slope, 200
Mean, 37–39
Mean and standard deviation
of a discrete probability
distribution, 78
Median, 40–41
One-Way Analysis of Variance
(ANOVA), 168–169
 p -chart, 218–219
Paired t test, 153–154
Poisson distribution, 85–86
Pooled-variance t test for the
difference in two means,
146–147
Quartiles, 45
Range (R) chart, 224–225
Range, 46
Regression measures of
variation, 192–193

- Slope, 188–191
 - Standard deviation, 46–47
 - Standard error of the estimate, 195
 - t* test for the slope, 199
 - Variance, 46–47
 - \bar{X} chart, 225–226
 - Y intercept, 188–191
 - Z scores, 50
 - Z test for the difference in two proportions, 142–143
 - Event, 61–62, 270
 - Expected frequency, 160, 270
 - Expected value of a random variable, 75, 270
 - Experimental error, 167
 - Experiments, 6, 270
 - Explanatory variable (see independent variable)
- F**
- F* distribution 168–169, 271
 - F* distribution tables, 254–261
 - F* test statistic, 168–169
 - Factor, 166
 - Five-number summary, 52, 271
 - Frame, 8, 271
 - Frequency distribution, 24–25, 271
- H**
- Histogram, 25, 271
 - Hypothesis testing 129–132
 - Hypothesis testing steps, 129–132
- I**
- Independent events, 66, 271
 - Independent variable, 182, 271
 - Inferential statistics, 5, 271
- J**
- Joint event, 62
- L**
- Least-squares method, 184
 - Left-skewed, 50–51
 - Level of significance, 129, 271
- M**
- Mean, 37–40, 269, 271
 - Mean Squares, 271
 - Total (*MST*), 168–169
 - Among Groups (*MSA*), 168–170
 - Within Groups (*MSW*), 168–169
 - Measures of
 - central tendency, 37–45
 - variation, 45–50
 - Median, 38, 40–41, 271
 - Microsoft Excel settings, 233
 - Misusing graphs, 30–31
 - Mode, 41
 - Mutually exclusive, 64–65, 271
- N**
- Normal distribution, 87–93, 271
 - Normal distribution tables, 246–247
 - Normal probability plot, 94–95, 271
 - Null event, 63
 - Null hypothesis, 126, 272
 - Numerical variable, 3, 272
- O**
- Observed frequency, 165, 272
 - One-Way Analysis of Variance, 166–174, 269
 - Operational definition, 4
- P**
- p*-chart, 214–219, 272
 - p*-value, 131, 272
 - Paired *t* test, 150–155, 272
 - Parameter, 2, 272

- Pareto diagram, 20–21, 272
- Percentage distribution, 24–25, 272
- Pie chart, 19, 272
- PivotTable Wizard, 265–267
- Placebo, 6, 272
- Point estimate, 107
- Poisson distribution, 83–84, 272
- Pooled-variance *t* test, 143–150
- Population, 2, 272
- Power of the test, 129, 272
- Practical significance, 128
- Primary data sources, 5
- Probability, 62–64, 272
- Probability distribution for discrete random variables, 73–74, 272
- Probability sampling, 7, 273
- Published sources, 5, 273
- Q**
- Quartiles, 41–45
- R**
- Random variable, 62
- Range, 45–46, 273
- Range (*R*) chart, 221–226
- Red bead experiment, 219–221
- Region of nonrejection, 128
- Region of rejection, 128, 273
- Regression model prediction, 187
- Residual analysis, 196–197, 273
- Response variable (see dependent variable)
- Right-skewed, 51
- S**
- Sample, 2, 273
- Sampling, 7, 273
- Sampling distribution, 273
 - of the mean, 104–107
 - of the proportion, 107
- Sampling error, 109
- Sampling with replacement, 8, 273
- Sampling without replacement, 8, 273
- Scatter plot, 28–30, 273
- Secondary data sources, 5
- Shape, 50–55
- Simple linear regression, 182, 273
 - Assumptions, 195–196
- Simple random sampling, 7–8, 273
- Single-blind study, 6
- Six Sigma, 211–212, 273
- Skewness, 50–52, 274
- Slope, 183–184, 274
- Special causes of variation, 212–213, 274
- Spreadsheet Solutions,
 - Bar and pie charts, 20
 - Binomial probabilities, 82
 - Chi-square tests, 163
 - Confidence interval estimate for the mean (σ unknown), 116
 - Confidence interval estimate for the proportion, 118
 - Descriptive statistics, 43
 - Dot scale diagrams, 27–8
 - Entering data, 10
 - Frequency distributions and histograms, 26–27
 - Normal probabilities, 94
 - One-Way Analysis of Variance (ANOVA), 172
 - Assumptions, 174
 - Paired *t* test, 151
 - Pareto diagrams, 21–22
 - Poisson probabilities, 87
 - Pooled variance *t* test for the difference in two means, 146
 - Scatter plots, 30
 - Two-way tables, 24

- Z test for the difference in two proportions, 140
 - Standard deviation, 46–48, 274
 - Standard deviation of a random variable, 76
 - Standard error of the estimate, 194–195, 274
 - Standard (Z) scores, 49–50
 - Statistic, 3
 - Statistics, 1, 274
 - Subjective approach to probability, 68
 - Sum of Squares,
 - Error (SSE), 191, 270
 - Regression (SSR), 191, 273
 - Total (SST), 167, 168–169, 192, 274
 - Among Groups (SSA), 167, 168–169, 274
 - Within Groups (SSW), 167, 168–169, 274
 - Summary table, 17–18, 274
 - Surveys, 6, 274
 - Symmetric, 50, 274
- T**
- t* distribution, 112–115, 274
 - t* distribution tables, 248–251
 - Tables of the,
 - Chi-square distribution, 252–253
 - Control chart factors, 262
 - F* distribution, 254–261
 - Normal distribution, 246–247
 - t* distribution, 248–251
 - Test of hypothesis,
 - Chi-square test, 159–166
 - for the difference between two proportions, 137–143
 - for the difference between the means of two independent groups, 143–150
 - for the slope, 198–199
 - One-Way Analysis of Variance, 166–174
 - Paired *t* test, 150–155
 - Test statistic, 127–128, 274
 - Time-series plot, 28, 274
 - Total quality management, 209–210
 - Treatment effect, 167
 - Two-way cross-classification tables, 22–24, 275
 - Type I error, 129, 275
 - Type II error, 129, 275
- V**
- Variable, 3, 275
 - Variance, 46–48, 275
- X**
- \bar{X} chart, 221–226
- Y**
- Y* intercept, 183, 275
- Z**
- Z scores, 49–50, 275

Wouldn't it be great

if the world's leading technical
publishers joined forces to deliver
their best tech books in a common
digital reference platform?

They have. Introducing
InformIT Online Books
powered by Safari.

POWERED BY
Safari
TECH BOOKS ONLINE®

■ Specific answers to specific questions.

InformIT Online Books' powerful search engine gives you
relevance-ranked results in a matter of seconds.

■ Immediate results.

With InformIT Online Books, you can select the book
you want and view the chapter or section you need
immediately.

■ Cut, paste and annotate.

Paste code to save time and eliminate typographical
errors. Make notes on the material you find useful and
choose whether or not to share them with your work
group.

■ Customized for your enterprise.

Customize a library for you, your department or your entire
organization. You only pay for what you need.

Get your first 14 days FREE!

For a limited time, InformIT Online Books is offering
its members a 10 book subscription risk-free for
14 days. Visit [http://www.informit.com/online-](http://www.informit.com/online-books)
books for details.

informIT
Online Books

informit.com/onlinebooks



Register Your Book

at www.awprofessional.com/register

You may be eligible to receive:

- Advance notice of forthcoming editions of the book
- Related book recommendations
- Chapter excerpts and supplements of forthcoming titles
- Information about special contests and promotions throughout the year
- Notices and reminders about author appearances, tradeshows, and online chats with special guests

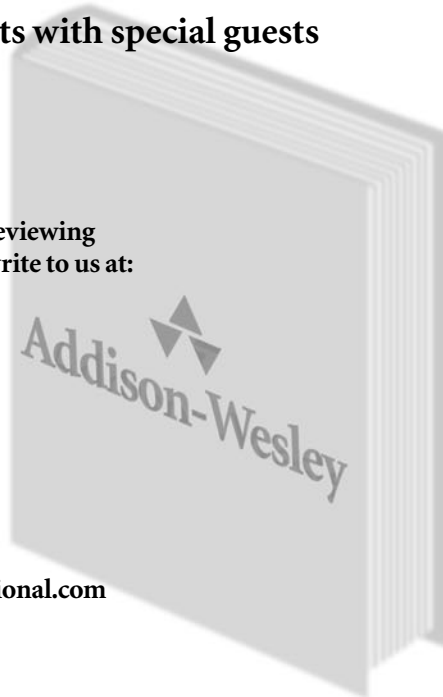


Contact us

If you are interested in writing a book or reviewing manuscripts prior to publication, please write to us at:

Editorial Department
Addison-Wesley Professional
75 Arlington Street, Suite 300
Boston, MA 02116 USA
Email: AWPro@aw.com

Visit us on the Web: <http://www.awprofessional.com>





www.informit.com

YOUR GUIDE TO IT REFERENCE



Articles

Keep your edge with thousands of free articles, in-depth features, interviews, and IT reference recommendations – all written by experts you know and trust.



Online Books

Answers in an instant from **InformIT Online Book's** 600+ fully searchable on line books. For a limited time, you can get your first 14 days **free**.



Catalog

Review online sample chapters, author biographies and customer rankings and choose exactly the right book from a selection of over 5,000 titles.

NATIONAL BEST-SELLER NOW IN PAPERBACK!

The Truth About Managing People

...And Nothing But the Truth

Stephen P. Robbins

BEST-SELLING MANAGEMENT AUTHOR
OVER 2,000,000 COPIES SOLD

Weirdos in the Workplace

Weirdos in the workplace. Every year, there are more of them and most people haven't a clue about working with them effectively. Some managers enforce conformity at all costs and lose the brilliant eccentrics who can deliver spectacular results. Others tolerate everything and alienate their most solid performers. This book offers a better path. HR consultant John Putzier shows how to determine which "weird" behavior should be harnessed for the greater good, and how to lead high-performance weirdos to greatness while weeding out those who do not add value.

ISBN 0131478990, © 2005, 224 pp., \$17.95

The Truth About Managing People

This is a management book that cuts through the soft opinion and conjecture books that have dominated the business shelves in recent years and shows what management researchers know actually works, or doesn't work, when it comes to managing people.

Contains over 60 proven "truths" that can transform how you manage people—and the results that are achieved.

ISBN 0131838474, © 2004, 240 pp., \$9.95

"John Putzier is as brilliant as he is irreverent. His musings not only make you LAUGH they make you THINK. This is a must-read for anyone in a leadership position or any leader wannabe, for that matter, in today's world of work."
—Keith J. Greene, SPHR, Director, Organizational Programs,
Society for Human Resource Management

Weirdos in the Workplace

THE NEW NORMAL...
THRIVING IN THE AGE OF THE INDIVIDUAL

John Putzier

Foreword by Libby Sartain, Chief People Officer, Yahoo! Inc.



For more information on our business titles, visit www.ft-ph.com