

LEARNING MADE EASY



2nd Edition

Business Statistics

for
dummies[®]

A Wiley Brand

Discover how statistics
and business go hand in hand

—
Grasp core principles and
methods of business statistics

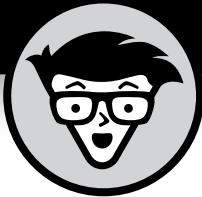
—
Learn how to use the
TI-84 Plus calculator



Alan Anderson, PhD

Business Statistics

for
dummies[®]
A Wiley Brand



Business Statistics

2nd Edition

By Alan Anderson, PhD

for
dummies[®]
A Wiley Brand

Business Statistics For Dummies®, 2nd Edition

Published by: John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, www.wiley.com

Copyright © 2024 by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Trademarks: Wiley, For Dummies, the Dummies Man logo, Dummies.com, Making Everything Easier, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: WHILE THE PUBLISHER AND AUTHORS HAVE USED THEIR BEST EFFORTS IN PREPARING THIS WORK, THEY MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES REPRESENTATIVES, WRITTEN SALES MATERIALS OR PROMOTIONAL STATEMENTS FOR THIS WORK. THE FACT THAT AN ORGANIZATION, WEBSITE, OR PRODUCT IS REFERRED TO IN THIS WORK AS A CITATION AND/OR POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE PUBLISHER AND AUTHORS ENDORSE THE INFORMATION OR SERVICES THE ORGANIZATION, WEBSITE, OR PRODUCT MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING PROFESSIONAL SERVICES. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR YOUR SITUATION. YOU SHOULD CONSULT WITH A SPECIALIST WHERE APPROPRIATE. FURTHER, READERS SHOULD BE AWARE THAT WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ. NEITHER THE PUBLISHER NOR AUTHORS SHALL BE LIABLE FOR ANY LOSS OF PROFIT OR ANY OTHER COMMERCIAL DAMAGES, INCLUDING BUT NOT LIMITED TO SPECIAL, INCIDENTAL, CONSEQUENTIAL, OR OTHER DAMAGES.

For general information on our other products and services, please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002. For technical support, please visit <https://hub.wiley.com/community/support/dummies>.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit www.wiley.com.

Library of Congress Control Number: 2023949253

ISBN 978-1-394-21992-6 (pbk); ISBN 978-1-394-21994-0 (ebk); 978-1-394-21993-3 (ebk)

Contents at a Glance

Introduction	1
Part 1: Getting Started with Business Statistics.....	5
CHAPTER 1: The Art and Science of Business Statistics	7
CHAPTER 2: Pictures Tell the Story: Graphical Representations of Data.....	21
CHAPTER 3: Identifying the Center of a Data Set.....	35
CHAPTER 4: Measuring Variation in a Data Set.....	53
CHAPTER 5: Measuring How Data Sets Are Related to Each Other	71
Part 2: Probability Theory and Probability Distributions.....	95
CHAPTER 6: Probability Theory: Measuring the Likelihood of Events	97
CHAPTER 7: Probability Distributions and Random Variables	115
CHAPTER 8: The Binomial and Poisson Distributions.....	127
CHAPTER 9: The Normal Distribution: So Many Possibilities!	145
CHAPTER 10: Sampling Techniques and Distributions.....	165
Part 3: Drawing Conclusions from Samples.....	185
CHAPTER 11: Confidence Intervals and the Student's t-Distribution	187
CHAPTER 12: Testing Hypotheses about the Population Mean	205
CHAPTER 13: Applications of the Chi-Square Distribution.....	245
CHAPTER 14: Applications of the F-Distribution	273
Part 4: More Advanced Techniques: Regression Analysis and Spreadsheet Modeling	287
CHAPTER 15: Simple Regression Analysis	289
CHAPTER 16: Key Statistical Techniques in Excel.....	317
Part 5: The Part of Tens	343
CHAPTER 17: Ten Common Errors That Arise in Statistical Analysis	345
CHAPTER 18: (Almost) Ten Key Categories of Formulas for Business Statistics.....	353
Index	363

Table of Contents

INTRODUCTION	1
About This Book.....	1
Foolish Assumptions.....	2
Icons Used in This Book	3
Beyond the Book.....	3
Where to Go from Here	3
PART 1: GETTING STARTED WITH BUSINESS STATISTICS.....	5
CHAPTER 1: The Art and Science of Business Statistics	7
Representing the Key Properties of Data.....	8
Analyzing data with graphs	8
Defining properties and relationships with numerical measures	11
Probability: The Foundation of All Statistical Analysis	13
Random variables	14
Probability distributions.....	15
Using Sampling Techniques and Sampling Distributions	17
Statistical Inference: Drawing Conclusions from Data.....	17
Confidence intervals	18
Hypothesis testing.....	18
Simple regression analysis.....	19
CHAPTER 2: Pictures Tell the Story: Graphical Representations of Data	21
Analyzing the Distribution of Data by Class or Category.....	22
Frequency distributions for quantitative data.....	23
Frequency distribution for qualitative values	27
Cumulative frequency distributions	28
Histograms: Getting a Picture of Frequency Distributions	29
Checking Out Other Useful Graphs	31
Line graphs: Showing the values of a data series.....	31
Pie charts: Showing the composition of a data set.....	32
Scatter plots: Showing the relationship between two variables..	33
CHAPTER 3: Identifying the Center of a Data Set	35
Looking at Methods for Finding the Mean.....	36
Arithmetic mean	36
Geometric mean	38
Weighted mean	40

Getting to the Middle of Things: The Median of a Data Set	42
Determining the Relationship Between the Mean and Median	44
Symmetrical	45
Negatively skewed.....	45
Positively skewed.....	46
Discovering the Mode: The Most Frequently Repeated Element.....	48
Computing the Mean, Median, and Mode with the TI-84 Plus Calculator	50
CHAPTER 4: Measuring Variation in a Data Set.....	53
Determining Variance and Standard Deviation	54
Finding the sample variance	54
Finding the sample standard deviation	55
Calculating population variance and standard deviation	59
Finding the Relative Position of Data	62
Percentiles: Dividing everything into hundredths	63
Quartiles: Dividing everything into fourths	64
Interquartile range: Identifying the middle 50 percent.....	66
Measuring Relative Variation.....	67
Coefficient of variation: The spread of a data set relative to the mean	67
Comparing the relative risks of two portfolios	68
Computing Measures of Dispersion with the TI-84 Plus Calculator	69
CHAPTER 5: Measuring How Data Sets Are Related to Each Other.....	71
Understanding Covariance and Correlation	72
Sample covariance and correlation coefficient.....	73
Population covariance and correlation coefficient.....	78
Comparing correlation and covariance	83
Interpreting the Correlation Coefficient.....	86
Showing the relationship between two variables.....	87
Application: Correlation and the benefits of diversification	89
Computing Covariance and Correlation with the TI-84 Plus Calculator	92
PART 2: PROBABILITY THEORY AND PROBABILITY DISTRIBUTIONS.....	95
CHAPTER 6: Probability Theory: Measuring the Likelihood of Events.....	97
Working with Sets	98
Membership.....	98
Subset	98

Union.....	99
Intersection	100
Complement.....	102
Betting on Uncertain Outcomes	103
The sample space: Everything that can happen.....	103
Event: One possible outcome	103
Computing probabilities of events	105
Looking at Types of Probabilities	106
Unconditional (marginal) probabilities: When events are independent.....	106
Joint probabilities: When two things happen at once	108
Conditional probabilities: When one event depends on another	108
Determining independence of events	109
Following the Rules: Computing Probabilities.....	110
Addition rule.....	110
Complement rule.....	112
Multiplication rule	113
CHAPTER 7: Probability Distributions and Random Variables	115
Defining the Role of the Random Variable	116
Assigning Probabilities to a Random Variable.....	119
Calculating the probability distribution	119
Visualizing a probability distribution with a histogram.....	121
Characterizing a Probability Distribution with Moments	121
Understanding the summation operator (Σ).....	122
Expected value.....	122
Variance and standard deviation	124
CHAPTER 8: The Binomial and Poisson Distributions	127
Looking at Two Possibilities with the Binomial Distribution.....	128
Checking out the binomial distribution	128
Computing binomial probabilities	129
Moments of the binomial distribution.....	134
Graphing the binomial distribution	135
Keeping the Time: The Poisson Distribution.....	137
Computing Poisson probabilities	138
Graphing the Poisson distribution	140
Computing Binomial and Poisson Probabilities with the TI-84 Plus Calculator.....	141
Computing binomial probabilities	141
Computing Poisson probabilities	142

CHAPTER 9: The Normal Distribution: So Many Possibilities!	145
Comparing Discrete and Continuous Distributions	146
Understanding the Normal Distribution	148
Graphing the normal distribution	149
Getting to know the standard normal distribution	151
Computing standard normal probabilities	152
Computing normal probabilities other than standard normal..	159
Computing Probabilities for the Normal Distribution with the TI-84 Plus Calculator	162
CHAPTER 10: Sampling Techniques and Distributions	165
Sampling Techniques: Choosing Data from a Population.....	166
Probability sampling	167
Nonprobability sampling	172
Sampling Distributions	174
Portraying sampling distributions graphically.....	175
Moments of a sampling distribution	177
The Central Limit Theorem	178
Converting \bar{X} to a standard normal random variable	179
PART 3: DRAWING CONCLUSIONS FROM SAMPLES	185
CHAPTER 11: Confidence Intervals and the Student's t-Distribution	187
Almost Normal: The Student's t-Distribution	188
Properties of the t-distribution	188
Degrees of freedom	189
Moments of the t-distribution	189
Graphing the t-Distribution	191
Probabilities and the t-Table	193
Point Estimates vs. Interval Estimates	194
Estimating Confidence Intervals for the Population Mean	195
Known population standard deviation.....	196
Unknown population standard deviation	199
Computing Confidence Intervals for the Population Mean with the TI-84 Plus Calculator	201
Population standard deviation is known	202
Population standard deviation is unknown.....	203

CHAPTER 12: Testing Hypotheses about the Population Mean	205
Applying the Key Steps in Hypothesis Testing for a Single Population Mean	206
Writing the null hypothesis	206
Coming up with an alternative hypothesis	206
Choosing a level of significance.....	209
Computing the test statistic.....	211
Comparing the critical value(s)	212
Using the decision rule	220
Testing Hypotheses About Two Population Means	223
Writing the null hypothesis for two population means.....	224
Defining the alternative hypotheses for two population means.....	224
Determining the test statistics for two population means	225
Testing Hypotheses about Population Means with the TI-84 Plus Calculator	235
Single population mean	235
Two population means.....	239
CHAPTER 13: Applications of the Chi-Square Distribution	245
Staying Positive with the Chi-Square Distribution	246
Representing the chi-square distribution graphically	247
Defining a chi-square random variable	248
Checking out the moments of the chi-square distribution.....	249
Testing Hypotheses about the Population Variance	250
Defining what you assume to be true: The null hypothesis	250
Stating the alternative hypothesis	251
Choosing the level of significance.....	253
Calculating the test statistic.....	253
Determining the critical value(s)	254
Practicing the Goodness of Fit Tests.....	258
Comparing a population to the Poisson distribution.....	259
Comparing a population to the normal distribution	265
Conducting a Goodness of Fit Test with the TI-84 Plus Calculator	270
CHAPTER 14: Applications of the F-Distribution	273
Getting to Know the F-Distribution.....	273
Defining an F random variable	275
Measuring the moments of the F-distribution	276
Testing Hypotheses about the Equality of Two Population Variances	278
The null hypothesis: Equal variances	279
The alternative hypothesis: Unequal variances	279

The test statistic.....	280
The critical value(s)	280
The decision about the equality of two population variances ..	282
Testing Hypotheses about Two Population Variances with the TI-84 Plus Calculator	283
PART 4: MORE ADVANCED TECHNIQUES: REGRESSION ANALYSIS AND SPREADSHEET MODELING.....	287
CHAPTER 15: Simple Regression Analysis.....	289
The Fundamental Assumption: Variables Have a Linear Relationship	290
Defining a linear relationship	291
Using scatter plots to identify linear relationships.....	292
Defining the Population Regression Equation	295
Estimating the Population Regression Equation.....	297
Testing the Estimated Regression Equation	303
Using the coefficient of determination (R^2).....	303
Computing the coefficient of determination.....	305
The t-test.....	306
Using Statistical Software.....	311
Assumptions of Simple Linear Regression	313
Conducting Simple Regression Analysis with the TI-84 Plus Calculator	314
CHAPTER 16: Key Statistical Techniques in Excel.....	317
Implementing Excel Functions.....	317
Checking Out Excel's Key Statistical Functions	318
Measures of central tendency.....	319
Measures of dispersion	321
Measures of association.....	322
Discrete probability distributions	324
Continuous probability distributions.....	326
Confidence intervals	331
Regression analysis.....	333
Going Deeper with the Analysis ToolPak.....	334
Computing covariance and correlation	335
Computing descriptive statistics	337
Regression analysis.....	338
Hypothesis testing.....	340

PART 5: THE PART OF TENS.....	343
CHAPTER 17: Ten Common Errors That Arise in Statistical Analysis.....	345
Designing Misleading Graphs	346
Drawing the Wrong Conclusion from a Confidence Interval	347
Misinterpreting the Results of a Hypothesis Test.....	348
Placing Too Much Confidence in the Coefficient of Determination (R^2)	348
Assuming Normality	349
Thinking Correlation Implies Causality.....	349
Drawing Conclusions from a Regression Equation When the Data Do Not Follow the Assumptions	350
Using Regression Analysis to Make Predictions About Values Outside the Range of Sample Data	350
Placing Too Much Confidence in Forecasts.....	351
Using the Wrong Distribution	351
CHAPTER 18: (Almost) Ten Key Categories of Formulas for Business Statistics.....	353
Summary Measures of a Population or a Sample	353
Probability	355
Discrete Probability Distributions.....	356
Continuous Probability Distributions.....	357
Sampling Distributions.....	357
Confidence Intervals for the Population Mean.....	358
Testing Hypotheses about Population Means	359
Testing Hypotheses about Population Variances.....	361
Using Regression Analysis	362
INDEX.....	363

Introduction

H ave you always been scared to death of statistics? You and just about everyone else! The equations are extremely intimidating, and the terminology sounds so . . . boring.

Why, then, is statistics so important? All business disciplines can be analyzed with statistical principles. Statistics make it possible to analyze real-world problems with actual data so that we can understand if our marketing strategy is really working, or how much a company should charge for its products, or any of a million other practical business questions.

Without a formal framework for analyzing these types of situations, it would be impossible to have any confidence in our results. This is where the science of statistics comes in. Far from being an overbearing collection of equations, it is a logical framework for analyzing practical business problems with real-world data.

This book is designed to show you how to apply statistics to practical situations in a step-by-step manner so that by the time you're done, you'll know as much about statistics as people with far more education in this area!

About This Book

All business degrees require at least some statistics courses, and there's a good reason for that! All business disciplines are empirical by nature, meaning that they need to analyze actual data to be successful. The purpose of this book is to:

- » Give you the principles on which statistical analysis is based
- » Provide you with many worked-out examples of these principles so that you can master them
- » Improve your understanding of the circumstances in which each statistical technique should be used

As a *For Dummies* title, this book is organized into modules; you can skip around and learn about various statistical techniques in the order that suits you. In cases where the contents of a chapter are based on previous readings, you are guided

back to the original material. Along the way are many helpful tips and reminders so that you get the most out of each chapter. I explain each equation in great detail, and all key terms are explained in depth.

In this updated Second Edition, I show you how to use the Texas Instruments TI-84 Plus and TI-84 Plus CE calculators to obtain results quickly and easily for just about every problem you will encounter in this book. I also added a new chapter that shows how easily statistical problems can be solved using Microsoft Excel.

This book can't make you an expert in statistics, but it provides you with a way of improving your knowledge very quickly so that you can use statistics in practical settings right away.

Foolish Assumptions

I am willing to make the following assumptions about you as the reader of this book:

- » You need to use the techniques in this book in a practical setting and have little or no previous experience with statistics.

OR

- » You're a student who feels overwhelmed by a traditional statistics course and feels the need for more background. You can benefit from seeing more examples of the material; statistics is a science that can be learned through practice!

OR

- » You're simply interested in improving your knowledge of this field.

In all of these cases, you're extremely well motivated and can put as much effort into learning statistics as you need. Congratulations! Your reward for reading this book will be a greater understanding of business statistics.

Icons Used in This Book

The following icons are designed to help you use this book quickly and easily. Be sure to keep an eye out for them.



REMEMBER



TIP



WARNING



TECHNICAL STUFF

The Remember icon points to information that's especially important to remember for exam purposes.

The Tip icon presents information like a memory acronym or some other aid to understanding or remembering material.

When you see this icon, pay special attention. The information that follows may be somewhat difficult, confusing, or harmful.

The Technical Stuff icon is used to indicate detailed information; for some people, it might not be necessary to read or understand.

Beyond the Book

In addition to the informative, clever, and (if I may say so) well-written material you're reading right now, this product also comes with some access-anywhere help and information online at Dummies.com. No matter how well you know statistics by the end of this book, a little extra information is always helpful. Check out this book's online Cheat Sheet to learn more about describing populations and samples, random variables, probability distributions, hypothesis testing, and more. Just go to www.dummies.com and search for "Business Statistics For Dummies Cheat Sheet."

Where to Go from Here

When you've become more adept at statistical analysis, you may want to tackle a full-blown statistical package, such as SPSS or SAS. These will eliminate a great deal of the computational burden, freeing you to concentrate on the analysis of the

results. You may also be interested in obtaining further education in this area. For example, you may want to pursue a graduate degree, such as an MBA (Master of Business Administration). This is an extremely important credential that can open a large number of doors in the business world. You'll need your statistical skills in order to earn this degree, since it is heavily used throughout the curriculum.

If you're not ready for graduate school, you may simply want to explore some college-level statistics courses at your local university. The most important thing is to continue using your statistical skills, as you'll only become adept at using them through constant practice.

1 **Getting Started with Business Statistics**

IN THIS PART . . .

Use histograms to provide a visual representation of the distribution of elements in a data set. A histogram can show which values occur most frequently, the smallest and largest values, and how spread out these values are.

Create graphs that reflect non-numerical data, such as colors, flavors, brand names, and so on. Graphs are used where numerical measures are difficult or impossible to compute.

Identify the center of a data set by using the mean (the average), median (the middle), and mode (the most commonly occurring value). These are known as the measures of central tendency.

Use formulas for computing covariance and correlation for both samples and populations; a scatter plot is used to show the relationship (if there is one) between two variables.

IN THIS CHAPTER

- » Looking at the key properties of data
- » Understanding probability's role in business
- » Sampling distributions
- » Drawing conclusions based on results

Chapter **1**

The Art and Science of Business Statistics

Statistical analysis is widely used in all business disciplines. For example, marketing researchers analyze consumer spending patterns to properly plan new advertising campaigns. Organizations use management consulting to determine how efficiently resources are being used. Manufacturers use quality control methods to ensure the consistency of the products they are producing. These types of business applications and many others are heavily based on statistical analysis.

Financial institutions use statistics for a wide variety of applications. For example, a pension fund may use statistics to identify the types of securities that it should hold in its investment portfolio. A hedge fund may use statistics to identify profitable trading opportunities. An investment bank may forecast the future state of the economy to determine which new assets it should hold in its own portfolio.

Whereas statistics is a quantitative discipline, the ultimate objective of statistical analysis is to explain real-world events. This means that in addition to the rigorous application of statistical methods, there is always a great deal of room for judgment. As a result, you can think of statistical analysis as both a science and an art; the art comes from choosing the appropriate statistical technique for a given situation and correctly interpreting the results.

In this chapter, I provide a brief introduction to the concepts that are covered throughout the book. I introduce several important techniques that help you to measure and analyze the statistical properties of real-world variables, such as stock prices, interest rates, corporate profits, and so on.

Representing the Key Properties of Data

The word *data* refers to a collection of *quantitative* (numerical) or *qualitative* (non-numerical) values. Quantitative data may consist of prices, profits, sales, or any variable that can be measured on a numerical scale. Qualitative data may consist of colors, brand names, geographic locations, and so on. Most of the data encountered in business applications are quantitative.



TECHNICAL
STUFF

The word *data* is actually the plural of *datum*; *datum* refers to a single value, while *data* refers to a collection of values.

You can analyze data with graphical techniques or numerical measures. I explore both options in the following sections.

Analyzing data with graphs

Graphs are a visual representation of a data set, making it easy to see patterns and other details. Deciding which type of graph to use depends on the type of data you're trying to analyze. Here are some of the more common types of graphs used in business statistics:

- » **Histograms:** A histogram shows the distribution of data among different intervals or categories, using a series of vertical bars.
- » **Line graphs:** A line graph shows how a variable changes over time.
- » **Pie charts:** A pie chart shows how data is distributed between different categories, illustrated as a series of slices taken from a pie.
- » **Scatter plots (scatter diagrams):** A scatter plot shows the relationship between two variables as a series of points. The pattern of the points indicates how closely related the two variables are.

Histograms

You can use a histogram with either quantitative or qualitative data. It's designed to show how a variable is distributed among different categories. For example,

suppose a marketing firm surveys 100 consumers to determine their favorite color. The responses are

Red:	23
Blue:	44
Yellow:	12
Green:	21

The results can be illustrated with a histogram, with each color in a single category. The heights of the bars indicate the number of responses for each color, making it easy to see which colors are the most popular (see Figure 1-1).

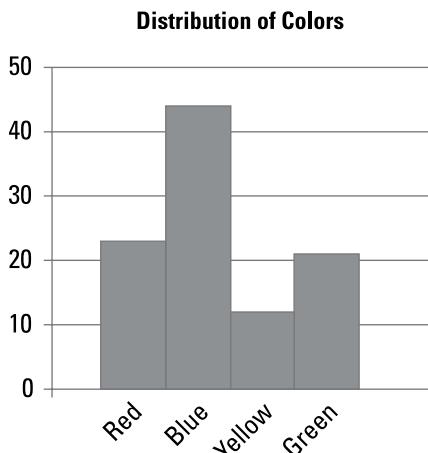


FIGURE 1-1:
A histogram for
preferred colors.

Based on the histogram, you can see at a glance that blue is the most popular choice, while yellow is the least popular choice.

Line graphs

You can use a line graph with quantitative data. It shows the values of a variable over a given interval of time. For example, Figure 1-2 shows the daily price of gold between August 1, 2023 and September 29, 2023.

With a line graph, it's easy to see trends or patterns in a data set. These types of graphs may be used by investors to identify which assets are likely to rise in the future based on their past performance.

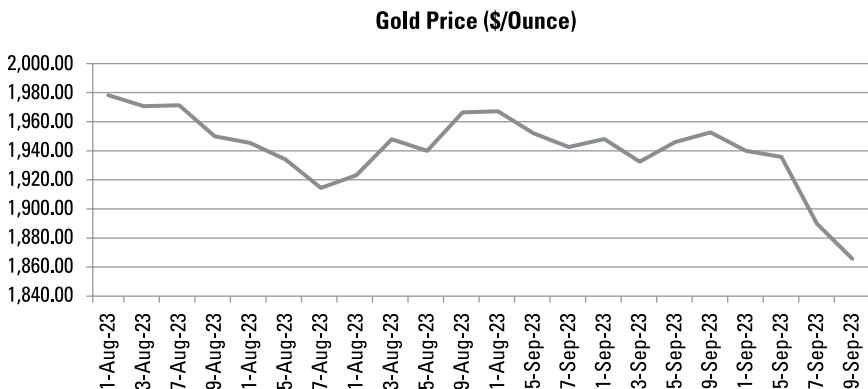


FIGURE 1-2:
A line graph of
gold prices.

Pie charts

Use a pie chart with quantitative or qualitative data to show the distribution of the data among different categories. For example, suppose that a chain of coffee shops wants to analyze its sales by coffee style. The styles that the chain sells are French Roast, Breakfast Blend, Brazilian Rainforest, Jamaica Blue Mountain, and Espresso. Figure 1-3 shows the proportion of sales for each style.

Distribution of Sales by Style

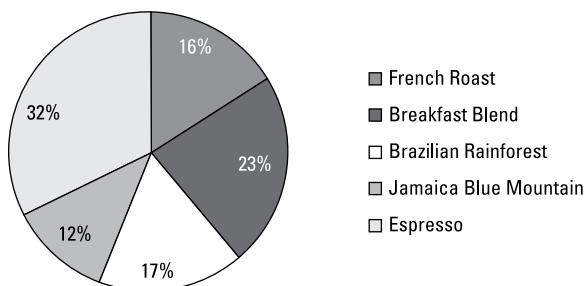


FIGURE 1-3:
A pie chart for
coffee sales.

The chart shows that Espresso is the chain's best-selling style, while Jamaica Blue Mountain accounts for the smallest percentage of the chain's sales.

Scatter plots

A scatter plot is designed to show the relationship between two quantitative variables. For example, Figure 1-4 shows the relationship between a corporation's sales and profits over the past 20 years.

Each point on the scatter plot represents profit and sales for a single year. The pattern of the points shows that higher levels of sales tend to be matched by higher levels of profits, and vice versa. This is called a positive relationship between the two variables.

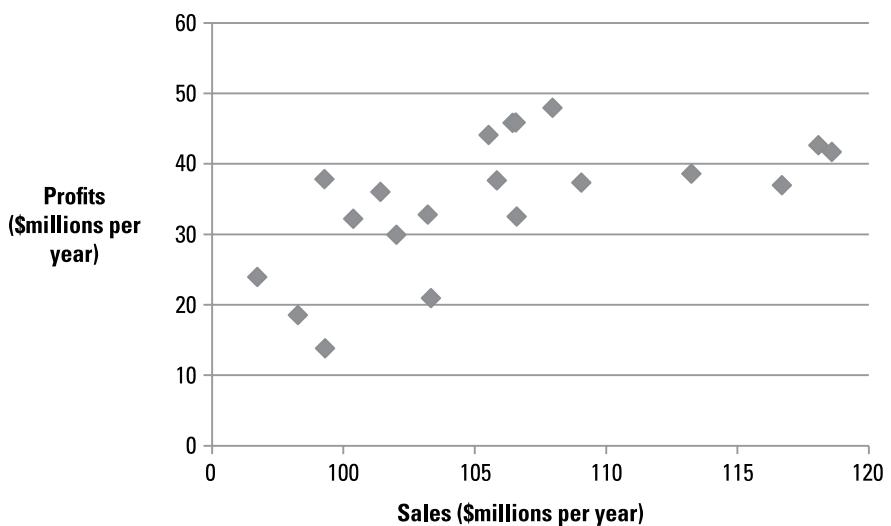


FIGURE 1-4:
A scatter plot
showing sales
and profits.

Defining properties and relationships with numerical measures

A *numerical measure* is a value that describes a key property of a data set. For example, to determine whether the residents of one city tend to be older than the residents in another city, you can compute and compare the average or *mean* age of the residents of each city. Some of the most important properties of interest in a data set are the *center* of the data and the *spread* among the observations.

Finding the center of the data

To identify the center of a data set, you use measures that are known as *measures of central tendency*; the most important of these are the mean, median, and mode.

The *mean* represents the average value in a data set, while the *median* represents the midpoint. The median is a value that separates the data into two halves; half of the elements in the data set are *less than or equal* to the median, and the remaining half are *greater than or equal* to the median. The *mode* is the most commonly occurring value in the data set.

The mean is the most widely used measure of central tendency, but it can give deceptive results if the data contain any unusually large or small values, known as *outliers*. In this case, the median provides a more representative measure of the center of the data. For example, median household income is usually reported by government agencies instead of mean household income. This is because mean household income is inflated by the presence of a small number of extremely wealthy households. As a result, median household income is thought to be a better measure of how standards of living are changing over time.

The mode can be used for either quantitative or qualitative data. For example, it may be used to determine the most common number of years of education among the employees of a firm. It may also be used to determine the most popular flavor sold by a soft drink manufacturer.

Measuring the spread of the data

Measures of dispersion identify how spread out a data set is, relative to the center. This provides a way of determining if the members of a data set tend to be very close to each other or if they tend to be widely scattered. Some of the most important measures of dispersion are

- » Variance
- » Standard deviation
- » Percentiles
- » Quartiles
- » Interquartile range (IQR)

The *variance* is a measure of the average squared difference between the elements of a data set and the mean. The larger the variance, the more “spread out” the data is. Variance is often used as a measure of risk in business applications; for example, it can be used to show how much uncertainty there is over the returns on a stock.

The *standard deviation* is the square root of the variance, and is more commonly used than the variance (because the variance is expressed in squared units). For example, the variance of a series of gas prices is measured in squared dollars, which is difficult to interpret. The corresponding standard deviation is measured in dollars, which is much more intuitively clear.

Percentiles divide a data set into 100 equal parts, each consisting of 1 percent of the total. For example, if a student’s score on a standardized exam is in the 80th percentile, then the student scored as well as or better than 80 percent of the other

students who took the exam. A *quartile* is a special type of percentile; it divides a data set into four equal parts, each consisting of 25 percent of the total. The first quartile is the 25th percentile of a data set, the second quartile is the 50th percentile, and the third quartile is the 75th percentile. The *interquartile range* identifies the middle 50 percent of the observations in a data set; it equals the difference between the third and the first quartiles.

Determining the relationship between two variables

For some applications, you need to understand the relationship between two variables. For example, if an investor wants to understand the risk of a portfolio of stocks, it's essential to properly measure how closely the returns on the stocks track each other. You can determine the relationship between two variables with two measures of association: covariance and correlation.

Covariance is used to measure the tendency for two variables to rise above their means or fall below their means at the same time. For example, suppose that a bioengineering company finds that increasing research and development expenditures typically leads to an increase in the development of new patents. In this case, R&D spending and new patents would have a positive covariance. If the same company finds that rising labor costs typically reduce corporate profits, then labor costs and profits would have a negative covariance. If the company finds that profits are not related to the average daily temperature, then these two variables will have a covariance that is very close to zero.

Correlation is a closely related measure. It's defined as a value between -1 and 1 , so interpreting the correlation is easier than the covariance. For example, a correlation of 0.9 between two variables would indicate a very strong positive relationship, whereas a correlation of 0.2 would indicate a fairly weak but positive relationship. A correlation of -0.8 would indicate a very strong negative relationship; a correlation of -0.3 would indicate a weak negative relationship. A correlation of 0 would show that two variables have no linear relationship between them.

Probability: The Foundation of All Statistical Analysis

Probability theory provides a mathematical framework for measuring uncertainty. This area is important for business applications since much statistical analysis is strongly related to probability theory. Understanding probability theory provides fundamental insights into all the statistical methods used in this book.

Probability is heavily based on the notion of *sets*. A set is a collection of objects. These objects may be numbers, colors, flavors, and so on. This chapter focuses on sets of numbers that may represent prices, rates of return, and so forth. Several mathematical operations may be applied to sets — union, intersection, and complement, for example.

The union of two sets is a new set that contains all the elements in the original two sets. The intersection of two sets is a set that contains only the elements contained in *both* of the two original sets (if any). The complement of a set is a set containing elements that are *not* in the original set. For example, the complement of the set of black cards in a standard deck is the set containing all red cards.

Probability theory is based on a model of how random outcomes are generated, known as a *random experiment*. Outcomes are generated in such a way that all *possible* outcomes are known in advance, but the *actual* outcome isn't known. The following rules help you determine the probability of specific outcomes occurring:

- » The addition rule
- » The multiplication rule
- » The complement rule

You use the addition rule to determine the probability of a union of two sets. The multiplication rule is used to determine the probability of an intersection of two sets. The complement rule is used to identify the probability that the outcome of a random experiment will *not* be an element in a specified set.

Random variables

A *random variable* assigns numerical values to the outcomes of a random experiment. For example, when you flip a coin twice, you're performing a random experiment because

- » All possible outcomes are known in advance.
- » The actual outcome isn't known in advance.

The experiment consists of two *trials*. On each trial, the outcome must be a “head” or a “tail.”

Assume that a random variable X is defined as the number of “heads” that turn up during the course of this experiment. X assigns values to the outcomes of this experiment as follows:

Outcome	X
{TT}	0
{HT, TH}	1
{HH}	2

where:

T represents a tail on a single flip

H represents a head on a single flip

TT represents two consecutive tails

HT represents a head followed by a tail

TH represents a tail followed by a head

HH represents two consecutive heads

X assigns a value of 0 to the outcome TT because no heads turned up. X assigns a value of 1 to both HT and TH because one head turned up in each case. Similarly, X assigns a value of 2 to HH because two heads turned up.

Probability distributions

A *probability distribution* is a formula or a table used to assign probabilities to each possible value of a random variable X. A probability distribution may be *discrete*, which means that X can assume one of a finite (countable) number of values, or *continuous*, in which case X can assume one of an infinite (uncountable) number of different values.

For the coin-flipping experiment from the previous section, the probability distribution of X can be a simple table that shows the probability of each possible value of X, written as P(X):

X	P(X)
0	0.25
1	0.50
2	0.25

The probability that $X = 0$ (that no heads turn up) equals 0.25 because this experiment has four equally likely outcomes: HH, HT, TH, and TT and in only one of those cases will there be no heads. You compute the other probabilities in a similar manner.

Discrete probability distributions

Several specialized discrete probability distributions are useful for specific applications. For business applications, two frequently used discrete distributions are:

- » Binomial
- » Poisson

You use the *binomial distribution* to compute probabilities for a process where only one of two possible outcomes may occur on each trial. You can use the *Poisson distribution* to measure the probability that a given number of events will occur during a given time frame.

Continuous probability distributions

Many continuous distributions may be used for business applications; one of the most widely used is the normal distribution. The *normal distribution* is useful for a wide array of applications in many disciplines. In business applications, variables such as stock returns are often assumed to follow the normal distribution. The normal distribution is characterized by a *bell-shaped curve*, and areas under this curve represent probabilities. The bell-shaped curve is shown in Figure 1-5.

The Normal Distribution

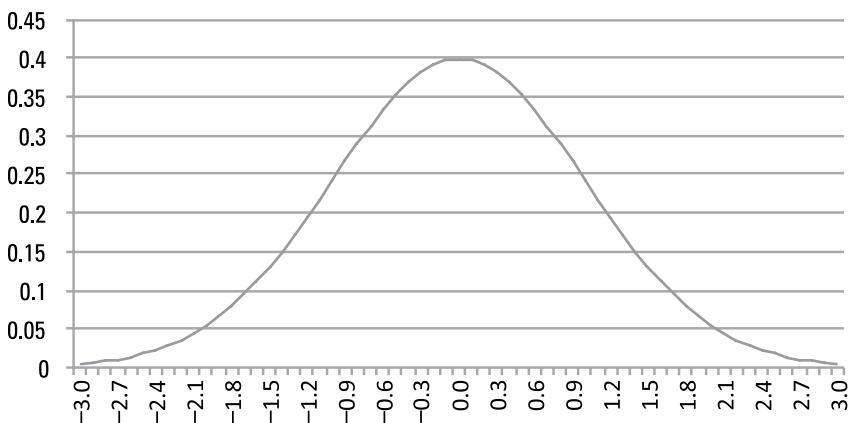


FIGURE 1-5:
The bell-shaped
curve of the
normal
distribution.

The normal distribution has many convenient statistical properties that make it a popular choice for statistical modeling. One of these properties is known as *symmetry*, the idea that the probabilities of values below the mean are matched by the probabilities of values that are equally far above the mean.

Using Sampling Techniques and Sampling Distributions

Sampling is used to determine the estimated properties of a population from sample data. A *population* is a collection of data that someone has an interest in studying. A sample is a selection of data randomly chosen from a population. For example, if a university is interested in analyzing the distribution of grade point averages (GPAs) among its MBA students, the population of interest would be the GPAs of every MBA student at the university; a sample would consist of the GPAs of a set of randomly chosen MBA students.

Several approaches can be used for choosing samples; a sample is a *subset* of the underlying population.

A *statistic* is a summary measure of a sample, while a *parameter* is a summary measure of a population. The properties of a statistic can be determined with a *sampling distribution* — a special type of probability distribution that describes the properties of a statistic.

The *central limit theorem* (CLT) gives the conditions under which the mean of a sample follows the normal distribution:

- » The underlying population is normally distributed.
- » The sample size is “large” (at least 30).

Statistical Inference: Drawing Conclusions from Data

Statistical inference refers to the process of drawing conclusions about a population from randomly chosen samples. In the following sections, I discuss two techniques used for statistical inference: confidence intervals and hypothesis testing.

Confidence intervals

A *confidence interval* is a set of values that's expected to contain the value of a population parameter with a specified level of confidence (such as 90 percent, 95 percent, 99 percent, and so on). For example, you can construct a confidence interval for the population mean by following these steps:

1. Estimate the value of the population mean by calculating the mean of a randomly chosen sample (known as the sample mean).
2. Calculate the lower limit of the confidence interval by subtracting a margin of error from the sample mean.
3. Calculate the upper limit of the confidence interval by adding the same margin of error to the sample mean.

The margin of error depends on the size of the sample used to construct the confidence interval, whether the population standard deviation is known, and the level of confidence chosen.

The resulting interval is known as a confidence interval. A confidence interval is constructed with a specified level of probability. For example, suppose you draw a sample of stocks from a portfolio, and you construct a 95 percent confidence interval for the mean return of the stocks in the entire portfolio:

$$(\text{lower limit}, \text{upper limit}) = (0.02, 0.08)$$

The returns on the entire portfolio are the population of interest. The mean return in each sample drawn is an *estimate* of the population mean. The sample mean will be slightly different each time a new sample is drawn, as will the confidence interval. If this process is repeated 100 times, 95 of the resulting confidence intervals will contain the true population mean.

Hypothesis testing

Hypothesis testing is a procedure for using sample data to draw conclusions about the characteristics of the underlying population.

The procedure begins with a statement, known as the *null hypothesis*. The null hypothesis is assumed to be true unless sufficient evidence against it is found. An *alternative hypothesis* — the result accepted if the null hypothesis is rejected — is also stated.

You calculate a *test statistic*, and you compare it with a *critical value* (or values) to determine whether the null hypothesis should be rejected. The specific test

statistic and critical value(s) depend on which population parameter is being tested, the size of the sample being used, and other factors.

If the test statistic is too extreme (for example, it's too large compared with the critical value[s]) the null hypothesis is rejected in favor of the alternative hypothesis; otherwise, the null hypothesis is *not* rejected.



TECHNICAL STUFF

If the null hypothesis isn't rejected, this doesn't necessarily mean that it's true; it simply means that there is not enough evidence to justify rejecting it.

Hypothesis testing is a general procedure and can be used to draw conclusions about many features of a population, such as its mean, variance, standard deviation, and so on.

Simple regression analysis

Regression analysis uses sample data to estimate the strength and direction of the relationship between two or more variables. *Simple regression analysis* estimates the relationship between a dependent variable (Y) and a single independent variable (X).

For example, suppose you're interested in analyzing the relationship between the annual returns of the Standard & Poor's (S&P) 500 Index and the annual returns of Apple stock. You can assume that the returns of Apple stock are related to the returns to the S&P 500 because the index is a reflection of the overall strength of the economy. The returns of Apple stock may be treated as the dependent variable (Y) and the returns of the S&P 500 may be treated as the independent variable (X). You can use regression analysis to measure the numerical relationship between the S&P 500 and Apple stock.

Simple regression analysis is based on the assumption that a linear relationship occurs between X and Y . A linear relationship takes this form:

$$Y = mX + b$$

Y is the dependent variable, X is the independent variable, m is the slope, and b is the intercept. The slope tells you how much Y changes due to a specific change in X ; the intercept tells you what the value of Y would be if X had a value of zero.

The goal of regression analysis is to find a line that best fits or explains the data. The population regression line is written as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

In this equation, Y_i is the dependent variable, X_i is the independent variable, β_0 is the intercept, β_1 is the slope, and ϵ_i is an error term.

A *sample* regression line, estimated from the data, is written as follows:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Here, \hat{Y}_i is the estimated value of Y_p , $\hat{\beta}_0$ is the estimated value of β_0 , $\hat{\beta}_1$ is the estimated value of β_1 and X_i is the independent variable.

The sample regression line shows the estimated relationship between Y and X; you can use this relationship to determine how Y responds to a given change in X. You can also use it to *forecast* future values of Y based on assumed values of X.

After estimating the sample regression line, the results are subjected to a series of tests to determine whether the equation is valid. If the equation isn't valid, you reject the results and try a new model.

IN THIS CHAPTER

- » Describing the properties of data with a frequency distribution
- » Illustrating frequency distributions with histograms
- » Tracking trends with line graphs, pie charts, and scatter plots

Chapter 2

Pictures Tell the Story: Graphical Representations of Data

Much of statistical analysis is based on numerical techniques, such as confidence intervals, hypothesis testing, regression analysis, and so on. In many cases, these techniques are based on assumptions about the data being used. One way to determine whether the data conform to these assumptions is to analyze a graph of the data, as a graph can provide many insights into the properties of a data set. For example, a graph may be used to show

- » How frequently a value occurs in a data set.
- » The average value of the elements in a data set.
- » Whether the elements in a data set are increasing or decreasing over time.
- » Whether the elements in two different data sets are related to each other.

Graphs are particularly useful for non-numerical data, such as colors, flavors, brand names, and more, where numerical measures are difficult or impossible to compute. In this chapter, I explain how to organize data in a convenient form so you can easily analyze it. I introduce charts and graphs — from histograms to line graphs to pie charts and scatter plots — that can help you visualize the most important properties of a data set.

Analyzing the Distribution of Data by Class or Category

To graph *quantitative* (numerical) data, you start by organizing the data into *classes* (also known as *intervals*). For example, suppose that the government is conducting a study that measures the salary ranges for employees in the software industry in the United States. Here's one possible set of classes:

- \$0 to \$24,999 per year
- \$25,000 to \$49,999 per year
- \$50,000 to \$74,999 per year
- \$75,000 to \$99,999 per year
- \$100,000 and more per year

By counting the number of employees that fall into each class, you can easily see how salaries are distributed in the software industry. If you make the data into a graph, you can then easily compare this information with salaries in other industries.

Qualitative (non-numerical) data may be organized into *categories*. For example, suppose that a marketing firm is studying the spending habits of consumers and wants to determine the most popular colors for a new line of watches. In this case, the colors are the relevant categories. What type of graph you use for analyzing a set of data depends on the type of data (quantitative or qualitative) you are using and the type of analysis you are performing. The following sections introduce several important types of graphs.

I also introduce the concept of a *frequency distribution*. This is a list of classes and the number of elements that belong to each class (known as *frequencies*). I cover the steps required to construct a frequency distribution, and I show two related types of distribution: relative frequency distribution and cumulative frequency distribution.

This section covers several widely used types of graphs, including histograms, pie charts, line graphs, and scatter plots. Histograms represent frequency distributions as a series of bars. Pie charts show what proportion of the elements of a data set belongs to various categories. A line graph shows how the value of a variable changes over time. Scatter plots are used to show the relationship between two variables.

Frequency distributions for quantitative data

Quantitative data consists of numerical values, such as prices, weights, distances, and so on. To graphically analyze quantitative data, you first have to organize them into a *frequency distribution* — a table that shows the number of observations that fall into each class within the data set.

For example, suppose that the following values represent the price of gasoline (dollars per gallon) at 20 randomly selected gas stations:

\$4.42	\$4.34
\$4.17	\$3.73
\$3.92	\$3.56
\$4.49	\$3.65
\$3.91	\$3.58
\$4.46	\$4.12
\$4.27	\$4.21
\$3.92	\$3.85
\$3.57	\$4.10
\$4.10	\$3.63

Now suppose that you organize the data into four classes, as follows:

\$3.50 to \$3.74

\$3.75 to \$3.99

\$4.00 to \$4.24

\$4.25 to \$4.49

Table 2-1 shows the frequency distribution for these.

TABLE 2-1

Frequency Distribution of Prices for 20 Gas Stations

Gas Prices (\$/Gallon)	Number of Gas Stations
\$3.50–\$3.74	6
\$3.75–\$3.99	4
\$4.00–\$4.24	5
\$4.25–\$4.49	5

Table 2-1 shows that the distribution of gas prices among these classes is very nearly equal. Seeing how the prices are distributed with a frequency distribution is much easier than inspecting the raw (original) data, which in this case is a list of 20 gas prices.

When you're constructing a frequency distribution, one of the most important considerations is the width of the classes. The class width equals the difference between the largest value that may be included in the class and the smallest. In Table 2-1, the class widths are \$0.25. Usually, the class widths will be equal.

Deciding how many classes to use depends on how much data you have and how detailed you need the results to be. For example, if the class width is too large, it can disguise the distribution of values within each class. If the class width is too small, then several classes may contain no elements or very few elements, which makes analyzing the results more cumbersome.



TIP

As a rule of thumb, the optimal number of classes in a frequency distribution is between 5 and 15.

Figuring the class width

In the gas station example, each class has a width of \$0.25. In general, you can determine the class width by subtracting the smallest value from the largest value and dividing by the total number of desired classes:

$$\text{Class width} = \frac{\text{Largest value in raw data} - \text{Smallest value in raw data}}{\text{Number of classes}}$$

Referring to the raw data (the list of 20 gas prices), you see that the largest price in the sample is \$4.49 and the smallest is \$3.56. To construct a frequency distribution with four classes, the width of each interval should be

$$\text{Class width} = \frac{\$4.49 - \$3.56}{4} = \frac{\$0.93}{4} = \$0.2325$$

So the class width is equal to approximately \$0.25. Although the class width could be kept at \$0.2325, using a width of \$0.25 is intuitively easier to follow (since prices can't be expressed in quarters of a cent).



TIP

- When you construct a frequency distribution, remember these key points:
- » The classes must not overlap. For example, if the frequency distribution refers to gasoline prices, it would be incorrect to have a class for \$1.00 to \$2.00 and another class for \$2.00 to \$3.00, because both contain \$2.00. It would be unclear which class contains prices of \$2.00.
 - » The classes must cover all elements in the data set being analyzed.
 - » Ideally, the classes should have equal widths; otherwise, analyzing the results is much more difficult.
 - » Class widths should ideally be "round" numbers, such as \$0.50, \$1.00, \$10.00, and so on, compared with numbers such as \$0.43, \$1.87, and \$2.15. These numbers are more difficult to grasp intuitively. For the gas station example, the widths are \$0.25, and this is preferable to \$0.2325, because \$0.2325 isn't a round number.

Observing relative frequency distributions

A frequency distribution shows the number of elements in a data set that belong to each class. In a relative frequency distribution, the value assigned to each class is the *proportion* of the total data set that belongs in the class. For example, suppose that a frequency distribution is based on a sample of 200 supermarkets. It turns out that 50 of these supermarkets charge a price between \$8.00 and \$8.99 for a pound of coffee. In a relative frequency distribution, the number assigned to this class would be 0.25 (50/200). In other words, that's 25 percent of the total.

Here's a handy formula for calculating the relative frequency of a class:

$$\frac{\text{class frequency}}{n}$$

Class frequency refers to the number of observations in each class; *n* represents the total number of observations in the entire data set. For the supermarket example in this section, the total number of observations is 200.

The relative frequency may be expressed as a proportion (fraction) of the total or as a percentage of the total. See Table 2-2, which gives both types of relative frequency based on the gas station data in Table 2-1.

TABLE 2-2**Relative Frequencies for Gas Station Prices**

Gas Prices (\$/Gallon)	Number of Gas Stations	Relative Frequency (fraction)	Relative Frequency (percent)
\$3.50–\$3.74	6	$6/20 = 0.30$	30%
\$3.75–\$3.99	4	$4/20 = 0.20$	20%
\$4.00–\$4.24	5	$5/20 = 0.25$	25%
\$4.25–\$4.49	5	$5/20 = 0.25$	25%

With a sample size of 20 gas stations, the relative frequency of each class equals the actual number of gas stations divided by 20. The result is then expressed as either a fraction or a percentage. For example, you calculate the relative frequency of prices between \$3.50 and \$3.74 as $6/20$ to get 0.30 (30 percent). Similarly, the relative frequency of prices between \$3.75 and \$3.99 equals $4/20 = 0.20 = 20$ percent.



TIP

One of the advantages of using a relative frequency distribution is that you can compare data sets that don't necessarily contain an equal number of observations. For example, suppose that a researcher is interested in comparing the distribution of gas prices in New York and Connecticut. Because New York has a much larger population, it also has many more gas stations. The researcher decides to choose 1 percent of the gas stations in New York and 1 percent of the gas stations in Connecticut for the sample. This turns out to be 800 in New York and 200 in Connecticut. The researcher puts together a frequency distribution as shown in Table 2-3.

TABLE 2-3**Frequency Distribution of Gas Prices in New York and Connecticut**

Price (\$/Gallon)	Number of New York Gas Stations	Number of Connecticut Gas Stations
\$3.00–\$3.49	210	48
\$3.50–\$3.99	420	96
\$4.00–\$4.49	170	56

Based on this frequency distribution, it's awkward to compare the distribution of prices in the two states. By converting this data into a relative frequency distribution, the comparison is greatly simplified, as seen in Table 2-4.

TABLE 2-4

Relative Frequency Distribution of Gas Prices in New York and Connecticut

Price (\$/Gallon)	New York Gas Stations	Relative Frequency	Connecticut Gas Stations	Relative Frequency
\$3.00-\$3.49	210	210/800 = 0.2625	48	48/200 = 0.2400
\$3.50-\$3.99	420	420/800 = 0.5250	96	96/200 = 0.4800
\$4.00-\$4.49	170	170/800 = 0.2125	56	56/200 = 0.2800

The results show that the distribution of gas prices in the two states is nearly identical. Roughly 25 percent of the gas stations in each state charge a price between \$3.00 and \$3.49; about 50 percent charge a price between \$3.50 and \$3.99; and about 25 percent charge a price between \$4.00 and \$4.49.

Frequency distribution for qualitative values

In this section, I use a qualitative data set to illustrate frequency distributions. Suppose that a data set consists of *qualitative* (non-numerical) values. In this example, consumers were asked to identify their favorite color on a survey. The 20 responses are listed as follows:

blue	blue	blue	black
black	black	black	black
white	blue	white	blue
red	red	red	red
silver	silver	black	white

In this case, the categories are colors. The frequency distribution of these data is as follows:

Color	Number of Responses
Black	6
Blue	5
Red	4
Silver	2
White	3

Table 2–5 shows the relative frequency distribution.

TABLE 2-5 Relative Frequency Distribution of Favorite Colors

Color	Number of Responses	Relative Frequency (fraction)	Relative Frequency
Black	6	$6/20 = 0.30$	30%
Blue	5	$5/20 = 0.25$	25%
Red	4	$4/20 = 0.20$	20%
Silver	2	$2/20 = 0.10$	10%
White	3	$3/20 = 0.15$	15%

You can see from the table that the most popular choice is black, and the least popular is silver.

Cumulative frequency distributions

Cumulative frequency refers to the total frequency of a given class and all prior classes. For example, Table 2–6 lists the cumulative frequencies for the gas station data from the earlier section “Frequency distributions for quantitative data.”

TABLE 2-6 Cumulative Frequency of Prices at 20 Gas Stations

Gas Prices (\$/Gallon)	Number of Gas Stations	Cumulative Frequency	Cumulative Frequency
\$3.50–\$3.74	6	6	30%
\$3.75–\$3.99	4	$6 + 4 = 10$	50%
\$4.00–\$4.24	5	$6 + 4 + 5 = 15$	75%
\$4.25–\$4.49	5	$6 + 4 + 5 + 5 = 20$	100%

To figure out the cumulative frequency of the \$3.75 to \$3.99 class, you add its class frequency (4) to the frequency of the previous class (\$3.50 to \$3.74, which is 6), so $6+4 = 10$. This result shows you that ten gas stations’ prices are between \$3.50 and \$3.99. Because 20 gas stations were used in the sample, the percentage of all gas stations with prices between \$3.50 and \$3.99 is $10/20$ or 50 percent of the total.

Therefore, it can easily be determined what percentage of the gas stations have prices less than or equal to a given value. Based on this table, the percentage of gas stations with prices of \$3.99 or less is 50 percent.

Histograms: Getting a Picture of Frequency Distributions

You can illustrate a frequency distribution, a relative frequency distribution, or a cumulative frequency with a special type of graph known as a histogram. (See the previous section, “Analyzing the Distribution of Data by Class or Category.”) With histograms, you list classes or categories on the horizontal axis and frequencies on the vertical axis. A bar represents each class or category.

A histogram’s job is to provide a visual of the distribution of elements in a data set. The histogram can show which values in a data set occur most frequently, the smallest and largest values in the data set, how “spread out” these values are, and so on. Figure 2-1 shows a histogram of the frequency distribution for the gas station prices from the previous section.

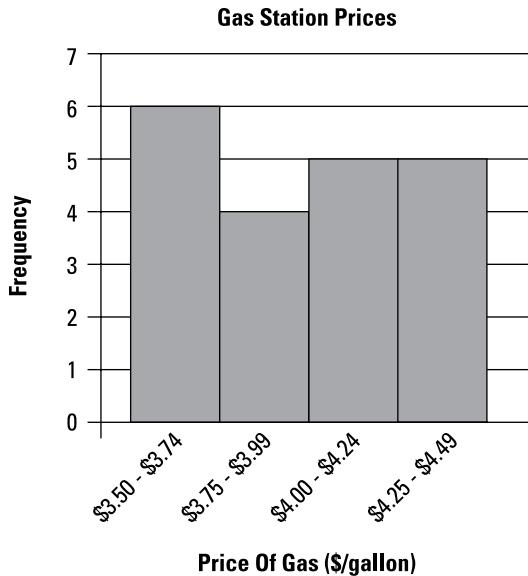


FIGURE 2-1:
Frequency distribution of gas station prices.

Figure 2–2 shows the relative frequency distribution.

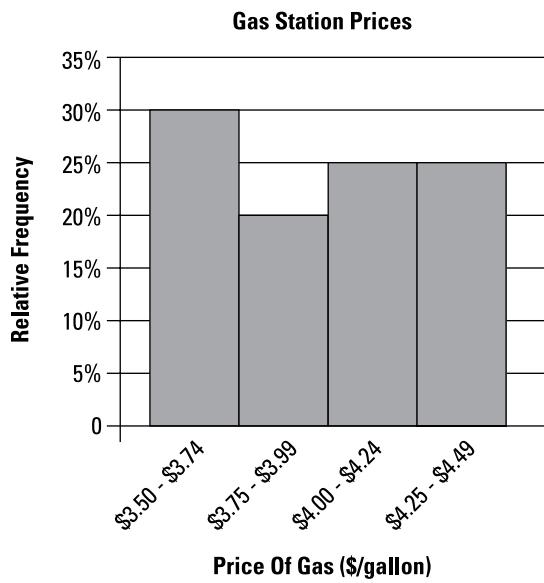


FIGURE 2-2:
Relative
frequency
distribution of
gas station prices.

Figure 2–3 shows the cumulative frequency distribution.

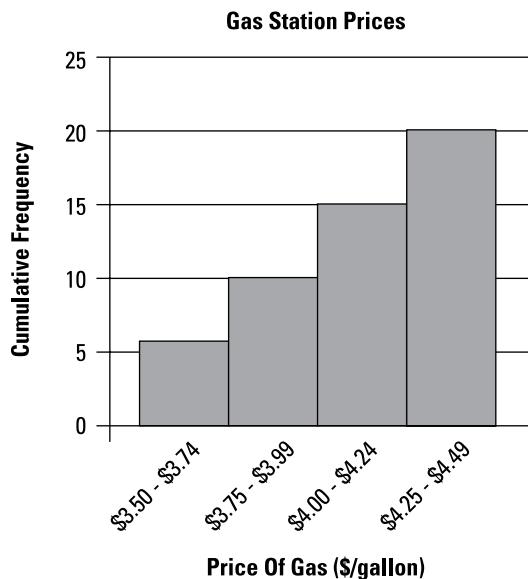


FIGURE 2-3:
Cumulative
frequency
distribution of
gas station prices.

Checking Out Other Useful Graphs

In addition to histograms, several other types of graphs can illustrate the properties of a data set. This section introduces you to some of the more common types of graphs you're likely to encounter and use.

Line graphs: Showing the values of a data series

A *line graph* is useful for showing how the value of a variable changes over time. With a line graph, the vertical axis represents the value of the variable, and the horizontal axis represents time. Each point on the graph represents the value of the variable at a single point in time, and a line connects the points. This line shows any trends in the data, such as whether the variable increases or decreases over time.

The following shows the price of gold (dollars per ounce) during the first six months of 2023:

Month	Gold Price (\$/Ounce)
January 2023	1,945.30
February 2023	1,836.70
March 2023	1,986.20
April 2023	1,999.10
May 2023	1,982.10
June 2023	1,929.40

The line chart shown in Figure 2–4 illustrates how the price of gold changed during this time period, based on the data shown in the table.

Using a line chart to detect patterns in the data is much easier than looking at the original data.



FIGURE 2-4:
Line graph showing how the price of gold fluctuated over six months' time.

Pie charts: Showing the composition of a data set

A *pie chart* is a circle graph that's divided into slices to represent the distribution of values in a data set. The area of each slice is proportional to the number of values in a given class or category. For example, suppose that a bank has 100 branches throughout the country; the following is the geographical distribution of these branches:

Branch Location	Number of Branches
Northeast	44
Northwest	32
Southeast	15
Southwest	9

The pie chart in Figure 2-5 illustrates these results.

The area of each slice in the pie chart indicates the proportional number of branches in each region. With this chart, you can easily see that the majority of the branches are in the northeast, with the fewest in the southwest.

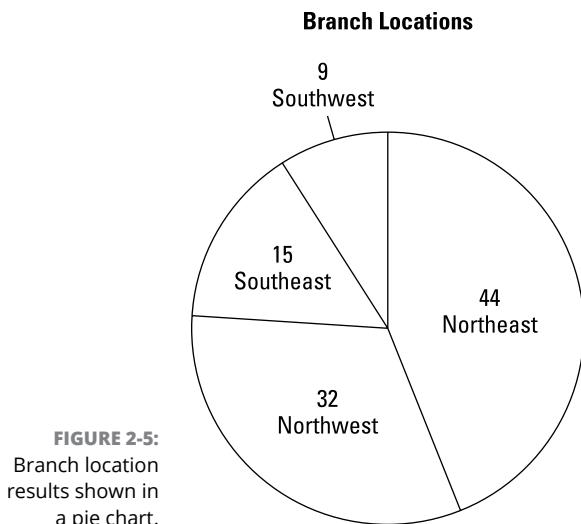


FIGURE 2-5:
Branch location
results shown in
a pie chart.

Scatter plots: Showing the relationship between two variables

A scatter plot (also known as a *scatter diagram*) shows the relationship between two quantitative (numerical) variables. These variables may be positively related, negatively related, or unrelated.

» **Positively related variables** indicate that

When one variable increases, the other variable tends to increase.

When one variable decreases, the other variable tends to decrease.

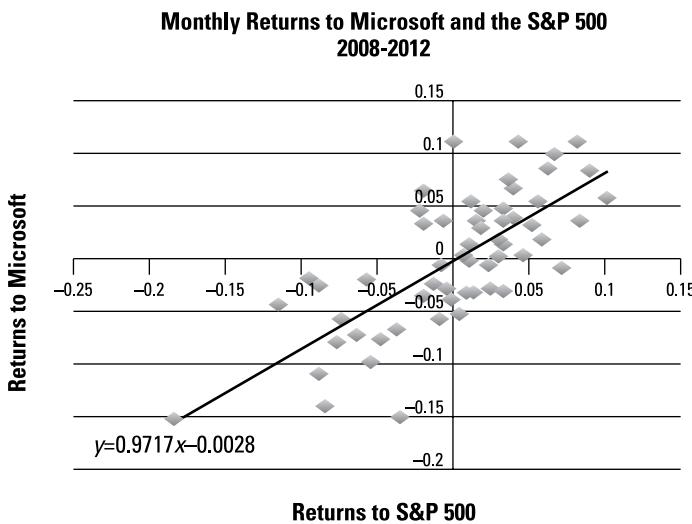
» **Negatively related variables** indicate that

When one variable increases or decreases, the other variable tends to do the opposite.

» **Unrelated variables** indicate that

No relationship is seen between the changes in the two variables.

The scatter diagram in Figure 2–6 shows the relationship between the monthly returns to Microsoft stock and the Standard & Poor’s (S&P) 500 Index from 2008 to 2012:



Each point on the graph represents the return to Microsoft stock and the return to the S&P 500 Index during a single month. The general direction of these points is from the lower-left corner of the graph to the upper-right corner, indicating that the two variables have a positive relationship.

The graph contains a *trend line*, which is a straight line designed to come as close as possible to all the points in the diagram. If two variables are positively related, the trend line has a positive slope; similarly, if two variables are negatively related, the trend line has a negative slope. If two variables are unrelated to each other, the trend line has a zero slope (that is, the trend line will be *flat*).

In the case of Microsoft and the S&P 500 Index, the equation of the trend line is

$$y = -0.0028 + 0.917x$$

In this equation, -0.0028 is the *intercept* (where the trend line crosses the vertical axis) and the *slope* is 0.917 (how much y changes due to a one-unit change in x).

Because the slope of the trend line is positive (0.917), the relationship between the returns to Microsoft stock and the S&P 500 Index is positive. The value of the slope also shows that each 1 percent increase in the returns to the S&P 500 Index occurs at the same time as an increase in the return to Microsoft by 0.917 percent, and that each 1 percent decrease in the returns to the S&P 500 Index occurs at the same time as a decrease in the return to Microsoft by 0.917 percent.

IN THIS CHAPTER

- » Computing the mean, median, and mode of a data set
- » Noting the specific characteristics of the mean, median, and mode

Chapter 3

Identifying the Center of a Data Set

The center of a data set (sample or population) provides useful information in many business applications. For example, it may be extremely important for a marketing firm to determine the average age of the customers who buy a specific product. Understanding the average household income of a company's customers would also be extremely useful in determining which types of new products to introduce. Portfolio managers at a pension fund are extremely interested in knowing the average rate of return of various stocks that they may be thinking about buying.

This chapter focuses on the techniques you use to find the center of a data set. There are several different ways to define the center: the average value, the middle value, the most frequently occurring value, and so on. Three of the most important measures of the center, formally known as *measures of central tendency*, are the mean, median, and mode.

The *mean* is the most commonly used measure of the center; it has the advantage of being easy to compute and interpret. In statistics, the word mean is used interchangeably with *average*. The median and mode are mainly used in situations where the mean is likely to give misleading results. This can happen if the data set contains any extremely large or small values, known as outliers.



TIP

An *outlier* is a value that's substantially different from the other elements in a data set. Outliers may have a dramatic impact on the accuracy of your calculations.

The *median* is the middle value of a data set (just like a median divides a highway into two equal halves). The *mode* is the most frequently occurring value in a data set. Each of these measures has its own unique set of advantages and disadvantages.

Looking at Methods for Finding the Mean

You can calculate the mean of a data set in several ways; the appropriate choice depends on the type of data and the application. This section explains how to find the three most common types of mean.

Arithmetic mean

The *arithmetic mean* is what most people think of when they hear the word *mean*. This type of mean is the easiest to calculate; it's the sum of the elements in a data set divided by the number of elements.

You use different formulas for computing the arithmetic mean for a *population* and a *sample*. A population is a collection of data that you're interested in studying; a sample is a selection chosen from a population. For example, if a government is interested in the distribution of household incomes, the population of interest would be the incomes of every household. A sample would be a set of incomes for households randomly chosen from the population.

Calculating the sample arithmetic mean

The formula for finding the sample arithmetic mean is

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

The key terms in this formula are:

- » \bar{X} (pronounced "X bar") = the sample mean
- » n = the number of elements in the sample
- » i = an *index*, which assigns a number to each sample element, ranging from 1 to n

- » X_i = a single element in the sample
- » Σ = the uppercase Greek letter sigma, known as the *summation operator*, which indicates that a sum is being computed

The summation operator is shorthand notation for adding a set of numbers. For example, if a data set contains five elements, the summation operator tells you to perform the following calculations:

$$\sum_{i=1}^5 X_i = X_1 + X_2 + X_3 + X_4 + X_5$$

Each of the Xs in this formula is *indexed* by a number ranging from 1 to n , where n is the size of the data set. In this example, n is 5.

Suppose that an investor wants to compute the arithmetic mean return of the stock of Omega Airlines, Inc. The investor takes a sample of annual returns — the period from 2019 to 2023.

Year	Omega Airlines Annual Return (percent)
2019	2
2020	-1
2021	3
2022	5
2023	1

To find the arithmetic mean, follow these steps:

1. Assign an index to each return in the sample:

$$X_1 = 2, X_2 = -1, X_3 = 3, X_4 = 5, X_5 = 1$$

Here, X_1 represents the return in 2019; X_2 is the return in 2020, and so on.

2. Compute the sum of the returns:

$$\sum_{i=1}^5 X_i = X_1 + X_2 + X_3 + X_4 + X_5 = 2 - 1 + 3 + 5 + 1 = 10$$

3. Divide the sum of the returns by the number of returns in the sample:

$$\bar{X} = \frac{\sum_{i=1}^5 X_i}{5} = \frac{10}{5} = 2$$

This result shows that the average return of this stock is 2 percent per year.

Calculating the population arithmetic mean

When you calculate the arithmetic mean of a population, the calculation is the same as for the arithmetic mean of a sample, but the notation is slightly different. Here's the formula for computing the arithmetic mean of a population:

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

The new term in this formula is μ , the lowercase Greek letter mu, which replaces \bar{X} from the sample arithmetic mean formula in the previous section. The μ represents the mean of a population. (Note that n is used to represent the size of a sample and N is used to represent the size of a population. Also, note that the actual calculations are the same for the sample mean and population mean.)



TIP

In statistics, it's common to use Greek letters to represent population measures and Latin letters (that is, the alphabet that you use every day) to represent sample measures.

Geometric mean

The main difference between arithmetic and geometric means is that the arithmetic mean is based on sums, while the geometric mean is based on products.

For the Omega Airlines example in the previous section, the arithmetic mean doesn't reflect the fact that the size of an investment in this stock grows over time and so it underestimates the true rate of return during the five-year sample period. This underestimation is one of the major drawbacks of the arithmetic mean. Based on the arithmetic mean return of 2 percent per year, the investor would have earned a cumulative return of 10 percent: $2 + 2 + 2 + 2 + 2 = 10$ percent from 2019 to 2023.

In fact, the cumulative return was approximately 10.3 percent. To illustrate this return, assume that an investor started with \$100,000 at the beginning of 2019. Table 3-1 shows the value of this investment from 2019 to 2023.

In each year, the starting balance is multiplied by the *gross return* (one plus the rate of return) during the year to get the ending balance. Each year's starting balance equals the previous year's ending balance.

TABLE 3-1**Computing the Return to Omega Airlines Stock**

Year	Omega Airlines Annual Return (percent)	Starting Balance	Ending Balance
2019	2	\$100,000.00	\$100,000.00(1.02) = \$102,000.00
2020	-1	\$102,000.00	\$102,000.00(0.99) = \$100,980.00
2021	3	\$100,980.00	\$100,980.00(1.03) = \$104,009.40
2022	5	\$104,009.40	\$104,009.40(1.05) = \$109,209.87
2023	1	\$109,209.87	\$109,209.87(1.01) = \$110,301.97

The ending balance in 2023 equals \$110,301.97. The cumulative rate of return during this period is the ratio of the ending balance to the starting balance minus one:

$$\begin{aligned}\text{Cumulative rate of return} &= \left(\frac{\text{Ending balance}}{\text{Starting balance}} \right) - 1 \\ &= \left(\frac{110,301.97}{100,000.00} \right) - 1 \\ &= 1.1030197 - 1 \\ &= 0.1030197 \\ &= 10.30197 \text{ percent}\end{aligned}$$

The cumulative return over period 2019–2023 is 10.30197 percent, more than the 10 percent implied by the arithmetic mean. In this case, the geometric mean provides a more accurate result than the arithmetic mean because the geometric mean takes into account the increasing size of the investment, while the arithmetic mean doesn't.

Because the geometric mean is based on products, for a sample or a population, you multiply the gross returns for each year to get the cumulative five-year return:

$$\begin{aligned}(1 + r_{2019})(1 + r_{2020})(1 + r_{2021})(1 + r_{2022})(1 + r_{2023}) \\ = (1.02)(0.99)(1.03)(1.05)(1.01) = 1.1030197\end{aligned}$$



One is added to each return to ensure that each term is positive; this is required when computing the geometric mean.

TIP

The returns are multiplied in order to indicate that *each year's return* is applied to the cumulative value of the investment, not the original value.

Because this sample has five returns, the next step is to raise the final result 1.1030197 to the *one-fifth* power:

$$(1.1030197)^{(1/5)} = 1.0198039$$

Raising a number to the one-fifth power is also known as taking the *fifth root* of the number. This corresponds to dividing by five when computing the arithmetic mean.



TIP

You can determine any exponent on a calculator with the exponentiation key; for most calculators, this key appears as Y^{x} or X^{y} .

Subtracting 1 from the example's result gives you $1.0198039 - 1 = 0.0198039 = 1.98039$ percent per year. If the investor earns this return each year for five years, the five-year return is computed as follows. First, the annual return plus one is multiplied by itself five times:

$$(1.0198039)(1.0198039)(1.0198039)(1.0198039)(1.0198039)$$

Subtracting one gives the cumulative five year return:

$$\begin{aligned}&= (1.0198039)^5 - 1 \\&= 0.1030197 \\&= 10.30197 \text{ percent}\end{aligned}$$

(Note that there can be slight differences in the results due to rounding.)



TIP

Weighted mean

Sometimes a data set contains a large number of repeated values. In these situations, you can simplify the process of computing the mean by using weights — the frequencies of a value in a sample or a population.

Calculating the weighted arithmetic mean

The formula for computing a weighted arithmetic mean for a sample or a population is

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

Here, w_i represents the *weight* associated with element X_i ; this weight equals the number of times that the element appears in the data set.

The *numerator* (the top half of the formula) tells you to multiply each element in the data set by its weight and then add the results together, as shown here:

$$\sum_{i=1}^n w_i X_i = w_1 X_1 + w_2 X_2 + w_3 X_3 + \dots + w_n X_n$$

The *denominator* (the bottom half of the formula) tells you to add the weights together:

$$\sum_{i=1}^n w_i = w_1 + w_2 + w_3 + \dots + w_n$$

You find the weighted arithmetic mean by dividing the numerator by the denominator.

As an example, suppose that a marketing firm conducts a survey of 1,000 households to determine the average number of TVs each household owns. The data show a large number of households with two or three TVs and a smaller number with one or four. Every household in the sample has at least one TV and no household has more than four. Here's the sample data for the survey:

Number of TVs per Household	Number of Households
1	73
2	378
3	459
4	90

Because many of the values in this data set are repeated multiple times, you can easily compute the sample mean as a weighted mean. Doing so is quicker than summing each value in the data set and dividing by the sample size.

Follow these steps to calculate the weighted arithmetic mean:

1. Assign a weight to each value in the data set:

$$X_1 = 1, w_1 = 73$$

$$X_2 = 2, w_2 = 378$$

$$X_3 = 3, w_3 = 459$$

$$X_4 = 4, w_4 = 90$$

2. Compute the numerator of the weighted mean formula.

Multiply each element in the sample by its weight and then add the products together:

$$\begin{aligned}\sum_{i=1}^4 w_i X_i &= w_1 X_1 + w_2 X_2 + w_3 X_3 + w_4 X_4 \\&= (73)(1) + (378)(2) + (459)(3) + (90)(4) \\&= 2,566\end{aligned}$$

3. Compute the denominator of the weighted mean formula by adding the weights together:

$$\begin{aligned}\sum_{i=1}^4 w_i &= w_1 + w_2 + w_3 + w_4 \\&= 73 + 378 + 459 + 90 \\&= 1,000\end{aligned}$$

4. Divide the numerator by the denominator:

$$\frac{\sum_{i=1}^4 w_i X_i}{\sum_{i=1}^4 w_i} = \frac{2,566}{1,000} = 2.566$$

The mean number of TVs per household in this sample is 2.566.

Getting to the Middle of Things: The Median of a Data Set

The *median* is a value that divides a sample or a population in half. In other words:

- » Half of the elements in the data set are *less than or equal* to the median.
- » Half of the elements in the data set are *greater than or equal* to the median.

For example, the sample of returns of Omega Airlines stock from 2019 to 2023 is shown here:

Year	Omega Airlines Annual Return (percent)
2019	2
2020	-1

Year	Omega Airlines Annual Return (percent)
2021	3
2022	5
2023	1

You can compute the median of this sample, using the following steps:

1. Sort the elements from the smallest to the largest.

Original data:

2, -1, 3, 5, 1

Sorted data:

-1, 1, 2, 3, 5

2. Identify the *middle* observation.

Because the sample contains five elements, the median is the third largest element (ensuring that two elements are below the median and two are above). The resulting value of the median is 2.

-1, 1, **2**, 3, 5

Note: If the sample contains an even number of elements, then no element exists in the middle of the data. Instead, you calculate the median as the *average* of the middle two elements.

Here's another example. This list is a sample of the returns from Epsilon Railways stock from 2018 to 2023:

Year	Epsilon Railways Annual Return (percent)
2018	0
2019	2
2020	3
2021	6
2022	1
2023	4

1. Sort the elements from smallest to largest.

Original data:

0, 2, 3, 6, 1, 4

Sorted data:

0, 1, 2, 3, 4, 6

2. Identify the middle observation.

In this example, there are six sample elements. Because 6 is an even number, you compute the median as the average of the third and fourth elements:

0, 1, **2, 3**, 4, 6

$(2 + 3)/2 = 2.5$

Note that three sample elements are below 2.5, and three elements are above 2.5.



TIP

The procedure for computing the median of a sample is the same as for computing the median of a population.

Determining the Relationship Between the Mean and Median

The relationship between the mean and median of a data set determines whether the data set is symmetrical about the mean, negatively skewed, or positively skewed.

In some data sets, the mean and median may equal each other. When this occurs, the data set is said to be *symmetrical about the mean*, meaning that values below the mean balance the values above the mean. A data set may also be *negatively skewed*, indicating the presence of extreme values below the mean. Likewise, a data set may be *positively skewed*, indicating the presence of extreme values above the mean.



TIP

If a data set is skewed, the mean and median won't equal each other; instead, the relationship between them determines the direction of the skew.

Symmetrical

A data set is symmetrical if the mean equals the median. Mathematically, this is expressed as

$$\text{mean} = \text{median}$$

The histogram in Figure 3-1 shows the frequency distribution for the daily returns of a stock with the following mean and median:

$$\text{mean} = 0.00 \text{ percent}$$

$$\text{median} = 0.00 \text{ percent}$$

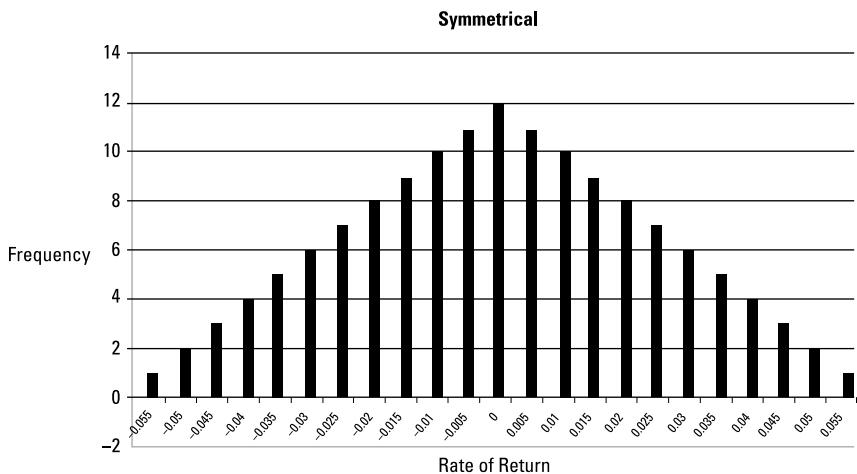


FIGURE 3-1:
Symmetrical
sample data.

The histogram shows that the left and right *tails* balance each other so that positive and negative values that are equal distances from the center are equally likely. (The left tail represents the smallest observations and the right tail represents the largest observations in the data set.) The left-hand side of this distribution is a mirror image of the right-hand side, showing that this distribution is symmetrical about the mean.

Negatively skewed

A data set is negatively skewed if the mean is less than the median. Mathematically, you can express this relationship as

$$\text{mean} < \text{median}$$

The histogram in Figure 3–2 shows the frequency distribution for the daily returns to a stock with the following mean and median:

mean = -0.95 percent

median = -0.75 percent

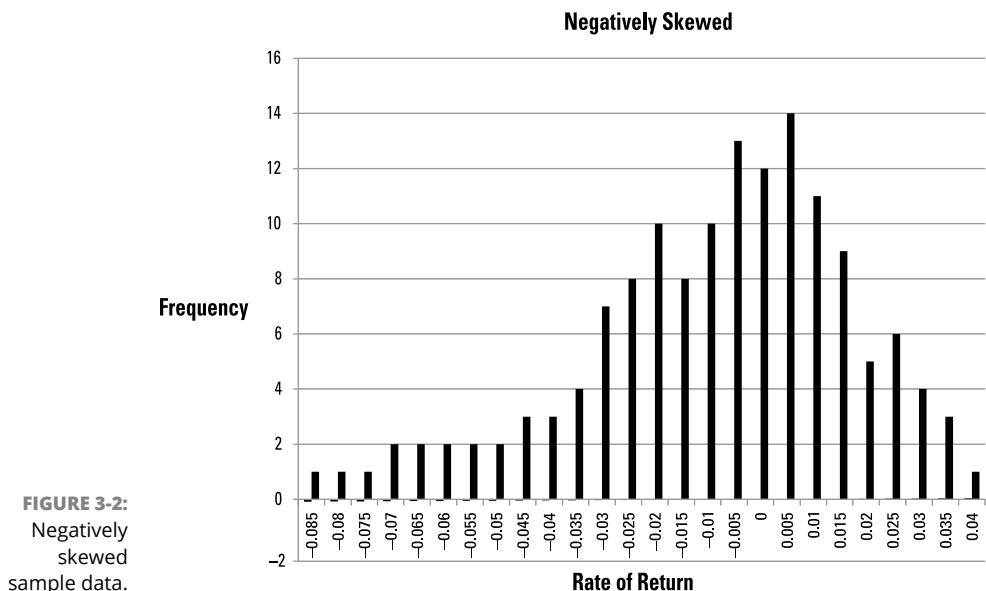


FIGURE 3-2:
Negatively
skewed
sample data.

The histogram shows a long *left tail*, which results from extreme negative values in the data set.



REMEMBER

The data in the left tail could be positive and the data set would still be negatively skewed as long as the mean is less than the median.

Positively skewed

A data set is positively skewed if the mean is greater than the median. Mathematically, this relationship looks like this:

mean > median

The histogram in Figure 3–3 shows the frequency distribution for the daily returns to a stock with the following mean and median:

mean = 1.55 percent

median = 0.70 percent

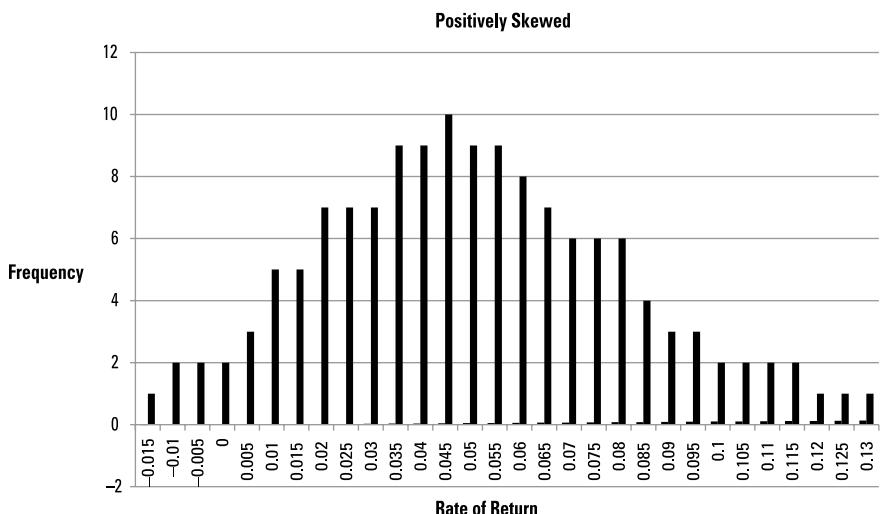


FIGURE 3-3:
Positively skewed
sample data.

The graph shows a *long right tail*, which results from extreme positive values in the data set.

PROS AND CONS OF THE MEAN AND MEDIAN

The mean is the most commonly used measure of the center of a data set. Under some conditions, though, the median (or even the mode) may be more representative of the center of the data set.

If a data set is symmetrical, the mean and the median are equal, so both are equally useful measures. When a data set is skewed, the median is likely to be a more representative measure of the center of the data than the mean because the median isn't as affected by extreme outcomes as much as the mean.

Discovering the Mode: The Most Frequently Repeated Element

The *mode* is the most frequently occurring value in a sample or a population. For example, suppose that a bank chooses a sample of 20 of its branches in New York City, and for each branch, the number of ATMs in the lobby is recorded as follows:

Three branches have two ATMs.

Six branches have three ATMs.

Eight branches have four ATMs.

Three branches have five ATMs.

Because most branches have four ATMs, 4 is the mode in this sample.



REMEMBER

One of the most unusual features of the mode is that it isn't necessarily unique; a data set can have two or more modes. It's also possible that a data set has no mode — that is, no values are repeated.

For example, suppose that the same bank chooses a sample of 20 of its branches in Connecticut. For each branch, the number of ATMs in the lobby is recorded. The results are given as follows:

Three branches have two ATMs.

Eight branches have three ATMs.

Eight branches have four ATMs.

One branch has five ATMs.

In this sample, more branches have three or four ATMs than any other number. Because the number of branches with three ATMs equals the number of banks with four ATMs, the mode of this sample is *both* 3 and 4.



TIP

The mode is most useful when a data set contains qualitative data (that is, non-numerical data). This type of data can include colors, flavors, brand names, and so on. With qualitative data, calculating a mean or a median is impossible, but you can still find the mode. With quantitative (numerical) data, the mean and the median are typically more useful than the mode.

As an example, suppose that a marketing firm conducts a survey to determine which color consumers would likely choose for a new car. The survey responses are as follows:

blue	red	blue
black	blue	black
blue	blue	black
blue	black	blue
white	silver	blue

Because this data is qualitative, calculating the mean or the median is impossible. But you can determine the mode by tabulating the frequency of the 15 responses. Because blue appears in the survey eight times, black, four times, white, red, and silver, once each, the mode is blue. Consumers in this survey prefer blue to other colors.

The distribution of colors is shown in Figure 3–4. In this example, the histogram shows colors on the horizontal axis and the corresponding frequencies on the vertical axis. Because blue occurs most frequently in this sample, it's the sample's mode.

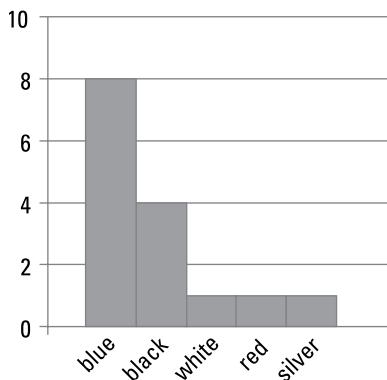


FIGURE 3-4:
Distribution of
colors chosen by
consumers.

Computing the Mean, Median, and Mode with the TI-84 Plus Calculator

Many functions on the Texas Instruments TI-84 Plus and Plus CE calculators require data to be entered in a list. To access lists, follow these steps:

1. Press [STAT], select EDIT and then 1>Edit, and then press [ENTER].



TIP

Press each key on the calculator one at a time; keys are never pressed simultaneously.

This shows a screen of columns, with headers L1, L2, L3, and so on. Each column represents a single list. If the chosen list already contains data, the best approach is to clear out the list before entering new data.

2. Use the arrow keys to move to the header of the list (not the top row) and press [CLEAR] and then [ENTER] to clear the list.

The arrow keys are found in the upper-right corner of the calculator.

3. Enter the new data, pressing the down arrow key or [ENTER] after each value is entered into the list.

Note that this data remains in the list until it is erased.

4. When the list is complete, press the [STAT] button to return to the STAT main menu, where any number of statistical calculations may be performed.

As an example, suppose that a researcher is interested in analyzing the mean, median, and mode of the price of chicken in New York City supermarkets. A sample of six supermarkets is chosen, with the results shown in the following table:

Store	Price of Chicken (\$/pound)
1	7.99
2	6.99
3	8.14
4	7.69
5	6.79
6	7.19

In this case, the store numbers do not need to be entered into a list; only the prices are needed. Once the prices have been entered into a list the mean, median, and mode are computed as follows:

1. Press the [STAT] button and then choose CALC and 1:1-Var Stats from the resulting menus, followed by the [ENTER] key.

1-Var Stats produces the following menu:

List:

FreqList:

Calculate

List: refers to the list containing the data (for example, L1). For L1, the name of the list is entered by pressing the [2nd] button followed by the [1] button.

(Note that the lists are the “second functions” of the buttons [1] through [9].) In this example, FreqList: (Frequency List) can be ignored, as none of the values in the data set are repeated.

2. Select Calculate with the arrow keys and then press [ENTER] to produce a list of results.

The first entry shows the value of the sample mean (\bar{x}) as 7.465. Scrolling down through the list of entries, the sample median is shown as “Med,” which in this case equals 7.44.

The 1-Var Stats function does not compute the mode of a data set. There is a workaround for this, though.

3. To sort the data in a list, press [STAT], choose EDIT and then 2:SortA, and then press [ENTER].

This shows on the screen as SortA(. Enter L1 using the buttons [2nd] followed by [1], followed by [)] (right parenthesis) and then [ENTER]. The message “Done” appears on the screen.

4. Press [STAT], choose EDIT and then 1>Edit, and then press [ENTER] to return to the lists.

The data in L1 have been sorted from low to high. This makes it easier to spot whether the data contains a mode (or possibly two or more modes). In this example, there is no mode since no price is repeated in the sample.

A weighted average can be computed by using the FreqList entry in the 1:1-Var Stats menu. For example, suppose that a survey is done of 40 small towns and the number of traffic lights in each town is recorded. The results are shown in the following table:

Traffic Lights	Towns
1	4
2	6
3	8
4	12
5	10

The table shows the number of towns with 1, 2, 3, 4, and 5 traffic lights. There are no towns in the survey without any traffic lights, and there are no towns in the survey with more than five traffic lights. Computing the average number of traffic lights among these towns is carried out as a weighted average.

To compute the weighted average, enter the number of traffic lights and the towns in two separate lists (such as L1 and L2). Once the data have been entered, press [STAT] and then choose CALC and 1:1-Var Stats. The following information should then be entered into the resulting menu:

List: L1

FreqList: L2

Calculate

Note that the number of traffic lights is stored in L1 and the number of towns (the frequencies) is stored in L2. After selecting Calculate, press [ENTER]. The weighted average is shown as \bar{x} , which is 3.45. This indicates that the average number of traffic lights per town is 3.45.

IN THIS CHAPTER

- » Computing variance and standard deviation
- » Finding the relative position of data: percentiles and quartiles
- » Measuring relative variation: the coefficient of variation

Chapter 4

Measuring Variation in a Data Set

One of the most important properties of a data set (a sample or population) is how “spread out” the data are from the center. (Techniques for measuring the center of a data set are covered in Chapter 3.) You can use several numerical measures, known as *measures of dispersion*, to calculate the spread of a data set.

This chapter covers the techniques used to compute the variance and standard deviation of a sample and a population. (Samples and populations are defined in Chapter 1.) Techniques for determining the relative position of an element within a sample or a population are also explained in detail; these include percentiles and quartiles. Finally, the coefficient of variation is introduced as a measure of *relative variation*; this enables a direct comparison of the properties of two samples or two populations.

Thanks to the standard deviation and the mean (covered in Chapter 3), you can calculate relative variation, which has many handy applications.

Determining Variance and Standard Deviation

Variance and standard deviation are the two most widely used measures of dispersion in statistics. They're both based on the average squared distance between the elements of a data set and the mean.

Standard deviation and variance are usually better than some other measures of dispersion, such as the range. The range is the difference between the largest and smallest elements in a data set. Interesting, but not that great. The range suffers from the drawback that it's only based on two values, so it doesn't measure the spread among the remaining values.

The variance indicates the size of the average *squared* difference between the elements of a data set and the mean of the data set. And here's what you need to know: The larger is the variance, the greater the spread among the elements of a data set.

Variance is often used as a measure of uncertainty or risk in business applications. For example, an investor may use variance to determine the degree of risk associated with owning a share of stock. If returns of the stock fluctuate significantly over time, it's a risky investment. Variance provides a method for assigning a numerical value to this fluctuation. The greater the stock's variance, the riskier it is.

Standard deviation is the *square root* of the variance. It's more commonly used than variance as a measure of risk because the variance is expressed in *squared units*. For example, the variance of a sample of prices is expressed as *dollars squared*, which is difficult to visualize. On the other hand, the standard deviation of prices is measured in dollars, which is much easier to interpret.

Finding the sample variance

Use the following formula to figure out the variance of a sample:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

Here's what each term means:

- » s^2 = the sample variance
- » \bar{X} (pronounced "X bar") = the sample mean (the average value of the sample elements)
- » n = the number of elements in the sample
- » i = an *index*, assigning a number to each sample element ranging from 1 to n
- » X_i = a single element in the sample
- » Σ = the uppercase Greek letter sigma, which indicates a sum is being computed

The *numerator* (the top half) of the sample variance formula is:

$$\sum_{i=1}^n (X_i - \bar{X})^2 = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2$$

This expression tells you to perform the following three calculations:

1. For each sample element, subtract the sample mean.
2. Square the result.
3. Compute the sum of these squares.

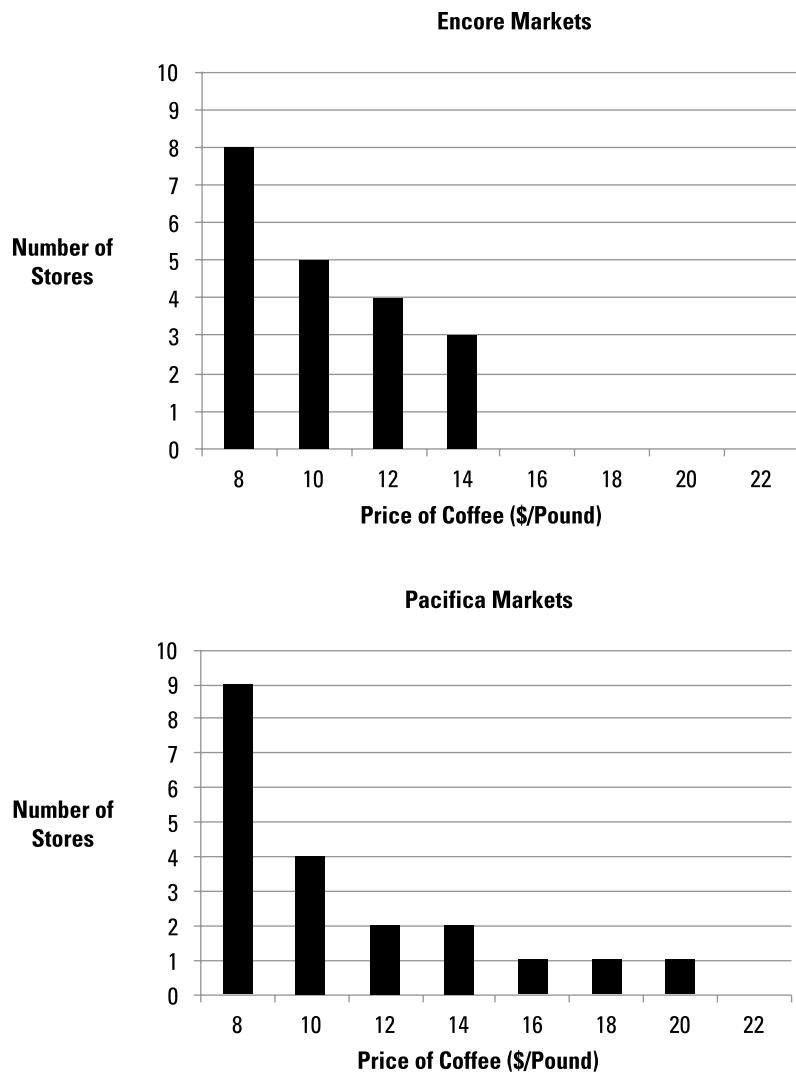
The *denominator* (the bottom half) of the sample variance formula is $n - 1$ (the sample size minus 1). Then, you find the sample variance by dividing the numerator by the denominator.

Finding the sample standard deviation

The sample standard deviation is the *square root* of the sample variance:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

Here's an example: Suppose you choose a sample of coffee prices from 20 stores in two supermarket chains: Encore Markets and Pacifica Markets. Figure 4-1a shows the distribution of prices at Encore Markets, and Figure 4-1b shows the distribution of prices at Pacifica Markets. The price of coffee per pound is shown on the horizontal (X) axis, while the number of stores that charge a given price are shown on the vertical (Y) axis.



**FIGURE 4-1
(A AND B):**
Distribution of
coffee prices at
Encore Markets
and Pacifica
Markets.

These graphs show that the prices are much more spread out at Pacifica's stores than at Encore's. In other words, Pacifica has greater *dispersion* among its prices. The range of possible prices at Pacifica's stores is much greater (at least one store charges \$19 per pound!), while at Encore, no store charges more than \$14. The dispersion among coffee prices is measured by the standard deviation, which is \$3.6631 at Pacifica's stores and \$2.1637 at Encore's stores. These numbers confirm what Table 4-1 and Table 4-2 show: There's more spread among Pacifica's prices than Encore's prices.

Tables 4-1 and 4-2 show the prices at 20 stores in each of the two chains.

TABLE 4-1**Sample Coffee Prices at Encore Markets (\$/Pound)**

8	10	11	8
8	9	8	8
13	8	9	14
12	8	12	14
10	12	8	9

TABLE 4-2**Sample Coffee Prices at Pacifica Markets (\$/Pound)**

15	17	9	7
13	7	7	9
9	8	7	7
9	13	7	11
19	11	7	7

The first step is to compute the sample mean coffee price. In this example, the sample mean for Encore is computed as follows:

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

The numerator is the sum of the coffee prices in the sample, which is 199. The denominator is the sample size, which is 20. The ratio of these two values is the sample mean, 9.95.

To compute the sample variance, subtract the sample mean from each sample coffee price, and square the results. The sum of these terms is the numerator of the sample variance formula. This is shown in Table 4-3.

TABLE 4-3**Calculations for the Sample Variance at Encore Markets**

$(8 - 9.95)^2 = 3.8025$	$(10 - 9.95)^2 = 0.0025$	$(11 - 9.95)^2 = 1.1025$	$(8 - 9.95)^2 = 3.8025$
$(8 - 9.95)^2 = 3.8025$	$(9 - 9.95)^2 = 0.9025$	$(8 - 9.95)^2 = 3.8025$	$(8 - 9.95)^2 = 3.8025$
$(13 - 9.95)^2 = 9.3025$	$(8 - 9.95)^2 = 3.8025$	$(9 - 9.95)^2 = 0.9025$	$(14 - 9.95)^2 = 16.4025$
$(12 - 9.95)^2 = 4.2025$	$(8 - 9.95)^2 = 3.8025$	$(12 - 9.95)^2 = 4.2025$	$(14 - 9.95)^2 = 16.4025$
$(10 - 9.95)^2 = 0.0025$	$(12 - 9.95)^2 = 4.2025$	$(8 - 9.95)^2 = 3.8025$	$(9 - 9.95)^2 = 0.9025$

The sum of these terms is 88.95. The sample variance is, therefore:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{88.95}{19} = 4.6816$$

Now, at last! Take the square root. The sample standard deviation is:

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} = \sqrt{4.6816} = 2.1637$$

Compute the sample variance and sample standard deviation for Pacifica Markets the same way. Table 4-4 shows the calculations for the numerator of the sample variance formula.

TABLE 4-4 Calculations for the Sample Variance at Pacifica Markets

$(15 - 9.95)^2 = 25.5025$	$(17 - 9.95)^2 = 49.7025$	$(9 - 9.95)^2 = 0.9025$	$(7 - 9.95)^2 = 8.7025$
$(13 - 9.95)^2 = 9.3025$	$(7 - 9.95)^2 = 8.7025$	$(7 - 9.95)^2 = 8.7025$	$(9 - 9.95)^2 = 0.9025$
$(9 - 9.95)^2 = 0.9025$	$(8 - 9.95)^2 = 3.8025$	$(7 - 9.95)^2 = 8.7025$	$(7 - 9.95)^2 = 8.7025$
$(9 - 9.95)^2 = 0.9025$	$(13 - 9.95)^2 = 9.3025$	$(7 - 9.95)^2 = 8.7025$	$(11 - 9.95)^2 = 1.1025$
$(19 - 9.95)^2 = 81.9025$	$(11 - 9.95)^2 = 1.1025$	$(7 - 9.95)^2 = 8.7025$	$(7 - 9.95)^2 = 8.7025$

The sum of these terms is 254.95. The sample variance is, therefore:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{254.95}{19} = 13.4184$$

The sample standard deviation is:

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} = \sqrt{13.4184} = 3.6631$$

These numbers confirm what Figure 4-1a and Figure 4-1b show: There's more spread among Pacifica's prices than Encore's prices: \$3.6631 compared to \$2.1637.



TIP

Although you can use graphs to inspect the dispersion of different samples or populations, comparing standard deviations is usually easier, and you don't have to examine the entire data set.

The standard deviation is a more useful measure of dispersion than variance. Again, variance is expressed in *squared* units (percent squared, dollars squared, and so on) because it's taken from the sum of *squared* differences between the elements in a data set and the mean of the data set. That's not as handy as standard deviation.

For example, Table 4-5 compares the variance and standard deviation of the Encore and Pacifica stores.

TABLE 4-5

Variance and Standard Deviation of Sample Stores

	Encore	Pacifica
Standard deviation (\$/pound)	2.1637	3.6631
Variance (\$ ² /pound)	4.6816	13.4184

Table 4-5 shows that the variance of coffee prices at Encore is \$4.6816 *squared* per pound, while the variance of coffee prices at Pacifica is \$13.4184 *squared* per pound. *Dollars squared* is a difficult concept to interpret — prices are never expressed in terms of dollars squared! So people most often use the standard deviation rather than the variance to show dispersion.

Calculating population variance and standard deviation

Unlike the mean, median, and mode, the variance and the standard deviation are calculated slightly differently for *samples* and *populations*. The following section shows the appropriate formulas for computing the variance and standard deviation of a population.

Finding the population variance

When you're calculating the variance for a population, use the following formula:

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

The parameters are:

- » σ^2 = population variance (σ is the lowercase Greek letter sigma)
- » μ = the population mean (μ is the Greek letter mu)
- » N = population size



TIP

Σ is the uppercase Greek letter sigma, which represents summation. σ is the lowercase sigma, which represents the population standard deviation.

The *numerator* (the top half) of the population variance formula is:

$$\sum_{i=1}^N (X_i - \mu)^2 = (X_1 - \mu)^2 + (X_2 - \mu)^2 + \dots + (X_n - \mu)^2$$

Use this formula and do the following calculations:

1. For each population element, subtract the population mean.
2. Square the result.
3. Compute the sum of the squares.

The *denominator* (the bottom half) of the population variance formula is N (the population size). You find the population variance by dividing the numerator of the population variance formula by the denominator.

Finding the population standard deviation

After you figure out the population variance, you can get the population standard deviation by taking the square root of the population variance:

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

For example, suppose an investor wants to analyze the dispersion of Alpha, Inc.'s, sales from one year to the next. Table 4-6 shows the sample of annual sales the investor takes (measured in millions of dollars per year) from 2018 to 2023. These results are considered to be a population since Alpha's first year of business was 2018.

TABLE 4-6

Alpha, Inc. Sales 2018–2023

Year	Sales (\$ million)
2018	18
2019	22
2020	31
2021	29
2022	42
2023	50

You find the population variance by following these steps:

1. Find the population mean.

The formula for calculating the sample mean is

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

Plug in the numbers from Table 4-6:

$$\mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{18 + 22 + 31 + 29 + 42 + 50}{6} = 32$$

The average annual sales during this period was \$32 million.

2. Work through the numerator of the sample variance formula.

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

The calculations are shown in Table 4-7.

In the third column ($(X_i - \mu)$), subtract the mean return from the actual return for each year. In the fourth column ($(X_i - \mu)^2$), square the result from the third column. The sum of the fourth column is the numerator of the sample variance formula; this equals 730.

TABLE 4-7

Calculations of Population Variance for Alpha, Inc.

Year	Alpha, Inc. Sales(\$ million)	$(X_i - \mu)$	$(X_i - \mu)^2$
2018	18	$18 - 32 = -14$	$(-14)^2 = 196$
2019	22	$22 - 32 = -10$	$(-10)^2 = 100$
2020	31	$31 - 32 = -1$	$(-1)^2 = 1$
2021	29	$29 - 32 = -3$	$(-3)^2 = 9$
2022	42	$42 - 32 = 10$	$(10)^2 = 100$
2023	50	$50 - 32 = 18$	$(18)^2 = 324$
		Sum	730

3. Solve the denominator of the population variance formula.

The denominator is 6. Because six elements are in this population, $N = 6$.

4. Substitute these values into the population variance formula.

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} = \frac{730}{6} = 121.667$$

The population variance of Alpha's sales is \$121.667 dollars squared.

Finding the population standard deviation

After you figure out the population variance, you get the population standard deviation by taking the square root of the population variance:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}} = \sqrt{121.667} = 11.030$$

The population standard deviation of Alpha's sales is \$11.030 million.

Finding the Relative Position of Data

Identifying the location or position of a value in a data set can be immensely useful, whether you're talking about business profitability, rates of return to stocks, or scores on school tests. You use three related measures known as *percentiles*, *quartiles*, and the *interquartile range*.

A percentile is a value that divides a sample or population into two groups, with a specified percentage in each group. For example, on a standardized exam, the 10th percentile is the score such that

- » 10 percent of the scores are equal to or below it.
- » 90 percent of the scores are equal to or above it.

Quartiles are closely related to percentiles; they subdivide a sample or a population into four equal parts. The interquartile range identifies the middle 50 percent.

Percentiles: Dividing everything into hundredths

Percentiles split up a data set into 100 equal parts, each consisting of 1 percent of the values in the data set.

For example, suppose that a corporation is analyzing the annual sales of its franchise owners. Those franchises whose sales belong to the 90th percentile will get an award. Being in the 90th percentile means that

- » 90 percent of the franchises have sales equal to or below this value.
- » 10 percent of the franchises have sales equal to or above this value.

As a result, 10 percent of the franchises will receive the award. When you hear someone say that they are in the “top 10 percent,” you can take that to mean that they are at or above the 90th percentile.

Percentiles provide a *relative ranking* for an element of a data set. For example, suppose that the corporation’s New York franchise has sales of \$1 million during the year. Judging whether this franchise is successful without knowing how this value compares with the other franchises is difficult. If it turns out that \$1 million places the New York franchise in the 10th percentile, then 90 percent of the other franchises outperformed it this year. On the other hand, if \$1 million places the New York franchise in the 80th percentile, then only 20 percent of the other franchises outperformed it this year.



TIP

The 50th percentile of a data set is the median because half of the values are less than or equal to the median, and half are greater than or equal to the median.

Suppose the Federal Reserve Bank of New York conducts a survey of the assets of the savings banks in its district. A sample of ten banks is chosen; the results (in hundreds of millions of dollars) are:

2, 3, 5, 7, 6, 4, 8, 9, 1, 2

To compute percentiles, first sort the elements from the smallest value to the largest. In this example, the sorted values are:

1, 2, 2, 3, 4, 5, 6, 7, 8, 9

There are several possible approaches to computing percentiles. One of them is to apply the following formula to compute an *index*. This index represents the location of the appropriate percentile.

$$\frac{P}{100}n + 0.5$$

Here, P is the percentile of interest (30th, 40th, and so on), and n is the size of the sample or population. You round the number to the nearest integer (whole number). The percentile equals the corresponding value in the data set.



TECHNICAL STUFF

When rounding a number with a fractional part, if the fractional part is 0.5 or greater, round *up* to the next higher integer; otherwise, round *down* to the next lower integer. For example, you round 3.4 down to 3, and 3.5 up to 4.

For example, in order to find the 30th percentile of a set of ten, the index is

$$\frac{P}{100}n + 0.5 = \frac{30}{100}(10) + 0.5 = 3.5$$

Round 3.5 up to 4 to see that the fourth smallest value, the number 3 in this example, is the 30th percentile.

1, 2, 2, **3**, 4, 5, 6, 7, 8, 9

Similarly, you find the 70th percentile of a set of ten as follows:

$$\frac{P}{100}n + 0.5 = \frac{70}{100}(10) + 0.5 = 7.5$$

Don't forget to round 7.5 up to 8, which shows that the eighth smallest value, or the number 7 in this example, is the 70th percentile.

1, 2, 2, 3, 4, 5, 6, **7**, 8, 9



TECHNICAL STUFF

Microsoft Excel uses a somewhat different approach to computing percentiles. If you use the PERCENTILE function, you will get 2.7 for the 30th percentile and 6.3 for the 70th percentile.

Quartiles: Dividing everything into fourths

Quartiles split up a data set into four equal parts, each consisting of 25 percent of the sorted values in the data set. Quartiles are related to percentiles like so:

First quartile (Q_1) = 25th percentile

Second quartile (Q_2) = 50th percentile

Third quartile (Q_3) = 75th percentile



Because the second quartile is the 50th percentile, it's also the *median* of a data set. The fourth quartile usually isn't used because its value is greater than every element in a data set, so what's the point?

One commonly used approach for calculating quartiles follows these two steps:

1. **Split the data into a lower half and an upper half (leaving out the median).**
2. **Compute the median of the lower half and the upper half.**

After you've split the data into lower and upper halves, you figure out the quartiles as follows:

Q_1 = the median of the lower half

Q_2 = the median of the entire data set

Q_3 = the median of the upper half

The following data represent a sample of eight stock returns for Gamma Industries:

5, 7, 6, 3, 0, -2, 4, 3

The sorted values are:

-2, 0, 3, 3, 4, 5, 6, 7

In this example, you have eight elements. Because 8 is an even number, the median is the average of the fourth and fifth elements (-2, 0, 3, 3, 4, 5, 6, 7):

$$(3 + 4)/2 = 3.5$$

Therefore, the second quartile (Q_2) is 3.5.

The values below the median constitute the lower half of the sorted sample:

-2, 0, 3, 3

The values above the median constitute the upper half of the sorted sample:

4, 5, 6, 7

Both the lower and upper halves have four sample elements. Because four is an even number, the median is the average of the second and third elements.

For the lower half, the median is: $(0 + 3)/2 = 1.5$. This is the *average* value of the two middle elements. Therefore, the first quartile (Q_1) is 1.5.

For the upper half, the median is $(5 + 6)/2 = 5.5$. Therefore, the third quartile (Q_3) is 5.5.



TECHNICAL STUFF

As with percentiles, Microsoft Excel uses a different approach to computing quartiles; if you use the QUARTILE function, you will get 3.5 for Q_2 , but you will also get 2.25 for Q_1 (instead of 1.5) and 5.25 for Q_3 (instead of 5.5).

Interquartile range: Identifying the middle 50 percent

The interquartile range (IQR) is the difference between the third quartile and the first quartile: $IQR = Q_3 - Q_1$. The IQR represents the middle 50 percent of the data set. For the Gamma Industries example, the IQR is $Q_3 - Q_1 = 5.5 - 1.5 = 4$.



REMEMBER

An advantage of the IQR is that it isn't greatly affected by *outliers* — values within a data set that are substantially different than the other elements in the data set. In fact, the IQR can help identify outliers within a data set.

You can find the outliers in a data set in several ways. One of the simpler approaches is to create a lower bound and an upper bound. What this means is that if any elements are below the lower bound or above the upper bound, they're outliers. You set these bounds based on quartiles and the interquartile range:

Lower bound: $Q_1 - 1.5(IQR)$

Upper bound: $Q_3 + 1.5(IQR)$

Based on the Gamma Industries data, the lower bound = $1.5 - 1.5(4) = -4.5$, and the upper bound = $5.5 + 1.5(4) = 11.5$. Because no value in this sample is below -4.5 or above 11.5, the sample has no outliers.

Measuring Relative Variation

Relative variation refers to the spread of a sample or a population as a proportion of the mean. Relative variation is useful because it can be expressed as a percentage, and is independent of the units in which the sample or population data is measured.

For example, you can use a measure of relative variation to compare the uncertainty or variation associated with the temperature in two different countries, even if one country uses Fahrenheit temperatures and the other uses Celsius temperatures. As another example, a measure of relative variation can be useful for comparing the returns earned by two portfolio managers. It wouldn't make any sense to compare the mean returns achieved by two different managers without explicitly considering the levels of risk that they have incurred. A measure of relative variation provides a number that considers both the risk and the return of a portfolio, so that it can be determined which portfolio is riskier relative to the return.

You can use several different types of measures of relative variation. One of the most popular is known as the coefficient of variation.

Coefficient of variation: The spread of a data set relative to the mean

The *coefficient of variation* (CV) indicates how "spread out" the members of a sample or population are relative to the mean. The coefficient of variation is measured as a percentage, so it's independent of the units in which the mean and standard deviation are measured. This enables the relative variation of different samples or populations to be compared directly to each other.

For example, the coefficient of variation can express the risk of an investment portfolio *per unit of return*. This means you can compare the performance of different portfolios to see which one offers the least amount of risk per unit of return.

Here's the formula for finding the coefficient of variation for either samples or populations:

$$CV = \left(\frac{\text{standard deviation}}{\text{mean}} \right) * 100$$

Suppose a corporation requires the services of a consulting firm to improve its accounting systems. The corporation has determined that the two best choices are

Superior Accounting, Inc., and Data Services Corp. The corporation has done some research about the pricing practices of these two firms. The average price charged per hour, along with the standard deviation, are shown in Table 4–8.

TABLE 4-8

Comparative Prices Charged by Superior Accounting and Data Services

	Superior Accounting	Data Services
Mean price (per hour)	\$200	\$175
Standard deviation (per hour)	\$80	\$75

Based on this data, the coefficient of variation for the prices charged by each firm are the following:

$$\text{Superior Accounting: } CV = \frac{\$80}{\$200} * 100 = 40.00 \text{ percent}$$

$$\text{Data Services: } CV = \frac{\$75}{\$175} * 100 = 42.86 \text{ percent}$$

These results show that although the prices charged by Superior Accounting have a larger standard deviation than Data Services, the relative variation of Data Services is greater (42.86 percent compared with 40.00 percent). This indicates that the relative uncertainty associated with Data Services' prices is greater than for Superior Accounting's prices.

Comparing the relative risks of two portfolios

Suppose a portfolio manager is responsible for an insurance company's equity portfolio and bond portfolio. The manager wants to know which portfolio is riskier in absolute and relative terms. The manager takes a sample of returns from the past ten years and computes the mean and standard deviation. See Table 4–9 for the results.

TABLE 4-9

Comparative Performance of Bond and Equity Portfolios

	Bond Portfolio	Equity Portfolio
Mean return	8%	20%
Standard deviation of returns	16%	30%

These results show that the equity portfolio offers a higher average (mean) return than the bond portfolio and that the equity portfolio is *riskier* in absolute terms than the bond portfolio (as the standard deviation of returns is greater for the equity portfolio).

Because the two portfolios offer different returns and different levels of risk, it's impossible to compare them directly without using a measure of relative risk, which shows how risky a portfolio is relative to its return. So you need to find the coefficient of variation for the two portfolios, using the CV formula:

$$\text{Bond: } CV = \frac{16 \text{ percent}}{8 \text{ percent}} * 100 = 200 \text{ percent}$$

$$\text{Equity: } CV = \frac{30 \text{ percent}}{20 \text{ percent}} * 100 = 150 \text{ percent}$$

The bond portfolio offers a level of risk that's 200 percent of the average return, while the equity portfolio offers a level of risk that's 150 percent of the average return. So while the equity portfolio is riskier in *absolute* terms (due to the higher standard deviation) the bond portfolio is riskier in *relative* terms (due to the higher coefficient of variation).

Computing Measures of Dispersion with the TI-84 Plus Calculator

You can compute measures of dispersion, including the variance and standard deviation, with the Texas Instruments TI-84 Plus and Plus CE calculators. As an example, enter the following data provided for chicken prices in New York City from Chapter 3 into a list:

Store	Price of Chicken (\$/pound)
1	7.99
2	6.99
3	8.14
4	7.69
5	6.79
6	7.19

To enter this data into a list, follow these steps:



1. Press [STAT], select EDIT and then 1>Edit, and then press [ENTER].

Old data can be erased by moving to the header of a list with the arrow keys and then pressing [CLEAR] followed by [ENTER].

2. Enter the new data and then press [STAT], select CALC and then 1:1-Var Stats.

1-Var Stats produces the following menu:

List:

FreqList:

Calculate

3. Assuming that the data have been entered into L1, enter L1 for List and ignore the FreqList because no values are repeated.

4. Select Calculate and then press [ENTER] to produce a long list of statistical measures:

S_x = the sample standard deviation, which is 0.55475

σ_x = the population standard deviation, which is 0.50642

Because the data in this example is a sample, the appropriate measure is $S_x = 0.55475$. The sample variance can then be obtained by squaring the sample standard deviation:

$$S_x^2 = (0.55475)^2 = 0.30775$$

The range is not listed, but instead the minimum and maximum values in the data set are shown as $\text{min}X = 6.79$ and $\text{max}X = 8.14$. The range is the difference between these two values, or $8.14 - 6.79 = 1.35$.

The first and third quartiles are shown as $Q_1 = 6.99$ and $Q_3 = 7.99$. The interquartile range (IQR) equals $Q_3 - Q_1 = 7.99 - 6.99 = 1.00$. (Remember that the second quartile Q_2 is the same as the median, or 7.44.)

The coefficient of variation equals the ratio of the standard deviation to the mean multiplied by 100 (the mean is listed in the output as $\bar{x} = 7.465$). The coefficient of variation therefore equals:

$$CV = \left(\frac{s}{\bar{x}} \right) 100 = \left(\frac{0.55475}{7.465} \right) 100 = 7.43\%$$

IN THIS CHAPTER

- » Working with measures of association: covariance and correlation
- » Determining the correlation coefficient

Chapter 5

Measuring How Data Sets Are Related to Each Other

A measure of association is a numerical value that reflects the tendency of two variables to move in the same direction or in opposite directions. For example, it makes sense that corporate profits and sales would both tend to increase when the economy is strong and decrease when the economy is weak. A measure of association is used to assign a numerical value to the strength and direction of this type of relationship.

The two most widely used measures of association are known as *covariance* and *correlation*. Measures of association can help answer questions such as “Do stock prices tend to rise during a period of falling interest rates?” and “Does the unemployment rate tend to increase during periods of rising oil prices?”

In this chapter, you see formulas for computing covariance and correlation for both samples and populations. The relationship between two variables is illustrated with a type of graph known as a *scatter plot*, which is useful for seeing the relationship that exists (if any) between two variables. (I cover several types of graphs such as the scatter plot in Chapter 2.) This chapter concludes by

illustrating how the risks of a portfolio of stocks may be diversified if the stocks have low or negative correlations between them.

Understanding Covariance and Correlation

Two of the most widely used measures of association are known as *covariance* and *correlation*. These are closely related to each other. You can think of correlation as a standardized version of covariance. Correlation is easier to interpret because its value is always between -1 and 1. For example, a correlation of 0.9 indicates a very strong relationship in which two variables nearly always move in the same direction; a correlation of -0.1 shows a very weak relationship in which there is a slight tendency for two variables to move in opposite directions. With covariance, there is no minimum or maximum value, so the values are more difficult to interpret. For example, a covariance of 50 may show a strong or weak relationship; this depends on the units in which covariance is measured.



TECHNICAL STUFF

Correlation is a measure of the strength and direction of two *linearly related* variables. Two variables are said to be linearly related if their relationship can be expressed with the following equation:

$$Y = mX + b$$

X and Y are variables; m and b are constants. m is the slope and b is the intercept. For example, suppose that the relationship between two variables is:

$$Y = 3X + 4$$

3 is the *coefficient* of X; this indicates that an increase of X by 1 is associated with an increase of Y by 3. Equivalently, a decrease of X by 1 is associated with a decrease of Y by 3. The 4 in this equation indicates that Y equals 4 when X equals 0.

Covariance and correlation show that variables can have a positive relationship, a negative relationship, or no relationship at all. With covariance and correlation, there are three cases that may arise:

- » **If two variables increase or decrease at the same time, the covariance and correlation between them is positive.** For example, the covariance and correlation between the stock prices of two oil companies is positive because many of the same factors affect the stock prices in the same way.
- » **If two variables move in opposite directions, the covariance and correlation between them is negative.** For example, the covariance and correlation between interest rates and new home sales is negative because rising interest rates increase the cost of purchasing a new home, so that

the demand for new home sales falls. The opposite occurs with falling interest rates.

- » **If two variables are unrelated to each other, the covariance and correlation between them is zero.** For example, the covariance and correlation between gold prices and new car sales is zero because the two have nothing to do with each other.

In the following sections, I introduce formulas for computing sample covariance, sample correlation, population covariance, and population correlation. These measures are illustrated with several examples.

Sample covariance and correlation coefficient

Sample covariance measures the strength and the direction of the relationship between the elements of two samples. (Recall from Chapter 1 that a sample is a randomly chosen selection of elements from an underlying population.)

The sample covariance between X and Y is

$$s_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Here's what each element in this equation means:

- » s_{XY} = the sample covariance between variables X and Y (the two subscripts indicate that this is the sample covariance, not the sample standard deviation).
- » \bar{X} ("X bar") = the sample mean for X.
- » \bar{Y} ("Y bar") = the sample mean for Y.
- » n = the number of elements in both samples.
- » i = an *index* that assigns a number to each sample element, ranging from 1 to n .
- » X_i = a single element in the sample for X.
- » Y_i = a single element in the sample for Y.
- » Σ = the uppercase Greek letter sigma that indicates that a sum is being computed.

The sample covariance may have any positive or negative value.

You calculate the *sample correlation* (also known as the *sample correlation coefficient*) between X and Y directly from the sample covariance with the following formula:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

The key terms in this formula are:

- » r_{XY} = sample correlation between X and Y
- » s_{XY} = sample covariance between X and Y
- » s_X = sample standard deviation of X
- » s_Y = sample standard deviation of Y

The formula used to compute the sample correlation coefficient ensures that its value ranges between -1 and 1 . For example, suppose you take a sample of stock returns from the Excelsior Corporation and the Adirondack Corporation from the years 2019 to 2023, as shown here:

Year	Excelsior Corp. Annual Return (percent) (X)	Adirondack Corp. Annual Return (percent) (Y)
2019	1	3
2020	-2	2
2021	3	4
2022	0	6
2023	3	0

What are the covariance and correlation between the stock returns? To figure that out, you first have to find the mean of each sample. (The sample mean is discussed in Chapter 3.) In this example, X represents the returns to Excelsior and Y represents the returns to Adirondack.

» The sample mean of X is

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^n X_i}{n} \\ &= \frac{(1-2+3+0+3)}{5} \\ &= \frac{5}{5} = 1\end{aligned}$$

You obtain the sample mean by summing all the elements of the sample and then dividing by the sample size. In this case, the sample elements sum to 5 and the sample size is 5. Dividing these numbers gives a sample mean of 1.

» The sample mean of Y is

$$\begin{aligned}\bar{Y} &= \frac{\sum_{i=1}^n Y_i}{n} \\ &= \frac{(3+2+4+6+0)}{5} \\ &= \frac{15}{5} = 3\end{aligned}$$

Table 5-1 shows the remaining calculations for the sample covariance:

TABLE 5-1 Computing the Sample Covariance

Year	Excelsior Corp Annual Return (percent)	Adirondack Corp Annual Return (percent)	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
2019	1	3	$1 - 1 = 0$	$3 - 3 = 0$	$(0)(0) = 0$
2020	-2	2	$-2 - 1 = -3$	$2 - 3 = -1$	$(-3)(-1) = 3$
2021	3	4	$3 - 1 = 2$	$4 - 3 = 1$	$(2)(1) = 2$
2022	0	6	$0 - 1 = -1$	$6 - 3 = 3$	$(-1)(3) = -3$
2023	3	0	$3 - 1 = 2$	$0 - 3 = -3$	$(2)(-3) = -6$
Mean	1	3		Sum	-4

The $(X_i - \bar{X})$ column represents the differences between each return to Excelsior in the sample and the sample mean; similarly, the $(Y_i - \bar{Y})$ column represents the same calculations for Adirondack. The entries in the $(X_i - \bar{X})(Y_i - \bar{Y})$ column equal the product of the entries in the previous two columns. The sum of the $(X_i - \bar{X})(Y_i - \bar{Y})$ column gives the numerator in the sample covariance formula:

$$\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = -4$$

The denominator equals the sample size minus one, which is $5 - 1 = 4$. (Both samples have five elements, $n = 5$). Therefore, the sample covariance equals

$$s_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$= \frac{-4}{4} = -1$$

To calculate the sample correlation coefficient, divide the sample covariance by the product of the sample standard deviation of X and the sample standard deviation of Y:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

You find the sample standard deviation of X by computing the sample variance of X and then taking the square root of the result (as I explain in Chapter 4). Table 5-2 shows the calculations for the sample variance of X.

TABLE 5-2

Computing the Sample Variance for Excelsior

Year	Excelsior Corp. Annual Return (percent)	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
2019	1	$1 - 1 = 0$	$(0)^2 = 0$
2020	-2	$-2 - 1 = -3$	$(-3)^2 = 9$
2021	3	$3 - 1 = 2$	$(2)^2 = 4$
2022	0	$0 - 1 = -1$	$(-1)^2 = 1$
2023	3	$3 - 1 = 2$	$(2)^2 = 4$
Mean	1	Sum	18

The $(X_i - \bar{X})$ column represents the differences between each return to Excelsior in the sample and the sample mean; the $(X_i - \bar{X})^2$ column represents the *squared* difference between each return to Excelsior and the sample mean. The sum of the $(X_i - \bar{X})^2$ column gives the numerator in the sample variance formula. You divide this number by the sample size minus one ($5 - 1 = 4$) to get the sample variance of X :

$$\begin{aligned}s_X^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \\&= \frac{18}{4} \\&= 4.5\end{aligned}$$

The sample standard deviation of X is the square root of 4.5, or $\sqrt{4.5} = 2.1213$.

Table 5–3 shows the calculations for the sample variance of Y .

TABLE 5-3

Computing the Sample Variance for Adirondack

Year	Adirondack Corp. Annual Return (percent)	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$
2019	3	$3 - 3 = 0$	$(0)^2 = 0$
2020	2	$2 - 3 = -1$	$(-1)^2 = 1$
2021	4	$4 - 3 = 1$	$(1)^2 = 1$
2022	6	$6 - 3 = 3$	$(3)^2 = 9$
2023	0	$0 - 3 = -3$	$(-3)^2 = 9$
Mean	3	Sum	20

Based on the calculations in Table 5–3, the sample variance of Y equals

$$\begin{aligned}s_Y^2 &= \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1} \\&= \frac{20}{4} \\&= 5\end{aligned}$$

The sample standard deviation of Y equals the square root of 5, or $\sqrt{5} = 2.2361$.

Substituting these values into the sample correlation formula gives you

$$\begin{aligned} r_{XY} &= \frac{s_{XY}}{s_X s_Y} \\ &= \frac{-1}{(2.1213)(2.2361)} \\ &= -0.2108 \end{aligned}$$

The negative result shows that there's a weak negative correlation between the stock returns of Excelsior and Adirondack. If two variables are *perfectly* negatively correlated (they *always* move in opposite directions), their correlation will be -1 . If two variables are *independent* (unrelated to each other), their correlation will be 0 . The correlation between the returns to Excelsior and Adirondack stock is -0.2108 , which indicates that the two variables show a slight tendency to move in opposite directions.

Population covariance and correlation coefficient

The population covariance measures the strength and the direction of the relationship between the elements of two populations. It's computed in a manner similar to the sample covariance.

You use the following formula to find the population covariance:

$$\sigma_{XY} = \frac{\sum_{i=1}^N (X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

The key terms here are:

- » σ_{XY} = the population covariance between variables X and Y
- » μ_X = the population mean for X
- » μ_Y = the population mean for Y
- » N = the number of elements in both populations
- » i = an *index* that assigns a number to each population element, ranging from 1 to N
- » X_i = a single element in the population for X
- » Y_i = a single element in the population for Y
- » Σ = the uppercase Greek letter sigma that indicates a sum is being computed

The population correlation coefficient is based on the population covariance. You use the following formula to find the population correlation coefficient:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

The key terms here are:

- » ρ_{XY} = the population correlation coefficient between variables X and Y
- » σ_{XY} = the population covariance between variables X and Y
- » σ_X = the population standard deviation of variable X
- » σ_Y = the population standard deviation of variable Y

For example, suppose that two new companies were created in 2019: Theta Corp. and Eta Corp. The returns to the two companies' stocks from 2019 to 2023 are shown in Table 5-4.

TABLE 5-4

Annual Returns to Theta and Eta

Year	Theta Corp. Annual Return (percent) (X)	Eta Corp. Annual Return (percent) (Y)
2019	11	6
2020	9	5
2021	4	1
2022	2	9
2023	5	12

Because these companies have been in business only since 2019, each set of returns represents a *population* (the entire history of returns). The population covariance and correlation between the returns to these stocks are computed as follows.

- » The population mean of X is

$$\begin{aligned}\mu_X &= \frac{\sum_{i=1}^N X_i}{N} \\ &= \frac{(11+9+4+2+5)}{5} \\ &= \frac{31}{5} = 6.2\end{aligned}$$

The population mean is obtained by summing all the elements of the population and then dividing by the population size. In this case, the 5 population elements sum to 31, and the population size is 5. Dividing these numbers gives a population mean of 6.2.

» The population mean of Y is

$$\begin{aligned}\mu_Y &= \frac{\sum_{i=1}^N Y_i}{N} \\ &= \frac{(6 + 5 + 1 + 9 + 12)}{5} \\ &= \frac{33}{5} = 6.6\end{aligned}$$

Table 5–5 shows the remaining calculations for the population covariance.

TABLE 5-5 Computing the Population Covariance

Year	Theta Corp. Annual Return (percent) (X)	Eta Corp. Annual Return (percent) (Y)	$(X_i - \mu_X)$	$(Y_i - \mu_Y)$	$(X_i - \mu_X)(Y_i - \mu_Y)$
2019	11	6	$11 - 6.2 = 4.8$	$6 - 6.6 = -0.6$	$(4.8)(-0.6) = -2.88$
2020	9	5	$9 - 6.2 = 2.8$	$5 - 6.6 = -1.6$	$(2.8)(-1.6) = -4.48$
2021	4	1	$4 - 6.2 = -2.2$	$1 - 6.6 = -5.6$	$(-2.2)(-5.6) = 12.32$
2022	2	9	$2 - 6.2 = -4.2$	$9 - 6.6 = 2.4$	$(-4.2)(2.4) = -10.08$
2023	5	12	$5 - 6.2 = -1.2$	$12 - 6.6 = 5.4$	$(-1.2)(5.4) = -6.48$
Mean	6.2	6.6		Sum	-11.60

The sum of the $(X_i - \mu_X)(Y_i - \mu_Y)$ column gives the numerator in the population covariance formula:

$$\sum_{i=1}^N (X_i - \mu_X)(Y_i - \mu_Y) = -11.60$$

The denominator equals the population size, which is 5. Therefore, the population covariance equals

$$\begin{aligned}\sigma_{XY} &= \frac{\sum_{i=1}^N (X_i - \mu_X)(Y_i - \mu_Y)}{n} \\ &= \frac{-11.60}{5} \\ &= -2.32\end{aligned}$$

To calculate the population correlation coefficient, divide the population covariance by the product of the population standard deviation of X and the population standard deviation of Y:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

You find the population standard deviation of X by computing the population variance of X and then taking the square root of the result (as I explain in Chapter 4). Table 5–6 shows the calculations for the population variance of X.

TABLE 5-6

Computing the Population Variance for Theta

Year	Theta Corp. Annual Return (%) (X)	$(X_i - \mu_X)$	$(X_i - \mu_X)^2$
2019	11	$11 - 6.2 = 4.8$	$(4.8)^2 = 23.04$
2020	9	$9 - 6.2 = 2.8$	$(2.8)^2 = 7.84$
2021	4	$4 - 6.2 = -2.2$	$(-2.2)^2 = 4.84$
2022	2	$2 - 6.2 = -4.2$	$(-4.2)^2 = 17.64$
2023	5	$5 - 6.2 = -1.2$	$(-1.2)^2 = 1.44$
Mean	6.2	Sum	54.80

The sum of the $(X_i - \mu_X)^2$ column gives the numerator in the population variance formula. You divide this number by the population size to get the population variance of X:

$$\begin{aligned}\sigma_X^2 &= \frac{\sum_{i=1}^N (X_i - \mu_X)^2}{N} \\ &= \frac{54.8}{5} \\ &= 10.96\end{aligned}$$

The population standard deviation of X is the square root of 10.96, or $\sqrt{10.96} = 3.3106$.

Table 5–7 shows the calculations for the population variance of Y.

TABLE 5-7

Computing the Population Variance for Eta Corporation

Year	Eta Corp. Annual Return (percent) (Y)	$(Y_i - \mu_Y)$	$(Y_i - \mu_Y)^2$
2019	6	$6 - 6.6 = -0.6$	$(-0.6)^2 = 0.36$
2020	5	$5 - 6.6 = -1.6$	$(-1.6)^2 = 2.56$
2021	1	$1 - 6.6 = -5.6$	$(-5.6)^2 = 31.36$
2022	9	$9 - 6.6 = 2.4$	$(2.4)^2 = 5.76$
2023	12	$12 - 6.6 = 5.4$	$(5.4)^2 = 29.16$
Mean	6.6	Sum	69.2

Based on the calculations in Table 5–7, the population variance of Y equals

$$\begin{aligned}\sigma_Y^2 &= \frac{\sum_{i=1}^N (Y_i - \mu_Y)^2}{N} \\ &= \frac{69.2}{5} \\ &= 13.84\end{aligned}$$

The population standard deviation of Y equals the square root of 13.84, or $\sqrt{13.84} = 3.7202$.

Substituting these values into the population correlation formula gives you:

$$\begin{aligned}\rho_{XY} &= \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \\ &= \frac{-2.32}{(3.3106)(3.7202)} \\ &= -0.1884\end{aligned}$$

The negative result shows that there's a weak negative correlation between the stock returns of Theta and Eta.

Comparing correlation and covariance

When trying to find the relationship between two variables, you see that the correlation coefficient has several advantages over the covariance, including the following:

- » The covariance has no lower or upper limits, whereas the correlation coefficient ranges between -1 and 1 , making it easier to interpret its meaning.

In the example with the returns to Excelsior and Adirondack stock (in the earlier section “Sample covariance and correlation”), the covariance is -1 . Although this negative number indicates a tendency for the stock returns to move in opposite directions, it’s difficult to judge the *strength* of this relationship. On the other hand, the correlation coefficient is -0.2108 ; because the correlation coefficient ranges from -1 to 1 , you can see that the relationship between the stock returns is negative but not very strong.

- » Unlike the covariance, the value of the correlation isn’t affected by the units in which X and Y are measured. For example, suppose that a sample of tuna is chosen from the catch of two different fishing boats. The covariance between the weights of the tuna caught by the two boats is computed. The value of the covariance is different if the weights are expressed in kilograms or in pounds; however, the correlation is the same whether weights are expressed in kilograms or pounds.

To illustrate the second point further, say you record a sample of the average temperatures (in Celsius and Fahrenheit) in two cities from 2019 to 2023 and come up with the following results:

Year	City 1 (Celsius)	City 2 (Celsius)	City 1 (Fahrenheit)	City 2 (Fahrenheit)
2019	0.0°C	-10.0°C	32.0°F	14.0°F
2020	20.0°C	15.0°C	68.0°F	59.0°F
2021	-8.0°C	22.0°C	17.6°F	71.6°F
2022	25.0°C	30.0°C	77.0°F	86.0°F
2023	14.0°C	25.0°C	57.2°F	77.0°F
Mean	10.2°C	16.4°C	50.4°F	61.5°F

Assume that X represents the temperature in City 1 and Y represents the temperature in City 2. Table 5–8 shows the calculations for the covariance between the temperatures in Celsius of both cities.

TABLE 5-8

Covariance between Celsius Temperatures in City 1 and City 2

Year	City 1 (Celsius)	City 2 (Celsius)	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
2019	0.0°C	-10.0°C	$0.0 - 10.2 = -10.2$	$-10.0 - 16.4 = -26.4$	$(-10.2)(-26.4) = 269.3$
2020	20.0°C	15.0°C	$20.0 - 10.2 = 9.8$	$15.0 - 16.4 = -1.4$	$(9.8)(-1.4) = -13.7$
2021	-8.0°C	22.0°C	$-8.0 - 10.2 = -18.2$	$22.0 - 16.4 = 5.6$	$(-18.2)(5.6) = -101.9$
2022	25.0°C	30.0°C	$25.0 - 10.2 = 14.8$	$30.0 - 16.4 = 13.6$	$(14.8)(13.6) = 201.3$
2023	14.0°C	25.0°C	$14.0 - 10.2 = 3.8$	$25.0 - 16.4 = 8.6$	$(3.8)(8.6) = 32.7$
Mean	10.2°C	16.4°C		Sum	387.6

The $(X_i - \bar{X})$ column represents the differences between each temperature in City 1 and the sample mean. The $(Y_i - \bar{Y})$ column represents the differences between each temperature in City 2 and the sample mean. The $(X_i - \bar{X})(Y_i - \bar{Y})$ column is simply the product of the $(X_i - \bar{X})$ column and the $(Y_i - \bar{Y})$ column. The sum of the $(X_i - \bar{X})(Y_i - \bar{Y})$ column gives the numerator in the sample covariance formula, which is 387.6.

$$\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = 387.6$$

The denominator equals the sample size minus one, which is $5 - 1 = 4$ (because both samples have five elements, $n = 5$). Therefore, the sample covariance equals

$$\begin{aligned}s_{XY} &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \\ &= \frac{387.6}{4} \\ &= 96.9\end{aligned}$$

You find the sample standard deviation of X by computing the sample variance of X and then taking the square root of the result (see Chapter 4). Table 5-9 shows the calculations for the sample variance of X (Celsius temperatures for City 1).

To finish the calculation for the sample variance of X , you divide the sum of the terms in the $(X_i - \bar{X})^2$ column by the sample size minus one, like so:

$$\begin{aligned}s_X^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \\ &= \frac{764.8}{4} \\ &= 191.2\end{aligned}$$

TABLE 5-9**Sample Variance of City 1**

Year	City 1 (Celsius)	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
2019	0.0°C	$0.0 - 10.2 = -10.2$	$(-10.2)^2 = 104.0$
2020	20.0°C	$20.0 - 10.2 = 9.8$	$(9.8)^2 = 96.0$
2021	-8.0°C	$-8.0 - 10.2 = -18.2$	$(-18.2)^2 = 331.2$
2022	25.0°C	$25.0 - 10.2 = 14.8$	$(14.8)^2 = 219.0$
2023	14.0°C	$14.0 - 10.2 = 3.8$	$(3.8)^2 = 14.4$
Mean	10.2°C	Sum	764.8

The sample standard deviation is the square root of the sample variance, or $\sqrt{191.2} = 13.8275$.

Following the same steps, you can find the sample variance of Y with the calculations in Table 5-10.

TABLE 5-10**Sample Variance of City 2**

Year	City 2 (C)	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$
2019	-10.0	$-10.0 - 16.4 = -26.4$	$(-26.4)^2 = 697.0$
2020	15.0	$15.0 - 16.4 = -1.4$	$(-1.4)^2 = 2.0$
2021	22.0	$22.0 - 16.4 = 5.6$	$(5.6)^2 = 31.4$
2022	30.0	$30.0 - 16.4 = 13.6$	$(13.6)^2 = 185.0$
2023	25.0	$25.0 - 16.4 = 8.6$	$(8.6)^2 = 74.0$
Mean	16.4	Sum	989.2

To get the sample variance, divide the sum of the terms in the $(Y_i - \bar{Y})^2$ column by the sample size minus one:

$$s_Y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$$

$$= \frac{989.2}{4}$$

$$= 247.3$$

The sample standard deviation is the square root of the sample variance, or $\sqrt{247.3} = 15.7258$.

Next, substitute these values into the sample correlation formula:

$$\begin{aligned} r_{XY} &= \frac{s_{XY}}{s_X s_Y} \\ &= \frac{96.9}{(13.8275)(15.7258)} \\ &= 0.4456 \end{aligned}$$

Repeating these same calculations for the temperatures in Fahrenheit, the covariance is 313.96 (compared with 96.9 when measured in Celsius) and the correlation remains at 0.4456. The covariance increases with Fahrenheit temperatures because the magnitude of the temperatures is greater, whereas the correlation isn't affected. The fact that the results depend on the units involved is one of the major drawbacks of using covariance instead of correlation.

Interpreting the Correlation Coefficient

Interpreting the correlation coefficient is easier than interpreting the covariance. Consider these examples:

- » A correlation of 0.9 (close to the maximum value of 1.0) indicates a strong positive relationship between X and Y ; when X increases, Y nearly always increases, and vice versa.
A correlation of 0.2 (close to zero) indicates a weak positive relationship; when X increases, Y is somewhat more likely to increase than decrease, and vice versa.
- » A correlation of -0.9 (close to the minimum value of -1.0) indicates a strong negative relationship between X and Y . Most of the time, when X increases, Y decreases; most of the time, when X decreases, Y increases.
A correlation of -0.2 (close to zero) indicates a weak negative relationship; when X increases, Y is somewhat more likely to decrease than increase, and vice versa.
- » A correlation of 0 indicates that X and Y are unrelated. When X increases or decreases, it has no direct effect on Y increasing or decreasing, and vice versa.

In the Fahrenheit and Celsius temperatures example in the previous section, the covariance was 96.9 for Celsius temperatures and 313.96 for Fahrenheit temperatures. Although the positive values indicate that the temperatures in both cities tend to increase or decrease at the same time, using the covariance measure alone makes it difficult to judge the *strength* of this relationship. On the other hand, the correlation for both Celsius and Fahrenheit temperatures was 0.4456, showing that a moderately strong, positive relationship exists between the temperatures in the two cities, whether measured in Celsius or Fahrenheit degrees.

In the following sections, you see a type of graph known as a *scatter plot* that is used to illustrate the relationship between two different variables. An extremely important application of correlation is introduced; correlation can be used to show the degree of diversification that is present in a portfolio of stocks. In other words, the correlation can be used to determine how much the addition of a stock to a portfolio will affect the overall risk of the portfolio.

Showing the relationship between two variables

As I discuss in detail in Chapter 2, a *scatter plot* is a special type of graph that shows the relationship between two variables X and Y. The values of X are shown on the horizontal axis, and the values of Y are shown on the vertical axis.

Suppose that X represents a corporation's sales and Y represents its profits. Then X and Y would normally have a positive correlation between them, because higher sales tend to be associated with higher profits and vice versa. Figure 5-1 shows the relationship between two variables with a strong positive correlation.

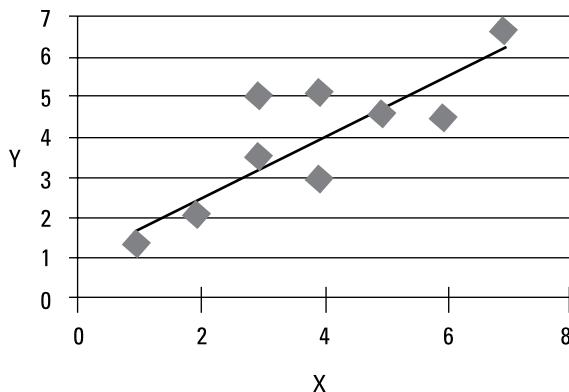
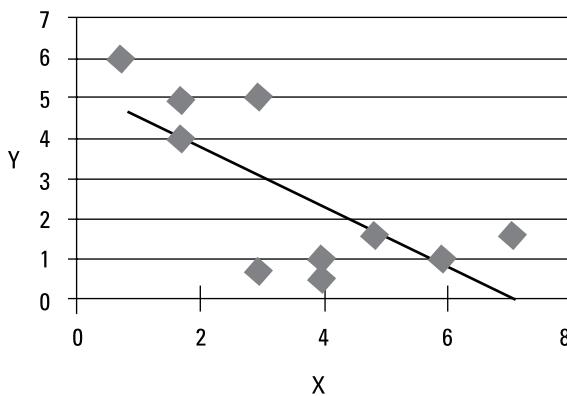


FIGURE 5-1:
Scatterplot
showing a strong
positive
relationship
between X and Y.

Each point on the graph represents a corporation's sales (X) and its profits (Y) during a given year. The graph shows that as X increases, there's a strong tendency for Y to also increase. The straight line is known as a *trend line*. A trend line shows the direction of the points on a scatter plot. It can have a positive slope, a negative slope, or a zero slope (which means that the line is perfectly flat). In this example, the trend line is positively sloped, which indicates that the correlation between X and Y is also positive. Because the points are extremely close to the trend line, the relationship between X and Y is very strong. With a weaker relationship, the points would be more scattered around the trend line.

Suppose that X represents a corporation's costs of production and Y represents its profits; then X and Y would normally have a negative correlation between them, because higher costs tend to be associated with lower profits and vice versa. Figure 5-2 shows the relationship between two variables with a strong negative correlation.

FIGURE 5-2:
Scatterplot
showing a strong
negative
relationship
between X and Y .



Each point on the graph represents a corporation's costs of production (X) and its profits (Y) during a given year. The graph shows that as X increases, there's a strong tendency for Y to decrease. The trend line has a negative slope, which indicates that the correlation between X and Y is negative.

By contrast, suppose that X represents the average daily temperature and Y represents a corporation's profits. Unless the corporation produces goods and services with a seasonal demand, these two variables are likely unrelated. Therefore, the correlation between X and Y will also be close to zero. Figure 5-3 shows the relationship between two unrelated variables.

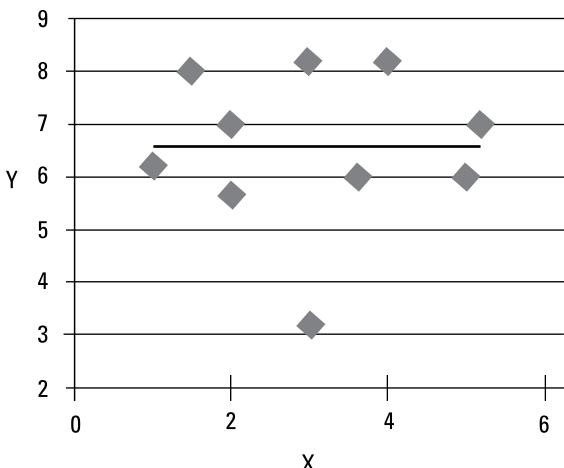


FIGURE 5-3:
Scatterplot
showing two
unrelated
variables.

Each point on the graph represents the average daily temperature (X) and a corporation's profits (Y) during a given year. The graph shows that as X increases, Y sometimes increases and sometimes decreases; no real pattern occurs. The trend line is almost perfectly flat, which indicates that the correlation between X and Y is very close to zero.

Application: Correlation and the benefits of diversification

You can measure the risk of a stock with the standard deviation of its returns. The greater the standard deviation, the further away the returns are from the mean on average (that is, the more “spread out” they are). This indicates more uncertainty over the actual return during a given year, so the risk is greater. You can measure the diversification benefits of adding a stock to a portfolio with the correlation coefficient. The lower the correlation coefficient between two stocks, the *greater* is the reduction in risk and therefore the greater are the benefits of diversification.

For a portfolio of stocks, the risk depends not only on the standard deviations of the individual stocks but also on the *correlations* between the stocks. With low or negative correlations, the portfolio can experience significant reductions in risk, which occurs because losses to some stocks tend to be offset by gains by other stocks at any given time. As a result, the variability of the portfolio's returns tends to be lower than the variability of the returns to the individual stocks.

The following data is a sample of returns to the stocks of Hilo, Inc., and Lohi Corp. over a ten-year span:

Year	Hilo	Lohi
2013	0.03	0.10
2014	0.06	0.10
2015	0.07	0.08
2016	0.09	0.05
2017	0.08	0.04
2018	0.10	0.07
2019	0.09	0.01
2020	0.04	0.02
2021	0.02	0.10
2022	0.06	0.13

Table 5–11 summarizes the sample mean, variance, standard deviation, and coefficient of variation of the stock returns.

TABLE 5-11

Summary Measures for Hilo and Lohi

	Hilo	Lohi
Mean	0.0640	0.0700
Variance	0.0007	0.0015
Standard deviation	0.0272	0.0392
Coefficient of variation (CV)	42.44 percent	55.94 percent

The sample covariance between the stocks is -0.0004 , and the sample correlation coefficient is -0.4179 .

Assume that an investor purchased \$100,000 of each stock for their portfolio at the start of 2013. The returns to the portfolio during this sample period are listed here:

Year	Portfolio
2013	0.065
2014	0.080
2015	0.075
2016	0.070
2017	0.060
2018	0.085
2019	0.050
2020	0.030
2021	0.060
2022	0.095

Because the portfolio is composed of 50 percent Hilo stock and 50 percent Lohi stock, you calculate the returns to the portfolio by multiplying the returns to each individual stock by 0.5 and combining the results, like so:

$$\text{Portfolio return} = 0.5(\text{return to Hilo}) + 0.5(\text{return to Lohi})$$

For example, in 2013, the portfolio return is computed as follows:

$$\text{Portfolio return} = 0.5(0.03) + 0.5(0.10) = 0.065$$

Table 5-12 summarizes the sample mean, variance, standard deviation, and coefficient of variation of the portfolio returns.

TABLE 5-12

Portfolio Summary Measures

Portfolio	
Mean	0.0670
Variance	0.0003
Standard deviation	0.0186
Coefficient of variation (CV)	27.74 percent

The mean return to the portfolio is halfway between the mean returns to Hilo (0.0640) and Lohi (0.0700). The risk of the portfolio, as measured by the standard deviation of the returns, is only 0.0186 compared with Hilo (0.0272) and Lohi (0.0392). As a result, the portfolio's coefficient of variation is only 27.74 percent compared with Hilo at 42.44 percent and Lohi at 55.94 percent.

This substantial reduction in risk is due to the fact that the portfolio is well diversified, as seen by the negative correlation (-0.4179) between the returns to the two stocks.

Computing Covariance and Correlation with the TI-84 Plus Calculator

You can use the Texas Instruments TI-84 Plus and Plus CE calculators to compute the two key measures of association: covariance and correlation. These measures are designed to measure the strength and direction of the relationship between two variables. The calculator does not directly show these measures, but instead provides the terms that are needed to compute covariance and correlation.

As an example, suppose that a stock market analyst is studying the relationship between a corporation's sales and profits. The following table shows the corporation's sales and profits (measured in millions of dollars) over the past five years:

Year	Sales	Profits
1	132	12
2	143	11
3	151	14
4	144	13
5	159	16

Because two variables are being analyzed, the sales data can be entered into List 1 (L1) and the profits data can be entered into List 2 (L2). To access the lists, press [STAT], select EDIT and then 1:Edit, and then press the [ENTER] button.

After clearing out old data by moving to the header of each list with the arrow keys and then pressing [CLEAR] followed by [ENTER], new data is entered into L1 and L2. The next step is to press [STAT] and then select CALC and 2:2-Var Stats.

Because two variables are being analyzed, the 2-Var Stats function is used instead of 1-Var Stats. The following menu appears:

XList:

YList:

FreqList:

Calculate

XList is the location of the data for the independent (X) variable and YList is the location of the data for the dependent (Y) variable. In this example, assume that profits depend on sales, so profits are the dependent (Y) variable and sales are the independent (X) variable. Because no data is repeated, the FreqList is not used.

After entering L1 as the XList and L2 as the YList, select Calculate and then press [ENTER]. This produces several important results. The terms needed to compute covariance and correlation are:

$$\bar{x} = 145.8$$

$$Sx = 10.03$$

$$n = 5$$

$$\bar{y} = 13.2$$

$$Sy = 1.92$$

$$\Sigma xy = 9687$$

\bar{x} is the average sales, while Sx is the sample standard deviation of the sales. \bar{y} is the average profits, while Sy is the sample standard deviation of the profits. n is the sample size (the number of years of data) and Σxy is the result of multiplying the sales by the profits for each year and then adding up the results. The sample covariance can then be computed as:

$$\begin{aligned}s_{xy} &= \left(\frac{1}{n-1} \right) \sum x_i y_i - \left(\frac{n}{n-1} \right) \bar{x} \bar{y} \\ s_{xy} &= \left(\frac{1}{5-1} \right) (9687) - \left(\frac{5}{5-1} \right) (145.8)(13.2) = \left(\frac{1}{4} \right) (9687) - \left(\frac{5}{4} \right) (1924.56) \\ &= 2421.75 - 2405.70 = 16.05\end{aligned}$$

The sample correlation is computed as:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{16.05}{(10.03)(1.92)} = 0.8334$$

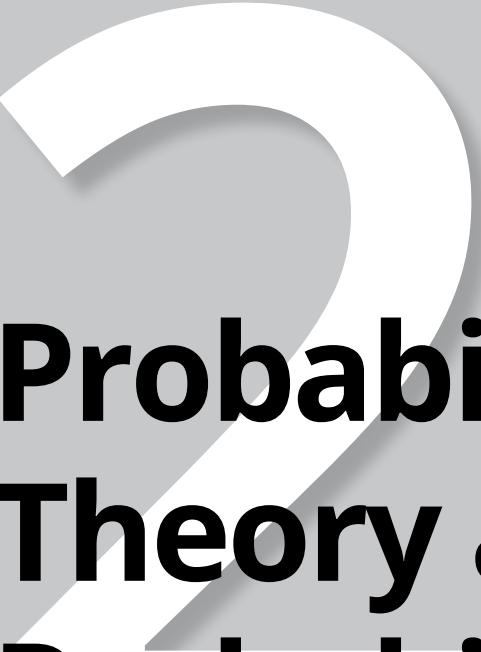
where:

r_{xy} = the sample correlation coefficient

s_{xy} = the sample covariance

s_x = the sample standard deviation of x

s_y = the sample standard deviation of y



Probability Theory and Probability Distributions

IN THIS PART . . .

Review the foundations of probability theory — the foundation of all statistical analysis.

Use random variables and probability distributions to determine if a random event will take place.

Use the binomial distribution to compute probabilities for processes where only one of two possible outcomes may occur. This could be something as simple as flipping a coin several times to see if the coin turns up heads or tails on each flip, or as complicated as a stock price increase.

Describe the rates of return to financial assets, the distribution of corporate profits, and the prices of key commodities (such as oil) using the normal distribution.

Understand two key areas of statistics: sampling and sampling distributions. Most statistical analysis is based on samples randomly drawn from a population.

IN THIS CHAPTER

- » Understanding sets and how they're related to each other
- » Determining the possible outcomes of an experiment
- » Applying types of probabilities
- » Using rules of probability

Chapter **6**

Probability Theory: Measuring the Likelihood of Events

Probability theory is a branch of mathematics that focuses on the analysis of random events and is the foundation of all statistical analysis. You can use probability theory to model a large number of situations that arise in practice. For example, you can use probability theory to estimate how likely it is that a new product will succeed in the marketplace, identify appropriate prices for an insurance company to charge its customers, and more.

This chapter reviews the mathematical foundations of probability theory, such as sets and events, defines types of probabilities, and introduces the rules of probability.

Working with Sets

Probability theory is based on the notion of a *set* — a collection of objects, such as numbers, letters, colors, names, and so on, individually called *elements*. You use mathematical operations, such as membership, subset, union, intersection, and complement, to create new sets from existing ones according to specific rules. For example, you use the operation union to combine two different sets into one new set that contains all the elements from both sets. I explore each of these operations in the following sections.

Membership

Membership indicates whether an element belongs to a set. For example, suppose that set A contains the elements 1 through 6 (the numbers on a die), which is shown mathematically as $A = \{1, 2, 3, 4, 5, 6\}$.

As you can see, the elements or *members* in a set are listed only once, are separated by commas, and are enclosed within *braces*: { }.

In this example, the element 3 belongs to set A . To indicate that an element is part of a set, you use the symbol \in . So $3 \in A$.

On the other hand, to indicate that an element is *not* part of a set, you use the symbol \notin . So in this case, the element 7 doesn't belong to set A , or $7 \notin A$, because it's not listed in the definition (between the braces) of set A .

Subset

A *subset* is a set that's completely contained within a larger set. For example, suppose that sets A and B are defined as follows:

$$A = \{1, 2, 3, 4, 5, 6\}$$

$$B = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

Set A represents the numbers on a die; set B represents the numbers from 1 to 10. In this example, set A is a *subset* of set B because every element of set A is also an element of set B . The symbol \supset represents that one set is a subset of another, as in $A \supset B$.

A *Venn diagram* is used to illustrate the relationship between sets. Sets are represented as circles so that it's easy to see how they're related to each other. If sets overlap, the area common to both sets is shaded. The Venn diagram in Figure 6-1

shows the relationship between sets A and B . The diagram shows that set A is completely contained within set B — that is, A is a subset of B . A is completely shaded because the area of overlap between A and B is A itself.

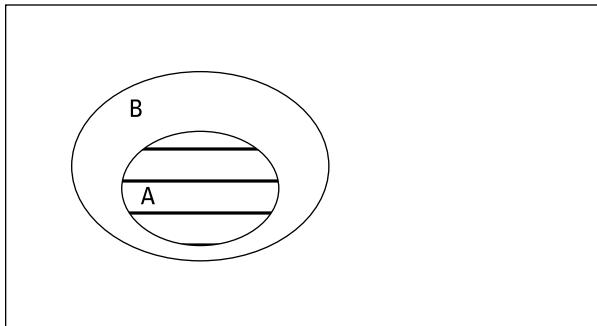


FIGURE 6-1:
Venn diagram
showing that set
 A is a subset
of set B .

As another example, suppose that set C contains the elements 1, 2, 3, 4, 5, and 6 (the numbers on a die), whereas set D contains the elements 1, 2, 3, and 7:

$$D = \{1, 2, 3, 7\}$$

$$C = \{1, 2, 3, 4, 5, 6\}$$

Set D is *not* a subset of set C because the element 7 belongs to set D but *not* to set C ; in mathematical terms, $D \not\subset C$.

Union

Two sets can be combined with a mathematical operation known as *union*. The union of two sets A and B is a set that contains the following:

- » All the elements in set A
- » All the elements in set B

This definition also includes the elements that belong to *both* sets. As an example, suppose that set A contains all the students at a university who are majoring in mathematics; set B contains all the students who are majoring in finance. The union of sets A and B contains all students who are majoring in math *and* all students who are majoring in finance *and* all students who are majoring in *both* (for example, double majors).

As another example, suppose that sets A and B are defined as follows:

$$A = \{2, 4, 6\}$$

$$B = \{1, 2, 3, 4\}$$

The union of these sets is all the numbers on the face of a die except 5:

$$A \cup B = \{1, 2, 3, 4, 6\}$$

The symbol \cup represents union.

The union shows all elements that appear in set A , set B , or both. Note that even though elements 2 and 4 appear in both sets A and B , they're not listed twice in the union; a set contains only *unique* values. The Venn diagram in Figure 6-2 shows the relationship between sets A and B . The shaded region in the diagram represents the union.

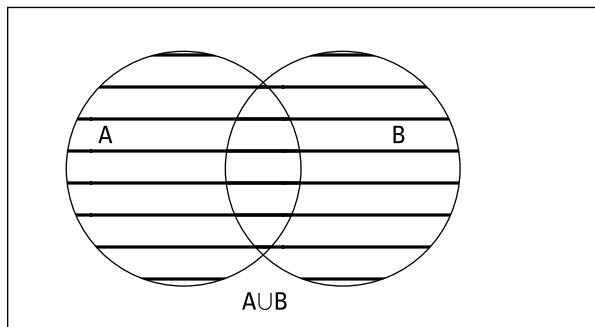


FIGURE 6-2:
Union of two sets.



TIP

The order in which you write the sets is irrelevant; for example, $B \cup A = A \cup B$.

Intersection

The *intersection* of two sets A and B is a set containing the elements that are in *both* sets. For example, suppose that sets A and B are defined as follows:

$$A = \{1, 3, 5, 7\}$$

$$B = \{3, 6, 7\}$$

The intersection of these sets is $A \cap B = \{3, 7\}$.

The intersection of A and B contains the elements 3 and 7 because these elements belong to *both* A and B . The symbol \cap represents intersection. The Venn diagram in Figure 6-3 shows the relationship between A and B . The shaded region in the diagram represents the intersection of these sets.

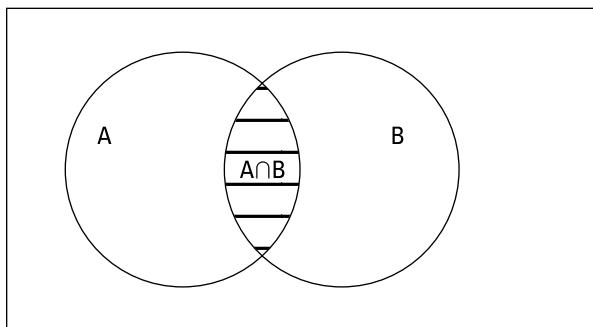


FIGURE 6-3:
Intersection of
two sets.

As another example, suppose that set C contains the elements 2, 4, 6:

$$A = \{1, 3, 5, 7\}$$

$$C = \{2, 4, 6\}$$

The intersection of these sets is $A \cap C = \{ \}$.

The intersection of sets A and C contains *no elements* because the sets don't have any of the same elements. The set containing no elements, or $\{ \}$, is known as an *empty set*. Two sets that have no elements in common are said to be *mutually exclusive*. The Venn diagram in Figure 6-4 shows the relationship between sets A and C . This diagram has no shaded region because the intersection of sets A and C contains no elements.

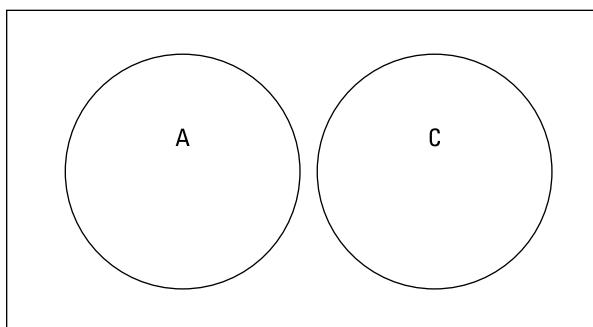


FIGURE 6-4:
An intersection
between two sets
containing no
elements.

Complement

The mathematical operation *complement* is based on the notion of a *universal set* or *sample space* — all the elements that a set may contain. For example, suppose that you roll a single die; the number that turns up may be any whole number between 1 and 6. Assume that set A contains the odd numbers that may turn up when you roll a die, and set B contains the even numbers:

$$A = \{1, 3, 5\}$$

$$B = \{2, 4, 6\}$$

In this case, the sample space contains all possible numbers that may turn up when you roll the die:

$$S = \{1, 2, 3, 4, 5, 6\}$$

The complement of set A is the set of all numbers that are elements of the sample space but *not* elements of A :

$$A^c = \{2, 4, 6\}$$

A^c is the set “ A complement.” It contains the elements 2, 4, and 6 because they *don’t* belong to set A , and they *do* belong to the sample space.

Note that elements such as 7, 8, 9, and so on aren’t elements of A^c because they’re not elements of set A , but they’re also not elements of the sample space. The complement of A is shown in the Venn diagram in Figure 6–5. The shaded region shows all the elements in the sample space that don’t belong to set A . Similarly, the complement of B is $B^c = \{1, 3, 5\}$.

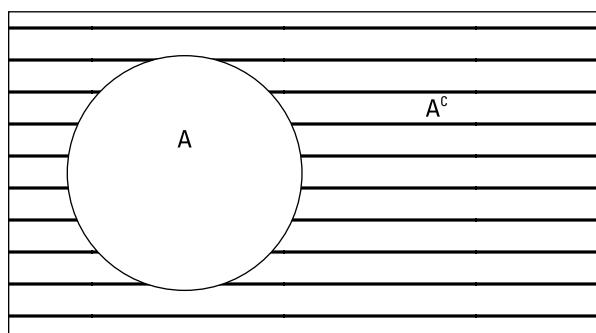


FIGURE 6-5:
Set A and its
complement A^c .

Betting on Uncertain Outcomes

Probability theory is based on the premise that a process generates uncertain (random) outcomes. This process is sometimes known as a *random experiment*, such as the following examples:

- » A roulette wheel is spun. The outcome can be a 0, a 00 (“double zero”), or any number between 1 and 36.
- » A lottery drawing results in a single winning number being chosen.
- » A futures contract trades throughout the day, resulting in a settlement price at the close of trading.

In each case, the outcome isn’t known in advance. Using probability, you can determine the likelihood of a specific outcome, such as the likelihood of getting an even number from a single spin of a roulette wheel. In this section, I introduce several key terms, along with an introduction to computing probabilities.

The sample space: Everything that can happen

A *sample space* is another name for the universal set (described in the earlier section “Complement”); it contains all the outcomes that can result from a random experiment. For example, suppose you flip a coin two times. The possible outcomes of this random experiment are:

- » Heads followed by heads (HH)
- » Heads followed by tails (HT)
- » Tails followed by heads (TH)
- » Tails followed by tails (TT)

The sample space for this random experiment is $S = \{HH, HT, TH, TT\}$. It includes all the possible outcomes.

Event: One possible outcome

An *event* is one possible outcome of a random experiment. More formally, it is a subset of the sample space. For example, in the coin-flipping experiment, suppose that the event $E = “2 tails turn up.”$ Event E is a set containing the element TT, or in mathematical terms, $E = \{TT\}$.

Event E is a subset of the sample space because it's completely contained within the sample space. As another example, suppose that the event F = "at least 1 head turns up." Event F is a set containing the elements HH, HT, TH, or $F = \{HH, HT, TH\}$.

In some cases, events may be related to each other. Two key ways in which events may be related to each other are known as mutually exclusive and independent. These are described in the following section.

Mutually exclusive events

Two events are said to be *mutually exclusive* if they can't both happen at the same time. Here are two events that are mutually exclusive:

A = The roll of a die is odd.

B = The roll of a die is even.

Clearly, the roll of a die must result in a number that is either odd or even; it can't be both. Therefore, events A and B are mutually exclusive.

As another example, based on the coin-flipping experiment, suppose that two events are defined:

G = Two heads turn up.

H = Two tails turn up.

It's impossible for *both* two heads to turn up *and* two tails to turn up. This means that G and H are mutually exclusive. This result can be demonstrated using sets as follows:

$G = \{HH\}$ and $H = \{TT\}$. These events have no elements in common; their intersection is the *empty set* $G \cap H = \{\}$.

The probability of the empty set is zero; therefore, the event that both G and H occur is *impossible*. This means that G and H are mutually exclusive.

Independent events

Two events A and B are said to be *independent* if the outcome of event A doesn't affect the outcome of event B and vice versa. For example, suppose that based on the coin-flipping experiment, event A is defined as the event that the first flip is a head, and event B is defined as the event that the second flip is a head. In other words:

$$A = \{\text{HH}, \text{HT}\}$$

$$B = \{\text{HH}, \text{TH}\}$$

Because the outcome of the first flip has no influence over the outcome of the second flip, events A and B are *independent* events. (See a more formal test of independence in the next section.) Note that A and B are not mutually exclusive; both A and B can occur.

Computing probabilities of events

If a sample space contains elements that are all equally likely to occur, then computing the probabilities of events is straightforward. For example, for the coin-flipping experiment in the earlier sections “The sample space: Everything that can happen” and “Event: One possible outcome,” these probabilities exist:

- » $P(\text{HH}) = 0.25$
- » $P(\text{HT}) = 0.25$
- » $P(\text{TH}) = 0.25$
- » $P(\text{TT}) = 0.25$

For example, the probability of getting two consecutive heads is $\frac{1}{4}$ (which equals 0.25). This is because HH is one of four possible outcomes when a coin is flipped twice. Furthermore, each outcome is equally likely to occur (because heads and tails are equally likely). Therefore, each outcome has a probability of $\frac{1}{4} = 0.25$.

$$\text{One possible outcome} \div 4 \text{ possibilities} = 0.25$$

As an example, suppose that the event K is defined as “at least one tail turns up.” Then event K contains the elements HT, TH and TT, or $K = \{\text{HT}, \text{TH}, \text{TT}\}$. When all outcomes of an experiment are equally likely, you find the probability of event K with this formula:

$$P(K) = \frac{\text{elements in } K}{\text{elements in } S}$$

Because event K contains three elements and the total number of elements in the sample space is four, $P(K) = 3/4 = 0.75$.

Based on this formula, the probability of the empty set is 0, and the probability of the entire sample space is 1. For example, suppose that event A is an impossible event. It is represented by a set containing no elements (the empty set). The sample space contains the elements 1, 2, and 3. The probability of A is, therefore,

$$P(A) = \frac{\text{elements in } A}{\text{elements in } S} = \frac{0}{3} = 0$$

The probability of S is:

$$P(S) = \frac{\text{elements in } S}{\text{elements in } S} = \frac{3}{3} = 1$$

Looking at Types of Probabilities

The three basic types of probabilities are:

- » Unconditional (marginal) probabilities: When events are independent
- » Joint probabilities: When two things happen at once
- » Conditional probabilities: When one event depends on another

In this section, you find out about each of these types of probabilities, and you also discover how you can use conditional probabilities to determine whether two events are independent.

Unconditional (marginal) probabilities: When events are independent

The *unconditional (marginal) probability* of an event is found as a row total or a column total in a joint probability table. As an example, Table 6–1 is a joint probability table, representing the distribution of students in a business school according to major and whether they’re working on a bachelor’s degree or a master’s degree. In this section, I show you how to use data like this to find unconditional probabilities.

Based on Table 6–1, the following events are defined:

- » B = pursuing a bachelor’s degree
- » M = pursuing a master’s degree
- » F = majoring in finance
- » A = majoring in accounting
- » T = majoring in marketing

TABLE 6-1

Joint Probability Table Showing the Distribution of Business Students

	Majoring in Finance	Majoring in Accounting	Majoring in Marketing	Total
Bachelor's degree	0.26	0.36	0.18	0.80
Master's degree	0.09	0.07	0.04	0.20
Total	0.35	0.43	0.22	1.00

You can find the unconditional probabilities of the following events directly from Table 6-1:

- » $P(B)$ = the probability of pursuing a bachelor's degree
- » $P(M)$ = the probability of pursuing a master's degree
- » $P(F)$ = the probability of majoring in finance
- » $P(A)$ = the probability of majoring in accounting
- » $P(T)$ = the probability of majoring in marketing

Suppose you want to find the probability that a randomly chosen business student is pursuing a bachelor's degree. In other words, you want to calculate $P(B)$. Referring to Table 6-1, you look at the first row (which refers to students pursuing their bachelor's degrees). The row total is 0.80. This is the probability that a randomly chosen student is pursuing a bachelor's degree.

Suppose you want to know the probability that a randomly chosen student is majoring in finance. In other words, you want to calculate $P(F)$. Referring to Table 6-1, you look at the first column (which refers to students majoring in finance). The column total is 0.35. This is the probability that a randomly chosen student is majoring in finance.

You can find the remaining unconditional probabilities in the same way. These are:

$$P(M) = 0.20$$

$$P(A) = 0.43$$

$$P(T) = 0.22$$

Joint probabilities: When two things happen at once

The probability that two different events occur at the same time is known as a *joint probability*. For example, the probability that a student is working on a bachelor's degree *and* is majoring in finance is a joint probability.

As you study Table 6-1, you can see that the intersection of two different events can determine joint probabilities. For example, to find the probability that a randomly chosen business student is pursuing a bachelor's degree *and* is majoring in finance, take the intersection of events B and F. This equals $P(B \cap F) = 0.26$.

You find the remaining joint probabilities in the same way:

$$P(B \cap A) = 0.36$$

$$P(B \cap T) = 0.18$$

$$P(M \cap F) = 0.09$$

$$P(M \cap A) = 0.07$$

$$P(M \cap T) = 0.04$$

Conditional probabilities: When one event depends on another

The *conditional probability* of an event is defined as the probability of an event *given that* another event has occurred. For example, the probability that a student is working on a bachelor's degree *given that* the student is majoring in accounting is a conditional probability. This is written as follows:

$$P(B|A)$$

The symbol “|” is used to indicate a conditional probability. (You pronounce this expression as “the probability of B *given* A.”)

To find the conditional probability of an event, you set up the ratio of a joint probability to an unconditional (marginal) probability (see previous sections on these types of probabilities). For example, say you want to find out what the probability is that a student who's known to be pursuing a bachelor's degree is majoring in marketing. Referring to Table 6-1, you first calculate the joint probability of pursuing a bachelor's degree and majoring in marketing, as follows:

$$P(B \cap T) = 0.18$$

Then you find that the unconditional probability of pursuing a bachelor's degree equals $P(B) = 0.80$. Therefore,

$$\frac{P(B \cap T)}{P(B)} = \frac{0.18}{0.80} = 0.225$$

As another example, to find the probability that an accounting major is pursuing a master's degree you take the joint probability of these two events:

$$P(M \cap A) = 0.07$$

The unconditional probability of majoring in accounting equals $P(A) = 0.43$. Therefore,

$$P(M | A) = \frac{P(M \cap A)}{P(A)} = \frac{0.07}{0.43} = 0.163$$

Determining independence of events

You can use conditional probabilities to determine whether two events are independent. Two events are independent if the probability of one event occurring doesn't influence the probability of the other occurring, and vice versa.

To prove independence, the following two conditions must be met:

$$P(A | B) = P(A)$$

$$P(B | A) = P(B)$$

Using the business students example from the earlier section "Joint probabilities: When two things happen at once" and referring to Table 6-1, you can determine whether the events "majoring in accounting" (A) and "pursuing a bachelor's degree" (B) are independent events.

The first step is to compute the conditional probabilities $P(A|B)$ and $P(B|A)$: The joint probability of events A and B is $P(A \cap B) = 0.36$. The unconditional probabilities of events A and B are

$$P(A) = 0.43$$

$$P(B) = 0.80$$

Therefore,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.36}{0.80} = 0.45$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.36}{0.43} = 0.84$$

because $P(A|B)$ must equal $P(A)$ and $P(B|A)$ must equal $P(B)$ for the two events to be independent. The results show that $P(A|B) = 0.45$, $P(A) = 0.43$, $P(B|A) = 0.84$, and $P(B) = 0.80$, so both conditions fail. Events A and B are *not* independent of each other; in other words, they're *dependent* on each other. Therefore, the decision to pursue a bachelor's or a master's degree appears to influence the choice of major.

Following the Rules: Computing Probabilities

In addition to computing joint, conditional, and unconditional probabilities (discussed in the previous sections), the following three rules can help you determine other probabilities:

- » The **addition rule** shows the probability of the union of two events.
- » The **complement rule** determines the probability of the complement of an event.
- » The **multiplication rule** identifies the probability of the intersection of events.

Addition rule

You use the addition rule to compute the probability of the union of two events. Mathematically speaking, for events A and B, the addition rule states that

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

This shows that the probability of the union of events A and B equals the sum of the probability of A and the probability of B, from which the probability of *both* events is subtracted. Subtracting the probability of both events is necessary to avoid the problem of *double-counting*. This is shown in the following example:

Suppose that event A contains the elements 1, 2, 3 and event B contains the elements 3, 4, 5. The sample space contains the elements 1, 2, 3, 4, 5.

$$\begin{aligned}A &= \{1, 2, 3\} \\B &= \{3, 4, 5\} \\S &= \{1, 2, 3, 4, 5\}\end{aligned}$$

Assuming that all elements are equally likely to be chosen, the corresponding probabilities are:

$$\begin{aligned}P(A) &= 3/5 \\P(B) &= 3/5 \\P(S) &= 5/5 = 1\end{aligned}$$

The union of A and B contains all the elements in the sample space:

$$A \cup B = \{1, 2, 3\} \cup \{3, 4, 5\} = \{1, 2, 3, 4, 5\} = S$$

As a result, the probability of A union B equals 1. (Recall that the sample space always has a probability of 1.) If you simply combine the probabilities of A and B , though, you will get a surprising result; they sum to $6/5$, which is greater than one.

$$P(A) + P(B) = 3/5 + 3/5 = 6/5$$

This result occurs because the element 3 appears in both A and B :

$$A \cap B = \{3\}$$

The probability of 3 was counted *twice*, one in set A and once in set B , which accounts for the sum of the probabilities being greater than one. By subtracting the probability of the element 3, the correct probability of one is found.

$$\begin{aligned}P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\P(A \cup B) &= 3/5 + 3/5 - 1/5 = 5/5 = 1\end{aligned}$$

Table 6-2 shows the distribution of coffees (measured in pounds) the Big Bean Corporation produces during a given day.

TABLE 6-2

Joint Probability Distribution for Coffee Styles

	Special Reserve Blend (S)	Kona Hawaii Blend (K)	Aromatic Blend (A)	Total
Decaffeinated (D)	0.12	0.08	0.22	0.42
Regular (R)	0.24	0.12	0.22	0.58
Total	0.36	0.20	0.44	1.00

If you choose a pound of coffee randomly from the daily output of the Big Bean Corporation, what's the probability that it's either the *Special Reserve Blend* (*S*) or the *Regular* (*R*) (or both)?

In this example, you use the addition rule because you're being asked to compute the probability of a union. You combine the probability of *S* with the probability of *R*, subtracting the intersection between them to avoid the problem of double-counting:

$$P(S \cup R) = P(S) + P(R) - P(S \cap R)$$

From Table 6–2, you can determine that $P(S) = 0.36$; that $P(R) = 0.58$; $P(S \cap R) = 0.24$. Therefore,

$$\begin{aligned} P(S \cup R) &= P(S) + P(R) - P(S \cap R) \\ &= 0.36 + 0.58 - 0.24 = 0.70 \end{aligned}$$

Seventy percent of the coffee produced by Big Bean is either the special reserve blend, regular, or both.

When two events *A* and *B* are *mutually exclusive* (that is, they can't both occur at the same time), the addition rule simplifies to

$$P(A \cup B) = P(A) + P(B) \text{ because } P(A \cap B) = 0.$$

For example, if you choose a pound of coffee randomly from the daily output of the Big Bean Corporation, what's the probability that it's either the *Kona Hawaii Blend* (*K*) or the *Aromatic Blend* (*A*)?

Because a pound of coffee can't be both the *Kona Hawaii Blend* and the *Aromatic Blend*, events *K* and *A* are mutually exclusive. This means that you can use the simplified version of the addition rule:

$$\begin{aligned} P(K \cup A) &= P(K) + P(A) \\ P(K \cup A) &= 0.20 + 0.44 = 0.64 \end{aligned}$$

Complement rule

The complement rule is expressed as follows:

$$P(A^C) = 1 - P(A)$$

A^C is the complement of event *A*.

Two events are said to be complements if they are mutually exclusive *and* their union equals the entire sample space. Here's an example: Suppose that an experiment consists of choosing a single card from a standard deck. Event A = "the card is red." Event B = "the card is black." Events A and B are complements because A and B are mutually exclusive (no card can be *both* red and black). The union of A and B is the sample space (the entire deck, because all cards must be either red or black, so the union of A and B equals the entire sample space).

In the Big Bean example from the previous section, the complement of event D (decaffeinated coffee) is event R (regular coffee) because all coffee must be either decaffeinated or regular, and no coffee can be *both*. You can find the probability of the complement of D as follows:

$$P(D^C) = 1 - P(D)$$

Referring to Table 6–2, you can see that $P(D) = 0.42$. Therefore, $P(D^C) = 1 - P(D) = 1 - 0.42 = 0.58$, which is equal to $P(R)$.

Multiplication rule

To figure out the probability of the intersection of two events, you use the multiplication rule. This is used to determine the probability that two events are *both* true. For example, suppose an experiment consists of choosing a card from a standard deck. Event A = "the card is red." Event B = "the card is a king." The multiplication rule can be used to determine the probability that the card is *both* red and a king (for example, a red king).

The multiplication rule can be written in two equivalent ways:

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

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

Note that these formulas are simply algebraic rearrangements of the definition of conditional probability:

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

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

Suppose the Omega Corporation has been the subject of takeover rumors for several months. The takeover is far more likely to occur if the economy rebounds next year. Omega's chief economist estimates that the likelihood of strong growth next year is 5 percent, the likelihood of weak growth is 35 percent, and the likelihood of negative growth is 60 percent. The likelihood of a takeover during a period

of strong growth is estimated to be 40 percent; during a period of weak growth, this falls to 20 percent; and during a period of negative growth, it's assumed to be only 5 percent. What is the probability that there is strong growth next year *and* Omega is taken over?

The following events are defined:

- » S = "strong growth"
- » W = "weak growth"
- » N = "negative growth"
- » T = "Omega is taken over"

The probability of the events S and T can be determined as follows:

$$P(T \cap S) = P(T|S)P(S)$$

Because there's a 5 percent chance of strong growth next year, $P(S) = 0.05$. The likelihood of a takeover during a period of strong growth is estimated to be 40 percent. Therefore, $P(T|S) = 0.40$. So the probability that there's strong growth next year *and* that Omega is taken over is

$$P(T \cap S) = P(T|S)P(S) = (0.40)(0.05) = 0.02$$

When two events A and B are *independent*, the multiplication rule simplifies to

$$\begin{aligned}P(A \cap B) &= P(A)P(B) \\P(A \cap B) &= P(B)P(A)\end{aligned}$$

This is because $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

IN THIS CHAPTER

- » Understanding the concept of the random variable
- » Describing the behavior of a random variable with a probability distribution
- » Summarizing the properties of a random variable with moments

Chapter 7

Probability Distributions and Random Variables

This chapter introduces two new concepts that are used to determine the probability that a random event takes place — random variables and probability distributions. These concepts are closely related to the notion of the *random experiment* (defined in Chapter 6). A random experiment is a *process* in which events unfold in an unpredictable way. A random variable is used to assign numerical values to all the possible outcomes of a random experiment. A probability distribution assigns probabilities to these numerical values.

In this chapter, I also define summary measures of a probability distribution, known as *moments*, such as expected value and variance. Random variables and probability distributions are used by economists, financial analysts, researchers, and others to model the behavior of economic and financial variables, such as interest rates, inflation rates, corporate earnings, and so on.

Defining the Role of the Random Variable

A random variable is based on a random experiment, a process that generates outcomes that aren't known in advance (see Chapter 6). For example, suppose that a game of chance consists of spinning a wheel with four colors — red, blue, green, and yellow — each color results in a prize ranging from \$1.00 to \$10.00. A random variable may be used to assign a prize value to each color. For example, you can define X to represent the prize that is received for each color, as follows:

red	$X = \$1$
blue	$X = \$2$
green	$X = \$5$
yellow	$X = \$10$

In this example, the random experiment consists of spinning the wheel. For each possible outcome (color), X assigns a numerical value that represents the prize received.

It may seem like a paradox, but a random variable is neither random nor a variable! In fact, a random variable is a *function*. It assigns a single numerical value to each outcome of a random experiment. Random variables may represent a large number of different financial and economic variables, including the following:

- » A corporation's profits during the upcoming quarter
- » The number of new customers resulting from a new advertising campaign
- » The value of the Dow Jones Industrial Average at the end of next year

As another example, suppose you conduct a simple random experiment by flipping a coin three times. The set of all possible outcomes, known as the sample space, consists of the following elements. (H represents a head turning up on a single flip of the coin, and T represents a tail turning up.)

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

S represents the sample space. Each element in the sample space is a single sequence of three flips; for example, HTH refers to a head followed by a tail followed by another head.

Because a head and a tail are equally likely to occur on each flip, each outcome of this random experiment is also equally likely to occur. For example, *HHT* is just as likely to happen as *THT*. With eight equally likely outcomes, each has a probability of $1/8$ or 0.125 .

An event is one outcome or a combination of outcomes of a random experiment. For example, suppose that you want to calculate the probability of the event E , where two or more heads turn up. This outcome can occur in four ways:

- » Three consecutive heads (*HHH*)
- » Two heads followed by one tail (*HHT*)
- » A head followed by a tail followed by another head (*HTH*)
- » A tail followed by two heads (*THH*)

You can express these possible outcomes more compactly with set notation:

$$E = \{\text{HHH}, \text{HHT}, \text{HTH}, \text{THH}\}$$

Since all the outcomes are equally likely, to compute the probability of the event E , you count the number of elements that correspond to event E and divide by the number of elements in the entire sample space (S):

$$P(E) = \frac{\text{elements in } E}{\text{elements in } S} = \frac{4}{8} = \frac{1}{2}$$

$P(E)$ is the probability of event E .

If the outcomes are not equally likely, then:

$$P(E) = P(\text{HHH}) + P(\text{HHT}) + P(\text{HTH}) + P(\text{THH})$$

This approach can be extremely cumbersome if the sample space contains a large number of elements. As an alternative, you can define a random variable to represent the number of heads that turn up during the random experiment. You can then determine the probability of event E from the probabilities of the different possible values of the random variable.

For example, let the random variable X equal the number of heads that turn up when a coin is flipped three times. X has a numerical value for each outcome of this experiment. Here are the outcomes of the experiment and the corresponding values of X .

Outcome	X
<i>HHH</i>	3
<i>HHT</i>	2
<i>HTH</i>	2
<i>THH</i>	2
<i>HTT</i>	1
<i>THT</i>	1
<i>TTH</i>	1
<i>TTT</i>	0

For example, *HHT* represents two heads followed by a tail; therefore, the value of X for *HHT* is 2. Similarly, for the outcome *TTH*, the value of X is 1.

Suppose that a marketing firm conducts a survey of customers to determine whether they're satisfied with the customer service received from the local cable company. Each customer answers yes or no. The survey yielded the following replies:

yes	no	no	yes
no	yes	yes	yes
yes	yes	yes	yes
yes	yes	no	yes
no	no	no	no

For the results, X is defined as follows:

$X = 0$: the customer reply is no

$X = 1$: the customer reply is yes

The results are shown in Table 7-1.

By organizing the results this way, you can easily see the proportion of the customers who are satisfied with their cable service.

TABLE 7-1**Survey Responses**

Number of Responses	X (0 = no, 1 = yes)
8	0
12	1

Assigning Probabilities to a Random Variable

Although random variables may provide useful information, their greatest advantage is that they simplify the calculation of probabilities. For example, in the case of the coin-flipping experiment in the previous section, computing probabilities directly from the values of a random variable is simpler than counting up all the ways in which an event can occur.

You can assign probabilities to each possible value of a random variable by using a *probability distribution* — a table or formula that shows these probabilities. A probability distribution has two important properties:

- » The probability of each value of a random variable is between 0 and 1.
- » The sum of the probabilities equals 1.

In the following sections, I show you how to construct a probability distribution. I also show you how to illustrate the properties of a probability distribution with a special type of graph known as a *histogram*.

Calculating the probability distribution

Based on the coin flip example in the earlier section, “Defining the Role of the Random Variable,” the range of possible values for X (the number of heads that turn up) is 0 to 3. Here is the number of ways in which each possible value of X may occur:

X	Outcomes
0	TTT
1	HTT, THT, TTH
2	HHT, HTH, THH
3	HHH

Because eight equally likely outcomes of this experiment can occur, the probability for each value of X equals the number of outcomes divided by the size of the sample space. Table 7–2 shows this probability distribution.

TABLE 7-2

Probability Distribution for the Coin-Flipping Experiment

X	P(X)
0	$1/8 = 0.125$
1	$3/8 = 0.375$
2	$3/8 = 0.375$
3	$1/8 = 0.125$

The probability distribution in Table 7–2 shows that

- » The probability of getting no heads ($X = 0$) is 0.125.
- » The probability of getting one head ($X = 1$) is 0.375.
- » The probability of getting two heads ($X = 2$) is 0.375.
- » The probability of getting three heads ($X = 3$) is 0.125.

Now, suppose that you want to calculate the probability of the event F , where two or more tails turn up. This outcome can occur in four ways:

- » Three consecutive tails (TTT)
- » Two tails followed by a head (TTH)
- » A tail followed by a head followed by another tail (THT)
- » A head followed by two tails (HTT)

The event F corresponds to a set containing four elements:

$$F = \{TTT, TTH, THT, HTT\}$$

For two or more tails to turn up, the experiment must result in either zero heads or one head. Therefore, you can calculate the probability of F as follows:

$$P(F) = P(X = 0) + P(X = 1) = 0.125 + 0.375 = 0.500$$

Visualizing a probability distribution with a histogram

You can express the probability distribution for the coin-flipping experiment graphically with a *histogram*. A histogram is a graph in which you place individual values or ranges of values on the horizontal axis and the frequency of occurrence for each value or range of values on the vertical axis.

The histogram for the probability distribution of the coin-flipping experiment is shown in Figure 7-1. The vertical axis shows the probability of X , and the horizontal axis shows the value of X (that is, the number of heads).

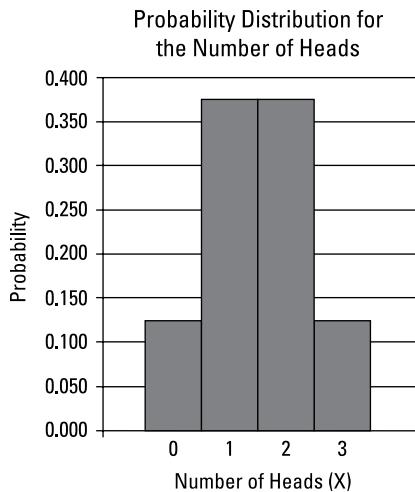


FIGURE 7-1:
Histogram showing the probability distribution of the number of heads.

The histogram shows that the two most likely outcomes of this experiment are one head or two heads ($X = 1$ or $X = 2$); these are equally likely to occur. The least likely outcomes are no heads or three heads ($X = 0$ or $X = 3$); these are also equally likely to occur.

Characterizing a Probability Distribution with Moments

Recall from Chapters 3 through 5 that the properties of samples and populations may be summarized in a convenient form with a series of numerical measures, including the mean, variance, standard deviation, and so on. The properties of a

probability distribution can also be summarized with a set of numerical measures known as *moments*.

In this section, I cover the most important of these moments: expected value (mean) and the variance. (The standard deviation isn't a separate moment; it's the square root of the variance.) First, though, I explain the role of the summation operator in calculating these moments.

Understanding the summation operator (Σ)

The summation operator is used to indicate that a set of values should be added together. (The summation operator was introduced in Chapter 3.) The formulas used to compute moments for a probability distribution are based on the summation operator. This is because each calculation must be repeated for each possible value of a random variable and the results must be summed.

As an example of the summation operator, suppose that a data set contains five elements. The summation operator tells you to perform the following calculations:

$$\sum_{i=1}^n X_i = X_1 + X_2 + X_3 + X_4 + X_5$$

X_i represents a single element in a data set; i is an *index*, and n is the number of elements to be summed.

Expected value

The *expected value* of a random variable X represents the average value of X that occurs if the random experiment is repeated a large number of times. You can think of the expected value as the *center* of the distribution.



REMEMBER

The expected value is a *weighted average* of its possible values, with weights equal to probabilities. The formula for computing expected value of X is

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

Here are the key terms in this formula:

- » $E(X)$ = the expected value of X
- » n = the number of possible values of X
- » i = an index
- » X_i = one possible value of X
- » $P(X_i)$ = the probability of X_i
- » Σ = the summation operator used to indicate that a sum is being computed

Suppose that a biopharmaceutical firm is planning to release several new drugs during the coming year, depending on whether or not the patents are approved. You can use the random variable X to represent the number of new drugs that will be released. Table 7-3 shows the probability distribution of these results.

TABLE 7-3

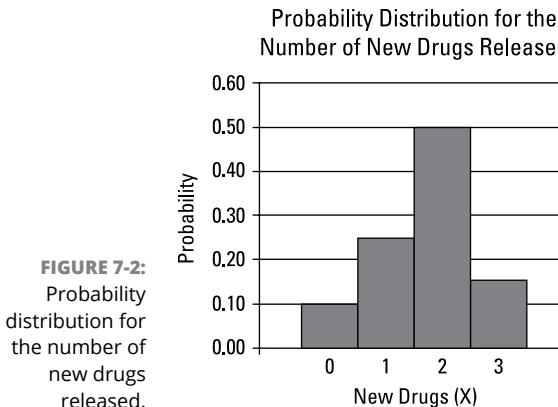
Probability Distribution for Release of New Drugs

X	P(X)
0	0.10
1	0.25
2	0.50
3	0.15

You can then use the probability distribution to determine the expected (average) value of X by setting up the possible values of X and the corresponding probabilities, like so:

$X_1 = 0$	$P(X_1) = 0.10$
$X_2 = 1$	$P(X_2) = 0.25$
$X_3 = 2$	$P(X_3) = 0.50$
$X_4 = 3$	$P(X_4) = 0.15$

The corresponding histogram is shown in Figure 7-2.



Next, you substitute these numbers into the expected value formula:

$$\begin{aligned} E(X) &= \sum_{i=1}^n X_i P(X_i) \\ &= X_1 P(X_1) + X_2 P(X_2) + X_3 P(X_3) + X_4 P(X_4) \\ &= (0)(0.10) + (1)(0.25) + (2)(0.50) + (3)(0.15) \\ &= 0.00 + 0.25 + 1.00 + 0.45 \\ &= 1.70 \end{aligned}$$

This result shows that the expected (average) number of new drugs that will be released during the coming year is 1.7. Although it's physically impossible to release 1.7 new drugs (since 1.7 is not an *integer* or whole number), if this experiment is repeated many times, the average number of new drugs released will be 1.7.

Variance and standard deviation

The *variance* of a random variable X is the average squared distance between the values of X and the expected value of X . In other words, variance is the amount of “spread” among the different values of X . The standard deviation is simply the square root of the variance. Note that the variance and standard deviation of a random variable are equivalent to the variance and standard deviation of a sample or population (discussed in Chapter 4).

The formula for computing the variance of X is

$$\sigma^2 = \sum_{i=1}^n [X_i - E(X)]^2 P(X_i)$$



REMEMBER

σ^2 represents the variance of X .

This expression tells you to perform the following calculations:

- » For each possible value of X (X_i), subtract the expected value of X .
- » Square the result.
- » Multiply this expression by the probability of X_i .
- » Compute the sum of these products.

For the example of the biopharmaceutical company (in the earlier section, “Understanding the summation operator [Σ]”) you compute the variance like so:

$$\begin{aligned}\sigma^2 &= \sum_{i=1}^n [X_i - E(X)]^2 P(X_i) \\ &= [0 - 1.7]^2 (0.1) + [1 - 1.7]^2 (0.25) + [2 - 1.7]^2 (0.5) + [3 - 1.7]^2 (0.15) \\ &= [2.89](0.1) + [0.49](0.25) + [0.09](0.5) + [1.69](0.15) \\ &= 0.2890 + 0.1225 + 0.0450 + 0.2535 \\ &= 0.71\end{aligned}$$

One of the major drawbacks to the variance is that it's measured in squared units, which makes interpretation difficult. In this example, the variance of the number of new drugs that will be released next year is 0.7100 drugs squared. It's hard to visualize what “drugs squared” actually means. As a result, the standard deviation is normally used in place of variance as a measure of spread. By taking the square root of 0.7100 drugs squared, you get a result of 0.8426 drugs, which is much more intuitively clear. For the example of the biopharmaceutical company, the standard deviation of the number of new drugs released next year equals

$$\begin{aligned}\sigma &= \sqrt{\sum_{i=1}^n [X_i - E(X)]^2 P(X_i)} \\ &= \sqrt{0.7100} \\ &= 0.8426\end{aligned}$$

The standard deviation is 0.8426 new drugs. You can think of the standard deviation as a measure of how much uncertainty is associated with the expected value.



REMEMBER

σ represents the standard deviation of X .

IN THIS CHAPTER

» Finding probabilities when only two things can happen with the binomial distribution

» Using the Poisson distribution to calculate the probability of events occurring during a given time frame

Chapter 8

The Binomial and Poisson Distributions

You can model many complex business problems by using probability distributions. These distributions help provide answers to questions such as, “What’s the likelihood that oil prices will rise during the coming year?” “What’s the probability of a stock market crash next month?” “How likely is it that a corporation’s earnings will fall below expectations this year?”

A probability distribution defines the statistical properties of a variable. Accurate modeling of financial variables requires that you pick the appropriate distribution for a given situation. Two of the more widely used probability distributions in business are the binomial and Poisson distributions. These are examples of discrete distributions, in which only a countable number of values are possible.

This chapter covers the key properties of the binomial and Poisson distributions and explains the circumstances under which you may apply them. For each distribution, I give you formulas for computing probabilities and also provide tables as alternatives to doing the computing yourself. This chapter also introduces summary measures of probability distributions, known as *moments*, which are closely related to the mean, variance, and standard deviation of samples and populations (described in Chapters 3 and 4). Then I wrap up the chapter by covering simplified formulas for computing the moments of the binomial and Poisson distributions.

Looking at Two Possibilities with the Binomial Distribution

You use the *binomial distribution* to compute probabilities for processes where only one of two possible outcomes may occur. (The fact that only two possible outcomes can occur is what gives the distribution its name.) Here are some examples of processes you can model with the binomial distribution:

- » When you flip a coin several times, the outcome of interest is whether the coin turns up heads or tails on each flip.
- » When you roll a die multiple times, the outcome of interest is whether the number that turns up on each roll is odd (1, 3, or 5) or even (2, 4, or 6).
- » When you look at the closing price of a stock each day for one year, the outcome of interest is whether the stock price increased or not.

As another example, suppose that you hold a portfolio of stocks. During the coming year, it's possible that some of these stocks may split. (A stock split results in additional shares being distributed to existing shareholders.) For each stock, only two possible outcomes may occur: The stock splits, or it doesn't split. As a result, you can use the binomial distribution to compute the probability of a given number of splits in your portfolio over the coming year.



WARNING

The binomial distribution is based on several specialized assumptions, which I explain in detail in the next section. If these assumptions aren't true, using the binomial distribution to compute probabilities for a given situation is likely to give inaccurate results.

Checking out the binomial distribution

You generate a binomial distribution by a special type of random experiment, known as a *binomial process*. This consists of a fixed number of repeated trials, each with only two possible outcomes and the following distinguishing features:

- » Each trial results in either a success or a failure. On each trial of a binomial process, two possible outcomes may take place — and they're designated as "success" and "failure." For example, if you're doing a series of coin flips, you may call the outcome of the coin landing with "heads" up a success and the outcome of "tails" up a failure.

- » The trials are independent of each other. Each trial of a binomial process is independent of previous trials; in other words, the outcome of one trial has no influence over the outcome of the other trials. For example, the probability of heads turning up on a coin flip doesn't depend on the outcomes of flips that have taken place in the past.
- » The probability of success remains constant for all trials. The probability of success in a binomial process doesn't change from one trial to the next; instead, it remains constant throughout the entire process. For example, the probability of a head turning up on a flip of a coin is always one-half (50 percent), no matter how many times the coin is flipped.

Computing binomial probabilities

You can compute the probability that a specified number of successes will occur during a fixed number of trials by using the binomial formula. For example, with this formula, you can determine the probability that five odd numbers turn up when a die is rolled ten times. The formula is:

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

Here's what each element of this formula means:

- » X = a binomial random variable whose value is determined by the number of successes that occur during a series of trials
- » x = the number of successes whose probability you are computing
- » n = the number of trials that take place
- » p = the probability of success on a single trial
- » $(1 - p)$ = the probability of failure on a single trial
- » $!$ = the factorial operator



REMEMBER

The capital X is a binomial (random variables are discussed in Chapter 7), and the lowercase x is a specific value, which refers to the number of successes whose probability you're calculating.

Factorial: counting how many ways you can arrange things

The exclamation point (!) doesn't just mean you're excited. The symbol is also the mathematical operator *factorial*. You pronounce $n!$ as "n factorial," which is the product of all positive integers less than or equal to n. For example:

$$0! = 1 \text{ (looks odd, but it's true)}$$

$$1! = 1$$

$$2! = (2)(1) = 2$$

$$3! = (3)(2)(1) = 6$$

$$4! = (4)(3)(2)(1) = 24$$

A general description is $n! = (n)(n - 1)(n - 2) \dots (2)(1)$. The factorial is a handy tool, but you can apply it only to 0 and positive integers.

You can use the factorial operator to count the number of ways you can arrange a group of objects. For example, suppose that a small bookshelf has enough room for three titles: *Algebra and Its Applications*; *Baseball: A History*; and *Chemistry in Everyday Life*. You can label these titles A, B, and C and then set up the possibilities for how many ways you can you arrange these books on the shelf like this:

ABC

ACB

BAC

BCA

CAB

CBA

This list covers every possibility. Each entry in the list is an *arrangement* of the three titles. Counting the number of elements in this list shows that you can arrange the books in six ways.

Fortunately, a much easier way to get this same result is to simply compute $3!$ (because three books are being arranged), giving a total number of arrangements of $3! = (3)(2)(1) = 6$.



TIP

Many calculators contain a built-in function for the factorial operator. It typically appears as $x!$ In Microsoft Excel, you can compute factorial with the function FACT.

Combinations: Counting how many choices you have

You use the combinations formula to count the number of *combinations* that can be created when choosing x objects from a set of n objects:

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

One distinguishing feature of a combination is that the order of objects is irrelevant.

For example, you can use this formula to count the number of ways you choose two elective classes from a set of eight for the upcoming semester. The order in which you choose the electives is immaterial; each possible selection is a *combination* of two objects.

As another example, suppose that you're painting your house with two colors from a set of four: green, blue, white, and yellow. Because the order in which you choose the colors is irrelevant, each pair of colors is a combination. How many different color schemes are possible with the given set of choices? You can answer this question by simply listing all the possible combinations:

- green, white
- green, blue
- green, yellow
- white, yellow
- white, blue
- blue, yellow

Note that choices such as green, white and white, green are not both listed as they represent the same combination.

This list shows that you have six possible choices of pairs of colors.

The quicker way to answer this question is to substitute these values into the combinations formula; in this case, x represents the number of colors to choose (2), and n represents the total number of colors you can choose from (4):

$$\frac{n!}{x!(n-x)!} = \frac{4!}{2!(4-2)!} = \frac{4!}{2!2!} = \frac{24}{(2)(2)} = 6$$



TIP

The formula for computing the number of combinations is sometimes expressed as

$$\binom{n}{x}$$

Read or say this expression as “ n choose x .” This function appears on many calculators as nCr . In Microsoft Excel, you can compute combinations with the function $COMBIN$.



TECHNICAL STUFF

When you’re selecting x objects from a group of n objects in such a way that the order of selection *does* matter, the choices are known as *permutations* instead of combinations.

Binomial formula: Computing the probabilities

Combinations are useful for computing binomial probabilities. You can find the probability of x successes during n trials with the binomial formula:

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

Here,

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

is the total number of ways you can get exactly x successes during n trials, and

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

is the probability of a sequence consisting of x successes and $(n - x)$ failures.

For example, suppose that 40 percent of all published books are fiction, so the remaining 60 percent are nonfiction. If you pick six books at random from a bookstore, what’s the probability that either none or one of them is fiction?

First, define fiction as a *success*. The probability of success on a single trial is $p = 0.4$, because 40 percent of all books are fiction. Each book you choose is a single trial of an experiment, so if you pick six books, you’re conducting $n = 6$ trials for this experiment. You then figure the probability of getting one or fewer fiction books by calculating the probabilities of getting none and one fiction book and then adding them together:

- » Based on the binomial formula, the probability of choosing *no* fiction books from a selection of six books is

$$P(X = 0) = \frac{6!}{0!(6-0)!} (0.40)^0 (0.60)^6$$

$$P(X = 0) = (1)(1)(0.0467)$$

$$P(X = 0) = 0.0467$$

- » Based on the binomial formula, the probability of choosing one fiction book from a selection of six is

$$P(X = 1) = \frac{6!}{1!(6-1)!} (0.40)^1 (0.60)^5$$

$$P(X = 1) = (6)(0.40)(0.07776)$$

$$P(X = 1) = 0.1866$$

Now add the probabilities together. The probability of getting either no fiction book or one is $0.0467 + 0.1866 = 0.2333$. Alternatively, you can get these results from a binomial table for six trials ($n = 6$), such as Table 8-1.

TABLE 8-1 Binomial Probabilities that Result from 6 Trials ($n = 6$)

	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$
$x = 0$	0.5314	0.2621	0.1176	0.0467	0.0156
$x = 1$	0.3543	0.3932	0.3025	0.1866	0.0938
$x = 2$	0.0984	0.2458	0.3241	0.3110	0.2344
$x = 3$	0.0146	0.0819	0.1852	0.2765	0.3125
$x = 4$	0.0012	0.0154	0.0595	0.1382	0.2344
$x = 5$	0.0001	0.0015	0.0102	0.0369	0.0938
$x = 6$	0.0000	0.0001	0.0007	0.0041	0.0156

Table 8-1 shows the probability of success (p) at the top of each column. In this example, because $p = 0.4$, the probability of choosing zero fiction books is $P(X = 0) = 0.0467$ (found in the $x = 0$ row and the $p = 0.4$ column). The probability of choosing one fiction book is $P(X = 1) = 0.1866$ (found in the $x = 1$ row and the $p = 0.4$ column). The probability of getting no fiction books or one fiction book is the sum of $0.0467 + 0.1866$, or 0.2333 .



TIP



TIP

Check out a binomial table with 19 values for n at www.statisticshowto.com/tables/binomial-distribution-table.

If you simply don't like using formulas or tables to compute binomial probabilities, or if you want to triple-check your numbers, you can also use a specialized calculator, such as the Texas Instruments TI-84 Plus, which contains built-in functions that compute these probabilities quickly and easily. (I show you how to compute binomial probabilities using the TI-84 Plus at the end of this chapter.) You can also use the function BINOM.DIST in Microsoft Excel.

Moments of the binomial distribution

Moments are summary measures of a probability distribution. The expected value represents the mean or average value of a distribution. The expected value is sometimes known as the *first moment* of a probability distribution. You calculate the expected value by taking each possible value of the distribution, weighting it by its probability, and then summing the results. The expected value is comparable to the mean of a population or sample (see Chapter 3).

The variance and standard deviation represent the dispersion among the possible values of a probability distribution. The variance and standard deviation of a probability distribution are equivalent to the variance and standard deviation of a population or sample. (The general formulas for computing moments for a discrete probability distribution are given in Chapter 7.) The variance is sometimes known as the *second central moment* of a probability distribution; the standard deviation isn't a separate moment, but simply the square root of the variance. Luckily, for the binomial distribution, you can reduce computation time by using a series of simplified formulas, which I discuss in the following sections.

Binomial distribution: Calculating the expected value

The *expected value* of a probability distribution is its average value. You get it by weighting each possible value by its probability of occurring. For the binomial distribution, the calculation of the expected value can be simplified to

$$E(X) = np$$

For example, suppose that 10 percent of all people are left-handed, and 90 percent are right-handed (which happens to be true). In a class of 40 students, what's the expected number of left-handed students? You can calculate the expected value by thinking of each student as a "trial," with a 10 percent chance of being left-handed (a "success") and 90 percent chance of being right-handed (a "failure").

Therefore, $n = 40$ and $p = 0.10$. The expected number of left-handed students in the class is $E(X) = np = (40)(0.10) = 4$.

Binomial distribution: Computing variance and standard deviation

The *variance* of a distribution is the average squared distance between each possible outcome and the expected value. For the binomial distribution, you may compute the variance with the following simplified formula:

$$\sigma^2 = np(1 - p)$$

The *standard deviation* of a distribution equals the square root of the variance. For the binomial distribution, you calculate the standard deviation as

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

For the example of left-handed students in the previous section,

- » The expected value is $E(X) = np = (40)(0.10) = 4$.
- » The variance is $\sigma^2 = np(1 - p) = 40(0.10)(0.90) = 3.6$.
- » The standard deviation is $\sqrt{3.6} = 1.9$.

Graphing the binomial distribution

You may want to illustrate the binomial distribution with a *histogram*. A histogram shows the possible values of a probability distribution as a series of vertical bars. The height of each bar reflects the probability of each value occurring. A histogram is a useful tool for visually analyzing the properties of a distribution, and (by the way) all discrete distributions may be represented with a histogram. (See Chapter 2 for more about histograms and other types of graphs.)

For example, suppose that a candy company produces both milk chocolate and dark chocolate candy bars. The product mix is 50 percent of the candy bars are milk chocolate and 50 percent are dark chocolate. Suppose that you choose ten candy bars at random, and choosing milk chocolate is defined as a success. The probability distribution of the number of successes during these ten trials with $p = 0.5$ is shown in Figure 8-1.

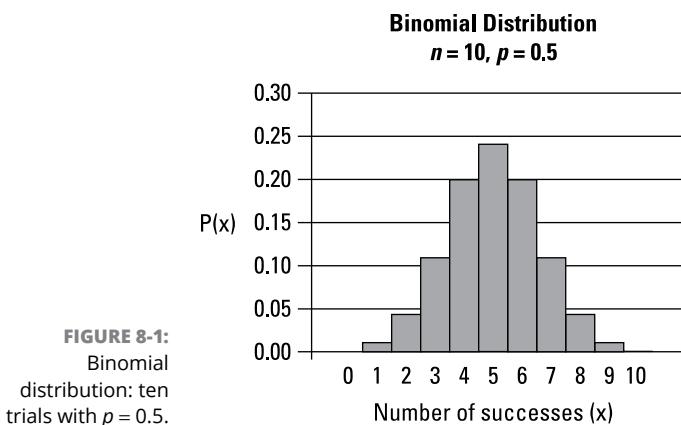


Figure 8-1 shows that when $p = 0.5$, the distribution is *symmetric* about its expected value of 5 ($np = 10[0.5] = 5$), where the probabilities of X being below the mean match the probabilities of X being the same distance above the mean.

For example, with $n = 10$ and $p = 0.5$,

$$P(X = 4) = 0.2051 \text{ and } P(X = 6) = 0.2051$$

$$P(X = 3) = 0.1172 \text{ and } P(X = 7) = 0.1172$$

If the probability of success is less than 0.5, the distribution is *positively skewed*, meaning probabilities for X are greater for values below the expected value than above it.

For example, with $n = 10$ and $p = 0.2$,

$$P(X = 4) = 0.0881 \text{ and } P(X = 6) = 0.0055$$

$$P(X = 3) = 0.2013 \text{ and } P(X = 7) = 0.0008$$

Figure 8-2 shows the probability distribution for $n = 10$ and $p = 0.2$.

If the probability of success is greater than 0.5, the distribution is *negatively skewed* – probabilities for X are greater for values above the expected value than below it.

For example, with $n = 10$ and $p = 0.8$,

$$P(X = 4) = 0.0055 \text{ and } P(X = 6) = 0.0881$$

$$P(X = 3) = 0.0008 \text{ and } P(X = 7) = 0.2013$$

Figure 8-3 shows the probability distribution for the same situation when $p = 0.8$.

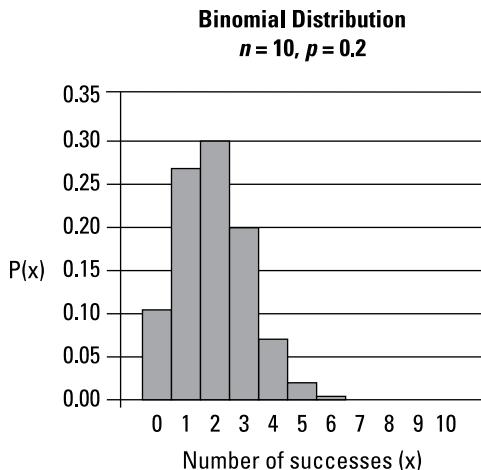


FIGURE 8-2:
Binomial distribution: ten trials with $p = 0.2$.

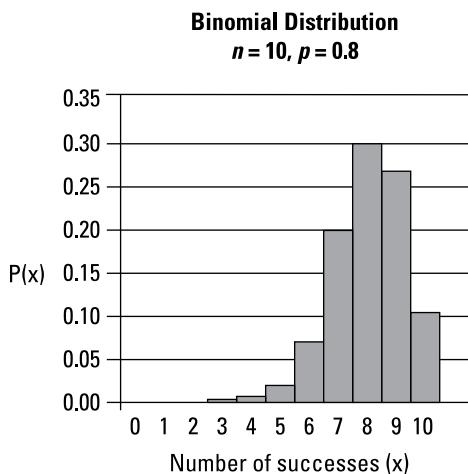


FIGURE 8-3:
Binomial distribution: ten trials with $p = 0.8$.

Keeping the Time: The Poisson Distribution

The *Poisson distribution* is useful for measuring how many events may occur during a given time horizon, such as the number of customers that enter a store during the next hour, the number of hits on a website during the next minute, and so forth. The *Poisson process* takes place over time instead of a series of trials; each interval of time is assumed to be *independent* of all other intervals.

For example, suppose that a bank counts the number of customers who enter each hour. If the number of customers that enter during a given hour is independent of the number that enter during all other hours (while the bank is open), you can use the Poisson distribution to find the probability that a specific number of customers enter the bank during the next hour.



TECHNICAL
STUFF

The Poisson distribution is named for Siméon Denis Poisson who was a French mathematician, physicist, and genius. He was wrong about only one major thing: He opposed the wave theory of light.

The following section shows you how to compute Poisson probabilities and how to compute moments for the Poisson distribution. Graphs are used to illustrate the key properties of the Poisson distribution.

Computing Poisson probabilities

You calculate Poisson probabilities with the following formula:

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

Here's what each element of this formula represents:

- » X = a Poisson random variable
- » x = number of events whose probability you are calculating
- » λ = the Greek letter "lambda," which represents the average number of events that occur per time interval
- » e = a constant that's equal to approximately 2.71828



TIP

e is a constant that's widely used in financial applications. One of the most important uses is in computing present values of sums of money when interest rates are *continuously compounded* — compounded an *infinite* number of times. Most calculators have a key labeled e^x that you can use to calculate the value of e raised to a specified power. In Excel, the appropriate function for determining the value of e is EXP.

For example, suppose that the number of messages a person receives on their cellphone averages one per hour and that the number of messages received each hour is independent of all other hours. What's the probability of a person receiving two messages in the next hour?

In this case, the value of lambda (λ) is equal to 1, because the average number of messages each hour equals 1. The probability of receiving two messages during the next hour is

$$P(X = 2) = e^{-1} \frac{1^2}{2!} = 0.1839$$

Alternatively, you can get results from a Poisson table set up like Table 8-2.

TABLE 8-2**Poisson Probabilities for Different Values of λ**

	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 1.5$	$\lambda = 2$	$\lambda = 2.5$	$\lambda = 3$
x = 0	0.6065	0.3679	0.2231	0.1353	0.0821	0.0498
x = 1	0.3033	0.3679	0.3347	0.2707	0.2052	0.1494
x = 2	0.0758	0.1839	0.2510	0.2707	0.2565	0.2240
x = 3	0.0126	0.0613	0.1255	0.1804	0.2138	0.2240
x = 4	0.0016	0.0153	0.0471	0.0902	0.1336	0.1680
x = 5	0.0002	0.0031	0.0141	0.0361	0.0668	0.1008
x = 6	0.0000	0.0005	0.0035	0.0120	0.0278	0.0504
x = 7	0.0000	0.0001	0.0008	0.0034	0.0099	0.0216
x = 8	0.0000	0.0000	0.0001	0.0009	0.0031	0.0081

Table 8-2 shows the Poisson probabilities for different values of λ . In the cell-phone example, because $x = 2$ and $\lambda = 1$, the appropriate probability $P(X = 2)$ is found in the $x = 2$ row and the $\lambda = 1$ column. The probability is 0.1839.



TIP

If you don't care for using formulas or a table, try a specialized calculator or Excel. The Excel function is POISSON.DIST.

The moments of the Poisson distribution are used to represent the average value of the distribution and the dispersion of the distribution. As with the binomial distribution, these moments may be computed with simplified formulas.

Poisson distribution: Calculating the expected value

As with the binomial distribution (discussed earlier in this chapter), you can use simple formulas to compute the moments of the Poisson distribution. The expected value of the Poisson distribution is

$$E(X) = \lambda$$

For example, say that on average three new companies are listed on the New York Stock Exchange (NYSE) each year. The number of new companies listed during a given year is independent of all other years. The number of new listings per year, therefore, follows the Poisson distribution, with a value of $\lambda = 3$. As a result, the expected number of new listings next year is $\lambda = 3$.

Poisson distribution: Computing variance and standard deviation

Compute the variance for the Poisson distribution as $\sigma^2 = \lambda$; the standard deviation (σ) equals $\sqrt{\lambda}$.

Based on the NYSE listing example in the previous section, the variance equals 3 and the standard deviation equals $\sqrt{3} = 1.732$.

Graphing the Poisson distribution

As with the binomial distribution, the Poisson distribution can be illustrated with a histogram. In Figures 8-4 through 8-6, the results are shown for three values of λ : 2 (Figure 8-4), λ : 5 (Figure 8-5) and λ : 7 (Figure 8-6).

For $\lambda = 2$ (Figure 8-4), the distribution is skewed to the right; for $\lambda = 5$ (Figure 8-5), the distribution is nearly symmetric about the mean of 5; for $\lambda = 7$ (Figure 8-6), the distribution is skewed to the left.

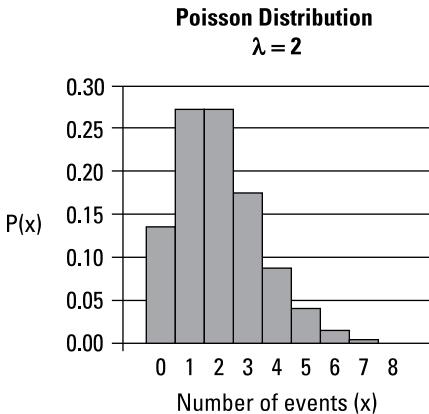


FIGURE 8-4:
Poisson distribution with $\lambda = 2$

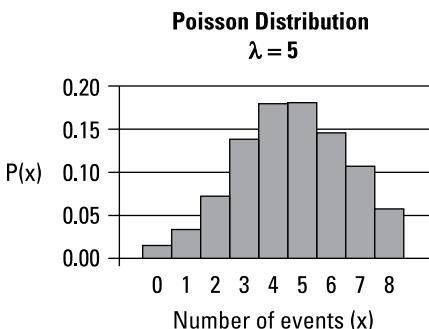
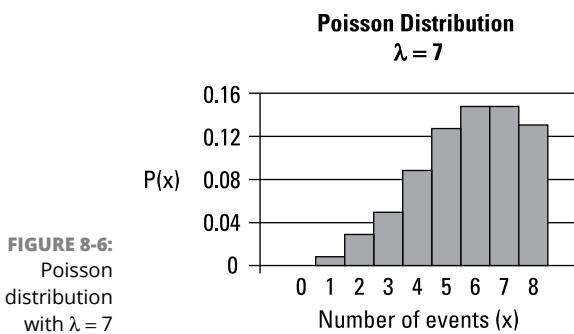


FIGURE 8-5:
Poisson distribution with $\lambda = 5$



Computing Binomial and Poisson Probabilities with the TI-84 Plus Calculator

You can use the Texas Instruments TI-84 Plus and Plus CE calculators to compute probabilities for the binomial and Poisson distributions.

Computing binomial probabilities

The binomial distribution is used for cases when an experiment is being conducted as a series of independent trials with only two possible outcomes: “success” and “failure.” The probability of success is unchanged throughout the experiment.

As an example, suppose that a coin is flipped ten times. The event of a head turning up is defined as a “success,” and the event of a tail turning up is defined as a “failure.” This experiment qualifies as a binomial experiment because each flip is independent of all other flips — each trial of the experiment can only result in one of two possible outcomes and the probability of a success on a single trial remains constant.

The probability of getting five heads is computed by pressing [2nd], [VARS], and then selecting A:binompdf. This produces the following menu:

trials:

p:

x-value:

Paste

Trials represents the number of trials of the experiment. p is the probability of success on a single trial. x -value is the number of successes of interest. Paste is used to perform the necessary calculations. In this example, the appropriate values are:

trials: 10

p : 0.5

x -value: 5

Paste

The number of trials is 10 because the coin is being flipped ten times. p equals 0.5 because the probability of obtaining a head on a single flip is 0.5 or 50 percent. The x -value is 5 because the desired probability is the probability of getting five heads in ten flips.

After entering this information, choosing Paste and then pressing the [ENTER] button twice produces the following result: 0.24609375. This shows that the probability of getting five heads with ten flips is about 24.6 percent.

A cumulative or “less than or equal to” probability can be computed by pressing [2^{nd}], [VARS], and then selecting B:binomcdf.

The inputs are the same as for A:binompdf. After choosing Paste and pressing the [ENTER] button twice, the result is 0.623046875. This shows that the probability of obtaining five or fewer heads from ten flips is about 62.3 percent.

Computing Poisson probabilities

The Poisson distribution is used for cases when a series of events occur during a specified time frame. For example, suppose that on average five cars drive over a bridge each hour. Then the probability of a specified number of cars driving over the bridge during the next hour can be computed using the Poisson distribution (assuming that the volume of traffic during a given time period is independent of the volume of traffic during other time periods).

In this case, the probability that there will be three cars passing over the bridge during the next hour is determined by pressing [2^{nd}], [VARS], and then selecting C:poissonpdf.



TIP

C:poissonpdf is the correct choice for the TI-84 Plus; note that on the TI-84 Plus CE, the DISTR menu contains more entries so the correct choice would be D:poissonpdf.

This produces the following menu:

λ :

x-value:

Paste

λ is the Greek letter “lambda”; it represents the average number of events that occur per unit of time. In this example, lambda equals five because on average, five cars pass over the bridge every hour. Note that the TI-84 Plus CE uses the Greek letter μ (“mu”) instead of λ . The x-value is the number of events of interest. In this example, the appropriate values are:

λ : 5

x-value: 3

Paste

After choosing Paste and pressing [ENTER] twice, the result is 0.1403738958. This shows that the probability that there will be three cars passing over the bridge during the next hour is about 14.0 percent.

A cumulative or “less than or equal to” probability can be computed by pressing [2^{nd}], [VARS], and then selecting D: poissoncdf. The inputs are exactly the same as for C: poissonpdf:

λ : 5

x-value: 3

Paste



TIP

On the TI-84 Plus CE the menu items are D: poissonpdf and E: poissoncdf.

After choosing Paste and pressing [ENTER] twice, the resulting probability is 0.2650259153. This shows that the probability that there will be three or fewer cars passing over the bridge during the next hour is about 26.5 percent.

IN THIS CHAPTER

- » Understanding the differences between discrete and continuous distributions
- » Discovering the properties of the normal distribution
- » Checking out normal probabilities

Chapter 9

The Normal Distribution: So Many Possibilities!

This chapter introduces one of the most important probability distributions in the field of statistics: the normal distribution. The normal distribution is especially important in business applications; it can be used to describe the behavior of many financial variables, such as the rate of return to an investment, a corporation's annual profits, consumer spending on new products, and so on. The normal distribution has one important feature that distinguishes it from discrete distributions, such as the binomial and Poisson distributions discussed in Chapter 8. It assigns probabilities to *ranges* of values instead of *individual* values.

The normal distribution is the most widely used distribution in business because you can use it to model many variables. For example, you can use the normal distribution to describe the rates of return to financial assets, the distribution of corporate profits, the prices of key commodities (such as oil), and so forth. Suppose that the returns to the stocks in the Standard and Poor's 500 (S&P 500) index

are normally distributed. The normal distribution can then be used to answer questions such as:

What is the probability that the S&P 500 will increase by at least 5 percent next year?

What is the probability that the S&P 500 will fall next year?

How much risk is associated with investing in the S&P 500?

In this chapter, I explain the differences between the two basic types of probability distributions: discrete and continuous. I also provide a detailed look at the properties of the normal distribution, including techniques for computing probabilities; however, because of the complexity of the normal distribution, I show you how to compute normal probabilities with standard tables instead of formulas. I also show you how to compute normal probabilities with the Texas Instruments TI-84 Plus and Plus CE calculators.

Comparing Discrete and Continuous Distributions

Discrete and continuous distributions are the two standard types of probability distributions that you use to compute probabilities for possible outcomes of a random experiment. (For more about random experiments and probability distributions, see Chapter 7.)

- » You use a discrete distribution with a random experiment that can generate a *finite* (countable) number of outcomes. (You see two examples of discrete distributions — binomial and Poisson — in Chapter 8.)
- » You use a continuous distribution with a random experiment that can generate an *infinite* (uncountable) number of outcomes.



TECHNICAL STUFF

Intuitively, a random experiment can generate a finite (countable) number of outcomes if it's possible to make up a list of all the possible outcomes of the experiment. For example, if a coin is flipped ten times and the variable of interest is how many times *heads turns up*, there are 11 possible outcomes: 0, 1, 2, . . . , 10. These outcomes can be easily listed. On the other hand, if an experiment consists of observing the length of time until the next phone call arrives, the number of possible times until the next phone call is *infinite* (*uncountable*). This is because the times are not restricted to whole numbers. The time may be 2.3 seconds, 1.41742 seconds, 8.19444212 seconds, and so on. A list containing all possible times until

the next phone call is impossible to construct, because there are an unlimited number of entries.



TIP

Computing probabilities for continuous distributions is more complex than for a discrete distribution; often, your best resources are tables or specialized calculators. For an example of an online calculator, visit www.solvemymath.com/online_math_calculator/statistics/continuous_distributions/index.php. At the end of this chapter, I show you how to compute normal probabilities with the TI-84 Plus and Plus CE calculators.

Aside from the number of possible outcomes, one of the most important differences between discrete and continuous distributions is this: With a continuous distribution, the probability that a random variable (X) equals a specific constant (x) is defined as zero. With an infinite number of possibilities, the likelihood of X being equal to a specific value is infinitesimally small.

For example, the probability of tomorrow's temperature at noon being exactly 72.141712987 degrees is pretty much zero. As a result, for any value x , $P(X \leq x)$ equals $P(X < x)$. A statement such as “the probability that the temperature at noon tomorrow will be less than or equal to 72 degrees” has the same interpretation as “the probability that the temperature at noon tomorrow will be less than 72 degrees.”

To demonstrate this statement mathematically, you can write $P(X \leq x)$ as $P(X < x) + P(X = x)$, because the probability that X is less than or equal to x consists of the sum of two different probabilities — the probability that X is strictly less than x and the probability that X is exactly equal to x . With a continuous distribution, $P(X = x) = 0$; therefore,

$$P(X \leq x) = P(X < x) + P(X = x)$$

$$P(X \leq x) = P(X < x) + 0$$

$$P(X \leq x) = P(X < x)$$

Based on this reasoning, $P(X \geq x) = P(X > x)$ is also true.



REMEMBER

With a discrete distribution, $P(X \leq x)$ does not equal $P(X < x)$, and $P(X \geq x)$ does not equal $P(X > x)$ unless $P(X = x) = 0$.

For example, suppose that a coin is flipped three times. The outcome of interest is whether a head turns up on each flip.

The probability that two or fewer heads turns up is computed as:

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

The probability that fewer than two heads turn up is computed as:

$$P(X < 2) = P(X = 0) + P(X = 1)$$

Therefore, unless $P(X = 2) = 0$, $P(X \leq 2)$ and $P(X < 2)$ gives different results.

In the continuous case, though, $P(X \leq 2)$ and $P(X < 2)$ are always equal.

Understanding the Normal Distribution

The normal distribution is a continuous probability distribution that can be used to describe a large number of different situations, not just in business applications but in a wide variety of other disciplines, such as psychology, sociology, biology, and so on. The normal distribution, sometimes called the Gaussian distribution, is named after scientist and mathematician Johann Carl Friedrich Gauss (1777–1855) who introduced the concept.

The normal distribution has several useful properties that can be used to describe real-world events. For example, under the normal distribution, there is a balance or *symmetry* between the likelihood of a value being below the mean of the distribution and being above the mean of the distribution.

As an example, suppose that researchers have determined that the heights of all men in a country are normally distributed with a mean of 69 inches and a standard deviation of 2 inches. Based on the normal distribution, the following events are equally likely:

A randomly chosen man is no more than 67 inches tall

A randomly chosen man is at least 71 inches tall

These events are equally likely because:

A height of 67 inches is one standard deviation below the mean ($69 - 1(2) = 67$)

A height of 71 inches is one standard deviation above the mean ($69 + 1(2) = 71$)

Similarly, the following events are equally likely:

A randomly chosen man is no more than 65 inches tall

A randomly chosen man is at least 73 inches tall

These events are equally likely because:

A height of 65 inches is two standard deviations below the mean ($69 - 2(2) = 65$)

A height of 73 inches is two standard deviations above the mean ($69 + 2(2) = 73$)

Because the normal distribution is a *continuous distribution*, it's defined for an infinite number of values. The normal distribution is defined for *all* values between negative infinity and positive infinity.

In the following sections, I show you how you can express the normal distribution graphically, I introduce you to the standard normal distribution, and I walk you through calculating probabilities for the normal distribution.

Graphing the normal distribution

The normal distribution can be graphed with a special type of curve, which is usually described as a *bell-shaped curve*. Normal probabilities can be determined by computing areas under this curve.

The bell-shaped curve has several key features. It's defined over the entire range of values between negative and positive infinity; it's *symmetrical* about the mean (for example, the area below the mean is a *mirror image* of the area above the mean); and most of the area under the normal distribution is close to the mean. The area declines rapidly for values that are several standard deviations away from the mean. As an example, the distribution of heights from the previous example is illustrated with a bell-shaped curve in Figure 9-1.

The mean of 69 inches is at the center of the distribution; the area to the left of the mean is a mirror image of the area to the right of the mean. Most of the area under the curve is close to the mean; the area falls off rapidly for large and small values of X. (The extreme right and left ends of the curve are known as the *tails* of the distribution.) Figure 9-2 shows that the probability of a randomly chosen man's height being between 67 inches and 71 inches is 68.27 percent.

The shaded region under the curve represents heights between 67 and 71 inches. This covers 68.27 percent of the area under the curve; therefore, the probability that a randomly chosen man's height is between 67 inches and 71 inches is 0.6827 or 68.27 percent.

The Normal Distribution

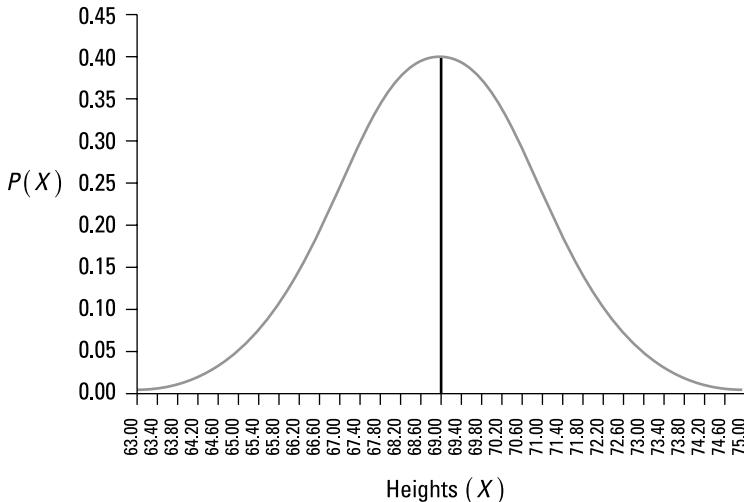


FIGURE 9-1:
The bell-shaped
curve of the
distribution of
heights.

The Normal Distribution

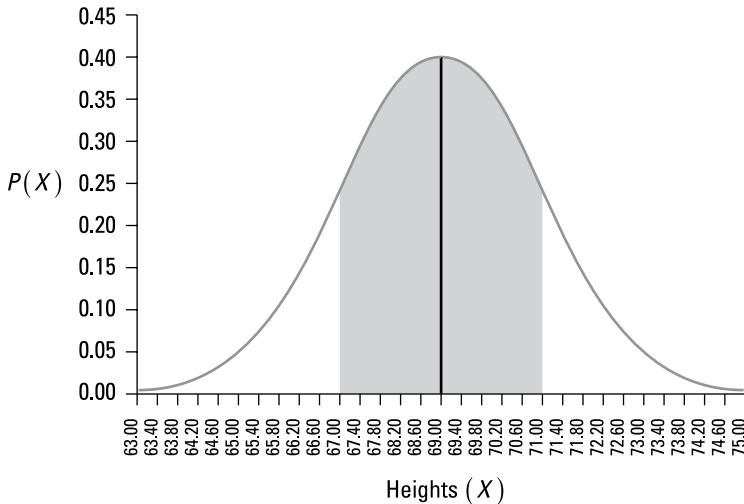


FIGURE 9-2:
The distribution
of heights
between
67 inches and
71 inches.

The normal distribution is uniquely characterized by two values:

- » The expected value (mean), represented by μ (the Greek letter "mu")
- » The standard deviation, represented by σ (the Greek letter "sigma")

There are an infinite number of different possible normal distributions, each with a different value of the mean and standard deviation.

THE NORMAL DISTRIBUTION IN STATISTICAL ANALYSIS

The normal distribution is used in conjunction with many statistical techniques. It plays a key role in a lot of applications, such as the following:

- Computing confidence intervals
- Testing hypotheses about the mean of a population
- Testing hypotheses about the means of two populations
- Regression analysis

In many business applications, variables are assumed to be normally distributed. For example, returns to stocks are often assumed to be normally distributed by investors, portfolio managers, financial analysts, risk managers, and so on. The assumption of normality is not only convenient, but many standard statistical techniques require it in order to generate valid results. For example, computing a confidence interval for the mean of a population may be based on the normal distribution. Many of the techniques used in regression analysis to check the validity of the results are based on the normal distribution. As a result, even when the assumption of normality is not perfectly accurate, the normal distribution is often used to perform statistical analyses due to its convenience.

Getting to know the standard normal distribution

The *standard normal distribution* is the special case where $\mu = 0$ and $\sigma = 1$. For example, suppose that the daily returns to a stock follow the standard normal distribution. The mean return over a single trading day is 0 percent, and the standard deviation is 1 percent; as a result:

- » The probability that tomorrow's return will be between -1 percent and +1 percent is 0.6827 or 68.27 percent. -1 percent represents one standard deviation below the mean, while +1 percent represents one standard deviation above the mean.
- » The probability that tomorrow's return will be between -2 percent and +2 percent is 0.9544 or 95.44 percent. -2 percent represents two standard deviations below the mean, while +2 percent represents two standard deviations above the mean.

- » The probability that tomorrow's return will be between -3 percent and +3 percent is 0.9973 or 99.73 percent. -3 percent represents three standard deviations below the mean, while +3 percent represents three standard deviations above the mean.



TECHNICAL STUFF

By convention, the letter Z represents a standard normal random variable, whereas the letter X represents any other normal random variable.

Computing standard normal probabilities

One approach to computing probabilities for the standard normal distribution is to use statistical tables. (For the mathematically inclined, the tables result from applying calculus to the normal distribution.) The standard normal table is designed to show *cumulative* probabilities; i.e., the probability that a standard normal random variable Z is less than or equal to a specified value, such as $P(Z \leq 2.50)$. Standard normal tables are divided into two parts; the first shows positive values for Z, and the second shows negative values for Z.

Computing other types of probabilities, such as $P(Z \geq 1.70)$, can be accomplished by using the properties of the standard normal distribution to rearrange these probabilities in a more convenient form.

The following sections illustrate how to compute normal probabilities using the standard normal tables.

Computing "less than or equal to" standard normal probabilities

Table 9-1 shows a portion of the standard normal table for positive values of Z. (The actual table typically shows Z values between 0 and 3.)

The table shows the probability that a standard normal random variable Z is *less than or equal* to a specific value. For example, to express the probability that Z is less than or equal to 1, you write $P(Z \leq 1.00)$. Here's how you find this probability:

1. Take the first digits before and after the decimal point (1.0 in 1.00) from the Z column, second row.
2. Take the second digit after the decimal point (0.00 in 1.00) from the corresponding column (0.00 in this case).
3. Find the appropriate probability at the intersection of this row and column.

Using this technique, the table shows that $P(Z \leq 1.00) = 0.8413$. Figure 9-3 shows this expression graphically.

TABLE 9-1**Standard Normal Table — Positive Values**

Z	0.00	0.01	0.02	0.03
0.9	0.8159	0.8186	0.8212	0.8238
1.0	0.8413	0.8438	0.8461	0.8485
1.1	0.8643	0.8665	0.8686	0.8708
1.2	0.8849	0.8869	0.8888	0.8907
1.3	0.9032	0.9049	0.9066	0.9082
1.4	0.9192	0.9207	0.9222	0.9236
1.5	0.9332	0.9345	0.9357	0.9370
1.6	0.9452	0.9463	0.9474	0.9484
1.7	0.9554	0.9564	0.9573	0.9582
1.8	0.9641	0.9649	0.9656	0.9664
1.9	0.9713	0.9719	0.9726	0.9732
2.0	0.9772	0.9778	0.9783	0.9788

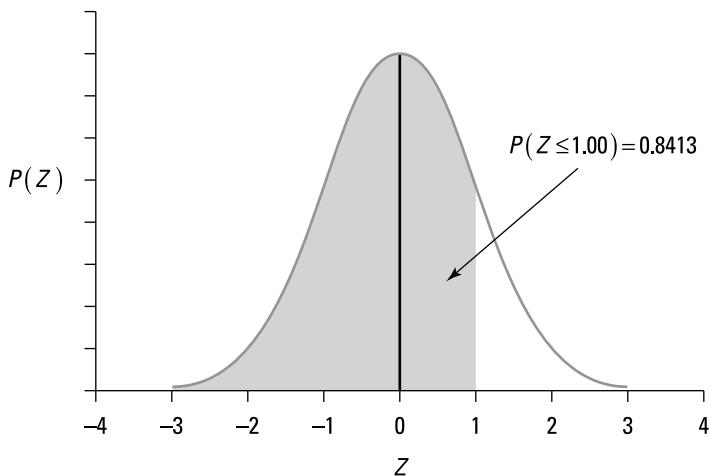
Standard Normal Probability Distribution

FIGURE 9-3:
Standard normal probability distribution where $P(Z \leq 1)$ equals 0.8413.

The shaded region to the left of 1 represents 84.13 percent of the area under the curve; therefore, $P(Z \leq 1.00) = 0.8413$ or 84.13 percent.

Negative probabilities also have a corresponding standard normal table. Take a look at Table 9-2. This shows several negative values for Z ; the actual table typically shows values ranging from 0 to -3.

TABLE 9-2**Standard Normal Table — Negative Values**

Z	0.00	0.01	0.02	0.03
-2.0	0.0228	0.0222	0.0217	0.0212
-1.9	0.0287	0.0281	0.0274	0.0268
-1.8	0.0359	0.0351	0.0344	0.0336
-1.7	0.0446	0.0436	0.0427	0.0418
-1.6	0.0548	0.0537	0.0526	0.0516
-1.5	0.0668	0.0655	0.0643	0.0630
-1.4	0.0808	0.0793	0.0778	0.0764
-1.3	0.0968	0.0951	0.0934	0.0918
-1.2	0.1151	0.1131	0.1112	0.1093
-1.1	0.1357	0.1335	0.1314	0.1292
-1.0	0.1587	0.1562	0.1539	0.1515
-0.9	0.1841	0.1814	0.1788	0.1762
-0.8	0.2119	0.2090	0.2061	0.2033
-0.7	0.2420	0.2389	0.2358	0.2327
-0.6	0.2743	0.2709	0.2676	0.2643
-0.5	0.3085	0.3050	0.3015	0.2981

Suppose you want to compute the probability that Z is less than -1.23, which you write as $P(Z \leq -1.23)$. The first digits before and after the decimal point (-1.2 in -1.23) are in the Z column, ninth row. The second digit after the decimal point (0.03 in -1.23) is in the far right column. You find the probability at the intersection of the row and column, so the table shows that $P(Z \leq -1.23) = 0.1093$. This is shown in Figure 9-4.

One of the drawbacks to using tables to compute standard normal probabilities is that they show only cumulative probabilities for Z; for example, Z is less than or equal to a specific value. But you can figure all other cases by combining the properties of the standard normal distribution with the tables.

Standard Normal Probability Distribution

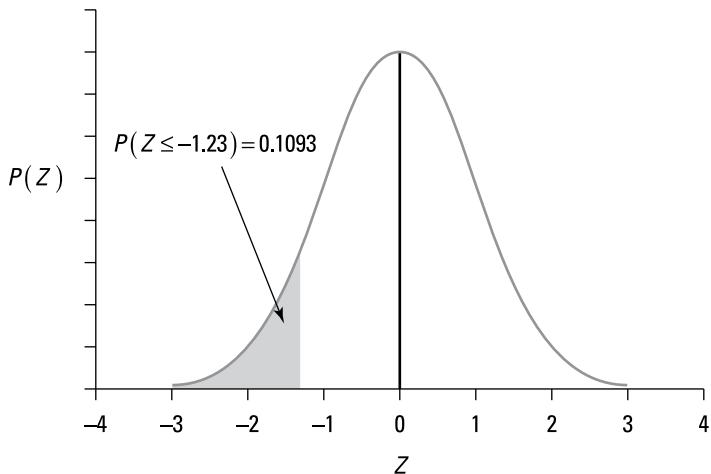


FIGURE 9-4:
Standard normal
probability
distribution
where
 $P(Z \leq -1.23)$
equals 0.1093.

Property 1: The area under the standard normal curve equals 1

The first of these properties is that the entire area under the standard normal curve equals 1. Because the curve covers the entire area between negative and positive infinity (∞), you can express this result as $P(-\infty \leq Z \leq \infty) = 1$. So the probability that a standard normal random variable Z falls between negative infinity and positive infinity is 1; in other words, Z will fall within this interval with *certainty*.



REMEMBER

When you consider all possible outcomes in any given situation, you can be certain that one outcome will occur. A probability of 1 indicates that an event will occur with *certainty*. A probability of 0 indicates that an event is *impossible*. All other probabilities fall between 0 and 1. (Probability theory is covered in Chapter 6.)

Property 2: The standard normal curve is symmetrical about the mean

The next key property of the standard normal distribution is *symmetry*, where the area to the left of the mean is a mirror image of the area to the right. As a result, the probability that Z is less than the mean is 0.5, and you write it as $P(Z \leq 0) = 0.5$ (because half of the area under this distribution is to the left of the mean, and half is to the right of the mean; the total area is 1), as shown in the Figure 9-5.

Because $P(Z \leq 0) = 0.5$, due to the symmetry of the standard normal probability distribution, it's also true that $P(Z \geq 0) = 0.5$, as illustrated in Figure 9-6.

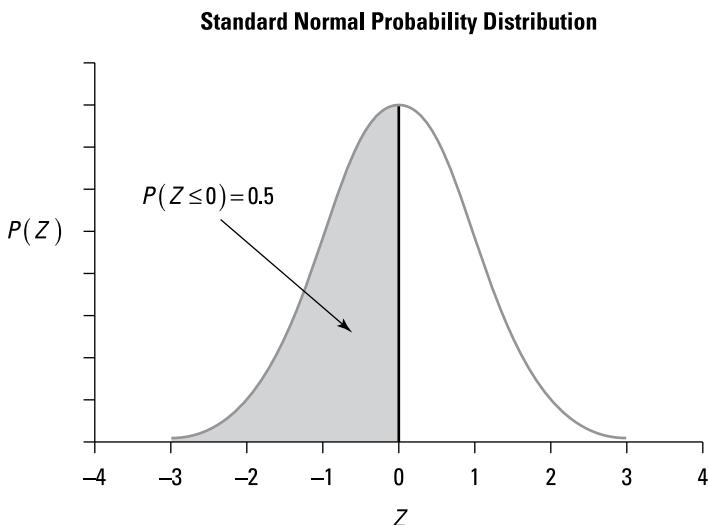


FIGURE 9-5:
Standard normal
probability
distribution
where
 $P(Z \leq 0) = 0.5$.

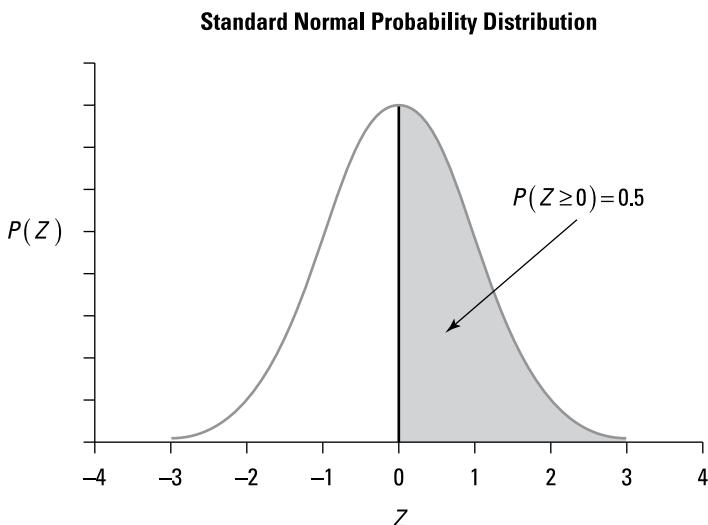


FIGURE 9-6:
Standard normal
probability
distribution
where
 $P(Z \geq 0) = 0.5$.

Other examples of symmetry include

$$P(Z \leq -1) = P(Z \geq 1) = 0.1587$$

$$P(Z \leq -2) = P(Z \geq 2) = 0.0228$$

Computing “greater than or equal to” standard normal probabilities

One type of probability you can’t compute directly from a table is the case where a standard normal random variable Z is *greater than or equal to* a specified value z : $P(Z \geq z)$. Instead, you rearrange the identity to yield a very useful result:

$$P(Z \leq z) + P(Z \geq z) = 1$$

This is a consequence of the first property of the standard normal distribution: The area under the standard normal curve equals 1.

Rearranging this equation gives you

$$P(Z \geq z) = 1 - P(Z \leq z)$$

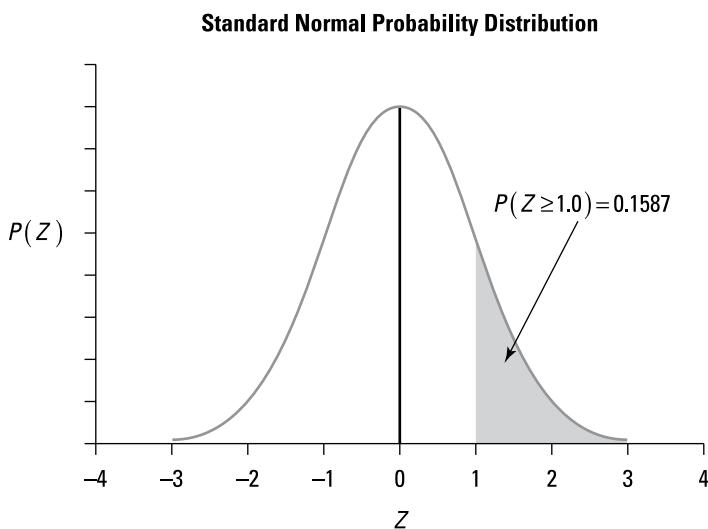
For example, to determine the probability that a standard normal random variable is greater than 1 (for example, $P(Z \geq 1)$), the first step is to rewrite the probability in a form that enables you to use the standard normal tables. This is shown as:

$$P(Z \geq 1) = 1 - P(Z \leq 1)$$

$$P(Z \leq 1) = 0.8413$$

$$\begin{aligned} P(Z \geq 1) &= 1 - 0.8413 \\ &= 0.1587 \end{aligned}$$

The result is shown in Figure 9-7.



Computing “in between” standard normal probabilities

Another type of probability that you can’t compute directly from a standard normal table is the case where a standard normal random variable Z is *between* two constants: c and d : $P(c \leq Z \leq d)$. But, lucky for you, you can work around this with the following identity:

$$P(c \leq Z \leq d) = P(Z \leq d) - P(Z \leq c)$$

You can now compute this probability by looking up $P(Z \leq c)$ and $P(Z \leq d)$ in the standard normal table and computing the difference between them. For example, suppose that you want to know the probability that Z is between one and two standard deviations above the mean. In this case, $c = 1.00$ and $d = 2.00$. This probability can be expressed as follows:

$$P(1.00 \leq Z \leq 2.00)$$

Algebraically, this can be rearranged in a form that involves two “less than or equal to” probabilities that can be looked up in the standard normal tables:

$$\begin{aligned} &P(1.00 \leq Z \leq 2.00) \\ &= P(Z \leq 2.00) - P(Z \leq 1.00) \end{aligned}$$

From the standard normal table (Table 9-1):

$$\begin{aligned} P(Z \leq 2.00) &= 0.9772 \\ P(Z \leq 1.00) &= 0.8413 \end{aligned}$$

As a result, you calculate the probability:

$$P(1.00 \leq Z \leq 2.00) = 0.9772 - 0.8413 = 0.1359$$

Figure 9-8 illustrates this probability.

Note that you can use this approach for negative values, too. For example, from the standard normal table (Table 9-2),

$$\begin{aligned} &P(-2.00 \leq Z \leq -1.00) \\ &= P(Z \leq -1.00) - P(Z \leq -2.00) \end{aligned}$$

$$\begin{aligned} P(Z \leq -2.00) &= 0.0228 \\ P(Z \leq -1.00) &= 0.1587 \end{aligned}$$

As a result:

$$\begin{aligned} &P(-2.00 \leq Z \leq -1.00) \\ &= P(Z \leq -1.00) - P(Z \leq -2.00) \\ &= 0.1587 - 0.0228 \\ &= 0.1359 \end{aligned}$$

Standard Normal Probability Distribution

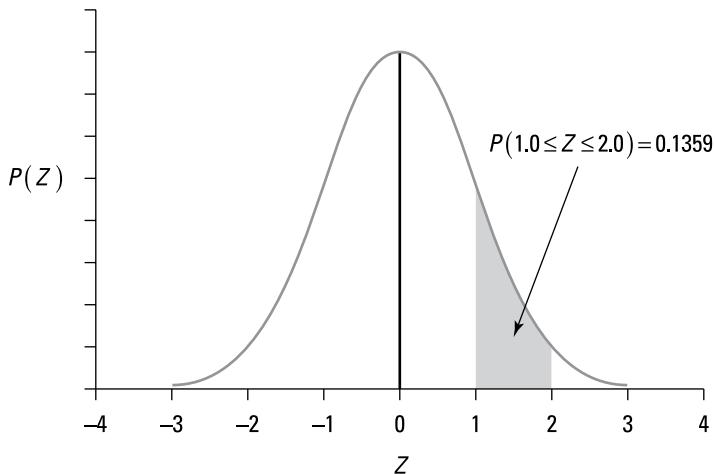


FIGURE 9-8:
Standard normal
probability
distribution
where $P(1.0 \leq Z \leq 2.0) = 0.1359$.

Computing normal probabilities other than standard normal

Many variables in business applications are assumed to be normally distributed, including rates of returns to stocks and other financial assets. Although these variables are normal, they're usually not *standard* normal. As a result, you can't compute probabilities for these variables from the standard normal tables without first transforming them into the equivalent standard normal form, as shown with the following formula:

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

In this expression, Z is a standard normal random variable, and X is a normal random variable with mean μ and standard deviation σ .

For example, suppose that the annual return of the stock of the Gamma Corporation is normally distributed with a mean of 5 percent and a standard deviation of 2 percent. What's the probability that the return from this stock over the coming year will be 4 percent or less?

Let X be a random variable that represents “the annual return for the stock of Gamma Corporation.” X is a normally distributed random variable with a mean (μ) of 0.05 and a standard deviation (σ) of 0.02. (Note that the percentages are converted into decimals for convenience.) X is *not* standard normal, because the mean isn't 0 and the standard deviation isn't 1.

To compute this probability, convert the rate of return X into a standard normal random variable Z as follows:

$$P(X \leq 0.04) = P\left(Z \leq \frac{X - \mu}{\sigma}\right) = P\left(Z \leq \frac{0.04 - 0.05}{0.02}\right) = P(Z \leq -0.5)$$

Based on the standard normal tables (refer to Tables 9-1 and 9-2 in the earlier section “Computing standard normal probabilities”), $P(Z \leq -0.5) = 0.3085$, so the probability that the stock’s return will be 4 percent or less is 0.3085 or 30.85 percent.

Similarly, you can determine the probability that the stock’s return next year will be 8 percent or more like so:

$$P(X \geq 0.08) = P\left(Z \geq \frac{X - \mu}{\sigma}\right) = P\left(Z \geq \frac{0.08 - 0.05}{0.02}\right) = P(Z \geq 1.5)$$

Recall from the earlier section “Computing ‘greater than or equal to’ standard normal probabilities” the following key property for the standard normal distribution:

$$P(Z \leq z) + P(Z \geq z) = 1$$

Rearranging this algebraically gives:

$$P(Z \geq z) = 1 - P(Z \leq z)$$

Therefore,

$$P(Z \geq 1.5) = 1 - P(Z \leq 1.5)$$

Based on the standard normal table (Table 9-1):

$$P(Z \leq 1.5) = 0.9332$$

Therefore,

$$\begin{aligned} &1 - P(Z \leq 1.5) \\ &= 1 - 0.9332 \\ &= 0.0668 \\ &= 6.68\% \end{aligned}$$

The probability that the stock's return next year will be between 7 percent and 8 percent can be computed as follows:

$$\begin{aligned}P(0.07 \leq X \leq 0.08) &= P\left(\frac{0.07 - 0.05}{0.02} \leq Z \leq \frac{0.08 - 0.05}{0.02}\right) \\&= P(1.00 \leq Z \leq 1.50) \\&= P(Z \leq 1.50) - P(Z \leq 1.00) \\&= 0.9332 - 0.8413 \\&= 0.0919 \\&= 9.19\%\end{aligned}$$

As another example, imagine that the scores on a standardized test are normally distributed with a mean score of 80 and a standard deviation of 10. If a student receives a score of 90, that student was outperformed by what proportion of all other students taking the test?

In other words, what is the probability of receiving a score of at least 90 on this test? Let X represent the random variable "score on the exam." X is a normally distributed random variable with a mean of 80 and a standard deviation of 10. Because X isn't a standard normal random variable, you must convert it:

$$P(X \geq 90) = P\left(Z \geq \frac{X - \mu}{\sigma}\right) = P\left(Z \geq \frac{90 - 80}{10}\right) = P(Z \geq 1.00)$$

Due to the symmetry of the standard normal distribution,

$$P(Z \geq 1.00) = 1 - P(Z \leq 1.00)$$

From the standard normal table (Table 9-1),

$$P(Z \leq 1.00) = 0.8413$$

Therefore,

$$\begin{aligned}P(Z \geq 1.00) &= 1 - P(Z \leq 1.00) \\&= 1 - 0.8413 \\&= 0.1587\end{aligned}$$

or only 15.87 percent of the students taking the exam scored 90 or better.

These techniques can be used to compute *any* normal probability, regardless of the mean and standard deviation of the distribution.

Computing Probabilities for the Normal Distribution with the TI-84 Plus Calculator

The Texas Instruments TI-84 Plus and Plus CE calculators contain the functions needed to compute any normal probability. The normal distribution is uniquely characterized by two values: the mean (μ) and the standard deviation (σ). The standard normal distribution is the special case where the mean is zero and the standard deviation is one.

As an example, suppose that it has been determined that the rate of return to a stock is normally distributed with a mean of 5 percent per year and a standard deviation of 2 percent per year. It is possible to compute the probability that next year's returns will fall between any two values, such as between 2 percent and 8 percent. (For the normal distribution, the probability that a variable is exactly equal to a specific value is defined to be equal to zero.)

Suppose that an investor is interested in the probability that the returns to the stock will be between 2 percent and 8 percent next year. This can be computed by pressing [2nd], [VARS], and then selecting 2:normalcdf.



REMEMBER

The following menu is produced when choosing 2:normalcdf:

lower:

upper:

μ :

σ :

Paste

Each normal probability refers to a range of values; lower refers to the lower limit of this range, and upper refers to the upper limit of this range. μ refers to the mean of the normal distribution and σ refers to the standard deviation of the normal distribution. In this example, using whole numbers to represent percentages, the appropriate values are:

lower: 2

upper: 8

μ : 5

σ : 2

Paste

where each entry is a percentage. Alternatively, the percentages can be expressed as decimals:

lower: 0.02

upper: 0.08

μ : 0.05

σ : 0.02

Paste

After choosing Paste and pressing the [ENTER] button twice, the resulting probability is 0.8663855426. This shows that the probability that the stock's returns will be between 2 percent and 8 percent next year is about 86.64 percent.

To compute a standard normal probability, the values of μ and σ are changed to 0 and 1, respectively. For example, the probability that a standard normal random variable will be between -1 and 1 is computed as:

lower: -1

upper: 1

μ : 0

σ : 1

Paste

After choosing Paste and pressing [ENTER] twice, the resulting probability is 0.6826894809. This shows that the probability that the variable will be between -1 and 1 is about 68.27 percent.



REMEMBER

When entering negative numbers, press the negative key (-) and then the number.



TIP

For some situations, the lower limit may be negative infinity or the upper limit may be positive infinity. The calculator does not have an infinity key; however, there are workarounds.

Negative infinity is implemented as: -2nd, 99

Typing in the negative sign (found next to the period) followed by pressing the [2nd] button, the comma (found above the 7 key), and then 99 shows up on the calculator as: -1E99 or -E99. This represents -1.0×10^{99} , which is the smallest value the calculator can express.

Positive infinity is implemented as: 2nd, 99

This is identical to negative infinity except that the negative sign is not used. This shows up on the calculator as 1E99 or E99. This represents 1.0×10^{99} , which is the largest value the calculator can express.

IN THIS CHAPTER

- » Getting familiar with sampling techniques
- » Using sampling distributions to estimate probabilities

Chapter **10**

Sampling Techniques and Distributions

A *population* is a collection of data that we are interested in studying; a *sample* is a selection of data randomly chosen from a population. The use of sample data is the basis for a wide variety of business applications. This is because obtaining information about an entire population is likely to be very time-consuming and costly. Instead, samples may be used to understand the behavior of the underlying population.

One of the requirements of using samples to draw conclusions about a population is that the samples accurately mirror the population; otherwise, any conclusions that are reached about the population are bound to be inaccurate. Several different types of sampling techniques have been developed to accurately capture the properties of a population. The choice of technique depends on several factors, such as:

What are the demographic characteristics of interest?

How easy will it be to obtain sample data?

How much data is needed to ensure accurate results?

For example, suppose that the New York State government wants to analyze the distribution of ages of everyone living in the state. This helps determine what type of funding is needed for various programs in the future. Although the ages of every single resident can be collected, this may be very time consuming and costly.

Instead, suppose that the government decides to randomly sample residents throughout the state and use this information to estimate the distribution of ages. Clearly, it makes no sense to focus only on high school students, because their ages are substantially lower than the overall population. Instead, samples are chosen that ideally match the demographic characteristics of the entire state. For example, questionnaires can be mailed to randomly chosen addresses throughout the state.

In this chapter, I introduce several types of sampling techniques that may be used for various types of studies. I also show you a special type of probability distribution, known as a *sampling distribution*. This is a special type of probability distribution that describes the properties of a *sample statistic*. (Sample statistics are summary measures of a sample; these include the sample mean, sample variance and sample standard deviation. Sample statistics are discussed in Chapters 3 and 4.) Due to its widespread use in statistical analysis, I focus on the sampling distribution of the *sample mean*.

Sampling Techniques: Choosing Data from a Population

Statistical inference is a methodology that lets you draw conclusions about a population from sample data. One of the most important challenges in statistical inference is choosing samples that accurately reflect the characteristics of the underlying population. Although you can choose from many sampling techniques, the appropriate technique depends on the type of information you're studying and your resources.

You can classify the two basic approaches to sampling as probability sampling and nonprobability sampling. Probability sampling is used when it is important to ensure that each member of a population has a chance of being chosen. Nonprobability sampling is a more subjective approach, and is often used when it would be difficult or impossible to use probability sampling. I explore both of these approaches in the following sections.

Probability sampling

When you use *probability sampling*, each member of the population has a chance of being chosen for the sample. In some of these techniques, each population member is *equally likely* to be chosen; in others, this is not the case. With probability sampling, it's possible to determine the probability that a given member of the population will be chosen. Within the category of probability sampling, you can choose from four types of sampling techniques, which I discuss in the following sections.

Simple random samples

In a simple random sample, each member in the population is equally likely to be chosen. There are several different ways in which population members may be chosen with equal probability. One approach is to assign a numerical value to each population member and then randomly choose numbers that correspond to these members. For example, suppose that a population consists of the following ten members of the finance faculty at a prestigious university:

- 1.** Benjamin Harrison
- 2.** Martin Van Buren
- 3.** John Tyler
- 4.** Millard Fillmore
- 5.** Grover Cleveland
- 6.** Chester Arthur
- 7.** James Polk
- 8.** Zachary Taylor
- 9.** James Buchanan
- 10.** Franklin Pierce

You would like to randomly choose five of these faculty members for a newly formed committee. You assign each faculty member a number from one to ten. (This can be done alphabetically or in any number of other ways.) To choose a simple random sample of five of these faculty members, you can use a random number generator.

A random number generator is a function that can be used to randomly choose numbers within a specified interval. As an example, you can use Microsoft Excel's `RANDBETWEEN` function; this generates whole numbers that are randomly chosen between any two values you specify.



TECHNICAL STUFF

For this example, you would need to generate a random number between 1 and 10. You would then enter `RANDBETWEEN(1,10)` into Excel and record the resulting number. You would repeat this process until you have five unique numbers. The faculty members associated with these numbers are then chosen for the new committee.

In this example, you don't want to choose the same number twice; if this happens, you simply discard the result and choose another random number until you have five unique numbers. The process you are using is known as sampling *without* replacement. If you are willing to choose the same number more than once, then no results would be discarded; the process that you would be using is known as sampling *with* replacement.

Suppose that the following sequence of random numbers is chosen:

`RANDBETWEEN(1,10) = 1`

`RANDBETWEEN(1,10) = 4`

`RANDBETWEEN(1,10) = 5`

`RANDBETWEEN(1,10) = 8`

`RANDBETWEEN(1,10) = 6`

Your simple random sample would then consist of the following faculty members:

1. Benjamin Harrison
4. Millard Fillmore
5. Grover Cleveland
8. Zachary Taylor
6. Chester Arthur

These are the lucky members of the new committee.

Systematic samples

With *systematic samples*, population members are assigned a numerical value, as is the case with simple random samples. Instead of using random numbers to choose population members, though, you will instead use a specific *sequence* of numbers.

For example, suppose an economist wants to study the distribution of 100 household incomes in a small town and wants to draw a sample size of ten. In this case, the economist draws every tenth population member (because the number of households divided by the sample size equals $100/10 = 10$). One way the economist

can draw every tenth member is to start with a random number (between 1 and 10) and then add ten to each number to get the desired sequence.

For example, you can use RANDBETWEEN(1,10) to obtain the starting value for the sequence. If this turns out to be a 3, then the appropriate sequence of random numbers would be:

3, 13, 23, 33, 43, 53, 63, 73, 83, 93

If instead the function RANDBETWEEN(1,10) generates a 5, then the appropriate sequence of random numbers would be:

5, 15, 25, 35, 45, 55, 65, 75, 85, 95

Other techniques can be used to randomly choose the first value, such as the flip of a coin, the roll of a die, and so on. Similarly, if a population contains 1,200 members and the economist wants a sample size of ten, the numbering sequence includes every 120th member ($1,200/10 = 120$). One way the economist can draw every 120th member is to start with a random number (between 1 and 120) and then add 120 to each number to get the sequence.

In this case, suppose that the function RANDBETWEEN(1,120) results in a value of 57; then the sequence would consist of the following values:

57, 177, 297, 417, 537, 657, 777, 897, 1017, 1137

As another example, suppose that a marketing firm wants to find out whether consumers are responding favorably to a newly launched advertising campaign. A researcher can choose a busy mall and ask every 20th customer that walks by how they feel about the new advertising campaign. In this case, though, the researcher wouldn't have a specific sequence of numbers, because it's impossible to determine in advance how many people are in the mall at any given time.

In this case, systematic samples are chosen based on incomplete knowledge of the underlying population. This approach is useful when the size of the entire population is not known.

Stratified samples

When using *stratified samples*, you divide a population into *strata* (levels or layers). The strata may reflect any of a wide variety of characteristics of the population data, such as ages, incomes, levels of education, and so on.

Basically, you choose a stratified sample in such a way that you ensure that the proportion of sample members in each stratum (singular of *strata*) matches the distribution found in the population.

For example, suppose that a college wants to conduct a survey of student attitudes toward the building of a costly new sports stadium as an alternative to expanding the current antiquated library. Instead of surveying every single student in the school, the college chooses stratified samples. It divides the entire student body by class: freshmen, sophomores, juniors, and seniors. (Assume for this example that the school doesn't offer any graduate programs, so all students belong to one of these four classes.) Here's how the classes break down:

Class	Number of Students
Freshmen	800
Sophomores	1,200
Juniors	1,000
Seniors	1,000

And the percentages of students in each class are as follows:

Class	Number of Students	Percentage of Total
Freshmen	800	20 percent
Sophomores	1,200	30 percent
Juniors	1,000	25 percent
Seniors	1,000	25 percent

If the college chooses a stratified sample of 200 students, the sample consists of the following:

40 freshmen (20 percent of 200)

60 sophomores (30 percent of 200)

50 juniors (25 percent of 200)

50 seniors (25 percent of 200)

Within each stratum, a simple random sample of the appropriate number of students is chosen. This selection method ensures that no class is under- or overrepresented in the sample data.

One of the advantages of the stratified sample approach is that you can draw conclusions about each individual stratum. For example, the college can analyze the attitudes of freshmen separately from the attitudes of sophomores, juniors, and seniors. On the other hand, one of the disadvantages of this approach is that you need more information about the characteristics of the population than with other approaches, such as the simple random sampling approach discussed earlier. In this example, you need to know the distribution of students among the freshman, sophomore, junior, and senior classes.

Cluster samples

With *cluster samples*, you subdivide a population into groups based on common characteristic (such as location, age, income level, and so forth). You choose groups randomly, and then you choose samples from those groups randomly.

Say you're a researcher conducting a national survey about attitudes toward proposed national legislation. You divide the entire voting age population of the United States into groups according to state of residency. You decide to choose a sample of eight states; you believe that this is sufficient to represent the entire country. In this case, you would first assign a number to each state in the United States. Next, you can use the function `RANDBETWEEN(1,50)` until you choose eight different states.

Within each selected state, voting age residents are randomly chosen using a simple random sample. This may be accomplished by assigning a number to each registered voter and then using a random number generator to randomly pick the desired number of voters.

Suppose that the following states are chosen:

- Wisconsin
- Rhode Island
- Michigan
- Utah
- Illinois
- South Carolina
- Arizona
- Oregon

Within each state, you choose simple random samples of voters.

The advantage of using cluster sampling is that it can be implemented more quickly and cheaply than stratified sampling. In this example, stratified sampling requires voters to be randomly chosen from each of the 50 states. The disadvantage of using cluster sampling is that it may not be as accurate as stratified sampling.

Nonprobability sampling

Unlike probability sampling, *nonprobability sampling* doesn't guarantee that each population member has a chance of being chosen. And with nonprobability sampling, you have to use subjective judgment. One of the major drawbacks to non-probability samples is that the results aren't as reliable for drawing conclusions about the overall population. It may be easier to get the samples, but there's a price — they're less useful than probability samples.

I discuss four of the nonprobability sampling techniques in the following sections, including convenience samples, quota samples, purposive samples, and judgment samples.

Convenience samples

When you choose population members primarily because they're accessible, you're using *convenience samples*. For example, if a marketing firm needs to study consumer attitudes toward new products, it may be forced to rely on the input of people who are willing to participate; they are not necessarily representative of the overall population.

Suppose for example a marketing firm decides to conduct a series of interviews at a mall to determine which new movies are likely to do well at the box office. The interviews are conducted at 3:00 in the afternoon on a Wednesday. Although there may be many volunteers who are willing to take part in the interviews, most or all of them are likely to be students and/or retirees, which doesn't reflect the overall population. Unless the marketing firm is only interested in the views of these groups, the results are not likely to be accurate.

Quota samples

Quota samples are closely related to stratified samples; in both cases, you divide population members into separate groups. The main difference is that with a quota sample, the number of sample members in each stratum may not exactly represent the numbers in the underlying population.

For example, suppose that a college is interested in comparing the GPA of its male and female students. Assume that the proportion of male students at this college is 60 percent, so the proportion of female students is 40 percent. A stratified sample would ensure that 60 percent of the sample members are male, and 40 percent are female. With a quota sample, any number of males and females may be chosen. Suppose that the college doesn't know the exact proportion of male and female students, so it decides to choose an equal mix of male and female students for the sample. Clearly, this doesn't reflect the proportions in the actual population.

Purposive samples

With *purposive samples*, you choose members of the population because they're *not* typical in some important way. For example, a company that produces a new product may be concerned that the product is too expensive for the average consumer to buy. The company may target students (who presumably have low incomes) to determine whether they'd consider buying the product. The logic is that if the product isn't too expensive to people with relatively low incomes, it won't be too expensive to people with higher incomes.

As another example, suppose that a snack foods company manually inspects all the potato chips that it produces before they are sold to the public. Any chips that appear to be burned are automatically discarded. This process is very time consuming and costly; the company wants to try a different approach.

Suppose that the smallest chips are most likely to be burned. Rather than inspecting every single potato chip, the company decides to save time by only inspecting the chips that appear to be unusually small. If these are not burned, the remaining chips are probably acceptable. The company is now using purposive samples to represent the entire population.

Judgment samples

When conducting a study with a *judgment sample*, you chose members based on your subjective judgment. You choose these members because they offer specific characteristics of interest. For example, suppose that half of the residents of a city are male (and, therefore, half are female). A handbag manufacturer wants to determine which features are most important to consumers in this city. If the company chooses to survey customers in the local mall, it may go out of its way to question a larger number of female customers (rather than male customers) because most handbags are purchased by women.

Sampling Distributions

A *statistic* is a summary measure of a sample, and a *parameter* is a summary measure of a population. (I discuss both summary measures of samples and populations in Chapters 3, 4, and 5.) The probability distribution of a statistic is known as a sampling distribution, which is what this section is all about.

Some examples of statistics include

- » Sample mean (\bar{X})
- » Sample variance (s^2)
- » Sample standard deviation (s)

Some examples of parameters are

- » Population mean (μ)
- » Population variance (σ^2)
- » Population standard deviation (σ)



REMEMBER

In many cases, a *population parameter* is costly and time-consuming to calculate. For example, figuring out the average age of everyone living in the United States would be very time-consuming! In these cases, the statistician uses sample statistics instead. The sample mean (\bar{X}) estimates the population mean (μ). The researcher can use a representative sample of U.S. residents to compute a sample mean, which would serve as an estimated value of the average age of all U.S. residents.

If you repeatedly draw samples from a population, the value of a statistic is most likely different for each sample. As a result, it's useful to think of a statistic as a *random variable* whose properties can be described with a *probability distribution*. (See Chapter 7 for details.)

In the following sections, I explore the characteristics of sampling distributions, including how to represent data from a sampling distribution graphically and how to compute the moments of a sampling distribution. The focus is on the sampling distribution of the sample mean \bar{X} .

Portraying sampling distributions graphically

As I explain in Chapter 2, a *histogram* is a graphical representation of data in which ranges of values, known as *classes*, appear on the horizontal axis (the x -axis) and probabilities on the vertical axis (the y -axis). Each class is shown as a single bar whose height equals the probability of that class.

A histogram shows at a glance how the values of a variable are distributed. In this section, histograms are used to describe the properties of the sampling distribution of \bar{X} .

One of the benefits of using histograms to analyze a sampling distribution is that it is easy to see if the sampling distribution is symmetrical about the mean, negatively skewed, or positively skewed.

A distribution is symmetrical about the mean if values below the mean occur as frequently as the values an equal distance above the mean. A negatively skewed distribution is one in which there are a small number of extremely small values; a positively skewed distribution is one in which there are a small number of extremely large values. (Skewness and symmetry are discussed in Chapter 3.)



A distribution is symmetrical about the mean if the mean equals the median. A distribution is negatively skewed if the mean is less than the median and positively skewed if the mean is greater than the median.

A histogram also shows at a glance the center or mean of a distribution, and how “spread out” are the members of the distribution. (Recall from Chapter 4 that the spread of a distribution is measured by its variance and its standard deviation.)

A histogram can be used to compare the properties of different sampling distributions or to observe the effect of different sample sizes on a sampling distribution. For example, suppose that a manufacturer of computer chips has found from experience that its assembly line produces two defective chips per hour, and that the number of defective chips produced during a given hour is independent of the number produced during any other hour. In other words, the distribution of defective chips follows the Poisson distribution with an average value of two per hour — in other words, $\lambda = 2$. (The Poisson distribution is discussed in detail in Chapter 8.) The distribution of defective chips is shown in Figure 10-1.

Suppose that a sample of five computer chips is randomly chosen, and the number of defective chips in each sample is recorded. This process is repeated 300 times. The resulting distribution consists of 300 sample means, ranging from a low of 0.6 to a high of 4.2. Figure 10-2 shows the distribution of the mean number of defective chips among the 300 samples of size 5.

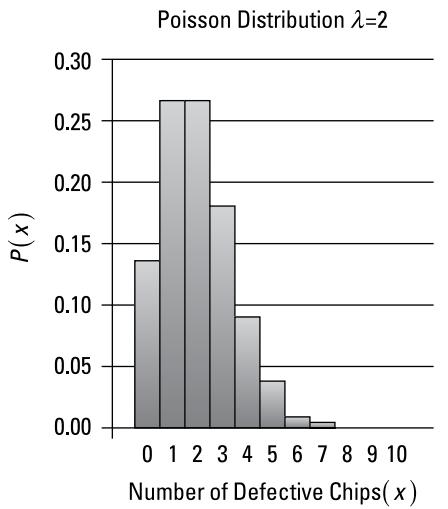


FIGURE 10-1:
Histogram for the distribution of defective chips.

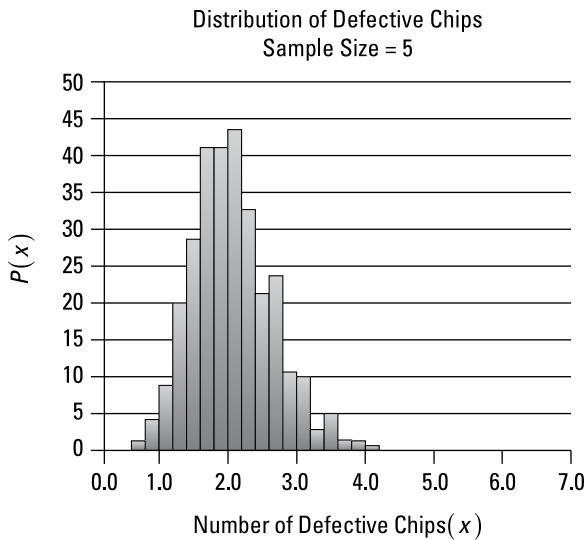


FIGURE 10-2:
Histogram of a sampling distribution of defective computer chips with a sample size of 5.

Note that the distribution of sample means with a sample size of 5 strongly resembles the Poisson distribution.

Suppose now that a sample of 30 computer chips is randomly chosen, and the number of defective chips in each sample is recorded. This process is repeated 300 times. The resulting distribution consists of 300 sample means, ranging from a low of 1.3 to a high of 3. Figure 10-3 shows the distribution of the mean number of defective chips among the 300 samples of size 30.

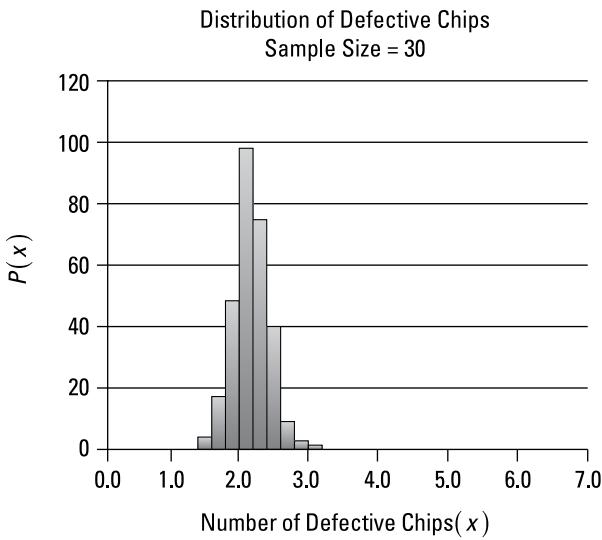


FIGURE 10-3:
Histogram of a sampling distribution of defective computer chips with a sample size of 30.

Note that the distribution of sample means with a sample size of 30 much more closely resembles the normal distribution than the Poisson distribution.

Figures 10-2 and 10-3 show that the sample mean remains centered on 2 regardless of the sample size, but the mean number of defectives is far less dispersed around the mean with a sample size of 30 compared with a sample size of 5. (You can tell that this is the case because the sample mean ranges from 0.6 to 4.2 with a sample size of 5, compared with 1.3 to 3 for a sample size of 30.)

In addition, the figures show that as the sample size grows from 5 to 30, the sampling distribution looks more like the normal distribution.

Moments of a sampling distribution

A sampling distribution is described by a series of summary measures known as *moments*, which include expected value (mean) and variance. The standard deviation is not a separate moment; it is the square root of the variance. The standard deviation of a sampling distribution is often referred to as the *standard error*.

For the sampling distribution of \bar{X} , the expected value is $\mu_{\bar{X}}$, which equals the mean of the underlying population (μ). The variance is $\sigma_{\bar{X}}^2$, and the standard deviation, also known as the *standard error*, is $\sigma_{\bar{X}}$.

The values of the variance and standard error depend on the relationship between the size of the sample (n) drawn from the population and the size of the population (N).

- » If the sample size is less than or equal to 5 percent of the population size, the sample is small, relative to the size of the population. In this case, the variance of \bar{X} equals

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

Here, σ^2 is the variance and σ is the standard deviation of the underlying population; n is the sample size.

The square root of the variance of \bar{X} is the standard error of \bar{X} :

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- » If the sample size is greater than 5 percent of the population size, the sample is large, relative to the size of the population. In this case, the standard error of \bar{X} equals

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

The term $\sqrt{\frac{N-n}{N-1}}$ is known as the *finite population correction factor*, which always assumes a value of less than or equal to 1 (it equals 1 only if the sample size is 1). You use the finite population correction factor to reduce the size of the standard error to reflect the fact that less variability from one sample mean to the next occurs when the sample size is large relative to the population.

The Central Limit Theorem

According to the central limit theorem, the sampling distribution of \bar{X} is normal if the underlying population is normal. If not, the sampling distribution of \bar{X} is at least approximately normal if the sample size is at least 30. Under these circumstances, you can use the normal distribution to determine the probability that the sample mean will fall within a specified range of values. (See Chapter 9 for techniques on using the normal distribution.)

For example, suppose you choose a sample of 50 gasoline prices from gas stations in a major city. You can use the normal distribution to determine the probability that the sample mean gas price is between \$3.50 and \$4.00 per gallon.

If the central limit theorem fails to hold, you can't use the normal distribution to compute probabilities for the sample mean; instead, you need to find an alternative probability distribution that closely resembles the population that you are studying.

Converting \bar{X} to a standard normal random variable

Based on the central limit theorem, if you draw samples from a population of $n \geq 30$, then \bar{X} is a normally distributed random variable. To determine probabilities for \bar{X} , you may use the standard normal probability tables. (These are discussed in Chapter 9.) Using the standard normal tables requires you to convert \bar{X} to a standard normal random variable.



REMEMBER

The standard normal distribution is the special case where the mean (μ) equals 0, and the standard deviation (σ) equals 1.

For any normally distributed random variable X with a mean μ and a standard deviation σ , you find the corresponding standard normal random variable (Z) with the following equation:

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

For the sampling distribution of \bar{X} , the corresponding equation is

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

As an example, suppose that there are 10,000 stocks trading each day on a regional stock exchange. It's known from historical experience that the returns to these stocks have a mean value of 10 percent per year, and a standard deviation of 20 percent per year.

An investor chooses to buy a random selection of 100 of these stocks for their portfolio. What's the probability that the mean rate of return among these 100 stocks is greater than 8 percent?

The investor's portfolio can be thought of as a sample of stocks chosen from the population of stocks trading on the regional exchange. The first step to finding this probability is to compute the moments of the sampling distribution.

» **Compute the mean:** $\mu_{\bar{X}} = \mu = 0.10$

The mean of the sampling distribution equals the population mean.

» **Determine the standard error:** This calculation is a little trickier because the standard error depends on the size of the sample relative to the size of the population. In this case, the sample size (n) is 100, while the population size (N) is 10,000. So you first have to compute the sample size relative to the population size, like so:

$$n/N = 100/10,000 = 0.01 = 1\%$$

Because 1 percent is less than 5 percent, you don't use the finite population correction factor to compute the standard error. Note that in this case, the value of the finite population correction factor is:

$$\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{10,000-100}{10,000-1}} = \sqrt{\frac{9,900}{9,999}} = 0.995$$

Because this value is so close to 1, using the finite population correction factor in this case would have little or no impact on the resulting probabilities. And because the finite population correction factor isn't needed in this case, the standard error is computed as follows:

$$\begin{aligned}\sigma_{\bar{X}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{0.20}{\sqrt{100}} \\ &= \frac{0.20}{10} \\ &= 0.02\end{aligned}$$

To determine the probability that the sample mean is greater than 8 percent, you must now convert the sample mean into a standard normal random variable using the following equation:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

To compute the probability that the sample mean is greater than 8 percent, you apply the previous formula as follows:

$$P(\bar{X} \geq 0.08) = P\left(Z \geq \frac{0.08 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right)$$

Because $\mu_{\bar{X}} = 0.10$ and $\sigma_{\bar{X}} = 0.02$, these values are substituted into the previous expression as follows:

$$P\left(Z \geq \frac{0.08 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) = P\left(Z \geq \frac{0.08 - 0.10}{0.02}\right) = P(Z \geq -1.00)$$

You can calculate this probability by using the properties of the standard normal distribution along with a standard normal table such as Table 10-1.

TABLE 10-1 Standard Normal Table — Negative Values

Z	0.00	0.01	0.02	0.03
-1.3	0.0968	0.0951	0.0934	0.0918
-1.2	0.1151	0.1131	0.1112	0.1093
-1.1	0.1357	0.1335	0.1314	0.1292
-1.0	0.1587	0.1562	0.1539	0.1515

Table 10-1 shows the probability that a standard normal random variable (designated Z) is *less than or equal to* a specific value. For example, you can write the probability that $Z \leq -1.00$ (one standard deviation below the mean) as $P(Z \leq -1.00)$. You find the probability from the table with these steps:

1. Locate the first digit before and after the decimal point (-1.0) in the first (Z) column.
2. Find the second digit after the decimal point (0.00) in the second (0.00) column.
3. See where the row and column intersect to find the probability:
 $P(Z \leq -1.00) = 0.1587$.

Because you're actually looking for the probability that Z is *greater than or equal to* -1, one more step is required.

Due to the *symmetry* of the standard normal distribution, the probability that Z is greater than or equal to a negative value equals one minus the probability that Z is less than or equal to the same negative value. For example,

$$P(Z \geq -2.0) = 1 - P(Z \leq -2.0)$$

This is because $Z \geq -2.00$ and $Z \leq -2.00$ are *complementary* events. (Complementary events are discussed in Chapter 6.) This means that Z must either be greater than or equal to -2 or less than or equal to -2 . Therefore,

$$P(Z \geq -2.0) + P(Z \leq -2.0) = 1$$

This is true because the occurrence of one of these events is *certain*, and the probability of a certain event is 1 . (Probability and certain events are covered in Chapter 6.)

After algebraically rewriting this equation, you end up with the following result:

$$P(Z \geq -2.0) = 1 - P(Z \leq -2.0)$$

For the portfolio example,

$$P(Z \geq -1.0) = 1 - P(Z \leq -1.0)$$

$$P(Z \geq -1.0) = 1 - 0.1587 = 0.8413$$

The result shows that there's an 84.13 percent chance that the investor's portfolio will have a mean return greater than or equal to 8 percent.

As another example, suppose that it is known that there are 120 surviving paintings by a well-known 19th century artist. These works have an average price of $\$1$ million and a standard deviation of $\$120,000$. Suppose that an art collector acquires a random selection of ten of these paintings. What's the probability that the mean price of these paintings is between $\$975,000$ and $\$1,025,000$?

In this case, the size of the population is $N = 120$. The sample size is $n = 10$. Therefore, the sample size represents $n/N = 10/120 = 0.08333$, which is 8.333 percent of the population. Because the sample size is *greater than* 5 percent, you use the finite population correction factor to compute the standard error, like so:

$$\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{120-10}{120-1}} = 0.96144$$

You then find the mean ($\mu_{\bar{X}} = \mu = 1,000,000$) and standard error of \bar{X} :

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{120,000}{\sqrt{10}} (0.96144) = 36,484$$

To calculate probabilities for \bar{X} , the first step is to convert the values of \bar{X} into standard normal random variables:

$$\begin{aligned} & P(975,000 \leq \bar{X} \leq 1,025,000) \\ &= P\left(\frac{975,000 - 1,000,000}{36,484} \leq Z \leq \frac{1,025,000 - 1,000,000}{36,484}\right) \\ &= P(-0.69 \leq Z \leq 0.69) = P(Z \leq 0.69) - P(Z \leq -0.69) \end{aligned}$$

The next step is to find the values of $P(Z \leq 0.69)$ and $P(Z \leq -0.69)$, and subtract one from the other. The art collector can get these values from standard normal tables, such as Table 10-2 and Table 10-3.

TABLE 10-2

Standard Normal Table — Positive Values

Z	0.06	0.07	0.08	0.09
0.5	0.7123	0.7157	0.7190	0.7224
0.6	0.7454	0.7486	0.7517	0.7549
0.7	0.7764	0.7794	0.7823	0.7852
0.8	0.8051	0.8078	0.8106	0.8133

TABLE 10-3

Standard Normal Table — Negative Values

Z	0.06	0.07	0.08	0.09
-0.8	0.1949	0.1922	0.1894	0.1867
-0.7	0.2236	0.2206	0.2177	0.2148
-0.6	0.2546	0.2514	0.2483	0.2451
-0.5	0.2877	0.2843	0.2810	0.2776

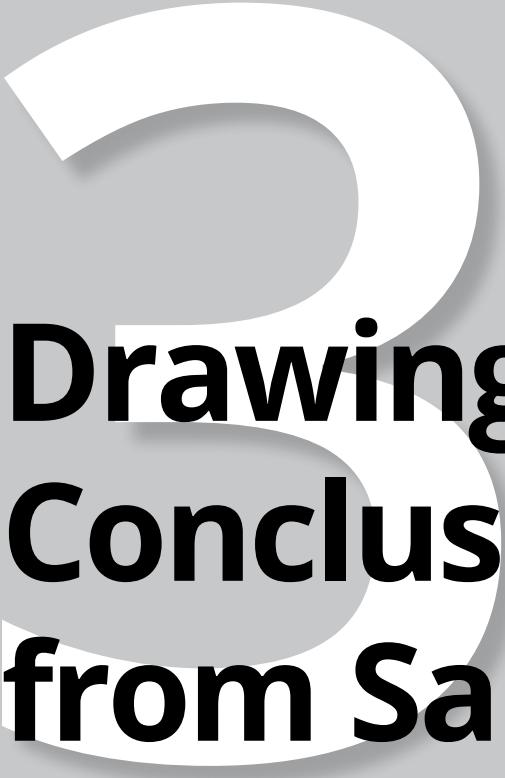
Table 10-2 shows that $P(Z \leq 0.69) = 0.7549$. This value is at the intersection of the 0.6 row for Z and the 0.09 column. Table 10-3 shows that $P(Z \leq -0.69) = 0.2451$. This value is at the intersection of the -0.6 row for Z and the 0.09 column.

Therefore,

$$P(Z \leq 0.69) - P(Z \leq -0.69) = 0.7549 - 0.2451 = 0.5098$$

The result is that there's a 50.98 percent chance that the sample mean falls somewhere between \$975,000 and \$1,025,000.

See Chapter 9 for instructions on using the Texas Instruments TI-84 Plus and Plus CE calculators to compute normal probabilities.



Drawing Conclusions from Samples

IN THIS PART . . .

Use confidence intervals to provide a range of possible values for a population parameter; these can be constructed for any population parameter: mean, variance, standard deviation, and so on.

Discover the techniques that are used to estimate confidence intervals for the population mean, including the standard normal distribution and the Student's t-distribution.

Draw conclusions about population properties — from a single population variance to multiple population variances — with hypothesis testing.

Use the chi-square distribution to test whether a population conforms to a specific probability distribution — a *goodness of fit* test.

Test hypotheses about two population variances with the F-distribution.

IN THIS CHAPTER

- » Getting familiar with the t-distribution
- » Developing techniques for constructing confidence intervals

Chapter **11**

Confidence Intervals and the Student's t-Distribution

A *confidence interval* is a range of numbers that's likely to contain the true value of an unknown population *parameter*, such as the population mean. (Parameters are numerical values that describe the properties of a population; they are discussed in Chapter 10.)

Here's an example. Suppose that you are asked to estimate how long it takes to commute to work each day. You respond by saying, "On average, it takes about 20 minutes to get to work." This estimate may be useful, but it doesn't give any indication how much your commuting time may vary from one day to the next.

Suppose that instead you respond by saying "Most days, it takes between 15 and 25 minutes to get to work." This range of values is more meaningful than the estimated average time of 20 minutes. With this interval, it's clear that the average commute time is 20 minutes (because this is halfway between 15 and 25 minutes). In addition, the numbers tell you that it'll be an unusual day if your commuting time is more than 25 minutes or fewer than 15 minutes.

This range of estimated values is known as a *confidence interval*. The starting point in constructing a confidence interval is the estimated mean or average, which is 20 minutes in this example. The next step is to construct a *margin of error*, which represents the degree of uncertainty associated with the estimated mean. In this example, the margin of error is five minutes.

Confidence intervals may be constructed for any population parameter: mean, variance, standard deviation, and so on. This chapter covers the techniques that are used to estimate confidence intervals for the population mean. These techniques are based on one of two probability distributions. One of these is the standard normal distribution (which I cover in detail in Chapter 9). The other is known as the *Student's t-distribution* (also known simply as the *t-distribution*) — which I introduce in this chapter.

Almost Normal: The Student's t-Distribution

The purpose of the t-distribution is to describe the statistical properties of sample means when the population standard deviation is unknown, while the standard normal distribution is used when the population standard deviation is known.

Properties of the t-distribution

The t-distribution shares a few key properties with the standard normal distribution. These properties are as follows:

- » They have a mean of 0.
- » They're *symmetric* about the mean (that is, the area below the mean is a mirror image of the area above the mean).
- » They can be described graphically with a bell-shaped curve.

Several key differences also exist between the two distributions, including the following:

- » The t-distribution has more area in the "tails," and less area near the mean than the standard normal distribution.
- » The variance and standard deviation of the t-distribution are larger than those of the standard normal distribution.

The larger variance and standard deviation in the t-distribution reflect that much more variability occurs when the population standard deviation is unknown as using the sample standard deviation introduces more uncertainty.

Degrees of freedom

As with the normal distribution, the t-distribution is an infinite family of distributions. Whereas the mean and standard deviation uniquely identify each normal distribution, each t-distribution is characterized by a value known as *degrees of freedom (df)*.

When you're estimating the sample mean, the number of degrees of freedom for the t-distribution equals the number of sample members that can *vary*. For example, if you choose a sample of size n to estimate the sample mean \bar{X} , the corresponding t-distribution has $n - 1$ degrees of freedom because the combination of $n - 1$ elements in the sample plus the sample mean uniquely identify the last element in the sample. Therefore, you have only $n - 1$ independent variables in the sample.

As an example, suppose that you choose a sample of three students to estimate the mean GPA of a university. If the sample mean, \bar{X} , equals 3.0, the first student's GPA is 2.5, the second student's GPA is 3.5, and the third student's GPA must be 3.0 because the sum of the GPAs must be 9.0 for the sample mean to be 3.0. As a result, the GPAs of any two students in this sample, along with the value of \bar{X} , uniquely determine the value of the third student's GPA. Therefore, the corresponding t-distribution has 2 degrees of freedom.

Moments of the t-distribution

A *moment* is a summary measure of a probability distribution (see Chapter 7 for a detailed explanation on moments). Probability distributions, including the t-distribution, have several moments, including:

- » The first moment of a distribution is the expected value, $E(X)$, which represents the mean or average value of the distribution.

For the t-distribution with v (the Greek letter "nu") degrees of freedom, the mean (or expected value) equals $\mu = E(X) = 0$. μ represents the mean of a population or a probability distribution, and v commonly designates the number of degrees of freedom of a distribution.

- » The second central moment is the variance (σ^2), and it measures the spread of the distribution about the expected value. The more spread out a distribution is, the more "stretched out" is the graph of the distribution. In other

words, the tails will be further from the mean, and the area near the mean will be smaller. For example, based on Figures 11-1 and 11-3 shown later in this chapter, it can be seen that the t-distribution with 2 degrees of freedom is far more spread out than the t-distribution with 30 degrees of freedom.

You use the formula $\sigma^2 = \frac{v}{v-2}$ to calculate the variance of the t-distribution.

As an example, with 10 degrees of freedom, the variance of the t-distribution is computed by substituting 10 for v in the variance formula:

$$\begin{aligned}\sigma^2 &= \frac{v}{v-2} \\ &= \frac{10}{10-2} \\ &= \frac{10}{8} \\ &= 1.25\end{aligned}$$

With 30 degrees of freedom, the variance of the t-distribution equals

$$\begin{aligned}\sigma^2 &= \frac{v}{v-2} \\ &= \frac{30}{30-2} \\ &= \frac{30}{28} \\ &= 1.07\end{aligned}$$

These calculations show that as the degrees of freedom increase, the variance of the t-distribution declines, getting progressively closer to 1.

- » The standard deviation is the square root of the variance (σ). (It is not a separate moment.)

For the t-distribution, you find the standard deviation with this formula:

$$\sigma = \sqrt{\frac{v}{v-2}}$$



TIP

For most applications, the standard deviation is a more useful measure than the variance because the standard deviation and expected value are measured in the same units while the variance is measured in *squared* units. For example, suppose you assume that the returns on a portfolio follow the t-distribution. You measure both the expected value of the returns and the standard deviation as a percentage; you measure the variance as a *squared* percentage, which is a difficult concept to interpret.

Graphing the t-Distribution

One of the interesting properties of the t-distribution is that the greater the degrees of freedom, the more closely the t-distribution resembles the standard normal distribution. As the degrees of freedom increases, the area in the tails of the t-distribution decreases while the area near the center increases. (The tails consists of the extreme values of the distribution, both negative and positive.) Eventually, when the degrees of freedom reaches 30 or more, the t-distribution and the standard normal distribution are extremely similar.

Figures 11-1, 11-2, and 11-3 illustrate the relationship between the t-distribution with different degrees of freedom and the standard normal distribution. Figure 11-1 shows the standard normal and the t-distribution with 2 degrees of freedom. Notice how the t-distribution is significantly more spread out than the standard normal distribution.

The graph in Figure 11-1 shows that the t-distribution has more area in the tails and less area around the mean than the standard normal distribution. (The standard normal distribution curve is shown with square markers.) As a result, more extreme observations (positive and negative) are likely to occur under the t-distribution than under the standard normal distribution.

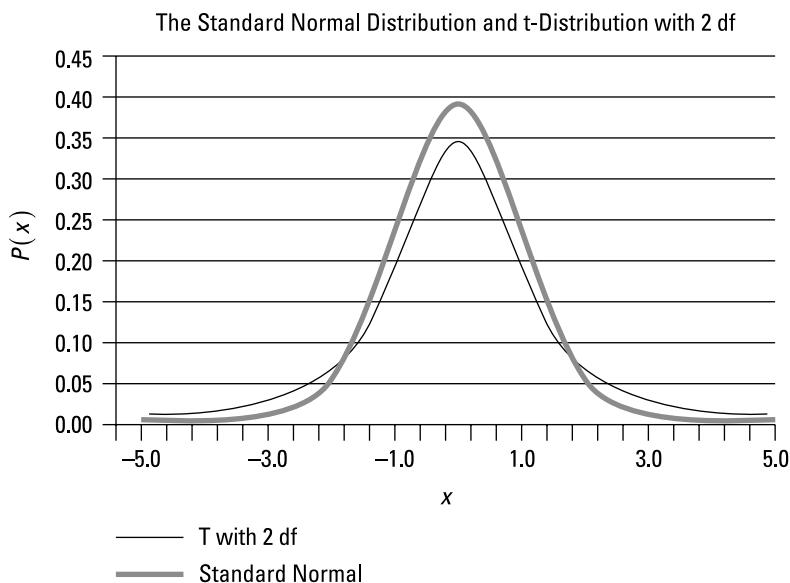
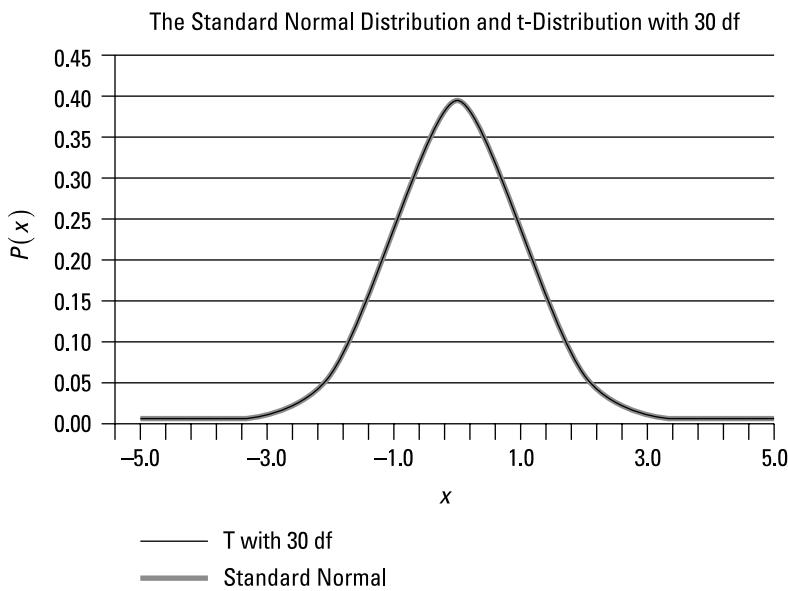
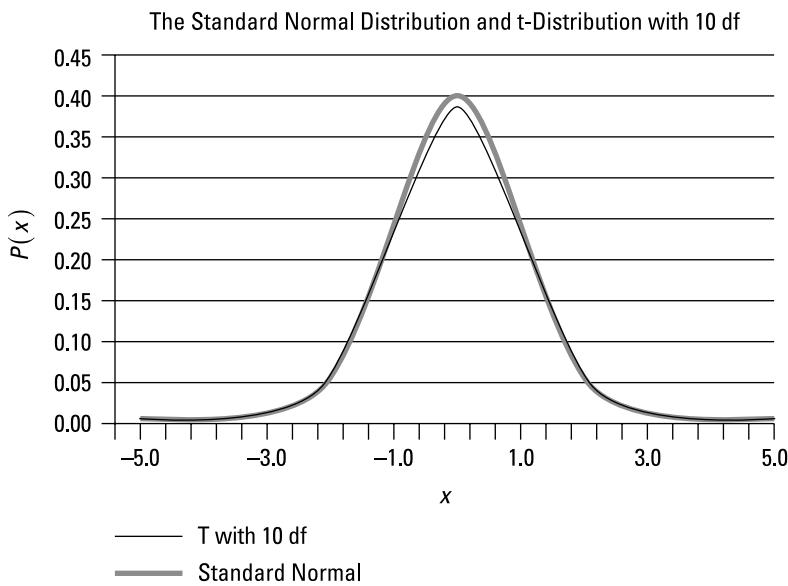


FIGURE 11-1:
The standard
normal and
t-distribution with
2 degrees of
freedom.

Figure 11-2 compares the standard normal distribution with the t-distribution with 10 degrees of freedom. The two are much closer to each other here than in Figure 11-1.

As you can see in Figure 11-3, with 30 degrees of freedom, the t-distribution and the standard normal distribution are almost indistinguishable.



Probabilities and the t-Table

The t-table is used to show probabilities for the t-distribution. The top row of the t-table lists different values of t_α , where the right tail of the t-distribution has a probability (area) equal to α ("alpha"). Table 11-1 is an excerpt from the full t-table.

TABLE 11-1

The t-Table

Degrees of Freedom (df)	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169

Table 11-1 shows that with 10 degrees of freedom and with $\alpha = 0.05$, $t_\alpha = 1.812$. So the right 5 percent tail of the distribution is located 1.812 standard deviations above the mean.

Alternatively, assume that X is a *random variable* that follows the t-distribution with 10 degrees of freedom. (Random variables are discussed in Chapter 7.) In this case, $P(X \geq 1.812) = 0.05$. This is equivalent to saying the area under the curve to the right of 1.812 is 0.05, or 5 percent of the total area (see Figure 11-4).

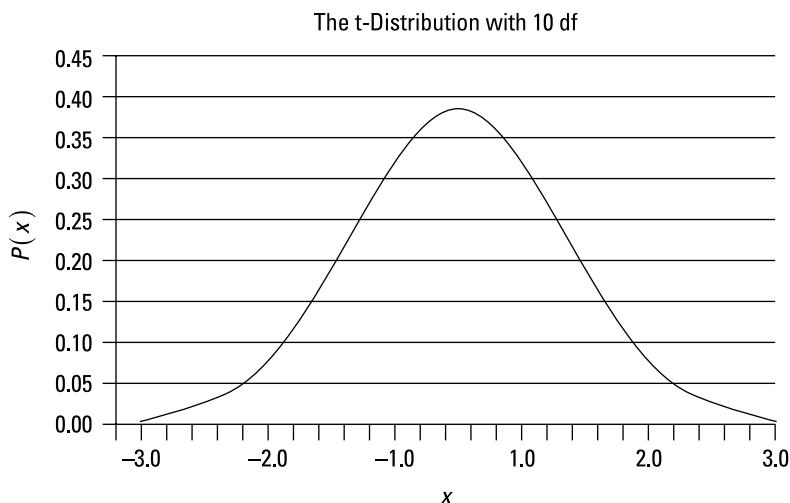


FIGURE 11-4:
The t-distribution
with 10 degrees
of freedom.

The shaded region starts at 1.812, which represents 1.812 standard deviations above the mean. The total area of the shaded region is 0.05 or 5 percent; therefore, the probability that a t-distributed random variable with 10 degrees of freedom exceeds 1.812 is 5 percent.

Point Estimates vs. Interval Estimates

When you don't know the mean, standard deviation, variance, and other summary measures of a population, you need to estimate them from a sample.

To estimate the mean of a population, you use the mean of a sample drawn from the population. You express the sample mean as \bar{X} ("X bar"). In a similar manner, to estimate the variance of a population, you use the sample variance, s^2 . And you estimate the standard deviation of a population with the sample standard deviation, s . (I cover techniques for estimating the sample variance and standard deviation in Chapter 4.)

These sample measures are formally known as *point estimators* — formulas that help estimate a population measure. For example, \bar{X} is a point estimator of the population mean μ . The numerical value of \bar{X} is a *point estimate*.



TIP

The distinction between *estimator* and *estimate* seems very subtle — an estimator is a formula, and an estimate is a numerical value.

The usefulness of a point estimator (formula) is limited by the fact that it produces only a single number. Suppose that a portfolio manager wants to estimate the mean annual return of a particular stock by choosing a sample of historical returns and calculating the sample mean. Suppose the sample mean turns out to be 8 percent. This information is useful, but it's difficult to judge how much the stock's returns may fluctuate from one year to the next based on this result.

Instead, suppose that the portfolio manager can estimate, with 95 percent certainty, that the return on the stock is between 6 and 10 percent, showing the stock's returns are relatively stable over time — the stock isn't extremely risky. The estimated range from 6 to 10 percent is an *interval estimate*.

In general, you compute an interval estimate with this formula:

$$\text{point estimate} \pm \text{margin of error}$$

This can be written as:

(point estimate – margin of error, point estimate + margin of error)



TIP

The symbol \pm indicates that two values exist: point estimate – margin of error, and point estimate + margin of error.

The margin of error depends on several factors, such as the type of point estimate being used, the size of the sample being used to construct the point estimate, and so forth. The margin of error is a measure of the degree of uncertainty associated with the point estimate.

Calculate an interval estimate of the population mean (μ) with this formula:

$$\bar{X} \pm \text{margin of error}$$

The margin of error is a measure of how much uncertainty is associated with the value of \bar{X} . Its value is closely related to the standard deviation of the underlying population and the size of the sample used to estimate \bar{X} .

Estimating Confidence Intervals for the Population Mean

A confidence interval is a specific type of interval estimate characterized by:

- » A confidence coefficient, expressed as $(1 - \alpha)$

α is known as the *level of significance*. For example, if you choose the level of significance to be 0.05, then the corresponding confidence coefficient equals $(1 - \alpha) = (1 - 0.05) = 0.95$.

- » A confidence level, expressed as $100(1 - \alpha)$

For example, if the confidence coefficient equals 0.95, then the corresponding confidence level equals $100(0.95) = 0.95 = 95$ percent.

Suppose that a 95 percent confidence interval is constructed for the population mean age in the United States based on the ages of people randomly chosen throughout the country. If this process is repeated 100 times (for example, 100 samples are drawn and a new confidence interval is estimated in each case), then you would expect that the true population mean age is contained in 95 of these 100 confidence intervals.

Two possible situations may arise when constructing a confidence interval for the population mean: A known population standard deviation and an unknown population standard deviation that you must estimate with the sample standard deviation(s). I discuss these situations in the following sections.

Known population standard deviation

If you know the population standard deviation, then the confidence interval is based on the *standard normal* distribution (which I discuss in detail in Chapter 9). Here's the formula for constructing this confidence interval:

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where:

\bar{X} = the sample mean

σ = the population standard deviation

n = the sample size

α = the level of significance

$Z_{\alpha/2}$ = a *quantile* or *critical value*, which represents the location of the right tail of the standard normal distribution with an area of $\alpha/2$

$Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ = the margin of error

The confidence interval can also be written as:

$$\left(\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

The two values contained in this interval are known as:

» The *lower limit* of the confidence interval: $\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

» The *upper limit* of the confidence interval: $\bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

For example, suppose that you want to construct a 95 percent confidence interval. This implies that $\alpha = 0.05$ (or 5 percent) so that $\alpha/2 = 0.025$ or 2.5 percent. You can find the value of $Z_{\alpha/2}$ from a standard normal probability table, such as shown in Table 11-2. The standard normal table shows probabilities less than or equal to a specific value. Because the area above $Z_{\alpha/2} = 0.025$, the area below $Z_{\alpha/2} = 1 - 0.025 = 0.975$ (because of the *symmetry* of the standard normal distribution).

By searching in the body of the standard normal table for the area 0.9750, you get the appropriate value of $Z_{\alpha/2}$ of 1.96. See Table 11-2 for this result.

TABLE 11-2

The Standard Normal Table

Z	0.05	0.06	0.07
1.7	0.9599	0.9608	0.9616
1.8	0.9678	0.9686	0.9693
1.9	0.9744	0.9750	0.9756
2.0	0.9798	0.9803	0.9808

Table 11-2 shows that the appropriate value of $Z_{\alpha/2}$ is 1.96. (You find the value 1.9 in the first [Z] column and the value 0.06 in the third [0.06] column.)



TIP

The most commonly used confidence intervals are a 90 percent confidence level, a 95 percent confidence level, and a 99 percent confidence level. In these three cases, the values of $Z_{\alpha/2}$ are listed in Table 11-3.

TABLE 11-3

Critical Z-Values

Confidence Level	$Z_{\alpha/2}$
90%	1.645
95%	1.960
99%	2.576

These results indicate that for a standard normal random variable Z , the following expressions are true:

$$P(Z \leq 1.645) = 0.9500$$

$$P(Z \geq 1.645) = 0.0500$$

$$P(Z \leq 1.960) = 0.9750$$

$$P(Z \geq 1.960) = 0.0250$$

$$P(Z \leq 2.576) = 0.9950$$

$$P(Z \geq 2.576) = 0.0050$$

The resulting confidence interval may then be expressed as follows:

$$P\left(\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

This expression shows that the population mean is contained within this interval with a level of confidence equal to $100(1 - \alpha)$.

For example, suppose that a hedge fund holds a portfolio consisting of 500 stocks. The standard deviation is 20 percent. If you choose a sample of 10 stocks and determine the sample mean to be 8 percent, you construct a 90 percent confidence interval by following these steps:

1. Calculate $\alpha/2$.

$$100(1 - \alpha) = 90 \text{ percent}$$

$$\alpha = 0.10$$

$$\alpha/2 = 0.05$$

2. Use the critical Z-values table (Table 11-3) to find the critical value: $Z_{\alpha/2} = 1.645$.

3. Compute the confidence interval.

$$\text{sample size: } n = 10$$

$$\text{population standard deviation: } \sigma = 0.20$$

$$\text{sample mean: } \bar{X} = 0.08$$

Therefore, the appropriate confidence interval is

$$\begin{aligned}\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} &= 0.08 \pm 1.645 \frac{0.20}{\sqrt{10}} \\ &= 0.08 \pm 0.104 \\ &= (-0.024, 0.184) \\ &= (-2.4\%, 18.4\%)\\\end{aligned}$$

For a 95 percent confidence interval, the only change you need to make is to the critical value, which you determine as follows:

$$100(1 - \alpha) = 95 \text{ percent}$$

$$\alpha = 0.05$$

$$\alpha/2 = 0.025$$

$$Z_{\alpha/2} = Z_{0.025} = 1.96$$

The 95 percent confidence interval is

$$\begin{aligned}\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} &= 0.08 \pm 1.96 \frac{0.20}{\sqrt{10}} \\ &= 0.08 \pm 0.124 \\ &= (-0.044, 0.204) \\ &= (-4.4\%, 20.4\%)\\\end{aligned}$$

Finally, you determine a 99 percent confidence interval with these adjustments:

$$\begin{aligned}100(1-\alpha) &= 99 \text{ percent} \\ \alpha &= 0.01 \\ \alpha/2 &= 0.005 \\ Z_{\alpha/2} &= Z_{0.025} = 2.576\end{aligned}$$

The 99 percent confidence interval is

$$\begin{aligned}\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} &= 0.08 \pm 2.576 \frac{0.20}{\sqrt{10}} \\ &= 0.08 \pm 0.163 \\ &= (-0.083, 0.243) \\ &= (-8.3\%, 24.3\%)\\\end{aligned}$$



TIP

As the level of confidence increases so does the *width* of the confidence interval because the only way to have more confidence that the interval actually contains the population mean is to include more values.

Unknown population standard deviation

If the population standard deviation is *not* known, then you compute an interval estimate for the population mean as follows:

$$\bar{X} \pm t_{\alpha/2}^{n-1} \frac{s}{\sqrt{n}}$$

where:

$t_{\alpha/2}^{n-1}$ = a quantile (critical value) that represents the location of the right tail of the t-distribution with $n - 1$ degrees of freedom with an area of $\alpha/2$

s = the sample standard deviation

In this case, you make the following changes to the formula:

- » You use the *sample* standard deviation (s) rather than the population standard deviation.
- » You use the t-distribution rather than the standard normal distribution because of the greater uncertainty associated with the sample standard deviation. $t_{\alpha/2}^{n-1}$ is a quantile or critical value taken from the t-distribution and represents the location of the right tail of the t-distribution with $n - 1$ degrees of freedom whose area equals $\alpha/2$.

As an example, suppose that $\alpha = 0.05$ so that $\alpha/2 = 0.025$. Also assume that the appropriate number of degrees of freedom is 9. You can get the value of $t_{\alpha/2}^{n-1}$ from a t-table, as in Table 11-1.

The appropriate column heading is $t_{0.025}$; with 9 degrees of freedom, the value of $t_{\alpha/2}^{n-1}$ is 2.262.

For example, a university has 10,000 students and wants to estimate the average GPA of the entire student body. It picks a sample of ten students, and the sample mean GPA is 3.10. The sample standard deviation is 0.25. You construct confidence intervals for the population mean as follows:

- » For a 90 percent confidence interval, the value of $\alpha/2$ is 0.05:

$$\begin{aligned}100(1-\alpha) &= 90 \text{ percent} \\ \alpha &= 0.10 \\ \alpha/2 &= 0.05\end{aligned}$$

With $n - 1 = 9$ degrees of freedom, based on the t-table (Table 11-1), $t_{\alpha/2}^{n-1} = 1.833$.

The sample size is $n = 10$, the population standard deviation is $s = 0.25$, and the sample mean is $\bar{X} = 3.10$. Therefore, the appropriate confidence interval is

$$\begin{aligned}\bar{X} \pm t_{\alpha/2}^{n-1} \frac{s}{\sqrt{n}} &= 3.10 \pm 1.833 \frac{0.25}{\sqrt{10}} \\ &= 3.10 \pm 0.1449 \\ &= (2.9551, 3.2449)\end{aligned}$$

- » For a 95 percent confidence interval, you follow similar calculations but change the critical value:

$$\begin{aligned}100(1-\alpha) &= 95 \text{ percent} \\ \alpha &= 0.05 \\ \alpha/2 &= 0.025\end{aligned}$$

With $n - 1 = 9$ degrees of freedom, based on the t-table (Table 11-1), $t_{\alpha/2}^{n-1} = 2.262$. Therefore, the appropriate confidence interval is

$$\begin{aligned}\bar{X} \pm t_{\alpha/2}^{n-1} \frac{s}{\sqrt{n}} &= 3.10 + 2.262 \frac{0.25}{\sqrt{10}} \\ &= 3.10 \pm 0.1788 \\ &= (2.9212, 3.2788)\end{aligned}$$

- » For a 99 percent confidence interval, you again change the critical value to

$$100(1 - \alpha) = 99 \text{ percent}$$

$$\alpha = 0.01$$

$$\alpha/2 = 0.005$$

With $n - 1 = 9$ degrees of freedom, based on the t-table (Table 11-1), $t_{\alpha/2}^{n-1} = 3.250$. Therefore, the appropriate confidence interval is

$$\begin{aligned}\bar{X} \pm t_{\alpha/2}^{n-1} \frac{s}{\sqrt{n}} &= 3.10 + 3.250 \frac{0.25}{\sqrt{10}} \\ &= 3.10 \pm 0.2569 \\ &= (2.8431, 3.3569)\end{aligned}$$

In each case, the confidence interval is wider than it would be when using the standard normal distribution.

Computing Confidence Intervals for the Population Mean with the TI-84 Plus Calculator

You can use the Texas Instruments TI-84 Plus and Plus CE calculators to compute confidence intervals for the population mean. The two basic cases for computing confidence intervals about the population mean are:

- » **The population standard deviation (σ) is known.** In this case, the confidence interval is based on the standard normal (Z) distribution.
- » **The population standard deviation (σ) is unknown.** In this case, the confidence interval is based on the t-distribution.

Population standard deviation is known

To construct a confidence interval for the population mean when the population standard deviation is known (so that the standard normal distribution will be used), follow these steps:

- 1. Press the [STAT] button.**
- 2. Use the arrow key to choose TESTS.**
- 3. Choose 7: ZInterval.**

This refers to a confidence interval based on the "Z" or standard normal distribution.

The first row (Inpt) offers the option of choosing the Data menu or the Stats menu (which you can do by using the arrow key to select the appropriate option and then pressing [ENTER]). Use the Data menu when the original data is entered into a list. (See Chapter 5 for instructions on entering data into a list.) Use the Stats menu when only summary statistics, such as the sample mean and sample size, are available.

When using the Data menu, the following information must be provided:

σ : The population standard deviation
List: The list containing the data for this problem
Freq: Set equal to one unless the data in the list will be repeated more than once
C-Level: The appropriate confidence level (for example, use 0.95 for a 95 percent confidence level)
Calculate: Used to compute the final results

When using the Stats menu, the following information must be provided:

σ : The population standard deviation
 X : The sample mean
 n : The sample size
C-Level: The appropriate confidence level (for example, use 0.95 for a 95 percent confidence level)
Calculate: Used to compute the final results

Population standard deviation is unknown

To construct a confidence interval for the population mean when the population standard deviation is unknown (so that the t-distribution will be used), follow these steps:

- 1. Choose the STAT menu.**
- 2. Use the arrow key to choose TESTS.**
- 3. Choose 8: TInterval.**

This refers to a confidence interval based on the t-distribution.

The first row (Inpt) offers the option of choosing the Data menu or the Stats menu.

When using the Data menu, the following information must be provided:

List: The list containing the data for this problem

Freq: Set equal to one unless the data in the list will be repeated more than once

C-Level: The appropriate confidence level (for example, use 0.95 for a 95 percent confidence level)

Calculate: Used to compute the final results

Note that the population standard deviation is not used because it is assumed that this value is unknown.

When using the Stats menu, the following information must be provided:

X: The sample mean

S_x: The sample standard deviation

n: The sample size

C-Level: The appropriate confidence level (for example, use 0.95 for a 95 percent confidence level)

Calculate: Used to compute the final results

IN THIS CHAPTER

- » Understanding the hypothesis testing process
- » Testing hypotheses about two population means

Chapter **12**

Testing Hypotheses about the Population Mean

Hypothesis testing is a multi-step statistical process which is used to test claims about a population measure, such as the mean. For example, you can use hypothesis testing on the following statements to determine whether they're likely to be true:

- » Mean income in the United States has risen over the past 25 years.
- » The average age of the population of Egypt is above 30.
- » The average return to the stocks in a portfolio is 10 percent.
- » The United States and Canada have average work weeks identical in length.
- » The average lifetime of brandy drinkers is 90.

You test hypotheses with a series of steps designed to show whether you can justify a claim. These steps apply to a lot of situations; for example, you can test claims about a population's mean, a population's variance, whether a population is normally distributed, and so forth. This chapter focuses on testing hypotheses about the mean value of a single population and the equality of the means of two different populations.

Applying the Key Steps in Hypothesis Testing for a Single Population Mean

Hypothesis testing requires sample data to draw conclusions about the characteristics of the underlying population. The necessary steps for any type of hypothesis test are outlined in the following sections.

Writing the null hypothesis

The *null hypothesis* is a statement that's assumed to be true unless strong contrary evidence exists. The null hypothesis can take several forms. You can use it to test statements about population measures, such as means and standard deviations, and to test statements about the relationship between two populations. An example of a null hypothesis is the mean age of Star Trek fans is higher than the mean age of Star Wars fans.

You write the null hypothesis for testing the value of a single population mean as

$$H_0: \mu = \mu_0$$

where H_0 stands for the null hypothesis, μ is the true population mean (whose value we do not know,) and μ_0 is the hypothesized value of the population, or the value that you *think* is true. For example, if you want to test the hypothesis that the mean number of runs scored per game in the American League is 4 you write the null hypothesis as $H_0: \mu = 4.0$.

If actual data shows that this is likely to be false, you *reject* the null hypothesis; otherwise, you *don't* reject the null hypothesis. (You never *accept* the null hypothesis; instead, you fail to reject it if there is not enough evidence against it.)

Coming up with an alternative hypothesis

Suppose that the null hypothesis is false. For example, you are testing the null hypothesis that the mean number of runs scored per game in the American League is 4. If data taken from actual games shows that this is likely to be false, it must be true that:

The number of runs scored is *more than* 4.

The number of runs scored is *less than* 4.

Prior to testing the null hypothesis, you must specify what alternative you accept if the null hypothesis is rejected. It turns out that there are actually three ways to express the alternative hypothesis:

The number of runs scored is *more than* 4.

The number of runs scored is *less than* 4.

The number of runs scored is *different from* 4.

The alternative that you choose depends on what type of action is taken as a result of the hypothesis test. For example, suppose that the commissioner decides that if the number of runs scored is *less than* 4, the league will encourage teams to shorten the distance to their outfield fences (which encourages more home runs.) You therefore use “the number of runs scored is *less than* 4” as your alternative hypothesis. This ensures that no action is taken unless it’s extremely clear that the number of runs is less than 4.

There are special names associated with the three types of alternative hypotheses:

- » Right-tailed test
- » Left-tailed test
- » Two-tailed test

A right-tailed test indicates that the actual population mean is *greater than* the hypothesized mean. A left-tailed test indicates that the actual population mean is *less than* the hypothesized population mean. A two-tailed test is a combination of the right-tailed and left-tailed tests; it indicates that the actual population mean is *different from* the hypothesized mean. (This combines the two alternative hypotheses that the actual population mean is *greater than* the hypothesized mean and the actual population mean is *less than* the hypothesized mean.)

Right-tailed test

A *right-tailed test* is a test to determine if the actual value of the population mean is *greater than* the hypothesized value.

Suppose that you’re testing a hypothesis about the mean of a population, and you’re interested in only strong evidence that the mean is *greater than* a specified value. In this case, you set up a right-tailed test. (“Right tail” refers to the largest values in a probability distribution.)

As an example of a right-tailed test, suppose that a department store wants to know whether the mean length of time its merchandise remains in inventory is 30 days. If the mean time is greater than 30 days, the store will overhaul its ordering procedures; if the mean is equal to or less than 30 days, the store will do nothing.

In this case, it's extremely important for the store to know whether the mean exceeds 30 days because a key decision depends on this information. The store doesn't want to spend time overhauling its procedures unless strong evidence shows that it's necessary; therefore, the most appropriate choice is a right-tailed test that shows the mean is greater than 30 days.

In general, you write the alternative hypothesis with a right-tailed test as

$$H_1: \mu > \mu_0$$

Here, H_1 represents the alternative hypothesis. In this example, you'd write the alternative hypothesis as $H_1: \mu > 30$.

Left-tailed test

A *left-tailed test* is a test to determine if the actual value of the population mean is *less than* the hypothesized value. (“Left tail” refers to the smallest values in a probability distribution.)

Suppose that you're testing a hypothesis about the mean of a population, and you're interested only in strong evidence that the mean is *less than* a specified value. In this case, you set up a left-tailed test.

For example, a pension fund wants to know whether any of its portfolio managers are earning an average return that falls short of the return to the Standard & Poor's 500 (S&P 500) stock index. (Assume this return is currently 8 percent.) If so, these managers won't receive the company's annual Christmas bonus.

In this situation, the fund is interested in knowing only which managers don't qualify for the Christmas bonus. As a result, the most appropriate choice for the alternative hypothesis is a left-tailed test that shows the mean return is less than 8 percent.

In general, you write the alternative hypothesis for a left-tailed test as:

$$H_1: \mu < \mu_0$$

In this example, you'd write the alternative hypothesis as $H_1: \mu < 0.08$ (0.08 is the decimal equivalent of 8 percent).

Two-tailed test

Building on the right-tailed test and the left-tailed test, consider the two-tailed test, which is used to determine if the actual value of the population mean is different from the hypothesized value; for example, greater than or less than. (A two-tailed test uses both the right tail and left tail of a probability distribution.)

Suppose that you're testing a hypothesis about the mean of a population, and you need to know whether the mean is different from a specified value. For example, a bottling company wants to be sure that the mean volume of its 1-liter bottles is actually 1 liter. Any value less than or more than this measurement can lead to significant problems. So the most appropriate choice is a two-tailed test that shows the mean volume is *not equal to* 1.

In general, you express the null and alternative hypotheses for a two-tailed test as

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

In this example, you'd write the alternative hypothesis as $H_1: \mu \neq 1$. This expression indicates that if the null hypothesis is false then either $H_1: \mu > 1$ or $H_1: \mu < 1$ will be accepted in its place.

In this case, a two-tailed test was conducted due to the extreme importance of determining immediately if the mean content of the bottles is *either* less than 1 or greater than 1. If overfilled bottles are a problem, but not underfilled bottles, you would use a right-tailed test. If underfilled bottles are a problem, but not over-filled bottles, you would use a left-tailed test.

Choosing a level of significance

To test a hypothesis, you must specify a *level of significance* — the probability of rejecting the null hypothesis when it's actually true.



TECHNICAL
STUFF

Rejecting the null hypothesis when it is actually true is known as a *Type I error*. By contrast, a *Type II error* occurs when you fail to reject the null hypothesis when it's actually false. The level of significance of a hypothesis test equals the probability of committing a Type I error. A Type I error is sometimes known as a “false positive”; a Type II error is sometimes known as a “false negative.”

In the process of testing a hypothesis, the following four results can take place. The two possible correct decisions are:

- » Rejecting the null hypothesis when it's false
- » Failing to reject the null hypothesis when it's true

The two possible incorrect decisions are:

- » Rejecting the null hypothesis when it's true
- » Failing to reject the null hypothesis when it's false

The probability of committing a Type I error is often designated with the letter α ("alpha"), and the probability of committing a Type II error is often designated with the letter β ("beta").

The larger is the probability of a Type I error that you choose for a hypothesis test, the smaller will be the probability of a Type II error, and vice versa. (One way to reduce both is to increase the sample size used for the hypothesis test.)



TECHNICAL STUFF

The probabilities of Type I and Type II errors do *not* add up to 1; they are *not* complementary events. (Complementary events are discussed in Chapter 6.)

When you're conducting a hypothesis test, you choose the value of α to find the right balance between avoiding Type I and Type II errors. In some types of applications, avoiding Type I errors is critically important; in other cases, Type I errors may not be as serious. In many hypothesis tests of a population value (such as the mean), the level of significance is often chosen to be 0.01, 0.05, or 0.10, with 0.05 being most common.

Although both Type I and Type II errors represent serious mistakes, in some situations, one mistake is far more important to avoid than the other. For example, in a jury trial, the null hypothesis is "the defendant is not guilty," which is assumed to be true unless strong contrary evidence suggests otherwise. The alternative hypothesis is that "the defendant is guilty."

In this situation, four outcomes can occur:

- » The jury reaches a correct decision by acquitting a defendant who is not guilty.
- » The jury reaches a correct decision by convicting a guilty defendant.
- » The jury commits a Type I error by wrongly convicting a defendant who is not guilty. (In this situation, the null hypothesis of being not guilty has been incorrectly rejected.)

- » The jury commits a Type II error by acquitting a guilty defendant (because the null hypothesis of being not guilty hasn't been rejected when it's actually false).

For a jury trial, avoiding a Type I error is far more important than avoiding a Type II error; as such, you set α equal to a very small value, which would imply a much larger value for β . (α would never be set equal to 0 because that would ensure that no one is ever convicted!)

Because a Type I error in this case indicates that a person who is not guilty has been convicted, and a Type II error indicates that a guilty person walks free, it's clearly imperative to avoid Type I errors even if it means more Type II errors.



TIP

Sir William Blackstone (1723–1780), the famous English judge and politician, once wrote that “It is better that ten guilty persons escape than that one innocent suffer.” A statistician might rephrase this in slightly less elegant terms: “It is extremely important to avoid Type I errors in a jury trial.”

Computing the test statistic

A *test statistic* is a numerical measure you construct to determine whether you should reject the null hypothesis. It shows how far the sample mean is from the hypothesized value of the population mean in terms of standard deviations. You calculate this value from a sample drawn from the underlying population.

For example, say you're testing a hypothesis about the mean age of the residents in a city. The city government wants to know whether the mean age is 40. You choose a sample of city residents, and you compute the mean age of the sample members. If the sample mean age is substantially different from 40, the null hypothesis will likely be rejected.

If you conduct a hypothesis test of the value of a single population mean, the form of the test statistic depends on whether the population standard deviation is known.



REMEMBER

When the population standard deviation is known, the test statistic is based on the Standard Normal (Z) distribution; otherwise, it is based on the t-distribution. (See Chapters 9 and 11 for discussions on the normal distribution and the Student's t-distribution.)

When the population standard deviation is unknown, the test statistic is

$$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

In this formula,

- » t indicates that this test statistic follows the Student's t-distribution.
- » \bar{X} is the sample mean.
- » μ_0 is the hypothesized population mean.
- » s is the sample standard deviation.
- » n is the sample size.
- » $\frac{s}{\sqrt{n}}$ is the *standard error* of the sample mean.

If the population standard deviation is known, the test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

The letter Z indicates that this test statistic follows the standard normal distribution. The standard normal distribution (see Chapter 9) is the special case of the normal distribution with mean (μ) of 0 and a standard deviation (σ) of 1.

Comparing the critical value(s)

After you calculate a test statistic, you compare it to one or two *critical values*, depending on the alternative hypothesis, to determine whether you should reject the null hypothesis. A critical value shows the number of standard deviations away from the mean of a distribution where a specified percentage of the distribution is above the critical value and the remainder of the distribution is below the critical value.

For example, based on the standard normal table (see Chapter 9), the probability that a standard normal random variable Z is less than 1.645 equals 0.95 or 95 percent. (On the standard normal table, a Z-score of 1.645 is found half-way between the Z-scores 1.64 and 1.65.) As a result, the probability that Z is greater than 1.645 is 0.05 or 5 percent. 1.645 is the critical value that divides the lower 95 percent of the distribution from the upper 5 percent of the distribution. Due to the symmetry of the standard normal distribution, -1.645 is the critical value that divides the lower 5 percent of the distribution from the upper 95 percent of the distribution.

This is shown in Figure 12-1. The shaded region is the upper 5 percent of the standard normal distribution, and the unshaded region is the lower 95 percent of the distribution.

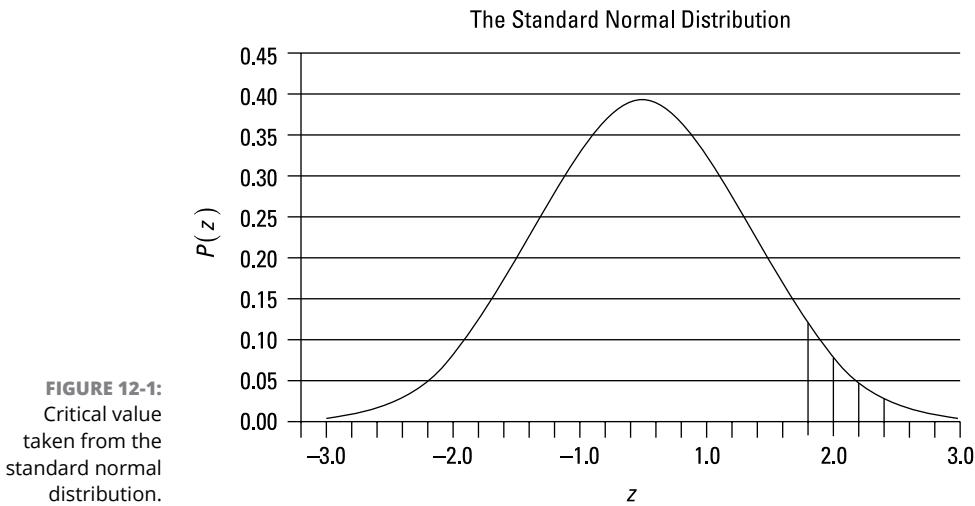


FIGURE 12-1:
Critical value
taken from the
standard normal
distribution.

The appropriate critical value depends on whether you are conducting a right-tailed test, a left-tailed test, or a two-tailed test, as follows:

- » A right-tailed test has one positive critical value.
- » A left-tailed test has one negative critical value.
- » A two-tailed test has two critical values, one positive and one negative.

Population standard deviation is unknown

When the population standard deviation is unknown, you take the resulting critical value or values from the Student's t-distribution, as follows:

- » Right-tailed test: critical value = t_{α}^{n-1}
- » Left-tailed test: critical value = $-t_{\alpha}^{n-1}$
- » Two-tailed test: critical value = $\pm t_{\alpha/2}^{n-1}$



REMEMBER

α is the level of significance, and n represents the sample size. You draw these critical values from the Student's t-distribution with $n - 1$ degrees of freedom (df). (See Chapter 11 for more on the Student's t-distribution as well as a detailed discussion of how degrees of freedom are determined.)

The number of degrees of freedom used with the t-distribution depends on the particular application. For testing hypotheses about the population mean, the appropriate number of degrees of freedom is one less than the sample size (that is, $n - 1$).

The critical value or values are used to locate the areas under the curve of a distribution that are too extreme to be consistent with the null hypothesis. For a right-tailed test, these are the large positive values, which are collectively known as the right tail of the distribution. For a left-tailed test, these are the large negative values, which are collectively known as the left tail of the distribution. In either case, the area in the tail equals the level of significance of the hypothesis test. For a two-tailed test, the value of the level of significance (α) is split in half; the area in the right tail equals $\alpha/2$, and the area in left tail equals $\alpha/2$, for a total of α .

RIGHT-TAILED TEST WITH THE T-DISTRIBUTION

As an example of a right-tailed test, suppose that the level of significance is 0.05 and the sample size is 10; then you get a single positive critical value:

$$t_{\alpha}^{n-1} = t_{0.05}^9$$

Refer to Table 12-1 to find the intersection of the row representing 9 degrees of freedom and the column headed $t_{0.05}$.

TABLE 12-1 The Student's t-distribution

Degrees of Freedom (df)	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947

The critical value is 1.833, or $t_{\alpha}^{n-1} = t_{0.05}^9 = 1.833$, as shown in Figure 12-2. The shaded region in the right tail represents the *rejection region*; if the test statistic falls in this area, the null hypothesis will be rejected.

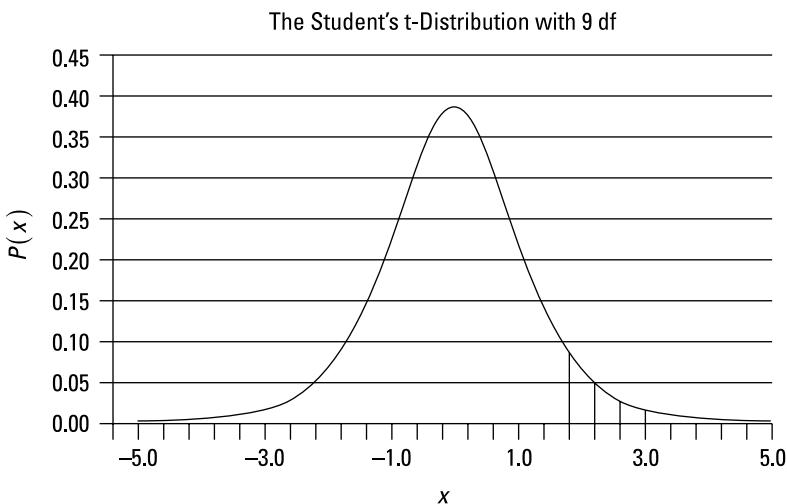


FIGURE 12-2:
Critical value
taken from the
t-distribution:
right-tailed test.

LEFT-TAILED TEST WITH THE T-DISTRIBUTION

As an example of a left-tailed test, suppose that the level of significance is 0.05 and the sample size is 10; then you get a single negative critical value:

$$-t_{\alpha}^{n-1} = -t_{0.05}^9$$

You get this number from the t-table (Table 12-1) at the intersection of the row representing 9 degrees of freedom and the $t_{0.05}$ column; the critical value is -1.833 , or $-t_{\alpha}^{n-1} = -t_{0.05}^9 = -1.833$, as shown in Figure 12-3. (The value is negative because it is in the left tail of the distribution.) The shaded region in the left tail represents the *rejection region*; if the test statistic falls in this area, the null hypothesis will be rejected.

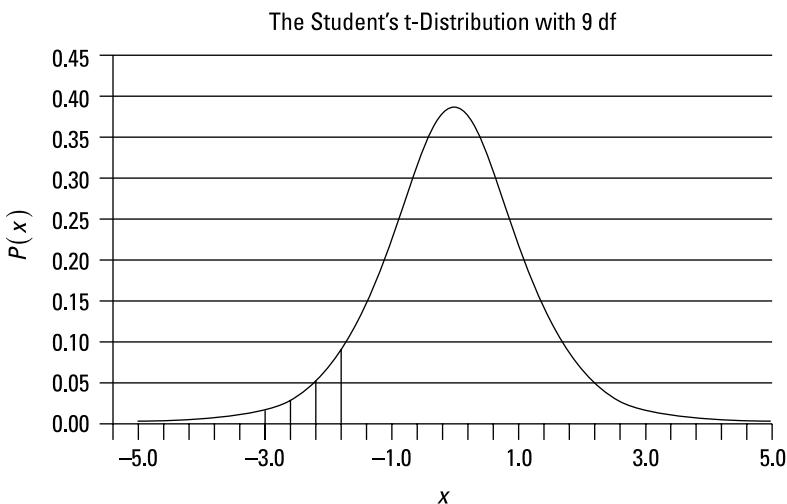


FIGURE 12-3:
Critical value
taken from the
t-distribution:
left-tailed test.

TWO-TAILED TEST WITH THE T-DISTRIBUTION

As an example of a two-tailed test, suppose that the level of significance is 0.05 and the sample size is 10; then you get a positive and a negative critical value:

$$\pm t_{\alpha/2}^{n-1} = \pm t_{0.025}^9$$

You can find the value of the positive critical value $t_{0.025}^9$ directly from Table 12-1.

In this case, you find the positive critical value $t_{0.025}^9$ at the intersection of the row representing 9 degrees in the *Degrees of Freedom (df)* column and the $t_{0.025}$ column. The positive critical value is 2.262; therefore, the negative critical value is -2.262. You represent these two values like so (as Figure 12-4 illustrates):

$$\pm t_{\alpha/2}^{n-1} = \pm 2.262$$

The shaded region in the two tails represents the *rejection region*; if the test statistic falls in either tail, the null hypothesis will be rejected.

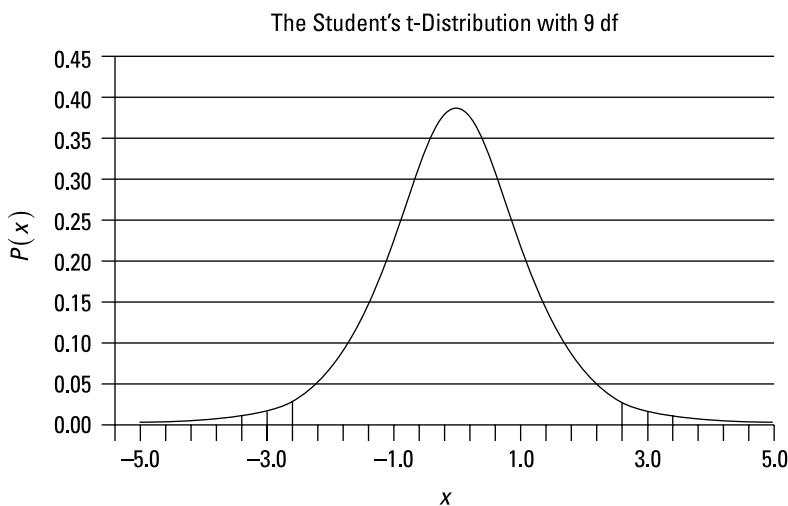


FIGURE 12-4:
Critical value
taken from the
t-distribution:
two-tailed test.

Population standard deviation is known

When the population standard deviation is known, you take the resulting critical value or values from the standard normal distribution, as follows:

- » Right-tailed test: critical value = Z_α
- » Left-tailed test: critical value = $-Z_\alpha$
- » Two-tailed test: critical values = $\pm Z_{\alpha/2}$

Because you draw these critical values from the standard normal distribution, you don't have to calculate degrees of freedom. Unlike the Student's t-distribution, the standard normal distribution isn't based on degrees of freedom. I walk you through how to find these critical values in the following sections.



TIP

TABLE 12-2

Common Critical Values of the Standard Normal Distribution

α	Right-Tailed Test	Left-Tailed Test	Two-Tailed Test
0.01	2.326	-2.326	± 2.576
0.05	1.645	-1.645	± 1.960
0.10	1.282	-1.282	± 1.645

RIGHT-TAILED TEST WITH THE Z-DISTRIBUTION

A *right-tailed* hypothesis test of the population mean with a level of significance of 0.05 has a single positive critical value: $Z_\alpha = Z_{0.05}$. You find the value by checking the body of Table 12-3 for a probability of $1 - \alpha$, which is 0.9500.

TABLE 12-3

Standard Normal Table — Positive Values

Z	0.04	0.05	0.06	0.07
1.5	0.9382	0.9394	0.9406	0.9418
1.6	0.9495	0.9505	0.9515	0.9525
1.7	0.9591	0.9599	0.9608	0.9616
1.8	0.9671	0.9678	0.9686	0.9693
1.9	0.9738	0.9744	0.9750	0.9756
2.0	0.9793	0.9798	0.9803	0.9808

Unfortunately, this exact value isn't in the table. The two closest values are 0.9495 and 0.9505, which you can find at the intersections of row 1.6 under the Z column and the 0.04 and 0.05 columns. The critical value is, therefore, halfway between 1.64 and 1.65; average it out to get 1.645, or $Z_\alpha = 1.645$, and see Figure 12-5 for a graphical depiction.

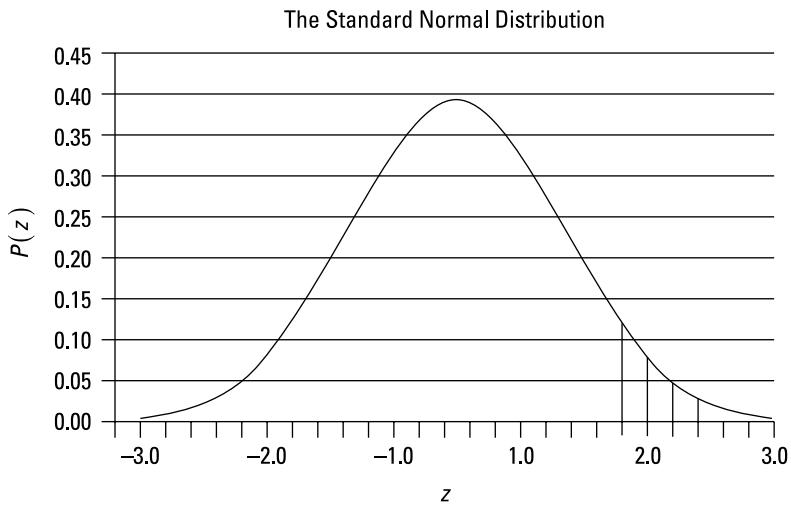


FIGURE 12-5:
Critical value taken from the standard normal distribution: right-tailed test.

LEFT-TAILED TEST WITH THE Z-DISTRIBUTION

A left-tailed hypothesis test with a level of significance of 0.05 has a single negative critical value $-Z_\alpha = -Z_{0.05}$, or simply -1.645 : $-Z_\alpha = -1.645$. Figure 12-6 represents this critical value graphically.

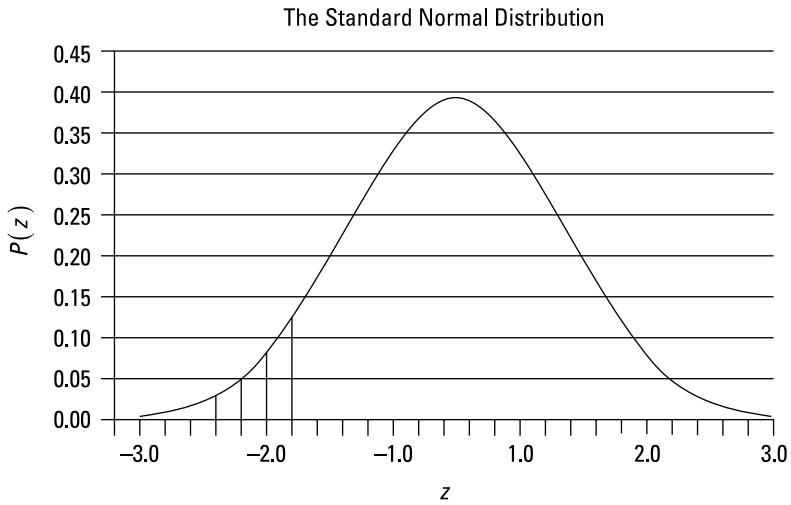


FIGURE 12-6:
Critical value taken from the standard normal distribution: left-tailed test.

TWO-TAILED TEST WITH THE Z-DISTRIBUTION

For a two-tailed hypothesis test of the population mean with a level of significance of 0.05, the two critical values are $\pm Z_{\alpha/2} = \pm Z_{0.025}$. You can find the positive critical value in a standard normal table, like Table 12-3.



TECHNICAL
STUFF

Finding critical values in a standard normal table is more complicated than finding critical values in a t-table. The body of the standard normal table contains probabilities, unlike in the t-table where the probabilities are contained in the column headings.

In this example, you find the positive critical value $Z_{\alpha/2} = Z_{0.025}$ by checking the body of the table for a probability of

$$(1 - \alpha / 2) = (1 - 0.05 / 2) = (1 - 0.025) = 0.9750$$

In other words, the positive critical value represents the number of standard deviations above the mean at which

- » 2.5 percent of the area under the standard normal curve is to the right of this point.
- » 97.5 percent of the area under the standard normal curve is to the left of this point.

Because the standard normal table shows areas to the left of specified values, you can find the positive critical value by locating the probability 0.9750, not 0.0250, in the body of the table (Table 12-3). You find this probability by following the row 1.9 under the Z column to the 0.06 column. Therefore, the critical value $Z_{\alpha/2} = Z_{0.025} = 1.96$. The corresponding negative critical value is -1.96 . You can write these critical values as $\pm Z_{\alpha/2} = \pm 1.96$. Figure 12-7 shows these values graphically.

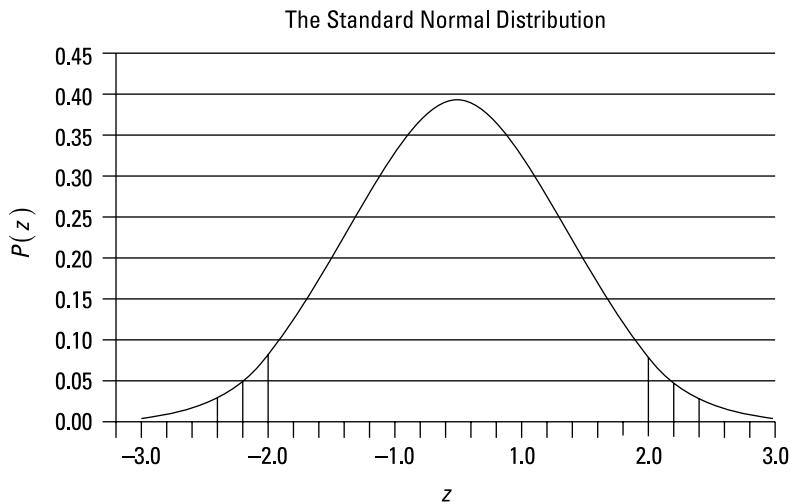


FIGURE 12-7:
Critical values taken from the standard normal distribution: two-tailed test.

Using the decision rule

You make the decision to reject the null hypothesis by looking at the relationship between the test statistic and the critical value(s), as follows:

- » **Right-tailed test:** If the test statistic is *greater than* the critical value, reject the null hypothesis $H_0: \mu = \mu_0$ in favor of the alternative hypothesis $H_1: \mu > \mu_0$; otherwise, don't reject the null hypothesis as there is insufficient evidence to show that the null hypothesis is false.
- » **Left-tailed test:** If the test statistic is *less than* the critical value, reject the null hypothesis $H_0: \mu = \mu_0$ in favor of the alternative hypothesis $H_1: \mu < \mu_0$; otherwise, don't reject the null hypothesis as there is insufficient evidence to show that the null hypothesis is false.
- » **Two-tailed test:** If the test statistic is *less than* the negative critical value or *greater than* the positive critical value reject the null hypothesis $H_0: \mu = \mu_0$ in favor of the alternative hypothesis $H_1: \mu \neq \mu_0$. Otherwise, don't reject the null hypothesis as there is insufficient evidence to show that the null hypothesis is false.

As an example, suppose that the government of a small country is interested in studying the characteristics of household incomes in the country. The government wants to know whether the mean household income is greater than \$25,000 per year. If so, the government will propose new types of taxes; otherwise, no new taxes will occur. The appropriate steps for testing the null hypothesis that the mean household income equals \$25,000 at the 5 percent level of significance are given as follows:

The null and alternative hypotheses are

$$H_0: \mu = 25,000$$

$$H_1: \mu > 25,000$$

In this example, the government uses a right-tailed test because it's looking for strong evidence that the mean household incomes are *greater than* \$25,000 per year. If true, the government will take an important action (for example, raise taxes).

Assume that the level of significance is 0.05. The government's chief statistician selects a sample of 100 households and computes the sample mean household income to be \$27,200 per year. The population standard deviation is known to be \$8,400. Because the population standard deviation is known, he uses the standard normal distribution to test this hypothesis. The appropriate test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

The value of the test statistic is, therefore,

$$\begin{aligned} Z &= \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \\ &= \frac{27,200 - 25,000}{8,400 / \sqrt{100}} \\ &= \frac{2,200}{840} \\ &= 2.62 \end{aligned}$$

The critical value is $Z\alpha = Z_{0.05} = 1.645$ (see Table 12-3).

Because the test statistic of 2.62 exceeds the critical value of 1.645, the government statistician rejects the null hypothesis in favor of the alternative hypothesis that the population mean exceeds \$25,000. As a result, there is strong evidence that mean household incomes exceed \$25,000 and the government will propose new taxes. Figure 12-8 shows this result graphically.

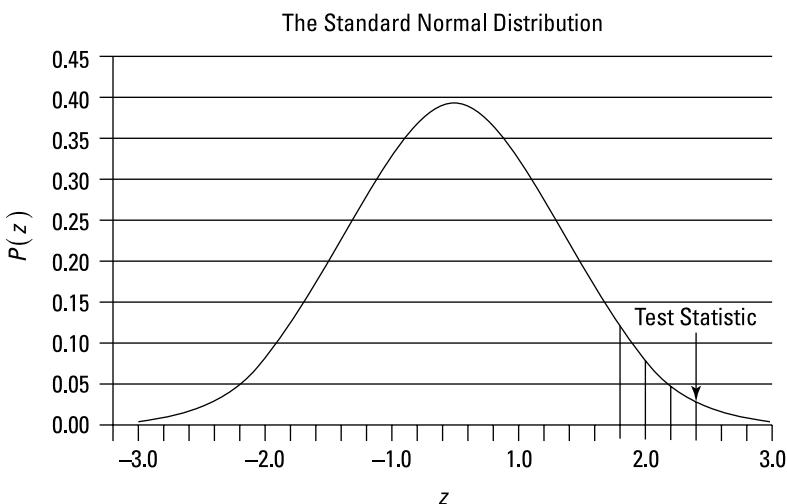


FIGURE 12-8:
Standard normal distribution: The null hypothesis is rejected.

As another example, suppose that the same government wants to study the average crop yields of its wheat farmers. The government wants to know whether the mean yield is equal to 10,000 bushels per year.

If the mean yield is below 10,000, the government will provide cash assistance to the farmers. If the mean yield is above 10,000, the government will export some of the surplus wheat to foreign countries. The government's chief statistician can test the null hypothesis that the mean crop yield equals 10,000 bushels. (Assume a 5 percent level of significance.)

The null and alternative hypotheses are

$$H_0: \mu = 10,000$$

$$H_1: \mu \neq 10,000$$

This example requires a two-tailed test, because the government is looking for strong evidence that mean crop yields are either *less than* or *greater than* 10,000 bushels per year. The government will undertake an important action in either case.

Assume that the level of significance is 0.05. The government statistician selects a sample of eight farms, and estimates the sample mean and standard deviation. The mean crop yield turns out to be 10,200 bushels. The sample standard deviation is 420 bushels.

Because the population standard deviation is unknown, the hypothesis test is based on the Student's t-distribution. The appropriate test statistic is, therefore,

$$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

The value of the test statistic is

$$\begin{aligned} t &= \frac{\bar{X} - \mu_0}{s / \sqrt{n}} \\ &= \frac{10,200 - 10,000}{420 / \sqrt{8}} \\ &= \frac{200}{148.49} \\ &= 1.35 \end{aligned}$$

The critical values are $\pm t_{\alpha/2}^{n-1}$.

With a sample size of $n = 8$, the appropriate number of degrees of freedom is $n - 1 = 7$. With a level of significance of 0.05, the value of $\alpha/2$ is 0.025. Therefore, you find the critical values in the t-table (Table 12-1) as follows:

$$\pm t_{\alpha/2}^{n-1} = \pm t_{0.025}^7 = \pm 2.365$$

With a two-tailed test, the decision rule is to

- » Reject the null hypothesis $H_0: \mu = \mu_0$ in favor of the alternative hypothesis $H_1: \mu \neq \mu_0$ if the test statistic is less than the negative critical value (-2.365) or greater than the positive critical value (2.365).
- » Fail to reject the null hypothesis if the test statistic is between the negative and positive critical values (-2.365 and 2.365).

Because the test statistic is 1.35, it's *greater than* the negative critical value of -2.365 , and *less than* the positive critical value of 2.365 . In other words, the test statistic is *not* in the rejection region, as shown in Figure 12-9. In this example, you do *not* reject the null hypothesis $H_0: \mu = \mu_0$. As a result, the government takes no action.

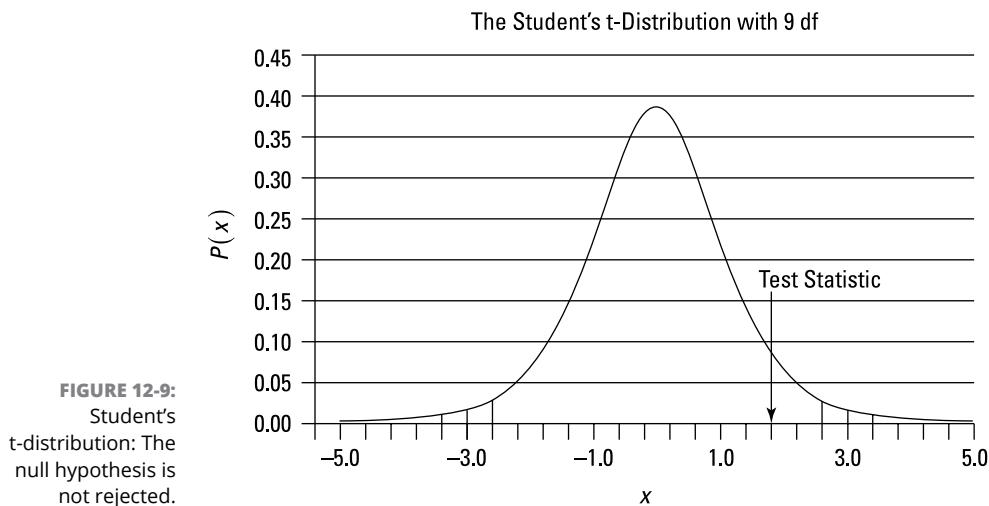


FIGURE 12-9:
Student's
t-distribution: The
null hypothesis is
not rejected.

Testing Hypotheses About Two Population Means

In addition to testing claims about the mean of a population, hypothesis testing can be used to compare the equality of two different population means. For example, you can use hypothesis testing on the following statements to determine whether they're true:

- » The mean price of gasoline per gallon is equal in New York and New Jersey.
- » The average life expectancy of men is the same in the United States and Canada.
- » The mean annual rainfall is equal in Washington and Oregon.
- » The length of the average flight delay is the same at Kennedy Airport and LaGuardia Airport.

The basic procedure for testing hypotheses about two population means is similar to the procedure for a single population mean (which I discuss in the section “Applying the Key Steps in Hypothesis Testing for a Single Population Mean”). The most important differences are the form of the test statistics you use for two population means and the calculation of the critical values. I outline the differences in the following sections.

Writing the null hypothesis for two population means

To test the equality of two population means, you write the null hypothesis as

$$H_0: \mu_1 = \mu_2$$

In this formula, H_0 is the null hypothesis, μ_1 is the mean of population 1, and μ_2 is the mean of population 2. Note that when testing hypotheses about two population means, one population is arbitrarily chosen to be “population 1” and the other becomes “population 2.”

Defining the alternative hypotheses for two population means

Just as you have an alternative hypothesis for testing a single population mean, when you test two population means, you also need an alternative hypothesis. If the null hypothesis is rejected, you must specify what other result will be accepted instead. This is the role of the alternative hypothesis.

The alternative hypothesis can take one of three forms:

- » **Right-tailed test:** $H_1: \mu_1 > \mu_2$
- » **Left-tailed test:** $H_1: \mu_1 < \mu_2$
- » **Two-tailed test:** $H_1: \mu_1 \neq \mu_2$

A right-tailed test is used to indicate if the mean of population 1 is *greater than* the mean of population 2. Similarly, a left-tailed test is used to show if the mean of population 1 is *less than* the mean of population 2. A two-tailed test is used to show if the mean of population 1 is *different from* the mean of population 2.

Determining the test statistics for two population means

When you're testing hypotheses about two population means, you can choose from several test statistics. The choice depends on:

- » whether the samples drawn from the two populations are independent of each other
- » whether the variances of the two populations are equal
- » whether the samples chosen from the two populations are large (at least 30) or small (less than 30)

Samples are *independent* if they're not related to each other. For example, samples of GPAs at two universities are independent samples, because none of the students in these samples attend both universities.

If you choose independent samples from two populations, you choose the test statistic and critical values based on the following questions:

- » Are the variances of the two populations equal?
- » If the variances are unequal, are the sample sizes large (at least 30)?

If the samples are *dependent*, the choice for test statistics and critical values are different. For example, suppose that medical researchers are conducting a study to determine whether a new cholesterol drug is effective in reducing LDL (bad cholesterol) in patients. If you chose a sample of LDL readings chosen from a set of patients prior to taking the drug and a sample of LDL readings among the same patients after taking the drug, these two samples would be closely related and, therefore, *dependent*. This type of hypothesis test requires a different procedure for constructing the test statistic and critical values than for independent samples. I explore using independent and dependent, or *paired*, samples in the following sections.

Using independent samples

When using independent samples, you first have to decide whether the populations being tested have equal variances (or if you have reason to believe that they're equal).

With equal population variances, the test statistic requires the calculation of a pooled variance — this is the variance that the two populations have in common. You use the Student's t-distribution to find the test statistic and critical values.

With unequal population variances, there are two possibilities:

- » You use the standard normal distribution for the test statistic and critical values if the samples are large (at least 30).
- » You use the t-distribution if at least one of the samples is small (less than 30).

The choice of distribution for the hypothesis test based on independent samples is summarized in Table 12-4.

TABLE 12-4

Choice of Probability Distribution for Independent Samples

Condition	Distribution
Equal variances	Student's t
Unequal variances: large samples	Standard Normal (Z)
Unequal variances: at least one small sample	Student's t

EQUAL POPULATION VARIANCES

If the variances of two populations are equal (or are assumed to be equal) the appropriate test statistic is based on the Student's t-distribution:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Here's what each term means:

- » \bar{x}_1 is the mean of the sample chosen from population 1.
- » \bar{x}_2 is the mean of the sample chosen from population 2.
- » μ_1 is the mean of population 1.
- » μ_2 is the mean of population 2.
- » $(\mu_1 - \mu_2)_0$ is the hypothesized difference between populations 1 and 2, which is 0 when the population means are hypothesized to be equal.
- » n_1 is the size of the sample chosen from population 1.
- » n_2 is the size of the sample chosen from population 2.
- » s_p^2 is the variance of the sample chosen from population 1.
- » s_p^2 is the variance of the sample chosen from population 2.

- » s_p^2 is the estimated common pooled variance of the two populations, or in mathematical terms:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

If you are conducting a hypothesis test of two population means with equal population variances, you take the critical values from the Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom, which gives you the following critical values:

- » Right-tailed test: $t_{\alpha}^{n_1+n_2-2}$
- » Left-tailed test: $-t_{\alpha}^{n_1+n_2-2}$
- » Two-tailed test: $\pm t_{\alpha/2}^{n_1+n_2-2}$

As an example, suppose that a marketing company is interested in determining whether men and women are equally likely to buy a new product. The company randomly chooses samples of men and women and asks them to assign a numerical value to their likelihood of buying the product (1 being the least likely, and 10 being the most likely).

Based on past experience, the population variances are assumed to be equal. The first step is to assign one group to be the first population ("population 1") and the other group to be the second population ("population 2"). The company designates men as population 1 and women as population 2.

The next step is to choose samples from both populations. (The sizes of these samples do not have to be equal.) Suppose that the company chooses samples of 21 men and 21 women. These samples are used to compute the sample mean and sample standard deviation for both men and women. (Sample means are covered in Chapter 3; sample standard deviations are covered in Chapter 4.)

Assume that the sample mean score of the men is 7.2; the sample mean score of the women is 6.7. Also assume that the sample standard deviation of the men is 0.4, and the sample standard deviation of the women is 0.3. With this data in place, the null hypothesis that the population mean scores are equal is tested by the marketing company at the 5 percent level of significance.

You can summarize the sample data like so:

$$\bar{x}_1 = 7.2 \text{ and } \bar{x}_2 = 6.7$$

$$s_1 = 0.4 \text{ and } s_2 = 0.3$$

$$n_1 = 21 \text{ and } n_2 = 21$$

The null hypothesis is $H_0: \mu_1 = \mu_2$. The alternative hypothesis is $H_1: \mu_1 \neq \mu_2$, because the marketing company just wants to know if the mean scores are the same for men and women.

To compute the test statistic, you first calculate the pooled variance:

$$\begin{aligned}s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\&= \frac{(21-1)(0.4)^2 + (21-1)(0.3)^2}{21+21-2} \\&= \frac{3.2+1.8}{40} \\&= 0.125\end{aligned}$$

You then substitute this result into the test statistic formula:

$$\begin{aligned}t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \\&= \frac{(7.2 - 6.7) - 0}{\sqrt{0.125 \left(\frac{1}{21} + \frac{1}{21} \right)}} \\&= \frac{0.5}{0.109} \\&= 4.587\end{aligned}$$

You can find the appropriate critical values from Table 12–5 (which is an excerpt from the Student's t-table, covered in Chapter 11). These are found as follows. The top row of the Student's t-table lists different values of t_α , where the right tail of the Student's t-distribution has a probability (area) equal to α ("alpha").

In this case, alpha (α) is 0.05; using a tail area of 0.025 ($\alpha/2$) and 40 degrees of freedom, you find that the critical values are:

$$\pm t_{\alpha/2}^{n_1+n_2-2} = \pm t_{0.05/2}^{21+21-2} = \pm t_{0.025}^{40} = \pm 2.021$$

TABLE 12-5

The Student's t-Distribution with a Large Number of Degrees of Freedom

Degrees of Freedom (df)	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660



TIP

Note that with a large number of degrees of freedom, the Student's t-distribution closely resembles the standard normal distribution (see Chapter 9 for more discussion of the normal distribution). For example, if you perform a two-tailed hypothesis test with $\alpha = 0.05$, the critical values drawn from the standard normal distribution are ± 1.96 , compared with ± 2.000 for the Student's t-distribution with 60 degrees of freedom.

Because the test statistic (4.587) exceeds the positive critical value (2.021), the null hypothesis $H_0: \mu_1 = \mu_2$ is rejected and the alternative hypothesis $H_1: \mu_1 \neq \mu_2$ is accepted. Men and women are not equally likely to buy the product.

UNEQUAL POPULATION VARIANCES: AT LEAST ONE SAMPLE IS SMALL

If the variances of two populations aren't equal (or you don't have any reason to believe that they're equal) and at least one sample is small (less than 30), the appropriate test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

In this case, you get the critical values from the t-distribution with degrees of freedom equal to

$$df = \frac{\left[\left(s_1^2/n_1\right) + \left(s_2^2/n_2\right)\right]^2}{\left[\frac{\left(s_1^2/n_1\right)^2}{(n_1-1)} + \frac{\left(s_2^2/n_2\right)^2}{(n_2-1)}\right]}$$



TIP

This value isn't necessarily equal to a whole number; if the resulting value contains a fractional part, you must round it to the next closest whole number.

For example, assume that Major League Baseball (MLB) is interested in determining whether the mean number of runs scored per game is higher in the American League (AL) than in the National League (NL). The population variances are assumed to be unequal.

The first step is to assign one group to represent the first population ("population 1") and the other group to represent the second population ("population 2"). MLB designates the American League as population 1 and the National League as population 2.

The next step is to choose samples from both populations. Suppose that MLB chooses a sample of 10 American League and 12 National League teams. The results

are used to compute the sample mean and sample standard deviation for both leagues. Assume that the sample mean for runs scored among the AL games is 8.1, whereas the sample mean for the NL games is 7.9. The sample standard deviation is 0.5 for AL games and 0.3 for NL games.

MLB tests the null hypothesis that the population mean scores are equal at the 5 percent level of significance.

Here's a summary of the sample data:

$$\bar{x}_1 = 8.1 \text{ and } \bar{x}_2 = 7.9$$

$$s_1 = 0.5 \text{ and } s_2 = 0.3$$

$$n_1 = 10 \text{ and } n_2 = 12$$

The null hypothesis is

$$H_0: \mu_1 = \mu_2$$

Because MLB is interested in determining whether the mean number of runs scored per game is higher in the American League than in the National League, you use a right-tailed test. The alternative hypothesis is $H_1: \mu_1 > \mu_2$.

In other words, the test is designed to find strong evidence that the mean of population 1 is *greater than* the mean of population 2. You then solve the test statistic as follows:

$$\begin{aligned} t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \\ &= \frac{(8.1 - 7.9) - 0}{\sqrt{\left(\frac{0.5^2}{10}\right) + \left(\frac{0.3^2}{12}\right)}} \\ &= \frac{0.2}{0.1803} \\ &= 1.109 \end{aligned}$$

And you find the degrees of freedom like so:

$$\begin{aligned} df &= \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{(n_2-1)}} = \frac{[(0.5^2/10) + (0.3^2/12)]^2}{\frac{(0.5^2/10)^2}{(10-1)} + \frac{(0.3^2/12)^2}{(12-1)}} \\ &= 0.00105625 / (0.000069444 + 0.00000511364) \\ &= 14.167 \sim 14 \end{aligned}$$

You round down the value of 14.167 to 14 because the degrees of freedom must be a whole number (or *integer*). With 14 degrees of freedom and a 5 percent level of significance, the critical value is $t_{0.05}^{14} = 1.761$.

This result is obtained from Table 12-1 by finding the column headed $t_{0.05}$ and the row corresponding to 14 degrees of freedom.

Because the test statistic (1.109) is below the critical value (1.761), the null hypothesis that $H_0: \mu_1 = \mu_2$ fails to be rejected. There's insufficient evidence to conclude that more runs are scored during American League games than National League games.

UNEQUAL POPULATION VARIANCES: BOTH SAMPLE SIZES ARE LARGE

If the variances of two populations *aren't* equal, and the size of both samples is 30 or greater, the appropriate test statistic is

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}}$$

This test statistic is based on the standard normal distribution.

As an example, say that a restaurant chain is interested in finding out whether the average sale per customer is the same in its domestic and foreign restaurants. The population variances are assumed to be unequal. The restaurant chooses a random sample of 40 domestic and 50 foreign restaurants, designating domestic restaurants as population 1 and foreign restaurants as population 2.

The sample mean spending per customer is \$5.14 in the domestic market and \$4.59 in the foreign market. The sample standard deviation is \$0.54 in the domestic market and \$0.38 in the foreign market. The null hypothesis that the population mean spending is equal in the two markets is tested at the 5 percent level of significance.

Here's a summary of this data:

$$\bar{x}_1 = 5.14 \text{ and } \bar{x}_2 = 4.59$$

$$s_1 = 0.54 \text{ and } s_2 = 0.38$$

$$n_1 = 40 \text{ and } n_2 = 50$$

The null hypothesis is $H_0: \mu_1 = \mu_2$.

Because this example requires a two-tailed test, the alternative hypothesis is $H_1: \mu_1 \neq \mu_2$.

You find the test statistic like so:

$$\begin{aligned} Z &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}} \\ &= \frac{(5.14 - 4.59) - 0}{\sqrt{\left(\frac{0.54^2}{40}\right) + \left(\frac{0.38^2}{50}\right)}} \\ &= \frac{0.55}{\sqrt{0.00729 + 0.002888}} \\ &= 5.452 \end{aligned}$$

The critical values are then $\pm Z_{\alpha/2} = \pm Z_{0.025} = \pm 1.96$ (see Table 12-3).

Because the test statistic (5.452) is greater than the positive critical value (1.96), the null hypothesis $H_0: \mu_1 = \mu_2$ is rejected.

Because this is a two-tailed test, you may reject the null hypothesis in favor of the alternative $H_1: \mu_1 \neq \mu_2$ (mean spending per customer in the domestic market is not equal to mean spending in the foreign market).

Working with dependent samples

You can choose samples to compare the mean of a population before and after a given event. In this case, the samples aren't independent; instead, they're dependent, or *paired* samples. Examples of paired samples include:

The cholesterol readings of randomly selected patients before taking a new drug and the cholesterol readings of the same patients after taking the drug

The grade point averages of randomly chosen students before being tutored and the grade point averages of the same students after being tutored

The productivity of a randomly selected group of employees prior to taking a new training course and the productivity of the same employees after taking the training course

With paired samples, the null hypothesis is based on the *differences* between the sample elements. Instead of stating that the population means are equal, the null hypothesis is that the difference between the population means equals 0.

When you're testing hypotheses about the equality of two population means with paired samples, you write the null hypothesis as

$$H_0: \mu_d = 0$$

where μ_d represents the mean difference between the two populations; it equals $\mu_d = \mu_1 - \mu_2$.

The three possible alternative hypotheses are

- » **Right-tailed test:** $H_1: \mu_d > 0$. In this case, the alternative hypothesis is that the mean of population 1 is *greater than* the mean of population 2.
- » **Left-tailed test:** $H_1: \mu_d < 0$. In this case, the alternative hypothesis is that the mean of population 1 is *less than* the mean of population 2.
- » **Two-tailed test:** $H_1: \mu_d \neq 0$. In this case, the alternative hypothesis is that the means of populations 1 and 2 *aren't* equal.

For paired samples, the test statistic is always based on the Student's t-distribution:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

Here, \bar{d} is the average difference between paired samples, and s_d is the standard deviation of the sample differences.

Compute the mean of the differences like this:

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n}$$

This formula indicates that you calculate the average difference between the paired samples by adding up all the individual differences and then dividing by the total number of elements in each sample.

Compute the standard deviation of the differences like this:

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$$

Note that this is the sample standard deviation formula (covered in Chapter 4).

With paired samples, you take the critical values from the Student's t-distribution with $n - 1$ degrees of freedom, where n is the number of paired observations.

For example, suppose that a pharmaceutical company is testing a new diet pill to determine whether taking it leads to weight loss. The company chooses a sample of eight volunteers. Table 12-6 shows the mean weights of these individuals before and after using the diet pill, along with the necessary calculations for computing the sample standard deviation:

TABLE 12-6

Paired Differences Between Two Samples

Subject	Weight Prior to Taking Diet Pill (x_1)	Weight After Taking Diet Pill (x_2)	$d_i = x_1 - x_2$	$(d_i - \bar{d})^2$
1	192	190	2	1.891
2	189	185	4	0.391
3	204	199	5	2.641
4	177	177	0	11.391
5	156	151	5	2.641
6	228	224	4	0.391
7	244	239	5	2.641
8	201	199	2	1.891
		Sum	27	23.875
		Mean	3.375	

The company tests the null hypothesis that weight remains unchanged after taking the diet pill at the 5 percent level of significance. The null hypothesis is $H_0: \mu_d = 0$.

Because the pharmaceutical company is looking for strong evidence that the weights of the volunteers *dropped* after taking the pill, it uses a right-tailed test. (In other words, the mean weights of the volunteers before taking the pill is *greater than* the mean weights of the volunteers after taking the pill.)

The alternative hypothesis is $H_1: \mu_d > 0$.

You work through the test statistic as follows:

1. Compute the mean of the differences:

$$\bar{d} = \frac{27}{8} = 3.375$$

2. Compute the sample standard deviation of the differences:

$$\begin{aligned}s_d &= \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}} \\&= \sqrt{\frac{23.875}{8-1}} \\&= 1.847\end{aligned}$$

3. Use these results to compute the test statistic:

$$\begin{aligned}t &= \frac{\bar{d}}{s_d / \sqrt{n}} \\&= \frac{3.375}{1.847 / \sqrt{8}} \\&= \frac{3.375}{0.653} \\&= 5.168\end{aligned}$$

The critical value is found in Table 12-1: $t_{\alpha}^{n-1} = t_{0.05}^7 = 1.895$.

Because the test statistic (5.168) exceeds the critical value (1.895), the null hypothesis is rejected in favor of the alternative hypothesis, which states that the difference between the weights prior to taking the pill and after taking the pill is positive. The results show that the pills are contributing to weight loss.

Testing Hypotheses about Population Means with the TI-84 Plus Calculator

You can use the Texas Instruments TI-84 Plus and Plus CE calculators to test hypotheses about a single population mean and about two population means.

Single population mean

When testing hypotheses about a single population mean, the standard normal (Z) distribution is used when the population standard deviation (σ) is known. The t-distribution is used when the population standard deviation is unknown.

To conduct a hypothesis test for the population mean with the standard normal distribution, follow these steps:

- 1. Press the [STAT] button.**
- 2. Use the arrow key to choose TESTS.**
- 3. Choose 1: ZTest.**

This refers to a hypothesis test for a single population mean based on the "Z" or standard normal distribution.

The first row (Inpt) offers the option of choosing the Data menu or the Stats menu for inputting data. Use the Data menu when the original data is entered into a list. Use the Stats menu when only summary statistics, such as the sample mean and sample size, are available.

When using the Data menu, the following information must be provided:

- » μ_0 : The hypothesized value of the population mean.
- » σ : The population standard deviation.
- » List: The list containing the data for this problem (L1, L2, L3, and so on).
- » Freq: Set equal to 1 unless the data in the list will be repeated; if the data are repeated, set Freq equal to the number of times the data are repeated.
- » μ : $\neq \mu_0 < \mu_0 > \mu_0$. This is where the alternative hypothesis is chosen. For a two-tailed test, choose $\neq \mu_0$; for a left-tailed test, choose $< \mu_0$; and for a right-tailed test, choose $> \mu_0$. Move to the correct choice with the arrow key and then press the [ENTER] key.
- » Calculate: Used to compute the final results.

When using the Stats menu, the following information must be provided:

- » μ_0 : The hypothesized value of the population mean.
- » σ : The population standard deviation.
- » \bar{X} : The sample mean.
- » n: The sample size.
- » μ : $\neq \mu_0 < \mu_0 > \mu_0$. This is where the alternative hypothesis is chosen. For a two-tailed test, choose $\neq \mu_0$; for a left-tailed test, choose $< \mu_0$; and for a right-tailed test, choose $> \mu_0$.
- » Calculate: Used to compute the final results.

When using Data for inputting data, the output consists of:

- » The alternative hypothesis (whether this is a left-tailed, right-tailed or two-tailed test).
- » The test statistic (Z).
- » The p-value (also known as the probability value). Recall that when the p-value is less than the level of significance (α), the null hypothesis is rejected; otherwise, it is not.
- » The sample mean.
- » The sample standard deviation.
- » The sample size.

When using Stats for inputting data, the output consists of:

- » The alternative hypothesis (whether this is a left-tailed, right-tailed or two-tailed test).
- » The test statistic (Z).
- » The p-value (also known as the probability value). Recall that when the p-value is less than the level of significance (α), the null hypothesis is rejected; otherwise, it is not.
- » The sample mean.
- » The sample size.

To conduct a hypothesis test for the population mean with the t-distribution, follow these steps:

1. Press the [STAT] button.
2. Use the arrow key to choose TESTS.
3. Choose 2: TTest.

This refers to a hypothesis test for a single population mean based on the t-distribution.

The first row (Inpt) offers the option of choosing the Data menu or the Stats menu for inputting data. Use the Data menu when the original data is entered into a list. Use the Stats menu when only summary statistics, such as the sample mean and sample size, are available.

When using the Data menu, the following information must be provided:

- » μ_0 : The hypothesized value of the population mean.
- » List: The list containing the data for this problem (L1, L2, L3, and so on).
- » Freq: Set equal to 1 unless the data in the list will be repeated; if the data are repeated, set Freq equal to the number of times the data are repeated.
- » μ : $\neq \mu_0 < \mu_0 > \mu_0$. This is where the alternative hypothesis is chosen. For a two-tailed test, choose $\neq \mu_0$; for a left-tailed test, choose $< \mu_0$, and for a right-tailed test, choose $> \mu_0$. It is necessary to move to the correct choice with the arrow key followed by pressing the [ENTER] key.
- » Calculate: Used to compute the final results.

When using the Stats menu, the following information must be provided:

- » μ_0 : The hypothesized value of the population mean.
- » \bar{X} : The sample mean.
- » Sx: The sample standard deviation.
- » n: The sample size.
- » μ : $\neq \mu_0 < \mu_0 > \mu_0$. This is where the alternative hypothesis is chosen. For a two-tailed test, choose $\neq \mu_0$; for a left-tailed test, choose $< \mu_0$, and for a right-tailed test, choose $> \mu_0$.
- » Calculate: Used to compute the final results.

When using Data for inputting data, the output will consist of:

- » The alternative hypothesis (whether this is a left-tailed, right-tailed or two-tailed test).
- » The test statistic (t).
- » The p-value (also known as the probability value). Recall that when the p-value is less than the level of significance (α), the null hypothesis is rejected; otherwise, it is not.
- » The sample mean.
- » The sample standard deviation.
- » The sample size.

When using Stats for inputting data, the output will consist of:

- » The alternative hypothesis (whether this is a left-tailed, right-tailed or two-tailed test).
- » The test statistic (t).
- » The p-value (also known as the probability value). Recall that when the p-value is less than the level of significance (α), the null hypothesis is rejected; otherwise, it is not.
- » The sample mean.
- » The sample standard deviation.
- » The sample size.



TIP

If the p-value is extremely small, the calculator will use scientific notation to represent it. For example, if the p-value is 1.423806916 E-5, this means $1.423806916 \times 10^{-5}$ or 0.00001423806916. Be sure to check for the E at the end of the p-value!

Two population means

To conduct a hypothesis test about two population means, you have a few possible choices. When the two populations are independent of each other, follow these steps:

1. Press the [STAT] button.
2. Use the arrow key to choose TESTS.
3. Choose either 3: 2SampZTest or 4: 2SampTTest:
 - a. Choose 3: 2SampZTest for cases when the standard normal distribution is used.
 - b. Choose 4: 2SampTTest for cases when the t-distribution is used.

Use 3: 2SampZTest when two independent populations are being studied, the populations do not have equal variances, and the samples chosen from both populations are large ($n \geq 30$).

Use 4: 2SampTTest for the following cases:

- » You have two independent populations with the same variance.
- » You have two independent populations with different variances and at least one small sample ($n < 30$) is chosen.

When choosing 3: 2SampZTest, the first row (Inpt) offers the option of choosing the Data menu or the Stats menu for inputting data. Use the Data menu when the original data is entered into two lists. Use the Stats menu when only summary statistics, such as the sample mean and sample size, are available.

When using the Data menu, the following information must be provided:

- » σ_1 : The standard deviation of population 1.
- » σ_2 : The standard deviation of population 1.
- » List1: The location of the data for population 1 (this is the list where the data have been stored, such as L1, L2, and so on).
- » List 2: The location of the data for population 2.
- » Freq1: The frequency of the data in List 1.
- » Freq2: The frequency of the data in List 2.
- » $\mu: \neq \mu_0 < \mu_0 > \mu_0$. This is where the alternative hypothesis is chosen. For a two-tailed test, choose $\neq \mu_0$; for a left-tailed test, choose $< \mu_0$; and for a right-tailed test, choose $> \mu_0$.
- » Calculate: Used to compute the final results.

When using the Stats menu, the following information must be provided:

- » σ_1 : The standard deviation of population 1.
- » σ_2 : The standard deviation of population 2.
- » \bar{X}_1 : The mean of the sample drawn from population 1.
- » n_1 : The size of the sample drawn from population 1.
- » \bar{X}_2 : The mean of the sample drawn from population 2.
- » n_2 : The size of the sample drawn from population 2.
- » $\mu: \neq \mu_0 < \mu_0 > \mu_0$. This is where the alternative hypothesis is chosen. For a two-tailed test, choose $\neq \mu_0$; for a left-tailed test, choose $< \mu_0$; and for a right-tailed test, choose $> \mu_0$.
- » Calculate: Used to compute the final results.

When using the Data menu for inputting data, the output will consist of:

- » The alternative hypothesis (whether this is a left-tailed, right-tailed or two-tailed test).
- » The test statistic (Z).
- » The p-value (also known as the probability value). Recall that when the p-value is less than the level of significance (α), the null hypothesis is rejected; otherwise, it is not.
- » The mean of sample 1.
- » The mean of sample 2.
- » The standard deviation of sample 1.
- » The standard deviation of sample 2.
- » The size of sample 1.
- » The size of sample 2.

When using the Stats menu for inputting data, the output will consist of:

- » The alternative hypothesis (whether this is a left-tailed, right-tailed or two-tailed test).
- » The test statistic (Z).
- » The p-value (also known as the probability value). Recall that when the p-value is less than the level of significance (α), the null hypothesis is rejected; otherwise, it is not.
- » The mean of sample 1.
- » The mean of sample 2.
- » The size of sample 1.
- » The size of sample 2.

When choosing 4: 2SampTTest, the first row (Inpt) offers the option of choosing the Data menu or the Stats menu for inputting data. Use the Data menu when the original data is entered into two lists. Use the Stats menu when only summary statistics, such as the sample mean and sample size, are available.

When using the Data menu, the following information must be provided:

- » List1: The location of the data for population 1 (this is the list where the data have been stored, such as L1, L2, and so on).
- » List 2: The location of the data for population 2.
- » Freq1: The frequency of the data in List 1.
- » Freq2: The frequency of the data in List 2.
- » μ : $\neq \mu_0 < \mu_0 > \mu_0$. This is where the alternative hypothesis is chosen. For a two-tailed test, choose $\neq \mu_0$; for a left-tailed test, choose $< \mu_0$; and for a right-tailed test, choose $> \mu_0$.
- » Pooled (No / Yes): If the two populations are assumed to have the same variance, choose Yes; otherwise, choose No.
- » Calculate: Used to compute the final results.

When using the Stats menu, the following information must be provided:

- » \bar{X}_1 : The mean of the sample drawn from population 1.
- » Sx_1 : The standard deviation of the sample drawn from population 1.
- » n_1 : The size of the sample drawn from population 1.
- » \bar{X}_2 : The mean of the sample drawn from population 2.
- » Sx_2 : The standard deviation of the sample drawn from population 2.
- » n_2 : The size of the sample drawn from population 2.
- » μ : $\neq \mu_0 < \mu_0 > \mu_0$. This is where the alternative hypothesis is chosen. For a two-tailed test, choose $\neq \mu_0$; for a left-tailed test, choose $< \mu_0$; and for a right-tailed test, choose $> \mu_0$.
- » Pooled (No / Yes): If the two populations are assumed to have the same variance, choose Yes; otherwise, choose No.
- » Calculate: Used to compute the final results.

When using the Data menu for inputting data, the output will consist of:

- » The alternative hypothesis (whether this is a left-tailed, right-tailed or two-tailed test).
- » The test statistic (t).

- » The p-value (also known as the probability value). Recall that when the p-value is less than the level of significance (α), the null hypothesis is rejected; otherwise, it is not.
- » The degrees of freedom.
- » The mean of sample 1.
- » The mean of sample 2.
- » The standard deviation of sample 1.
- » The standard deviation of sample 2.
- » If Pooled = Yes, the sample pooled variance appears here.
- » The size of sample 1.
- » The size of sample 2.

When using the Stats menu for inputting data, the output will consist of:

- » The alternative hypothesis (whether this is a left-tailed, right-tailed or two-tailed test).
- » The test statistic (t).
- » The p-value (also known as the probability value). Recall that when the p-value is less than the level of significance (α), the null hypothesis is rejected; otherwise, it is not.
- » The degrees of freedom.
- » The mean of sample 1.
- » The mean of sample 2.
- » If Pooled = Yes, the sample pooled variance appears here.
- » The size of sample 1.
- » The size of sample 2.



REMEMBER

When testing hypotheses about the means of two dependent populations, the differences between the samples are treated as a single sample. This means that you will use 2: TTest from the STAT – TESTS menu, as described earlier.

IN THIS CHAPTER

- » Introducing the chi-square distribution
- » Testing hypotheses about the variance of a single population
- » Implementing goodness of fit tests with the chi-square distribution

Chapter **13**

Applications of the Chi-Square Distribution

This chapter covers two types of hypothesis tests: tests about the population variance and *goodness of fit* tests. Goodness of fit tests determine whether a population follows a specified distribution, such as the normal distribution (for a thorough introduction to the normal distribution, see Chapter 9). Because many business applications rely on the assumption of normality, goodness of fit tests are particularly valuable.

To implement a goodness of fit test, you use a continuous distribution known as the *chi-square distribution*. This distribution has many interesting features, which I explain in detail and illustrate throughout this chapter; its properties are quite different from the normal distribution.

I also explain how to use a chi-square table to compute probabilities under the chi-square distribution, and I show you how to compute moments for the chi-square distribution. (*Moments* are summary measures of a probability distribution that provide a great deal of useful information in a very compact form.)

You can use the chi-square distribution to test hypotheses about the variance of a population. For example, you can use the chi-square distribution to determine the level of risk contained in a stock portfolio. (The process of testing a hypothesis about a population variance is closely related to other types of hypothesis tests, which I cover in Chapter 12.)

Staying Positive with the Chi-Square Distribution

The chi-square distribution (χ^2) is a *continuous* probability distribution, which means that it's defined for an infinite number of values. I introduce continuous probability distributions, including the normal, Student's t-, and F-distributions, in Chapters 9, 11, and 14, respectively. (To read about discrete probability distributions, check out Chapter 8.)

The chi-square distribution has several different applications. This section shows you how to use the chi-square distribution to:

- » Test hypotheses about the variance of a population
- » Carry out "goodness of fit" tests

Portfolio managers, financial analysts, traders, and so on regularly use continuous distributions in business applications to analyze the properties of financial variables. Two of the more widely used continuous distributions are the normal and Student's t-distributions (see Chapters 9 and 11, respectively). Many business situations can be described with the normal distribution, such as returns to stocks, corporate profits, and so on. The normal and Student's t-distribution can also be used to construct confidence intervals (described in Chapter 11) and test hypotheses about population means (described in Chapter 12).

As with the Student's t-distribution, the chi-square distribution is uniquely characterized by a value known as *degrees of freedom* (*df*). The number of degrees of freedom is based on the sizes of samples used to estimate population parameters, such as the mean or the variance.

Here are two important features of the chi-square distribution:

- » It's defined only for positive values.
- » It's *not* symmetrical about its mean; instead, it's positively skewed.

A distribution may be symmetrical about its mean, in which case the area below the mean is a mirror image of the area above the mean. For a symmetric distribution, the mean equals the median. (I discuss symmetry in Chapter 3.) A distribution may also be negatively skewed, where the mean is less than the median, or positively skewed, where the mean is greater than the median.

The chi-square distribution is positively skewed; graphically, it has a long right tail. The next section shows several graphs of the chi-square distribution with different numbers of degrees of freedom. The smaller the degrees of freedom, the more skewed the distribution is; with a larger number of degrees of freedom, the distribution becomes more symmetrical and begins to resemble the normal distribution.

Following the graphs of the chi-square distribution is a discussion of how to compute the moments of the chi-square distribution.

Representing the chi-square distribution graphically

Figures 13-1, 13-2, and 13-3 show the chi-square distribution with 5, 10, and 30 degrees of freedom. In each case, the horizontal axis represents different possible values of the chi-square distribution; the vertical axis represents the corresponding probabilities. With a continuous distribution such as the chi-square, probabilities correspond to areas under the curve.

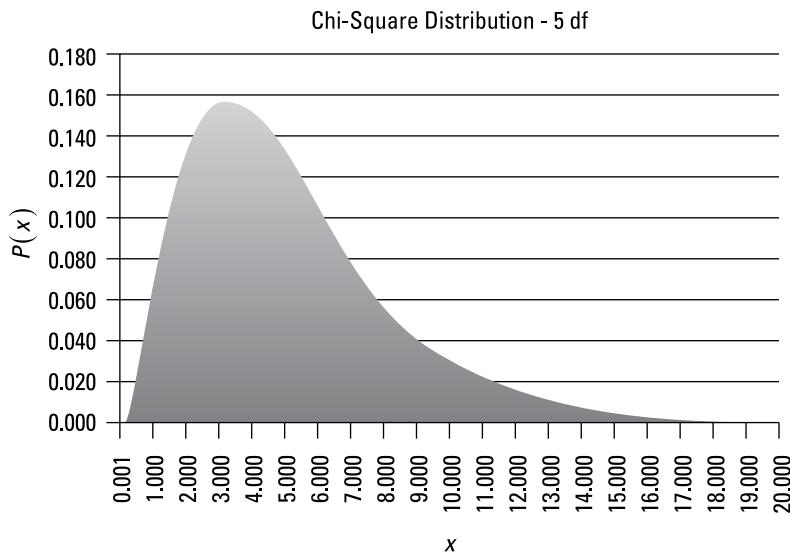


FIGURE 13-1:
The chi-square distribution with 5 degrees of freedom.

As you can see in each figure, the distribution isn't defined for negative values — that is, no negative values appear along the horizontal axis. Additionally, as the number of degrees of freedom increases, the distribution shifts to the right and begins to resemble the normal distribution (it has a long right tail and is skewed to the right).

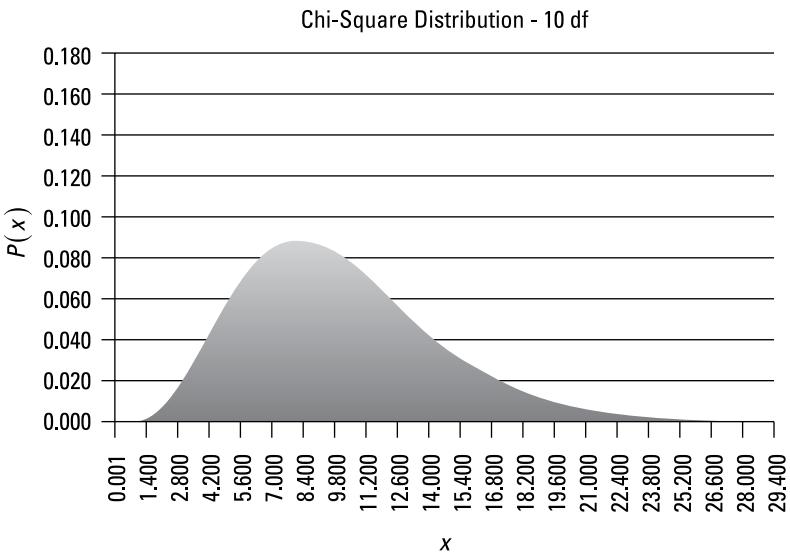


FIGURE 13-2:
The chi-square distribution with 10 degrees of freedom.

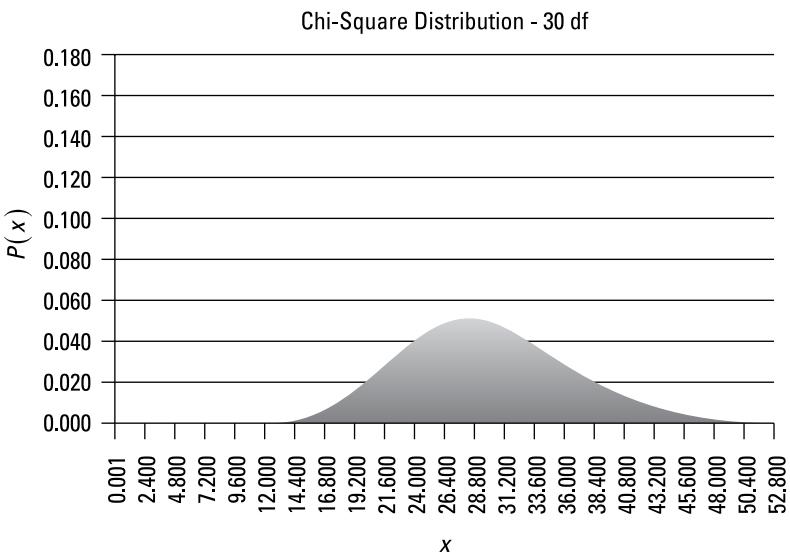


FIGURE 13-3:
The chi-square distribution with 30 degrees of freedom.

Defining a chi-square random variable

A chi-square random variable is composed of a sum of independent, squared standard normal random variables (Z^2) (see Chapter 7 for details about probability distributions and random variables and Chapter 9 for details about the normal distribution). The standard normal distribution is the special case of the normal

distribution where the mean (μ) equals 0 and the standard deviation (σ) equals 1. You can write the definition of a chi-square random variable mathematically as

$$\chi_v^2 = Z_1^2 + Z_2^2 + Z_3^2 + \dots + Z_v^2$$

Because each standard normal random variable is squared, the sum of these terms is guaranteed to be positive (which is why the chi-square distribution isn't defined for negative values).

The letter v (or "nu") represents the number of terms in this expression; here, v is the number of degrees of freedom of the distribution. For example, the chi-square distribution with 5 degrees of freedom is defined as follows:

$$\chi_5^2 = Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2$$

Checking out the moments of the chi-square distribution

Moments are summary measures of a probability distribution (see Chapter 8 for details) and include the *expected value* (or mean) and the *variance* (how spread out the values are). The *standard deviation* is the square root of the variance.

Each probability distribution has its own unique set of formulas for computing the expected value, variance, and standard deviation. For the chi-square distribution, these are given as follows:

- » The expected value equals the number of degrees of freedom (v) of the distribution:

$$E(X) = v$$

For example, in a chi-square distribution with 5 degrees of freedom, the expected value is 5.

- » The variance equals two times the number of degrees of freedom:

$$\sigma^2 = 2v$$

For example, for the chi-square distribution with 5 degrees of freedom, the variance is $2 \times 5 = 10$.

- » The standard deviation is the square root of the variance:

$$\sigma = \sqrt{2v}$$

For example, for the chi-square distribution with 5 degrees of freedom, the standard deviation is the square root of 10, which is approximately 3.16.



REMEMBER

Moments capture the key properties of a probability distribution. The expected value is another name for the average; the variance and standard deviation show how “spread out” the values of the distribution are relative to the expected value.

Testing Hypotheses about the Population Variance

In business, one of the most widely used applications of the chi-square distribution is to determine whether the variance of a population equals a specified value. The basic approach to testing a hypothesis about the population variance exactly mirrors the approach used for the population mean (which I cover in Chapter 12). The most important changes take place in the test statistic and critical values you use. In the following sections, I walk you through the steps to testing hypotheses about the population variance.

Defining what you assume to be true: The null hypothesis

The first step in the hypothesis testing procedure is writing *the null hypothesis*, which is a statement that’s assumed to be true unless strong contrary evidence exists against it. For example, suppose that a manufacturer is concerned that the variance of the computer chips that it produces exceeds 0.001, which would indicate that there’s a problem with the production process. The manufacturer can test this hypothesis by selecting a sample of computer chips and computing their sample variance.

The manufacturer may not want to make any changes to the production process unless clear evidence shows that it’s necessary. Therefore, it uses the null hypothesis that the variance equals 0.001. If this hypothesis is rejected, the alternative that the variance exceeds 0.001 is accepted instead. Unless the null hypothesis can be disproved with strong evidence, no changes are made to the production process. (Hypothesis testing is introduced in Chapter 12.)

For testing hypotheses about the population variance, the null hypothesis statement is based on the assumption that the population variance equals the *hypothesized value* of the population (σ_0^2). This assumption isn’t abandoned without strong contradictory evidence.

Mathematically, you write the null hypothesis as

$$H_0 : \sigma^2 = \sigma_0^2$$

The variance with a subscript of 0 (σ_0^2) is the hypothesized value of the variance. This is the value that you believe the population variance is equal to. The hypothesis test shows whether this belief is backed up by actual data.

For example, suppose that an economist wants to determine whether the variance of the inflation rate over the past 20 years equals 0.0001, in which case, you write the null hypothesis as $H_0 : \sigma^2 = 0.0001$. The economist continues to assume that this is the correct variance unless the hypothesis test provides strong evidence against this claim.

Stating the alternative hypothesis

Your second step in a hypothesis test is to specify the *alternative hypothesis*. If the statistical evidence against the null hypothesis is strong enough to reject it, you need an alternative statement to accept in its place.

The alternative hypothesis is a statement of what you accept to be true if the null hypothesis is rejected. For example, the economist in the previous section may want to know whether the actual variance is less than 0.0001, greater than 0.0001, or simply different from 0.0001 if the null hypothesis is rejected.

You can express the alternative hypothesis in three ways: as right-tailed, left-tailed, and two-tailed tests.

- » With a right-tailed test, you look for evidence that the actual population variance is *greater than* the hypothesized value.
- » With a left-tailed test, you look for evidence that the population variance is *less than* the hypothesized value.
- » With a two-tailed test, you look for evidence that the population variance is *either* less than or greater than the hypothesized value.

I explore each option in the following sections. (Right-tailed tests, left-tailed tests, and two-tailed tests are introduced in Chapter 12.)

Right-tailed test: Determining whether the hypothesized variance is too low

If you're interested in knowing only whether the population variance is greater than the hypothesized value, you use a right-tailed test. In this case, you express the alternative hypothesis (H_1) as

$$H_1 : \sigma^2 > \sigma_0^2$$

For example, suppose that a manufacturing company wants to keep the weights of its computer chips as uniform as possible. The company has determined from experience that the maximum variance the chips can tolerate is 0.0006 milligrams squared. (Variances are measured in terms of squared units, as I discuss in Chapter 4.) The manufacturing company can test the variance by choosing a sample from the assembly line and computing the sample variance. In this case, the company can test the hypothesis that the variance equals 0.0006 ($H_0 : \sigma^2 = 0.0006$). The alternative hypothesis is that the variance exceeds (or is greater than) 0.0006 ($H_1 : \sigma^2 > 0.0006$).

The results of this test show whether the manufacturing process is working correctly or whether it needs to be adjusted.

Left-tailed test: Determining whether the hypothesized variance is too high

If you're interested in knowing only whether the population variance is less than the hypothesized value, you use a left-tailed test. In this case, you express the alternative hypothesis as

$$H_1 : \sigma^2 < \sigma_0^2$$

For example, suppose that an equity analyst is studying the pattern of returns to U.S. stocks since the outbreak of the last financial crisis. The analyst wants to determine whether markets have begun to stabilize since the crisis began, which is indicated by a drop in the variances of the returns to U.S. stocks. The analyst believes that one stock is particularly representative of the performance of the overall economy. The analyst wants to see whether the variance of its returns has remained at 0.0004 or whether it's fallen below this level. In this case, the analyst can test the null hypothesis that the variance equals 0.0004 ($H_0 : \sigma^2 = 0.0004$). The alternative hypothesis is that the variance is less than 0.0004 ($H_1 : \sigma^2 < 0.0004$).

The results of this test show whether the variance of this stock has fallen below 0.0004. If so, the markets have stabilized since the outbreak of the financial crisis.

Two-tailed test: Determining whether the hypothesized variance is too low or too high

In some situations, it's extremely important for you to know whether the population variance is greater than or less than the hypothesized value. In this case, you use the two-tailed test, and write the alternative hypothesis as

$$H_1 : \sigma^2 \neq \sigma_0^2$$

For example, suppose that the variance of the returns to an investor's portfolio has historically been 0.0009; the investor wants to determine whether this number has increased or decreased over the past year. In this case, the investor can use a two-tailed hypothesis test. The null hypothesis is that the variance equals 0.0009 ($H_0 : \sigma^2 = 0.0009$), and the alternative hypothesis is that the variance doesn't equal 0.0009 ($H_1 : \sigma^2 \neq 0.0009$).

Choosing the level of significance

To test a hypothesis, you have to choose a *level of significance*. The level of significance, designated with α , refers to the probability of rejecting the null hypothesis when it's actually true, called a *Type I error*. (Chapter 12 provides details on Type I and Type II errors in hypothesis testing.)

You must choose the level of significance carefully. The greater the level of significance, the greater the likelihood of rejecting the null hypothesis when it's true — and the lower the likelihood of failing to reject the null hypothesis when it's false.

You choose the level of significance based on the relative importance of avoiding these errors. For many business applications, the level of significance is set to 0.05 (or 5 percent). Other commonly used values are 0.01 and 0.10.

Calculating the test statistic

To test hypotheses about the population variance, you must draw a sample from the underlying population so you can compute the sample variance. The sample variance is required to compute the test statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

This equation shows that the test statistic follows the chi-square distribution, with $n - 1$ degrees of freedom (n is the sample size); s^2 is the sample variance, and σ_0^2 is the hypothesized value of the population variance. This expression is used as a test statistic because it can be shown to follow the chi-square distribution with $n - 1$ degrees of freedom.

The purpose of the test statistic is to determine how extreme a sample statistic is (in this case, the sample variance) compared with the hypothesized value of the corresponding population parameter (here, the population variance). If the test statistic is too extreme (the value is an extremely large positive or negative number), it's highly unlikely that the null hypothesis is true, and it will be rejected. Otherwise, the null hypothesis won't be rejected.

To determine how extreme the test statistic is, you compare its value to one or two numbers known as *critical values*, depending on the alternative hypothesis. When testing hypotheses about the population variance, critical values are taken from the chi-square distribution. They represent the cutoff point between a specified area under the chi-square distribution.

For example, for the chi-square distribution with 10 degrees of freedom, a critical value of 18.30 is the cutoff point between the upper 5 percent of the chi-square distribution and the lower 95 percent of the chi-square distribution. In other words, for a chi-square random variable X ,

$$P(X \geq 18.30) = 0.05$$

$$P(X \leq 18.30) = 0.95$$

Determining the critical value(s)

To test a hypothesis about the variance of a population, the critical value(s) depends on the alternative hypothesis. Unlike critical values drawn from the standard normal distribution or the Student's t-distribution, the chi-square distribution has no negative critical values. Instead, you determine the critical values with the alternative hypothesis tests as explained in the following sections.

Right-tailed test: Testing hypotheses about the population variance

A right-tailed test has a single critical value because you're looking only for evidence that the test statistic is *too large* to be consistent with the null hypothesis. If you don't find this evidence, you won't reject the null hypothesis. The form of the critical value is

$$\chi^2_{\alpha, n-1}$$

In this expression,

$$\chi^2 = \text{a value chosen from the chi-square distribution}$$

α = the level of significance of the hypothesis test (for example, 0.01, 0.05, 0.10, and so on)

n = the sample size

The values of α and n uniquely identify the appropriate test statistic drawn from the chi-square distribution. This value represents the threshold of the right tail of the chi-square distribution with area α and $n - 1$ degrees of freedom. The area in the right tail is α . You can find this critical value in a chi-square table, such as Table 13-1.

TABLE 13-1**The Chi-Square Table**

df\Right-Tail Area	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209

For example, suppose that you conduct a right-tailed test with a level of significance of 0.05 (5 percent). You draw a sample of size 10. Plugging those numbers into the critical value, you get

$$\chi^2_{\alpha,n-1} = \chi^2_{0.05,9}$$

You then look at the chi-square table (Table 13-1). The top row represents areas in the right tail of the chi-square distribution. The first column represents the number of degrees of freedom.

In this example, you're looking for a right-tail area of 0.05 with 9 degrees of freedom ($n - 1 = 10 - 1 = 9$). By looking in the row corresponding to 9 degrees of freedom and the column corresponding to a right-tail area of 0.05, you see that the critical value is 16.919. Therefore,

$$\chi^2_{\alpha,n-1} = \chi^2_{0.05,9} = 16.919$$

As a result, if the test statistic is greater than 16.919, you reject the null hypothesis; otherwise, you don't reject the null hypothesis.

Left-tailed test: Testing hypotheses about the population variance

A left-tailed test has a single critical value because you're looking only for evidence that the test statistic is *too small* to be consistent with the null hypothesis.

If you don't find this evidence, you won't reject the null hypothesis. The form of the test statistic is

$$\chi^2_{(1-\alpha),n-1}$$

This value represents the threshold of the left tail of the chi-square distribution with area α and $n - 1$ degrees of freedom. The area in the right tail is therefore $1 - \alpha$.

Using the example in the previous section and referring to Table 13-1, if you do a left-tailed test with a level of significance of 0.05 and a sample of size 10, you find the appropriate critical value in the row with 9 degrees of freedom but a right-tail area of 0.95, which is

$$\chi^2_{(1-\alpha),n-1} = \chi^2_{0.95,9} = 3.325$$

As a result, if the test statistic is less than 3.325, you reject the null hypothesis; otherwise, you don't reject the null hypothesis.

Two-tailed test: Testing hypotheses about the population variance

A two-tailed test has two critical values. You're looking for evidence that the test statistic is too large or too small to be consistent with the null hypothesis. If you don't find this evidence, you won't reject the null hypothesis. The form of the critical values are

$$\begin{aligned}\chi^2_{(1-\alpha/2),n-1} \\ \chi^2_{(\alpha/2),n-1}\end{aligned}$$

The two-tailed test has a right tail and a left tail. Each has an area equal to $\alpha/2$. So, for example, if you do a two-tailed test with a level of significance of 0.05 and a sample of size 10, the appropriate critical values are 2.700 and 19.023 (see Table 13-1).

$$\begin{aligned}\chi^2_{(1-\alpha/2),n-1} &= \chi^2_{0.975,9} = 2.700 \\ \chi^2_{(\alpha/2),n-1} &= \chi^2_{0.025,9} = 19.023\end{aligned}$$

The boundary of the left 2.5 percent tail of the chi-square distribution is 2.700, and the boundary of the right 2.5 percent tail of the chi-square distribution is 19.023. Note that with the right-tailed test, the right tail has an area of 5 percent; with a left-tailed test, the left tail has an area of 5 percent. With a two-tailed test, the 5 percent area is split between the left and right tails; therefore, each has an area of 2.5 percent.

As a result, if the test statistic is less than 2.700 or greater than 19.023, you reject the null hypothesis; otherwise, you don't reject the null hypothesis.

Making the decision

You decide whether to reject the null hypothesis by looking at the relationship between the test statistic and the critical value(s). There are three possible cases: a right-tailed test, a left-tailed test, and a two-tailed test.

- » **Right-tailed test:** If the test statistic is greater than the critical value $\chi^2_{\alpha,n-1}$, you reject the null hypothesis $H_0: \sigma^2 = \sigma_0^2$ in favor of the alternative hypothesis $H_1: \sigma^2 > \sigma_0^2$. Otherwise, you don't reject the null hypothesis.
- » **Left-tailed test:** If the test statistic is less than the critical value $\chi^2_{(1-\alpha),n-1}$, you reject the null hypothesis $H_0: \sigma^2 = \sigma_0^2$ in favor of the alternative hypothesis $H_1: \sigma^2 < \sigma_0^2$. Otherwise, you don't reject the null hypothesis.
- » **Two-tailed test:** If the test statistic is less than the critical value $\chi^2_{(1-\alpha/2),n-1}$, you reject the null hypothesis $H_0: \sigma^2 = \sigma_0^2$ in favor of the alternative hypothesis $H_1: \sigma^2 \neq \sigma_0^2$.
If the test statistic is greater than the critical value $\chi^2_{(\alpha/2),n-1}$, you reject the null hypothesis $H_0: \sigma^2 = \sigma_0^2$ in favor of the alternative hypothesis $H_1: \sigma^2 \neq \sigma_0^2$.
Otherwise, you don't reject the null hypothesis.

As an example of the complete process used to test hypotheses about the population variance, suppose that an investor chooses a sample of 30 stocks from a portfolio. The investor calculates the standard deviation of the returns on these stocks (that is, their *volatility*) to be 23 percent on an annual basis. The investor wants to know whether the volatility of the entire portfolio is less than 25 percent on an annual basis at the 5 percent level of significance. (A volatility of 25 percent [0.25] translates into a variance of $0.25^2 = 0.0625$.) So the null hypothesis is $H_0: \sigma^2 = 0.0625$, and the alternative hypotheses is $H_1: \sigma^2 < 0.0625$.

Because the investor wants to know only whether the variance is less than 0.0625, you use a left-tailed test. The level of significance is $\alpha = 0.05$.

With a sample size of 30 and a sample variance of 0.23, the test statistic is

$$\begin{aligned}\chi^2 &= \frac{(n-1)s^2}{\sigma_0^2} \\ &= \frac{(30-1)(0.23)^2}{(0.25)^2} \\ &= 24.546\end{aligned}$$

Because this is a left-tailed test with $\alpha = 0.05$ and sample size = 30, the number of degrees of freedom = 29 ($30 - 1$). The critical value is, therefore, $\chi^2_{(1-\alpha),n-1} = \chi^2_{0.95,29}$. You can find the result in a chi-square table, such as Table 13-2.

TABLE 13-2

The Chi-Square Table with Larger Numbers of Degrees of Freedom

df\Right-Tail Area	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892

You find the critical value in the row corresponding to 29 degrees of freedom ($n - 1 = 30 - 1 = 29$) and the column with a right-tail area of 0.95 ($[1 - \alpha] = [1 - 0.05] = 0.95$). The result is 17.708.

To reject this hypothesis, the test statistic must be less than the critical value. In this case, the critical value is 24.546, and the test statistic is 17.708; therefore, the null hypothesis isn't rejected. There isn't enough evidence to conclude that the portfolio volatility is less than 25 percent.

Practicing the Goodness of Fit Tests

One of the most important applications of the chi-square distribution is to test whether a population conforms to a specific probability distribution. This type of test is called a *goodness of fit test*.

In this section, I show you examples of how to use sample data from a population to determine whether the population follows the Poisson distribution (covered in Chapter 8) or the normal distribution (discussed in Chapter 9). Note that these aren't the only possible applications of goodness of fit tests; in principle, you can compare any population to any probability distribution.

Comparing a population to the Poisson distribution

You use the Poisson distribution to describe the distribution of events occurring over a given interval of time. To test the hypothesis that a population follows the Poisson distribution, you express the null and alternative hypotheses as follows:

- » H_0 : The population follows the Poisson distribution.
- » H_1 : The population doesn't follow the Poisson distribution.

Alternatively, the null and alternative hypotheses may include an assumption about the parameter λ (the Greek letter “lambda”), which represents the expected number of events that occur during a given time frame.

For example, the null and alternative hypotheses could be

- » H_0 : The population follows the Poisson distribution with $\lambda = 1$.
- » H_1 : The population doesn't follow the Poisson distribution with $\lambda = 1$.

Use this approach if you have reason to believe that the value of $\lambda = 1$. In this case, the interpretation of the results is slightly different. If the null hypothesis that the population follows the Poisson distribution is rejected, the population actually follows a different distribution. If the null hypothesis that the population follows the Poisson distribution with $\lambda = 1$ is rejected, the population either doesn't follow the Poisson distribution or it follows the Poisson distribution but with a different value of λ .

One of the unusual features of a goodness of fit test is that you always implement the alternative hypothesis as a right-tailed test. Based on the construction of the test statistic, the null hypothesis that a population follows a specified distribution is rejected only if the test statistic is *too large*; therefore, a goodness of fit test is always right-tailed.

And you construct the test statistic in such a way as to see how closely the elements in a sample match up with the assumed probability distribution. To construct the test statistic, you choose sample data and arrange them into categories. For example, suppose that a bank manager wants to determine whether the distribution of customers who enter the bank during lunch hour (12 noon to 1 p.m.) follows the Poisson distribution. This information helps the manager determine the optimal number of tellers to use during this time period.

In this case, the population consists of the number of customers who enter the bank during lunch hour. Suppose that the manager chooses a random sample of 100 lunch hours from the past year and counts the number of customers who enter during each of those 100 hours. The manager then organizes the results as shown here:

Number of Customers per Hour	Number of Hours
0	9
1	12
2	15
3	20
4	27
5	12
6	5

According to these results, during each hour in the sample, the number of customers ranged from 0 to 6, so the manager organizes the data into a total of seven categories. The number of customers in each category is known as the *observed frequency* of the category. You must compare these numbers with the *expected frequencies* — the number of customers expected if the distribution of customers per hour really does follow the Poisson distribution.

In this example, you can find the expected frequencies for each category by computing the Poisson probabilities for each category and multiplying the result by the sample size. For example, suppose that the probability of three customers entering the bank each hour under the Poisson distribution is 0.2240, indicating that in a sample of 100 hours, the expected number of customers (or the expected frequency) is $0.2240 \times 100 = 22.40$ customers. (Of course, it's impossible for 22.40 customers to show up during lunch hour! This is simply an average.)

After you determine the expected frequency of each category, you compute the test statistic with this formula:

$$\chi^2 = \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j}$$



REMEMBER

Here, j is an index for the category being tested, k is the number of total categories, O_j is the observed frequency in category j , and E_j is the expected frequency in category j .

The closer the observed frequencies are to the expected frequencies, the smaller the value of the test statistic. A small value for this statistic indicates that the null hypothesis (which states that the population follows the Poisson distribution) should *not* be rejected.

Because the goodness of fit test is always right-tailed, it has a single critical value:

$$\chi^2_{\alpha, k-1-m}$$

Note that m is a parameter whose value equals 0 if the null hypothesis specifies a value of λ and 1 if the null hypothesis doesn't specify a value of λ .



TIP

Unlike hypothesis tests of the population variance, where the appropriate number of degrees of freedom is $n - 1$, with a goodness of fit test, the appropriate number of degrees of freedom is $k - 1 - m$.

When you determine the values of the test statistic and the critical value, the decision rule is to reject the null hypothesis if the test statistic exceeds the critical value; otherwise, don't reject the null hypothesis.

To test the hypothesis that the distribution of customers that enters the bank during lunch hour follows the Poisson distribution, the first step is to specify the null and alternative hypotheses:

- » H_0 : The population follows the Poisson distribution.
- » H_1 : The population doesn't follow the Poisson distribution.

Assume that the level of significance is 0.05 (5 percent).

Before you construct the table of observed and expected frequencies, you must estimate the value of λ from the sample data, because it isn't specified in the null hypothesis. In this case, λ represents the average number of bank customers per hour.

Because each possible number of bank customers is repeated many times in the sample, the average number of bank customers per hour can be computed as a weighted average (see Chapter 3). The formula is

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

Where:

\bar{X} = the sample mean

X_i = a single sample element

w_i = the *weight* associated with element X_i , which equals the number of times that the element appears in the sample

To compute the numerator of this formula, you multiply each number of customers per hour in the sample by the actual number of hours in which this number occurred. This is shown as follows:

Number of Customers per Hour	Number of Hours	Customers per Hour × Number of Hours
0	9	(0)(9) = 0
1	12	(1)(12) = 12
2	15	(2)(15) = 30
3	20	(3)(20) = 60
4	27	(4)(27) = 108
5	12	(5)(12) = 60
6	5	(6)(5) = 30
SUM		300

This results in a sum of 300. The denominator is the sum of the weights:

$$9 + 12 + 15 + 20 + 27 + 12 + 5 = 100$$

The average number of customers per hour is

$$\bar{X} = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i} = \frac{300}{100} = 3$$

You use this result as the value of lambda: $\lambda = 3$.

The next step is to compute the expected frequencies for each category. You find the probability of no customers entering the bank during the next hour when $\lambda = 3$ from the Poisson distribution with this formula:

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

where:

X = a Poisson random variable

x = number of events that occur per unit of time

λ = the average number of events that occur per unit of time

e = a constant equal to approximately 2.71828

$!$ = the “factorial” operator (introduced in Chapter 8)

The factorial operator can only be applied to positive whole numbers and zero. So $0!$ equals 1, as does $1!$, and $2!$ equals $(2)(1) = 2$; in other words, $2!$ equals itself times all smaller positive whole numbers. Based on this pattern, $3!$ equals $(3)(2)(1) = 6$, and $4!$ equals $(4)(3)(2)(1) = 24$. All remaining factorials are computed in the same way. The factorial operator may be used for several applications; one of these is to count the number of *arrangements* that may be formed from a collection of objects. For example, if three paintings are hung next to each other in the reading room of a library, the number of ways the paintings may be arranged equals $3! = 6$.

For the bank customer case, the probability of no customers entering the bank during the lunch hour is computed with the Poisson formula as follows:

$$\begin{aligned} P(X = 0) &= e^{-\lambda} \frac{\lambda^x}{x!} \\ &= e^{-3} \frac{3^0}{0!} \\ &= 0.0498 \end{aligned}$$

You do the same calculations with the probabilities for $X = 1$, $X = 2$ all the way up to $X = 6$. The probability that $X = 1$ is computed as follows:

$$\begin{aligned} P(X = 1) &= e^{-\lambda} \frac{\lambda^x}{x!} \\ &= e^{-3} \frac{3^1}{1!} \\ &= 0.1494 \end{aligned}$$

The probability that $X = 2$ is computed as follows:

$$\begin{aligned} P(X = 2) &= e^{-\lambda} \frac{\lambda^x}{x!} \\ &= e^{-3} \frac{3^2}{2!} \\ &= 0.2240 \end{aligned}$$

The probabilities for $X = 3, 4, 5$, and 6 are computed in a similar manner:

$$P(X=3) = 0.2240$$

$$P(X=4) = 0.1680$$

$$P(X=5) = 0.1008$$

$$P(X=6) = 0.0504$$

Because the sample size is 100, you multiply the probabilities by 100 to get the expected frequencies, as shown here:

$$X=0: \text{expected frequency} = 0.0498(100) = 4.98$$

$$X=1: \text{expected frequency} = 0.1494(100) = 14.94$$

$$X=2: \text{expected frequency} = 0.2240(100) = 22.40$$

$$X=3: \text{expected frequency} = 0.2240(100) = 22.40$$

$$X=4: \text{expected frequency} = 0.1680(100) = 16.80$$

$$X=5: \text{expected frequency} = 0.1008(100) = 10.08$$

$$X=6: \text{expected frequency} = 0.0504(100) = 5.04$$

Substitute these values into the test statistic formula:

$$\begin{aligned}\chi^2 &= \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j} \\&= \frac{(9 - 4.98)^2}{4.98} + \frac{(12 - 14.94)^2}{14.94} + \frac{(15 - 22.40)^2}{22.40} + \frac{(20 - 22.40)^2}{22.40} \\&\quad + \frac{(27 - 16.80)^2}{16.80} + \frac{(12 - 10.08)^2}{10.08} + \frac{(5 - 5.04)^2}{5.04} \\&= 13.08\end{aligned}$$

Then, you determine the critical value as follows:

$$\chi^2_{\alpha, k-1-m}$$

The first step is to identify the values of α , k , and m :

- » $\alpha = 0.05$ because you're using a level of significance of 0.05 (5 percent).
- » $k = 7$ because there are seven categories (the number of customers that enter the bank during lunch hour is 0, 1, 2, 3, 4, 5, or 6, for a total of 7 possibilities).
- » $m = 1$ because the null hypothesis doesn't specify a value of λ . (In other words, you computed the value of λ from the sample data.)

Therefore, $k - 1 - m = 7 - 1 - 1 = 5$.

You can find the critical value in Table 13-1 by finding the intersection of the 0.05 right-tail area column and the 5 df row:

$$\chi^2_{\alpha, k-1-m} = \chi^2_{0.05, 7-1-1} = \chi^2_{0.05, 5} = 11.070$$

The test statistic does exceed the critical value. Because this is a right-tailed test, the correct conclusion is that the null hypothesis is rejected. In other words, the number of customers entering the bank per hour does not follow the Poisson distribution.

Comparing a population to the normal distribution

Testing the hypothesis that a population follows the normal distribution is similar to testing the hypothesis that a population follows the Poisson distribution (see the previous section). The two most important differences are that you compute the expected frequencies from the normal distribution, and the definition of m is slightly different for the critical value. In this case, m is defined as follows:

- » $m = 0$ if the value of the mean (μ) and standard deviation (σ) are both specified in the null hypothesis.
- » $m = 1$ if the value of the mean or the standard deviation (but not both) is specified in the null hypothesis.
- » $m = 2$ if the value of neither the mean nor the standard deviation are specified in the null hypothesis.

As an example, suppose that a portfolio manager wants to determine whether the returns to a portfolio are normally distributed, with a mean of 5 percent and a standard deviation of 10 percent.

The observed frequencies are 22 for -15 to -5 percent returns, 29 for -5 to 5 percent returns, 37 for 5 to 15 percent returns, and 12 for 15 to 25 percent returns. The null and alternative hypotheses are

- » H_0 : The population is normally distributed with a mean of 5 percent and standard deviation of 10 percent.
- » H_1 : The population isn't normally distributed with mean of 5 percent and standard deviation of 10 percent.

Assume that the level of significance is 0.05 (5 percent).

You determine the expected frequencies from the standard normal distribution by following these steps:

1. Define X to be the return to a portfolio.

The mean return is 5 percent and the standard deviation of the return is 10 percent.

2. Assume that X is normally distributed.

To compute probabilities for X using the normal table, you must first convert it into a standard normal random variable (I show you how to do so in Chapter 9).

In this example, the returns are assumed to be normally distributed with a mean of 5 percent and a standard deviation of 10 percent. Next, you compute the probability that X is between -15 percent and -5 percent.

Because X is a normal random variable but not *standard normal*, you must convert X into the equivalent standard normal form in order to use a normal table to compute probabilities. If a calculator such as the TI-84 is used, this step is not necessary. (Recall that the standard normal distribution has a mean of 0 and a standard deviation of 1, as discussed in Chapter 9.) The appropriate formula is

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

where:

μ = the mean of X .

σ = the standard deviation of X .

By converting X into a standard normal random variable, it is now possible to compute probabilities for X , using the standard normal tables.

$$P(-15 \leq X \leq -5) = P\left(\frac{-15 - 5}{10} \leq Z \leq \frac{-5 - 5}{10}\right) = P(-2.00 \leq Z \leq -1.00)$$

The standard normal tables are set up to compute *cumulative* probabilities; in other words, the probability that Z is less than or equal to a specified value.

In this example, you're looking for the probability that Z is between -2.00 and -1.00. This can be computed from the standard normal tables by rewriting the expression in the equivalent form:

$$P(Z \leq -1.00) - P(Z \leq -2.00)$$

You can get these probabilities from the standard normal table. See Table 13–3 for a selection of probabilities associated with negative Z values.

TABLE 13-3

Selected Standard Normal Probabilities for Negative Z Values

Z	0.00	0.01	0.02	0.03
-2.0	0.0228	0.0222	0.0217	0.0212
-1.5	0.0668	0.0655	0.0643	0.0630
-1.0	0.1587	0.1562	0.1539	0.1515

You find the probability that Z is less than or equal to -1.00 at the intersection of the row for -1.0 under the Z column and the 0.00 column, which is 0.1587 . Likewise, you find the probability that Z is less than or equal to -2.00 at the intersection of the -2.0 row and the 0.00 column, which is 0.0228 .

Combining these values gives you

$$P(Z \leq -1.00) - P(Z \leq -2.00) = 0.1587 - 0.0228 = 0.1359$$

You determine the probability that X is between -5 percent and $+5$ percent as

$$P(-5 \leq X \leq 5) = P\left(\frac{-5-5}{10} \leq Z \leq \frac{5-5}{10}\right) = P(-1.00 \leq Z \leq 0.00)$$

Algebraically, this is equivalent to

$$P(Z \leq 0.00) - P(Z \leq -1.00)$$

One of the properties of the standard normal distribution is that the probability that Z is less than or equal to 0 is 0.5 because the entire area under the standard normal curve equals 1 and because the distribution is symmetrical about the mean of 0 . These statements imply the following:

$$P(Z \leq 0.00) = 0.5$$

$$P(Z \geq 0.00) = 0.5$$

Based on Table 13–3, the probability that Z is less than or equal to $-1.00 = 0.1587$. Therefore, $P(Z \leq 0.00) - P(Z \leq -1.00) = 0.5 - 0.1587 = 0.3413$.

You compute the probability that X is between $+5$ percent and $+15$ percent as

$$P(5 \leq X \leq 15) = P\left(\frac{5-5}{10} \leq Z \leq \frac{15-5}{10}\right) = P(0.00 \leq Z \leq 1.00)$$

You can rewrite this as $P(Z \leq 1.00) - P(Z \leq 0.00)$.

You can find the probability that Z is less than or equal to 1.00 in the standard normal table. Take a look at Table 13-4 to see a section of this table for positive Z values.

TABLE 13-4

Selected Standard Normal Probabilities for Positive Z Values

Z	0.00	0.01	0.02	0.03
1.0	0.8413	0.8438	0.8461	0.8485
1.5	0.9332	0.9345	0.9357	0.9370
2.0	0.9772	0.9778	0.9783	0.9788

You have already determined that the probability that Z is less than or equal to 0 equals 0.5. You can see the probability that Z is less than or equal to 1.00 by intersecting the row for 1.0 and the 0.00 column, which is 0.8413. Therefore, $P(Z \leq 1.00) - P(Z \leq 0.00) = 0.8413 - 0.5000 = 0.3413$.

You determine the probability that X is between +15 percent and +25 percent in a similar manner:

$$P(15 \leq X \leq 25) = P\left(\frac{15-5}{10} \leq Z \leq \frac{25-5}{10}\right) = P(1.00 \leq Z \leq 2.00)$$

or

$$P(Z \leq 2.00) - P(Z \leq 1.00) = 0.9772 - 0.8413 = 0.1359$$

Because the sample size equals 100, the expected frequency of each category equals the probability of each category times 100.

$$P(-15\% \leq X \leq -5\%) : 0.1359(100) = 13.59$$

$$P(-5\% \leq X \leq 5\%) : 0.3413(100) = 34.13$$

$$P(5\% \leq X \leq 15\%) : 0.3413(100) = 34.13$$

$$P(15\% \leq X \leq 25\%) : 0.1359(100) = 13.59$$

You can then combine the observed and expected returns into a single table, as Table 13-5 shows.

TABLE 13-5

Observed and Expected Frequencies

Returns	-15% to -5%	-5% to 5%	5% to 15%	15% to 25%
Observed frequency	22	29	37	12
Expected frequency	13.59	34.13	34.13	13.59

Based on this table, the test statistic is computed as follows:

$$\begin{aligned}\chi^2 &= \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j} \\ &= \frac{(22 - 13.59)^2}{13.59} + \frac{(29 - 34.13)^2}{34.13} + \frac{(37 - 34.13)^2}{34.13} + \frac{(12 - 13.59)^2}{13.59} \\ &= 6.40\end{aligned}$$

The critical value is determined as follows:

$$\chi_{\alpha, k-1-m}^2$$

The first step is to identify the values of α , k , and m .

- » $\alpha = 0.05$ because you're using a level of significance of 0.05 (5 percent).
- » $k = 4$ because there are four categories of returns: -15 percent to -5 percent, -5 percent to +5 percent, +5 percent to +15 percent, and +15 percent to +25 percent.
- » $m = 0$, because the value of the mean (μ) and standard deviation (σ) are both specified in the null hypothesis.

Therefore, $k - 1 - m = 4 - 1 - 0 = 3$.

You can find the critical value in Table 13-1 by finding the intersection of the 0.05 right-tail area column and the 3 df row:

$$\chi_{\alpha, k-1-m}^2 = \chi_{0.05, 3}^2 = 7.815$$

Because this is a right-tailed test, the test statistic must exceed the critical value to reject the null hypothesis that the population is normal with a mean of 5 percent and a standard deviation of 10 percent. Because the test statistic is 6.40 and the critical value is 7.815, you don't reject the null hypothesis. This indicates that there is insufficient evidence to reject the hypothesis that the population is normally distributed with a mean of 5 percent and a standard deviation of 10 percent.

Conducting a Goodness of Fit Test with the TI-84 Plus Calculator

You can use the Texas Instruments TI-84 Plus and Plus CE calculators to carry out a goodness of fit test with the chi-square distribution.

When conducting a goodness of fit test, you must compute the expected frequencies for the problem using the appropriate probability distribution. For example, if the goodness of fit test is used to determine if a population follows the normal distribution, you must calculate the necessary normal probabilities and use these to compute the expected frequencies for each category in the data set. Once this is done, you can store the observed frequencies and expected frequencies in two lists — for example, observed frequencies in List 1 (L1) and expected frequencies in List 2 (L2). (The steps used to enter data into lists can be found at the end of Chapter 3.)

For example, suppose that you are testing the portfolio example in this chapter; the observed number of returns in each category would be stored in L1 and the expected number of times each category of returns is expected to occur based on the normal distribution would be stored in L2.

Once you have stored the data in lists, one possible approach to computing the test statistic is to enter a formula into another list, such as List 3 (L3). To do this, follow these steps:

- 1. Navigate to the header of L3 with the arrow keys.**
- 2. Enter the following:**

$$(2^{\text{nd}} 1 - 2^{\text{nd}} 2)^2 / 2^{\text{nd}} 2$$

After entering the exponent 2, you must use the right arrow key to move one space to the right and then enter the remaining keystrokes; otherwise, the calculator will include everything following the 2 in the exponent.

This appears in the window as $(L_1 - L_2)^2 / L_2$.

Because the observed frequencies are stored in L1 and the expected frequencies are stored in L2, this is equal to $(O_j - E_j)^2 / E_j$.

- 3. Press the [ENTER] button to fill in the entire list.**

The sum of these terms is the test statistic. It can be obtained by manually adding up the terms in L3 or by following these steps:

1. Press [2nd STAT].
2. Use the right arrow key to select MATH.
3. Choose 5:sum(and then press [ENTER].
4. Type 2nd 3) (this represents L3) and press [ENTER].

This produces the output: sum(.

Using the example from earlier in this chapter of testing to see if returns to a portfolio are normal, this produces the sum 6.402866088.

Alternatively, once you have entered the observed and expected frequencies into lists, you can run the goodness of fit test as follows:

1. Press [STAT]
2. Use the right arrow key to select TESTS.
3. Choose D: χ^2 : GOF-Test (chi-square goodness of fit test) and press [ENTER].
4. You then need to provide the following information:

Observed:

Expected:

df:

Enter 2nd 1 (L1) for Observed and 2nd 2 (L2) for Expected. Recall that the degrees of freedom equals $k - 1 - m$; in the portfolio example earlier in this chapter, this equals 3.

5. After entering the degrees of freedom, choose Calculate and then press [ENTER].

This produces the following output:

$$\chi^2 = 6.402866088$$

$$P = 0.093572953$$

$$df = 3$$

$$CNTRB = \{5.204422369 \ 0.7710782303 \ 0.2413389979 \ 0.1860264901\}$$

This output shows the chi-square test statistic ($\chi^2 = 6.402866088$), followed by the p-value for the goodness of fit test. Because the p-value is well above the level of significance of 0.05, the null hypothesis fails to be rejected.

The last entry (CNTRB) shows the intermediate calculations used to produce the test statistic; the sum of these numbers is 6.402866088. (Note that because this is a very long line, you must continually use the right arrow to see all of these results.)

IN THIS CHAPTER

- » Introducing the F-distribution
- » Testing hypotheses about the equality of two population variances
- » Choosing a level of significance

Chapter **14**

Applications of the F-Distribution

This chapter introduces the *F-distribution*. The F-distribution is used for several applications in statistics, such as testing hypotheses about three or more population means, testing the quality of results in regression analysis, and testing hypotheses about two population variances. In this chapter, I introduce the key properties of the F-distribution and show you how to test hypotheses about two population variances.

Getting to Know the F-Distribution

The F-distribution is a *continuous* probability distribution, which means that it is defined for an *infinite* number of different values. (Continuous probability distributions, such as the normal distribution, are introduced in Chapter 9.) The F-distribution was named after the renowned statistician Sir Ronald Fisher (1890–1962).

The F-distribution shares one important property with the Student's t-distribution (introduced in Chapter 11): Probabilities are determined by a concept known as *degrees of freedom* (*df*). Unlike the Student's t-distribution, the F-distribution is characterized by two different types of degrees of freedom — *numerator* and *denominator* degrees of freedom.

The F-distribution has two extremely important properties:

- » It's defined only for positive values.
- » It's *not* symmetrical about its mean; instead, it's *positively skewed*.

A distribution is positively skewed if the mean is greater than the median. (The mean and the median are introduced in Chapter 3. The mean is the average value of a distribution, and the median is the midpoint; half of the values in a distribution are less than or equal to the median, and half are greater than or equal to the median.)

A good example of a positively skewed distribution is household incomes. Suppose that half of the households in a country have incomes less than or equal to \$50,000 and half have incomes greater than or equal to \$50,000; this indicates that the median household income is \$50,000. Among households with incomes below \$50,000, the smallest possible value is \$0. Among households with incomes above \$50,000, there may be incomes of several million dollars per year. This imbalance between incomes below the median and above the median causes the mean to be substantially higher than the median. Suppose for example that the mean income in this case is \$120,000. This shows that the distribution of household incomes is positively skewed.

Figure 14-1 shows a graph of the F-distribution for different combinations of numerator and denominator degrees of freedom. In each case, numerator degrees of freedom are listed first, and denominator degrees of freedom are listed second (for example, 1,5 indicates 1 numerator degree of freedom, and 5 denominator degrees of freedom). The *level of significance* in each case is 0.05.

A level of significance is used to test a *hypothesis*. (Hypothesis testing is covered in detail in Chapter 12.) A hypothesis test begins with a *null hypothesis*; this is a statement that's assumed to be true unless there is *strong* contrary evidence. There is also an *alternative hypothesis*; this is a statement that is accepted in place of the null hypothesis if there is sufficient evidence to reject the null hypothesis.

The level of significance, designated α (alpha), refers to the probability of incorrectly rejecting the null hypothesis when it is actually true. This is known as a *Type I error*. By contrast, a *Type II error* occurs when you fail to reject the null hypothesis when it's actually false. Therefore, with a level of significance of 0.05, there is a 5 percent chance of committing a Type I error.

Figure 14-1 shows that the distribution isn't defined for negative values (as you can see, no negative values appear along the horizontal axis). Additionally, as the number of degrees of freedom increases, the shape of the distribution shifts to the right. The distribution has a long right tail (more formally, it's *skewed to the right*, or *positively skewed*).

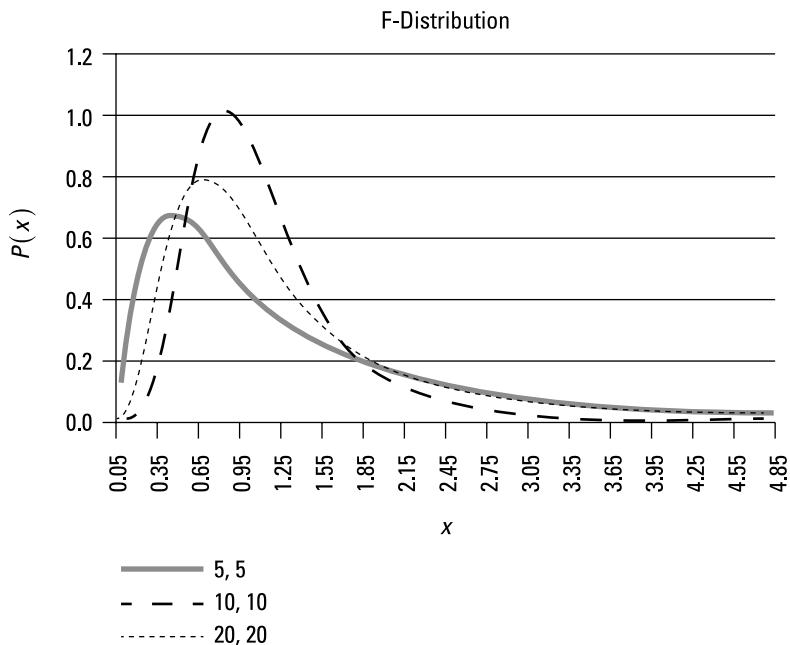


FIGURE 14-1:
The shape of the
F-distribution
varies with its
degrees of
freedom (df).

In the following sections, I go into even more detail about the F-distribution, such as the properties of the F random variable and show you how to compute the moments of the F-distribution.

Defining an F random variable

The F-distribution is defined in terms of the chi-square (χ^2) distribution (see Chapter 13 for details). The chi-square distribution is a continuous distribution that is characterized by its degrees of freedom, where the degrees of freedom are determined by the size of the samples used. Like the F-distribution, the chi-square distribution is only defined for positive values and is positively skewed.

The chi-square distribution has several different applications, including testing hypotheses about the variance of a population and testing hypotheses about the probability distribution followed by a population.

The following equation shows that an F random variable is the ratio of two *independent* chi-square random variables: χ_1^2 and χ_2^2 (where χ is the Greek letter chi, pronounced “ki”) and their respective degrees of freedom (v_1 and v_2):

$$F^{v_1, v_2} = \frac{\chi_1^2 / v_1}{\chi_2^2 / v_2}$$

F^{v_1, v_2} is a random variable that follows the F-distribution and has v_1 numerator degrees of freedom and v_2 denominator degrees of freedom (where v is the Greek letter nu, pronounced “new”).

Measuring the moments of the F-distribution

Moments are summary measures of a probability distribution and include the following:

- » The *expected value* is known as the first moment of a probability distribution and represents the mean or average value of a distribution.
- » The *variance* is the second central moment and shows how spread out or scattered the values of a distribution are around the expected value.
- » The *standard deviation* isn't a separate moment but is the square root of the variance.

For most applications, the standard deviation is more useful than the variance (because the standard deviation is measured in the same units as the expected value whereas the variance is measured in squared units). For the F-distribution, you use this formula to determine the expected value:

$$E(X) = \frac{v_2}{v_2 - 2}$$

$E(X)$ represents the expected value, and v_2 represents the denominator degrees of freedom (defined in the previous section).



The expected value formula requires the denominator degrees of freedom to be greater than 2. Otherwise, the expected value becomes negative or undefined.

REMEMBER

The expected value represents the *average* value of the F-distribution. For example, Figure 14-1 shows a graph of the F-distribution with 5 numerator degrees of freedom and 5 denominator degrees of freedom. The expected value equals:

$$\begin{aligned}E(X) &= \frac{\nu_2}{\nu_2 - 2} \\&= \frac{5}{5 - 2} \\&= \frac{5}{3} \\&= 1.67\end{aligned}$$

Figure 14-1 also shows a graph of the F-distribution with 20 numerator degrees of freedom and 20 denominator degrees of freedom. The expected value equals:

$$\begin{aligned}E(X) &= \frac{\nu_2}{\nu_2 - 2} \\&= \frac{20}{20 - 2} \\&= \frac{20}{18} \\&= 1.11\end{aligned}$$

This shows that the average value of the F-distribution with 20 numerator degrees of freedom and 20 denominator degrees of freedom is *less than* the average value of the F-distribution with 5 numerator degrees of freedom and 5 denominator degrees of freedom.

To compute the variance, you use this formula:

$$\sigma^2 = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$$



REMEMBER

The variance formula requires the denominator degrees of freedom to be greater than 4; otherwise, the variance becomes negative or undefined.

The standard deviation is the *square root* of the variance:

$$\sigma = \sqrt{\frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}}$$

The variance and the standard deviation are used as measures of how spread out the values of the F-distribution are compared with the expected value.

For example, for the F-distribution with 5 numerator degrees of freedom and 5 denominator degrees of freedom, the variance equals

$$\begin{aligned}\sigma^2 &= \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)} \\&= \frac{2(5^2)(5 + 5 - 2)}{5(5 - 2)^2(5 - 4)} \\&= \frac{2(25)(8)}{5(9)(1)} \\&= \frac{400}{45} \\&= 8.89\end{aligned}$$

The standard deviation equals the square root of 8.89, or 2.98.

For the F-distribution with 20 numerator degrees of freedom and 20 denominator degrees of freedom, the variance equals

$$\begin{aligned}\sigma^2 &= \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)} \\&= \frac{2(20^2)(20 + 20 - 2)}{20(20 - 2)^2(20 - 4)} \\&= \frac{2(400)(38)}{20(324)(16)} \\&= \frac{30,400}{103,680} \\&= 0.29\end{aligned}$$

The standard deviation equals the square root of 0.29, or 0.54.

In Figure 14-1, the F-distribution with 20 numerator degrees of freedom and 20 denominator degrees of freedom has a tail that falls off very rapidly (so that the distribution is less spread out) compared with the F-distribution with 5 numerator degrees of freedom and 5 denominator degrees of freedom; therefore, the distribution with 20 numerator and denominator degrees of freedom has a lower variance and standard deviation.

Testing Hypotheses about the Equality of Two Population Variances

Hypothesis testing for the equality of two population variances is based on the F-distribution with the assumption of independent samples.

The basic six-step process you use to test hypotheses about the equality of two population variances is the same as for testing hypotheses about a single population variance (which I explain in detail in Chapter 13). The main differences are the form of the null and alternative hypotheses and the calculation of the test statistic and critical values, which are based on the F-distribution instead of the chi-square distribution. In the following sections, I walk you through testing hypotheses for two population variances.

The null hypothesis: Equal variances

The first step in the hypothesis testing procedure is writing the *null hypothesis*, which is a statement that's assumed to be true unless strong contrary evidence exists against it.

In this case, the null hypothesis is written as follows:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

σ_1^2 is the variance of population 1, and σ_2^2 is the variance of population 2. The null hypothesis is that the two population variances are equal. This is accepted unless strong evidence indicates otherwise.

The alternative hypothesis: Unequal variances

The alternative hypothesis is a statement of what you will accept to be true if the null hypothesis is rejected. The alternative hypothesis can take one of three forms:

- » **Right-tailed test:** You use a right-tailed test if you're interested only in knowing whether the variance of population 1 is greater than the variance of population 2. In this case, the alternative hypothesis is

$$H_1 : \sigma_1^2 > \sigma_2^2$$

- » **Left-tailed test:** You use a left-tailed test if you're interested only in knowing whether the variance of population 1 is less than the variance of population 2. In this case, the alternative hypothesis is

$$H_1 : \sigma_1^2 < \sigma_2^2$$

- » **Two-tailed test:** You use a two-tailed test to determine whether the variances of population 1 and 2 are different. In this case, the alternative hypothesis is

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

The test statistic

For testing hypotheses about the equality of two population variances, the appropriate test statistic is

$$F = \frac{s_1^2}{s_2^2}$$

Here, F indicates that the test statistic follows the F-distribution, s_1^2 is the variance of the sample drawn from population 1, and s_2^2 is the variance of the sample drawn from population 2. Note that the test statistic requires that s_1^2 be greater than or equal to s_2^2 .

The critical value(s)

To test a hypothesis, you have to choose a *level of significance*. The level of significance, designated with α , refers to the probability of rejecting the null hypothesis when it's actually true (this is known as a Type I error).

To test a hypothesis about the equality of two population variances, you use the following critical values.

Right-tailed test for the F-distribution

A right-tailed test has a single critical value:

$$F_{\alpha}^{v_1, v_2}$$



TIP

v_1 is the numerator degrees of freedom of the F-distribution and equals $n_1 - 1$, where n_1 is the size of the sample drawn from population 1. v_2 is the denominator degrees of freedom of the F-distribution and equals $n_2 - 1$, where n_2 is the size of the sample drawn from population 2.

This critical value represents the threshold of the right tail of the F-distribution with v_1 and v_2 degrees of freedom; the area in the right tail is α . You can find this critical value in an F-table. Because each critical F-value requires two types of degrees of freedom, it's impossible to show both degrees of freedom and the level of significance together in the same table. Instead, you must dedicate an entire table to a single value of the level of significance. (You can see an excerpt of the F-table for a value of α equal to 0.05 in Table 14-1.)

For example, suppose that you conduct a right-tail test with a level of significance of 0.05 (5 percent). You draw a sample size of 5 from the first population and a sample size of 4 from the second population.

You compute the numerator degrees of freedom by subtracting 1 from the size of the sample drawn from population 1:

$$v_1 = n_1 - 1 = 5 - 1 = 4$$

You find the denominator degrees of freedom by subtracting 1 from the size of the sample drawn from population 2:

$$v_2 = n_2 - 1 = 4 - 1 = 3$$

You can find the appropriate critical value in Table 14-1.

TABLE 14-1 A Section of the F-Table with $\alpha = 0.05$

$v_2 \setminus v_1$	3	4	5	6	7	8	9
3	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	3.86	3.63	3.48	3.37	3.29	3.23	3.18

The top row represents the numerator degrees of freedom (v_1). The first column represents the denominator degrees of freedom (v_2). In this example, you're looking for a right-tail area of 5 percent with $v_1 = n_1 - 1 = 5 - 1$, which is 4 numerator degrees of freedom, and $v_2 = n_2 - 1 = 4 - 1$, which is 3 denominator degrees of freedom.

You find this critical value at the intersection of the 4 column and the row labeled 3 under the v_2/v_1 heading; it equals 9.12.

Left-tailed test for the F-distribution

A left-tailed test also has a single critical value, represented as

$$F_{\alpha}^{v_1, v_2}$$

This is a very unusual result. The critical value is the same for a right-tailed or a left-tailed test because the F-distribution is undefined for negative values. Also, the test statistic is set up with the larger sample variance in the numerator. The

null hypothesis is rejected when the ratio of the sample variances is substantially greater than 1. The test statistic can't be negative.

Two-tailed test for the F-distribution

A two-tailed test has a *single* critical value:

$$F_{\alpha/2}^{v_1, v_2}$$

The decision about the equality of two population variances

You make the decision whether to reject the null hypothesis by looking at the relationship between the test statistic and the critical value(s). Here, I break down the results of the three alternative hypothesis tests:

- » **Right-tailed test:** If the test statistic is greater than the critical value $F_{\alpha}^{v_1, v_2}$, you reject the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ in favor of the alternative hypothesis $H_1 : \sigma_1^2 > \sigma_2^2$, otherwise, you don't reject the null hypothesis.
- » **Left-tailed test:** If the test statistic is greater than the critical value $F_{\alpha}^{v_1, v_2}$, you reject the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ in favor of the alternative hypothesis $H_1 : \sigma_1^2 < \sigma_2^2$, otherwise, you don't reject the null hypothesis.
- » **Two-tailed test:** If the test statistic is greater than the critical value $F_{\alpha/2}^{v_1, v_2}$, you reject the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ in favor of the alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$, otherwise, you don't reject the null hypothesis.

As an example, suppose that an investor wants to determine whether two portfolios have the same volatility (that is, standard deviation). The investor takes a sample of ten stocks from each portfolio. The sample standard deviation of portfolio 1 is 26 percent, and the sample standard deviation of portfolio 2 is 24 percent.

The null hypothesis is $H_0 : \sigma_1^2 = \sigma_2^2$, and the alternative hypothesis is $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Assume that the level of significance is $\alpha = 0.05$ (5 percent). The test statistic is

$$F = \frac{s_1^2}{s_2^2}$$

with s_1^2 greater than or equal to s_2^2 .

Plugging in the numbers, you get the following result:

$$\begin{aligned} F &= \frac{s_1^2}{s_2^2} \\ &= \frac{(0.26)^2}{(0.24)^2} \\ &= 1.174 \end{aligned}$$

Note that the percentages were converted into decimals before computing the test statistic.

Because this is a two-tailed test with a 5 percent level of significance, with both samples having size 10, the numerator and denominator degrees of freedom both equal 9. The critical value is $F_{(0.025)}^{9,9}$ (that is, 4.03), as you find from the F-table with $\alpha = 0.025$ (see Table 14-2).

TABLE 14-2

A Section of the F-Table with $\alpha = 0.025$.

$v_2 \setminus v_1$	7	8	9	10
7	4.99	4.90	4.82	4.76
8	4.53	4.43	4.36	4.30
9	4.20	4.10	4.03	3.96
10	3.95	3.85	3.78	3.72

Because the test statistic is 1.174, which is well below the critical value of 4.03, you don't reject the null hypothesis. The investor concludes that there is insufficient evidence to show that the volatilities of the two portfolios are different.

Testing Hypotheses about Two Population Variances with the TI-84 Plus Calculator

You can use the Texas Instruments TI-84 Plus and Plus CE calculators to carry out the F-Test for testing the equality of two population variances. To do this, follow these steps:

1. Press the [STAT] button.
2. Use the right arrow key to select TESTS.
3. Choose E: 2-SampFTest and then press [ENTER].

The first row (Inpt) offers the option of choosing the Data menu or the Stats menu for inputting data. Use the Data menu when the original data is entered into a list (in this case, two lists). Use the Stats menu when only summary statistics, such as the sample mean and sample size, are available. (The steps needed to enter data into lists are covered at the end of Chapter 3.)

When using the Data menu, the following information must be provided:

- » List 1 and List 2: The lists containing the data for this problem.
- » Freq1 and Freq2: These are set equal to 1 unless the data in a list will be repeated; if the data are repeated, Freq is set equal to the number of times that the data are repeated.
- » $\sigma_1: \neq \sigma_2 < \sigma_2 > \sigma_2$: This is where the alternative hypothesis is chosen. For a two-tailed test, choose $\neq \sigma_2$; for a left-tailed test, choose $< \sigma_2$; and for a right-tailed test, choose $> \sigma_2$. Move to the correct choice with the arrow key and then press the [ENTER] key.
- » Calculate: Used to compute the final results.

When using the Stats menu, the following information must be provided:

- » Sx1: The standard deviation of sample 1.
- » n1: The size of sample 1.
- » Sx2: The standard deviation of sample 2.
- » n2: The size of sample 2.
- » $\sigma_1: \neq \sigma_2 < \sigma_2 > \sigma_2$: This is where the alternative hypothesis is chosen. For a two-tailed test, choose $\neq \sigma_2$; for a left-tailed test, choose $< \sigma_2$; and for a right-tailed test, choose $> \sigma_2$. Move to the correct choice with the arrow key and then press the [ENTER] key.
- » Calculate: Used to compute the final results.

Using the example from earlier in this chapter of an investor who wants to compare the volatilities of two stock portfolios to see if they are equal, the appropriate steps are:

1. Press the [STAT] button.
2. Use the right arrow key to select TESTS.
3. Choose E: 2-SampFTest and then press [ENTER].

4. Choose Stats mode and enter the following information:

Sx1: 0.26

n1: 10

Sx2: 0.24

n2: 10

$\sigma_1 \neq \sigma_2$

5. Choose Calculate and then press [ENTER].

This produces the following output:

$\sigma_1 \neq \sigma_2$

F = 1.173611111

p = 0.8154182316

Sx1 = 0.26

Sx2 = 0.24

n1 = 10

n2 = 10

Be sure that Sx1 is the larger sample standard deviation, and Sx2 is the smaller sample standard deviation.

This output shows that the F-statistic is 1.173611111. The p-value is 0.8154182316. Because this is well above the level of significance of 0.05, the null hypothesis that the population variances are equal is not rejected.

More Advanced Techniques: Regression Analysis and Spreadsheet Modeling

IN THIS PART . . .

Use the powerful technique of regression analysis to estimate the relationship between two variables, and discover how to test the validity of the results.

Find out how to implement functions in Microsoft Excel and perform complex statistical analyses with the Analysis ToolPak.

Discover Microsoft Excel's key statistical analysis functions including measures of central tendency, measures of dispersion, measures of association, discrete and continuous probability distributions, confidence intervals, and regression analysis.

IN THIS CHAPTER

- » Understanding the assumptions underlying regression analysis
- » Implementing the simple regression model
- » Interpreting the regression results

Chapter 15

Simple Regression Analysis

Regression analysis is a statistical methodology that helps you develop a model of the response of a dependent variable (Y) to a change in an independent variable (X). The two types of regression analysis are *simple regression analysis* (which I discuss in this chapter) and *multiple regression*. Simple regression analysis is used for cases where there is only one independent variable while multiple regression is used for cases with two or more independent variables.

For example, suppose that a researcher is interested in analyzing the relationship between the annual returns to the Standard & Poor's 500 (S&P 500) and the annual returns to Apple stock.



REMEMBER

The Standard and Poor's 500 (S&P 500) is a broad-based stock market index; it contains the 500 largest U.S. stocks, based on *market capitalization*. (The market capitalization of a stock equals the market price of the stock times the number of outstanding shares.) The returns to the S&P 500 are often used to represent the performance of the U.S. stock market.

The researcher assumes that the returns to Apple stock are at least partially explained by the returns to the S&P 500 because the S&P reflects overall activity in the economy. In other words, the researcher assumes that the returns to Apple stock depend on the returns to the S&P 500.

To analyze this relationship with simple regression analysis, you treat the returns to Apple stock as a dependent variable (Y) and the returns to the S&P 500 as an independent variable (X). Regression analysis makes it possible to determine *how much* the returns on Apple stock respond to changes in the returns to the S&P 500. (In other words, how strong is the relationship between Apple stock and the S&P 500.)

This chapter introduces the basic regression analysis framework, including the underlying assumptions and the formulas you need to estimate the relationships between different variables. I also cover techniques for testing the validity of the results in great detail.

The Fundamental Assumption: Variables Have a Linear Relationship

Simple regression analysis is based on the assumption that a linear relationship exists between X and Y . Intuitively, if two variables have a linear relationship between them, a graph of the two variables is a straight line. (For a more formal discussion of linear relationships, see the following section “Defining a linear relationship.”)

For example, suppose that an equity analyst at a prestigious investment bank wants to determine the relationship between a corporation’s sales and profits to help better estimate the proper value of the corporation’s stock. The analyst has reason to believe that the relationship between sales and profits is linear. Further, the analyst assumes that profits are the dependent variable in this relationship, while sales are the independent variable. Specifically, the analyst believes that each \$1,000 increase in sales triggers an increase in profits by \$200, while each \$1,000 decrease in sales has the opposite effect.

The analyst may use regression analysis to determine the actual relationship between these variables by looking at the corporation’s sales and profits over the past several years. The regression results show whether this relationship is valid. In addition to sales, the corporation’s profits may change as a result of changes in other factors. Further, it may be that there is no relationship between sales and profits. The regression line can be used to show how much on average profits change as a result of a \$1,000 change in sales.

In the following sections, I dig deeper into the linear relationship between the dependent and independent variables and show you how to represent this relationship graphically.

Defining a linear relationship

In terms of geometry, you can graph a linear relationship with a straight line. Algebraically, the general expression for a linear relationship is

$$Y = mX + b$$

X is the independent variable, Y is the dependent variable whose value is determined by the value of X, m is the slope coefficient (how much Y changes in response to a one-unit change in X), and b is the intercept (the value of Y if X equals 0).



You calculate the slope of a line (m) with this formula:

$$m = \frac{\Delta Y}{\Delta X}$$

Here, ΔY (“delta Y”) represents the change in Y, and ΔX (“delta X”) represents the change in X.

Think of the slope as a measure of how much Y changes due to a given one-unit change in X, or how *sensitive* the value of Y is to changes in X. A linear relationship is one in which the slope is a *constant*.

You see a linear relationship graphed as a straight line, with the dependent variable (Y) on the vertical axis and the independent variable (X) on the horizontal axis. See Figure 15-1 for the relationship between X and Y in the equation $Y = 2X + 3$.

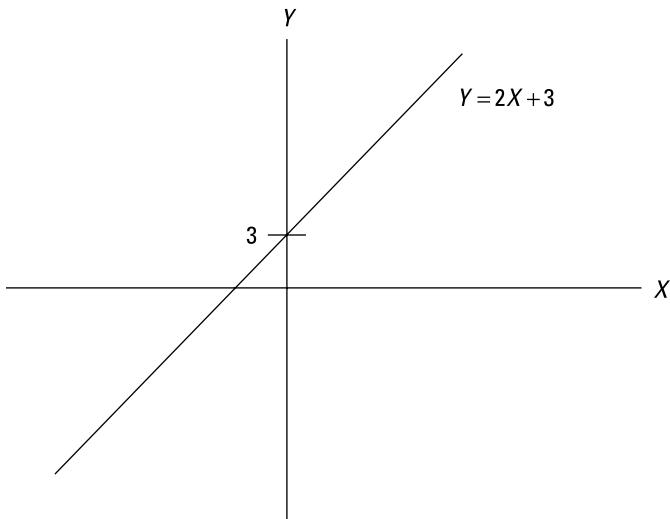


FIGURE 15-1:
Graph of a linear
relationship:
 $Y = 2X + 3$.

The equation of the line, $Y = 2X + 3$, tells you two important things:

- » The *slope* of the line is 2 (this is the constant that's multiplied by X), which shows that
 - For each increase in X by 1, Y increases by 2.
 - For each decrease in X by 1, Y decreases by 2.
- » The *intercept* of the line is 3, so if $X = 0$, the value of Y is 3. (In Figure 15-1, you see that 3 is the point where the line crosses the Y axis.)

Using scatter plots to identify linear relationships

A *scatter plot* is a special type of graph designed to show the relationship between two variables. (See Chapter 5 for an introduction to scatter plots.)

With regression analysis, you can use a scatter plot to visually inspect the data to see whether X and Y are linearly related. The following are some examples. Figure 15-2 shows a scatter plot for two variables that have a *nonlinear* relationship between them.

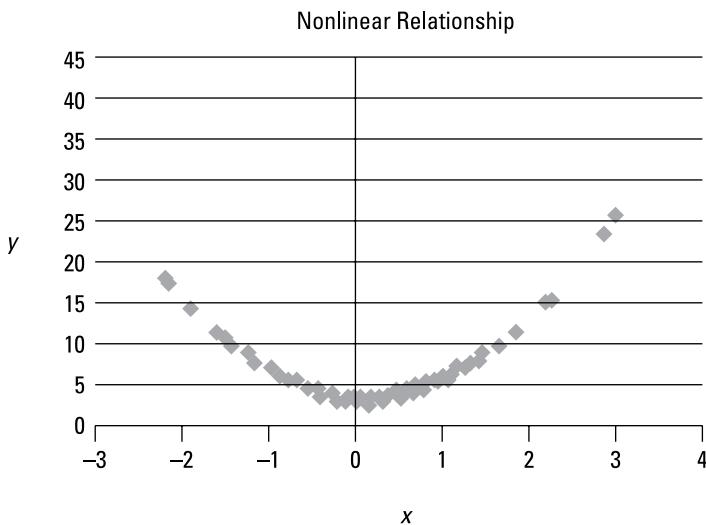


FIGURE 15-2:
Scatter plot
of a nonlinear
relationship.

Each point on the graph represents a single (X , Y) pair. Because the graph isn't a straight line, the relationship between X and Y is nonlinear. Notice that starting with negative values of X , as X increases, Y at first decreases; then as X continues to increase, Y increases. The graph clearly shows that the slope is continually changing; it isn't a constant. With a linear relationship, the slope never changes.

In this example, one of the fundamental assumptions of simple regression analysis is violated, and you need another approach to estimate the relationship between X and Y . One possibility is to transform the variables; for example, you can run a simple regression between $\ln(X)$ and $\ln(Y)$. (" \ln " stands for the natural logarithm.) This often helps eliminate nonlinearities in the relationship between X and Y . Another possibility is to use a more advanced type of regression analysis, which can incorporate nonlinear relationships.

Figure 15-3 shows a scatter plot for two variables that have a close positive linear relationship between them. The correlation between X and Y equals 0.9. (See Chapter 5 for an overview on correlation.)

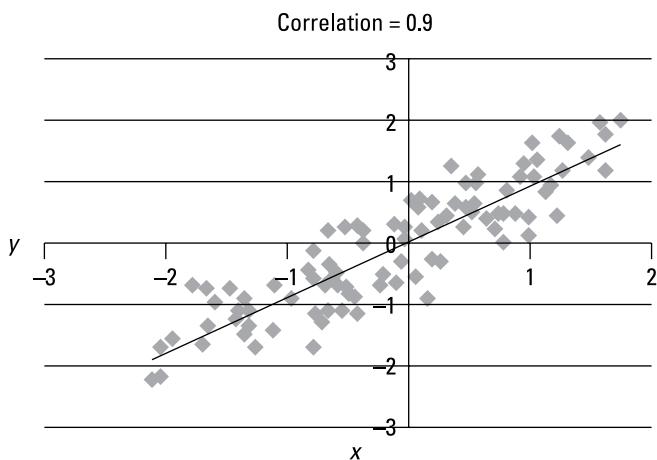


FIGURE 15-3:
Scatter plot
of a close
positive linear
relationship.

Figure 15-3 shows that X and Y typically rise above their means or fall below their means at the same time. The straight line is a *trend line*, designed to come as close as possible to all the data points. The trend line has a positive slope, which shows a positive relationship between X and Y . The points in the graph are tightly clustered about the trend line due to the strength of the relationship between X and Y . (Note: The slope of the line is *not* 0.9; 0.9 is the correlation between X and Y .)

Figure 15–4 shows a scatter plot for two variables that have a loose positive linear relationship between them; the correlation between X and Y equals 0.2. It shows a weaker connection between X and Y. Note that the points on the graph are more scattered about the trend line than in Figure 15–3, due to the weaker relationship between X and Y.

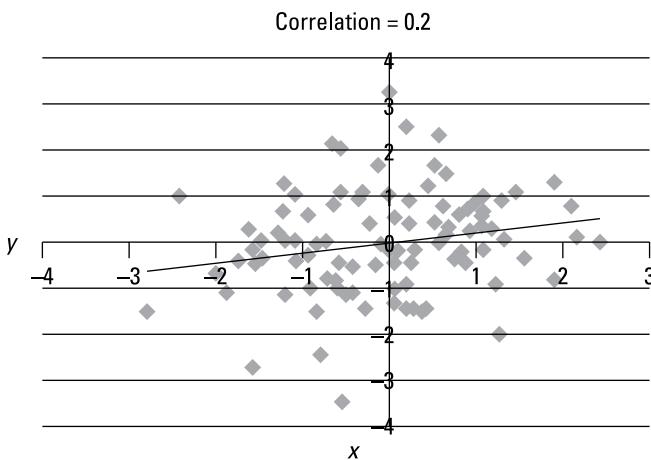


FIGURE 15-4:
Scatter plot of a
loose positive
linear
relationship.

Figure 15–5 is a scatter plot for two variables that have a close negative linear relationship between them; the correlation between X and Y equals -0.9 . It shows a very strong tendency for X and Y to rise above or fall below their means at opposite times. The trend line has a negative slope, which shows a negative relationship between X and Y. The points in the graph are tightly clustered about the trend line.

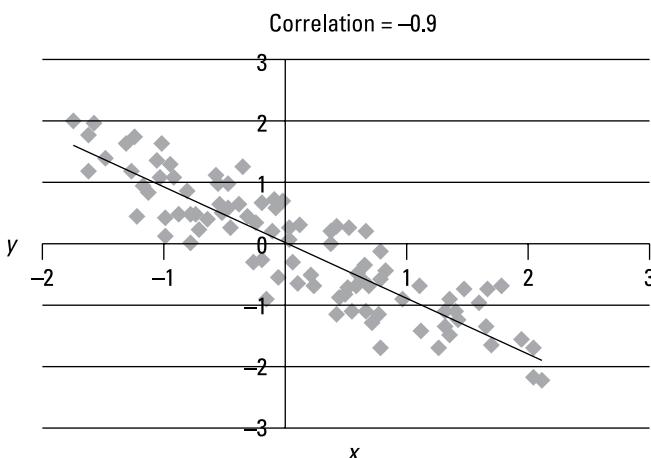


FIGURE 15-5:
Scatter plot
of a close
negative linear
relationship.

Figure 15–6 is a scatter plot for two variables that have a loose negative linear relationship between them. The correlation between X and Y equals -0.2 . It shows that there is a slight tendency for X and Y to rise above or fall below their mean at opposite times. Note that the points on the graph are more scattered about the trend line than in Figure 15–5 due to the looser relationship between X and Y.

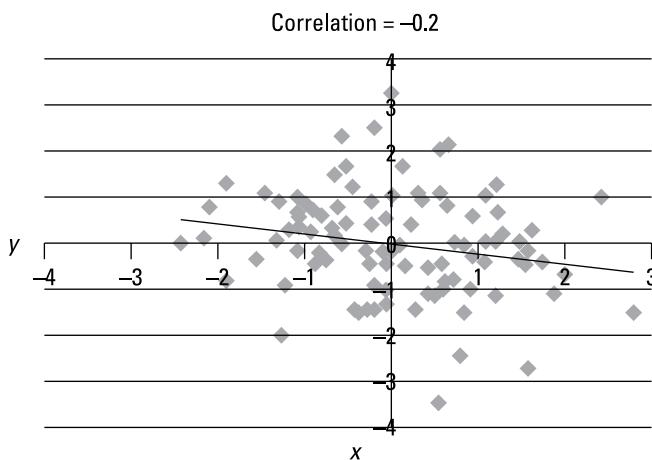


FIGURE 15-6:
Scatter plot
of a loose
negative linear
relationship.

Defining the Population Regression Equation

With regression analysis, you typically draw a sample of data from a population to estimate the relationship between X and Y. The equation that best explains the population data is known as the *population regression equation*, or *population regression line*:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



TECHNICAL
STUFF

The symbol β is the Greek letter “beta,” and the symbol ε is “epsilon.” β_0 and β_1 are known as *coefficients* of the regression line. β_1 is the slope coefficient and β_0 is the intercept coefficient (or simply the intercept). A coefficient is a constant that is multiplied by a variable.

Based on the assumption that the relationship between X and Y is linear, the regression line is designed to capture this relationship as closely as possible. Other key terms in the equation include the following:

- » i = an index used to identify the members of the population.
- » Y_i = a single value of Y , indexed by i , in a population of size n , with the values of Y expressed as $Y_1, Y_2, Y_3, \dots, Y_n$
- » X_i = a single value of X , indexed by i , in a population of size n , with the values of X expressed as $X_1, X_2, X_3, \dots, X_n$
- » ε_i = an “error term,” indexed by i ; each observation in the population (X_i, Y_i) has an error term associated with it.

Using the example of the equity analyst from the earlier section, “The Fundamental Assumption: Variables Have a Linear Relationship,” suppose that the analyst is studying the corporation’s sales and profits during the years 2013 to 2023. X_1 represents sales in 2013, and Y_1 represents profits in 2013. X_2 represents sales in 2014, and Y_2 represents profits in 2014. The analyst continues through 2023, where X_{10} is 2023 sales, and Y_{10} is 2023 profits. Each (X_i, Y_i) pair is a single observation chosen from the population.

The population regression equation has a slope and an intercept and one other term that you don’t normally find in the equation for a straight line — the *error term*. The error term is included because the population regression equation doesn’t perfectly capture the relationship between X and Y. For example, suppose that in the population regression line, $\beta_0 = 10$ (\$10 million) and $\beta_1 = 2$ (\$2 million). Assume that actual year 2013 sales (X_1) were 100 (\$100 million). The population regression line indicates that profits in 2013 should be

$$\begin{aligned}Y_1 &= \beta_0 + \beta_1 X_1 \\Y_1 &= 10 + (2)(100) \\&= \$210 \text{ million}\end{aligned}$$

Suppose that 2013 profits were actually \$200 million. The population regression line overstates actual 2013 sales by \$10 million. As a result, you compute the error term for 2013 (ε_1) as follows:

$$\begin{aligned}Y_1 &= 10 + 2X_1 + \varepsilon_1 \\ \varepsilon_1 &= Y_1 - 10 - 2X_1 \\ \varepsilon_1 &= 200 - 10 - 2(100) \\ \varepsilon_1 &= 200 - 10 - 200 \\ \varepsilon_1 &= -10\end{aligned}$$

Estimating the Population Regression Equation

In most situations, estimating the population regression line with the entire population is impractical because collecting the amount of required data can be expensive and time-consuming. Instead, you draw a sample from the underlying population that reflects the underlying population as closely as possible. You use the sample data to construct a *sample regression equation*, or *sample regression line*, which you then use as an estimate of the actual population regression equation. (Sampling techniques and sampling distributions are discussed in Chapter 10.)

The sample regression equation is expressed as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Here, \hat{Y}_i is the estimated value of Y_i , associated with X_i , $\hat{\beta}_0$ is the estimated value of β_0 , and $\hat{\beta}_1$ is the estimated value of β_1 .



TECHNICAL STUFF



TIP

Note that there is no estimated error term in this equation because the estimated value of Y_i is actually the average value of a probability distribution; thus, there is no error term associated with it.

The symbol \wedge often indicates an *estimated value*. The proper name for this punctuation mark is *caret*. Often, it's informally called a "hat." For example, you pronounce $\hat{\beta}_0$ as "beta zero hat."

You determine these estimated values for $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing the sum of the squared differences between the actually observed Y values contained in the sample data and those that have been *predicted* by the sample regression equation, as shown in the following equation:

$$\min \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Note: In this formula, *min* stands for "minimize" and tells you to choose values of $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the predicted values of Y are as close as possible to the actual values of Y. Think of each term

$$(Y_i - \hat{Y}_i)$$

as a potential mistake or error by the regression line. If this term is *positive*, the regression line has *underestimated* the true value of Y_i . If this term is *negative*, the regression line has *overestimated* the true value of Y_i . If this term equals zero, the regression line has correctly estimated the true value of Y_i .

The objective of regression analysis is to find the equation that minimizes the sum of these errors.



TECHNICAL STUFF

Note that the value being minimized is actually the sum of the *squared* values of $(Y_i - \hat{Y}_i)$. This is because the sum of these terms always equals zero if they are not squared.

The difference between the actual value of Y_i and the predicted value of \hat{Y}_i is known as a *residual* — an estimate of the corresponding error term in the population regression equation — and is expressed as follows:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

$\hat{\varepsilon}_i$ represents the residual associated with a single observation from the population (X_i, Y_i) .

As an example, suppose the quality control manager for a manufacturing company is interested in seeing the relationship between annual costs of production and total output for a specific product. The manager estimates a regression equation based on production data for the years 2016 to 2023. In this case, X_i represents the quantity produced during a given year, and Y_i represents total costs during the same year. X represents the quantity produced and Y represents the total costs because costs depend on output, not the other way around.

The manager assigns indexes to the years in the sample as follows: 2016 = Year 1, 2017 = Year 2, 2018 = Year 3, and so forth.

Based on the production data taken from the years 2016 to 2023, the estimated regression equation is

$$\hat{Y}_i = 3 - 1.5X_i$$

The diagram in Figure 15–7 shows the relationship between the actual value of Y, the predicted value of Y, the mean of Y, and the residual for Year 1 (2016). The variables in this diagram are:

X_1 is total output during Year 1.

Y_1 is total cost during Year 1.

\hat{Y}_1 is the estimated total cost during Year 1.

\bar{Y} is known as “Y bar” and is the average value of Y during the sample period.

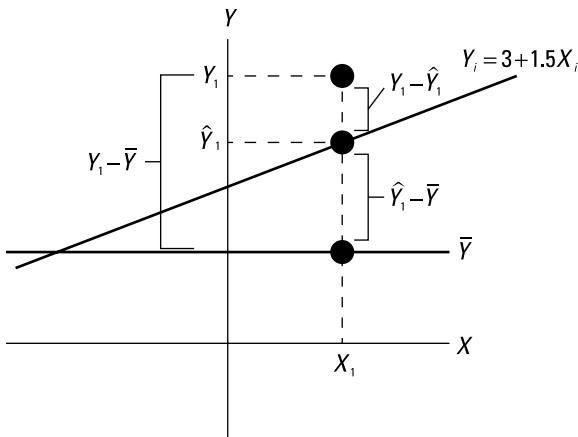


FIGURE 15-7:
Predicted value
of Y versus actual
value of Y .

Notice that the actual value of Y_1 is greater than the value estimated by the regression line. Both values are greater than the average or mean value of Y . (This information is used to construct a measure that explains how well the regression line matches the sample data in the later section “Computing the coefficient of determination.”)

For each year’s production data from 2016 to 2023,

- » $Y_i - \hat{Y}_i$ is the difference between the actual and estimated total cost in Year i .
- » $\hat{Y}_i - \bar{Y}$ is the difference between the estimated total cost in Year i and the average total cost during the sample period.
- » $Y_i - \bar{Y}$ is the difference between the total cost in Year i and the average total cost during the sample period.

Note that $Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$.

- » $(Y_i - \hat{Y}_i)$ is the size of the incorrect prediction (error) by the regression equation. It equals the difference between the actual value of Y and the value predicted by the regression equation.
- » $(\hat{Y}_i - \bar{Y})$ shows the benefit of using this regression equation to predict the value of Y_i instead of using an alternative, such as simply assuming that each value of Y_i equals \bar{Y} .

You estimate the regression equation with formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of the squared residuals:

$$\min \hat{\varepsilon}_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

The resulting equations for the slope of the estimated regression equation is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

And the equation for the intercept is

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$



TECHNICAL STUFF

These formulas are known as *ordinary least squares* (OLS) estimators. OLS is a methodology for estimating regression coefficients. Some of the more advanced versions include generalized least squares (GLS) and weighted least squares (WLS).



REMEMBER

\bar{X} is the mean or average value of X ; \bar{Y} is the mean or average value of Y . (The mean is covered in Chapter 3.)

As an example, suppose that X represents the monthly number of hours of studying by college students, and Y represents their corresponding grade point averages (GPAs). To conduct this study, you choose a sample of eight students and list their study hours and GPAs like so:

Y (GPA)	X (Monthly Hours of Studying)
3.5	16
3.2	14
3.0	12
2.6	11
2.9	12
3.3	15
2.7	13
2.8	11

Then you can create a scatter plot like Figure 15–8 to represent the data. Figure 15–8 shows that the relationship between these two variables is approximately linear. As a result, you can estimate the relationship between these two variables with simple regression analysis.

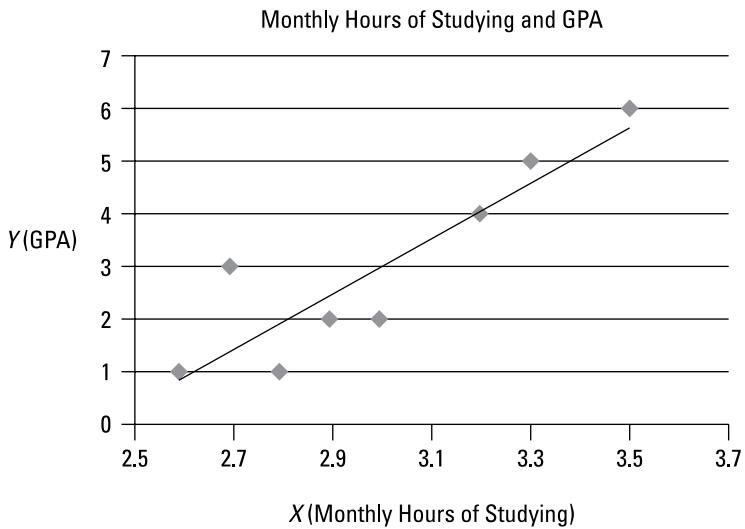


FIGURE 15-8:
Scatter plot of
monthly study
hours and GPA.

You compute the coefficients of the sample regression equation by following these steps:

1. Find the sample mean of X and Y :

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^n X_i}{n} \\ &= \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8}{n} \\ &= \frac{16 + 14 + 12 + 11 + 12 + 15 + 13 + 11}{8} \\ &= 13.0\end{aligned}$$

In this case, you add up the monthly hours of studying for the eight students in the sample and then divide by 8. This gives a sample mean of 13.0 hours for these students.

$$\begin{aligned}\bar{Y} &= \frac{\sum_{i=1}^n Y_i}{n} \\ &= \frac{Y_1 + Y_2 + Y_3 + Y_4 + Y_5 + Y_6 + Y_7 + Y_8}{n} \\ &= \frac{3.5 + 3.2 + 3.0 + 2.6 + 2.9 + 3.3 + 2.7 + 2.8}{8} \\ &= 3.0\end{aligned}$$

In this case, you add up the GPAs for the eight students in the sample and then divide by 8. This gives a sample mean of 3.0 for these students.

The results of the remaining steps are summarized in Table 15-1.

TABLE 15-1**Computing the Regression Slope and Intercept**

Y (GPA)	X (Monthly Hours of Studying)	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
3.5	16	3	9	0.5	1.5
3.2	14	1	1	0.2	0.2
3.0	12	-1	1	0.0	0.0
2.6	11	-2	4	-0.4	0.8
2.9	12	-1	1	-0.1	0.1
3.3	15	2	4	0.3	0.6
2.7	13	0	0	-0.3	0.0
2.8	11	-2	4	-0.2	0.4
Sum			24		3.6

2. To compute $(X_i - \bar{X})$, you subtract the mean of X from each value of X.
 3. To find the value of $(X_i - \bar{X})^2$, you square the value of $(X_i - \bar{X})$ for each result you found in the previous step.
 4. You calculate $(Y_i - \bar{Y})$ by subtracting the mean of Y from each value of Y.
 5. You compute $(X_i - \bar{X})(Y_i - \bar{Y})$ by multiplying the results in Steps 2 and 4.
- The sum in the $(X_i - \bar{X})^2$ column shows that $\sum_{i=1}^n (X_i - \bar{X})^2 = 24$. The sum in the $(X_i - \bar{X})(Y_i - \bar{Y})$ column shows that $\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = 3.6$.
6. Based on these results, you compute the values of the regression coefficients as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{3.6}{24} = 0.15$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 3 - (0.15)(13) = 1.05$$

7. You write the estimated (sample) regression equation as

$$\hat{Y}_i = 1.05 + 0.15X_i$$

The slope of this equation shows that for students who study between 11 and 16 hours per month (the range of values in the sample) each additional monthly hour of studying is associated with an average increase of 0.15 points in a student's GPA. The intercept may be interpreted to mean that a student who doesn't study at all (so that $X = 0$) will have a GPA of 1.05.

You can use the sample regression equation to estimate the GPA that results from a specified number of hours of studying. For example, if a student studies for 15 hours a month, the sample regression equation predicts a GPA of $\hat{Y}_i = 1.05 + 0.15 X_i = 1.05 + (0.15)(15) = 3.30$.



WARNING

When using a regression line to predict the value of Y for a given value of X, don't use any values of X that aren't contained in the sample data. In this example, the regression line is based on values of X between 11 and 16; the results of using a value of X outside of this range is subject to a great deal of uncertainty.

Testing the Estimated Regression Equation

After you estimate the regression line (see the earlier section "Estimating the Population Regression Equation"), you can do several tests to check the validity of the results. It may be the case that there is no real relationship between the dependent and independent variables; simple regression generates results even if this is the case. It is, therefore, important to subject the regression results to some key tests that enable you to determine if the results are reliable.

In the following sections, I introduce a statistic that is designed to check whether a regression equation makes sense. This is known as the *coefficient of determination*, also known as R^2 (R squared). This is used as a measure of how well the regression equation actually describes the relationship between the dependent variable (Y) and the independent variable (X).

The next technique that may be used to check regression results is a hypothesis test of the coefficients of the regression equation. The steps used to carry out this hypothesis test are similar to those found in Chapter 12, where hypothesis testing is introduced for the first time. This hypothesis test is sometimes known as the "t-test" because the test statistic and critical values are derived from the Student's t-distribution (discussed in Chapter 11). In this case, if the null hypothesis fails to be rejected, this calls into question the validity of the regression results.

Using the coefficient of determination (R^2)

The coefficient of determination, also known as R^2 , is a statistical measure that shows the percentage of *variation* explained by the estimated regression line. *Variation* refers to the sum of the squared differences between the values of Y and the mean value of Y, expressed mathematically as

$$\sum_{i=1}^n (Y_i - \bar{Y})^2$$

R^2 always takes on a value between 0 and 1. The closer R^2 is to 1, the better the estimated regression equation fits or explains the relationship between X and Y.

The expression $\sum_{i=1}^n (Y_i - \bar{Y})^2$ is also known as the *total sum of squares* (TSS). This sum can be divided into the following two categories:

- » **Explained sum of squares (ESS):** Also known as the *explained variation*, the ESS is the portion of total variation that measures how well the regression equation explains the relationship between X and Y.

You compute the ESS with the formula

$$ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

- » **Residual sum of squares (RSS):** This expression is also known as *unexplained variation* and is the portion of total variation that measures discrepancies (errors) between the actual values of Y and those estimated by the regression equation.

You compute the RSS with this formula:

$$RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

The smaller the value of RSS relative to ESS, the better the regression line fits or explains the relationship between the dependent and independent variable.

- » **Total sum of squares (TSS):** The sum of RSS and ESS equals TSS.

You compute TSS with this formula:

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

R^2 (the coefficient of determination) is the ratio of explained sum of squares (ESS) to total sum of squares (TSS):

$$R^2 = \frac{ESS}{TSS}$$

You can also use this formula:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Based on the definition of R^2 , its value can never be negative. Also, R^2 can't be greater than 1, so $0 \leq R^2 \leq 1$.



TECHNICAL STUFF

With simple regression analysis, R^2 equals the square of the correlation between X and Y.

Computing the coefficient of determination

The coefficient of determination is used as a measure of how well a regression line explains the relationship between a dependent variable (Y) and an independent variable (X). The closer the coefficient of determination is to 1, the more closely the regression line fits the sample data.

The coefficient of determination is computed from the sums of squares determined in the earlier section “Using the coefficient of determination (R^2).” These calculations are summarized in Table 15–2.

TABLE 15-2

Computing the Coefficient of Determination (R^2)

Y_i	X_i	$\hat{Y}_i = 1.05 + 0.15X_i$	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$\sum_{i=1}^n (Y_i - \bar{Y})^2$
3.5	16	3.45	0.0025	0.2025	0.25
3.2	14	3.15	0.0025	0.0225	0.04
3.0	12	2.85	0.0225	0.0225	0.00
2.6	11	2.70	0.0100	0.0900	0.16
2.9	12	2.85	0.0025	0.0225	0.01
3.3	15	3.30	0.0000	0.0900	0.09
2.7	13	3.00	0.0900	0.0000	0.09
2.8	11	2.70	0.0100	0.0900	0.04
Sum			0.1400	0.5400	0.68

To compute ESS, you subtract the mean value of Y from each of the estimated values of Y; each term is squared and then added together:

$$ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = 0.54$$

To compute RSS, you subtract the estimated value of Y from each of the actual values of Y; each term is squared and then added together:

$$RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 0.14$$

To compute TSS, you subtract the mean value of Y from each of the actual values of Y; each term is squared and then added together:

$$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2 = 0.68$$

Alternatively, you can simply add ESS and RSS to obtain TSS:

$$TSS = ESS + RSS = 0.54 + 0.14 = 0.68$$

The coefficient of determination (R^2) is the ratio of ESS to TSS:

$$R^2 = \frac{ESS}{TSS} = \frac{0.54}{0.68} = 0.7941$$

This shows that 79.41 percent of the variation in Y is explained by variation in X. Because the coefficient of determination can't exceed 100 percent, a value of 79.41 indicates that the regression line closely matches the actual sample data.

The t-test

Another important test of the results of regression analysis is to determine whether the slope coefficient (β_1) is different from 0. If the slope coefficient is close to 0, X provides little or no explanatory power for the value of Y. In such a case, you should replace X with another independent variable in the regression equation.

To determine whether β_1 is different from 0, you need to conduct a *hypothesis test*. (You find more about hypothesis testing in Chapter 12.) The name of the hypothesis test used in this case is the *t-test*, because the test statistic and critical values are based on the Student's t-distribution (covered in Chapter 11). You use this test to determine whether the slope coefficient (β_1) of the estimated regression equation is significantly different from 0. If $\beta_1 = 0$, X doesn't explain the value of Y, and the regression results are then meaningless. The t-test is conducted in several stages. These are detailed in the following sections.

Null and alternative hypotheses

The first is to specify the null hypothesis and the alternative hypothesis. A null hypothesis is a statement that is assumed to be true unless you find very strong evidence against it. An alternative hypothesis is a statement that is accepted instead of the null hypothesis if you reject the null hypothesis.

With the t-test, the null hypothesis is that the slope coefficient (β_1) equals 0:

$$H_0 : \beta_1 = 0$$

This hypothesis implies that the independent variable (X) doesn't explain the value of the dependent variable (Y). The t-test is a very conservative test; the burden of proof is to show that X *does* explain Y.

The alternative hypothesis is that the slope coefficient doesn't equal 0:

$$H_1 : \beta_1 \neq 0$$



REMEMBER

As discussed in Chapter 12, this type of alternative hypothesis is known as a *two-tailed test*.

This test is usually conducted as a two-tailed test because we are trying to determine if the slope coefficient equals zero or not; it is not necessary to know if the slope coefficient is positive or negative.

Level of significance

The level of significance of a hypothesis test is a measure of the likelihood of a specific type of error, known as a *Type I error*. This occurs when the null hypothesis is incorrectly rejected when it is actually true. A Type II error results when the null hypothesis is *not* rejected even though it is false. With a small level of significance, there is a very low chance of committing a Type I error, but a relatively large probability of committing a Type II error. As the level of significance is increased, the probability of a Type I error increases but the probability of a Type II error decreases.

The choice of level of significance is based on the importance of avoiding Type I errors. When you test hypotheses about regression coefficients, the level of significance (α) is often chosen to be 0.05 (5 percent).

Test statistic

A test statistic is a numerical value that is used to determine if the null hypothesis should be rejected. If the test statistic has a large value (positive or negative), the likelihood that the null hypothesis is rejected is also large.

For testing hypotheses about β_1 the appropriate test statistic is

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

This expression is known as a *t-statistic* because it follows the Student's t-distribution (covered in Chapter 11).

The term $s_{\hat{\beta}_1}$ is the *standard error* of $\hat{\beta}_1$, which you can think of as the standard deviation of $\hat{\beta}_1$. (Standard errors are covered in Chapter 10.)

In other words, $s_{\hat{\beta}_1}$ is the amount of *uncertainty* associated with the use of $\hat{\beta}_1$ to estimate β_1 . The larger the standard error of $\hat{\beta}_1$, the less likely you are to reject the null hypothesis that $\beta_1 = 0$.

You compute the test statistic for this hypothesis test as follows:

$$SEE = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}}$$

Also known as *standard error of the regression* (SER), the standard error of the estimate (SEE) is a measure of the dispersion of the sample values above and below the estimated regression line. Based on Table 15-2, SEE is computed as follows.

RSS is computed as:

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 0.14$$

With a sample size of 8, SEE equals:

$$\begin{aligned} SEE &= \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}} \\ &= \sqrt{\frac{0.14}{6}} \\ &= 0.15275 \end{aligned}$$

1. Calculate the standard error of $\hat{\beta}_1$:

$$S_{\hat{\beta}_1} = \frac{SEE}{\sqrt{\sum_{i=1}^n X_i^2 - n\bar{X}^2}}$$

SEE equals 0.15275. $\sum_{i=1}^n X_i^2$ represents the sum of the squared values of X. $n\bar{X}^2$

represents the sample size times the square of the sample mean. You can get the values of $\sum_{i=1}^n X_i^2$ and $n\bar{X}^2$ from Table 15-3.

The sample mean is obtained by adding up the values in the X_i column, then dividing the sum by the sample size of 8:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{16 + 14 + 12 + 11 + 12 + 15 + 13 + 11}{8} = 13$$

TABLE 15-3**The Standard Error of $\hat{\beta}_1$**

X_i	X_i^2
16	256
14	196
12	144
11	121
12	144
15	225
13	169
11	121

The sum of the squared values of X is obtained by squaring each value of X and then summing the results:

$$\sum_{i=1}^n X_i^2 = 256 + 196 + 144 + 121 + 144 + 225 + 169 + 121 = 1,376$$

The formula for computing the standard error of $\hat{\beta}_1$ is:

$$S_{\hat{\beta}_1} = \frac{SEE}{\sqrt{\sum_{i=1}^n X_i^2 - n\bar{X}^2}} = \frac{0.15275}{\sqrt{1,376 - (8)(13^2)}} = 0.03118$$

2. Calculate the test statistic:

$\hat{\beta}_1 = 0.15$ (see the earlier section “Estimating the Population Regression Equation”); therefore, combining this with the standard error of $\hat{\beta}_1$, the t-statistic for $\hat{\beta}_1$ is computed as

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{0.15}{0.03118} = 4.81$$

Critical values

A critical value shows the number of standard deviations away from the mean of a distribution where

- » a specified percentage of the distribution is above the critical value.
- » the remainder of the distribution is below the critical value.

To test a hypothesis, the test statistic is compared with one or two critical values. If the test statistic is more *extreme* than the relevant critical value(s), the null hypothesis is rejected. Otherwise, the null hypothesis fails to be rejected. (It's technically incorrect to say that a null hypothesis is accepted, because you don't know every value in the population being tested.)

With simple regression analysis, the critical values are taken from the Student's t-table with $n - 2$ degrees of freedom. (These are found in Table 15-4.)



Degrees of freedom refers to the number of *independent* values in a sample. When it's necessary to estimate two measures from a sample (in this case, $\hat{\beta}_0$ and $\hat{\beta}_1$) the number of degrees of freedom equals the sample size minus 2.



REMEMBER

The Student's t-distribution is a continuous distribution that has a mean of zero and a larger variance and standard deviation than the standard normal distribution (covered in Chapter 9). The distribution is sometimes described as having "fat tails" because it's more spread out.

The purpose of the t-distribution is to describe the statistical properties of sample means that are estimated from *small* samples; the standard normal distribution is used for *large* samples. (There's much more about the Student's t-distribution Chapter 11.)

In this case, say you choose the level of significance (α) to be 0.05. (This is a widely used value for testing hypotheses about regression coefficients.) Because the sample size (n) is 8, the appropriate critical values are

$$\pm t_{\alpha/2}^{n-2} = \pm t_{0.025}^6$$

You find these values in the Student's t-table, such as Table 15-4.

TABLE 15-4 The Student's t-Distribution

Degrees of Freedom (df)	t0.10	t0.05	t0.025	t0.01	t0.005
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169

You find the value of the positive critical value $t_{0.025}^6$ at the intersection of the row for 6 degrees of freedom and the column labeled $t_{0.025}$, which is 2.447. The value of the negative critical value $-t_{0.025}^6$ is then -2.447.

Decision rule

A decision rule is used to determine if the null hypothesis should be rejected. Because the alternative hypothesis is $H_1 : \beta_1 \neq 0$, there are two critical values: one positive, one negative. (These are shown to be -2.447 and 2.447 in the previous section.)

If the test statistic is either greater than 2.447 or less than -2.447, the null hypothesis will be *rejected*. This indicates that there is strong evidence that the slope coefficient β_1 is not equal to zero; in other words, the regression equation *does* explain the relationship between the dependent variable (GPA) and the independent variable (monthly hours of studying).

If the test statistic falls between these two values, the null hypothesis *fails* to be rejected. In this case, there is insufficient evidence to reject the hypothesis that β_1 equals zero. This shows that the regression equation *does not* explain the relationship between the dependent variable (GPA) and the independent variable (monthly hours of studying).

In this case, the test statistic is 4.81, which is greater than 2.447. Therefore, you reject the null hypothesis in favor of the alternative hypothesis, indicating that $\hat{\beta}_1$ is different from 0 (that is, it's *statistically significant*). Therefore, strong evidence shows that X (monthly hours of studying) does explain the value of Y (GPA).

This result does not imply that hours of studying is the *only* factor that explains GPA, but it is an important determinant of GPA.

You can also test whether the estimated intercept ($\hat{\beta}_0$) is statistically significant, but often doing so isn't necessary. The slope coefficient is the most important value in the regression equation.

Using Statistical Software

Many spreadsheet programs (such as Microsoft Excel) and specialized statistical packages (such as SPSS) allow you to generate the results you need for regression analysis. For example, you can use a spreadsheet program to get the results shown in Figure 15–9 for the GPA example from the “Estimating the Population Regression Equation” section earlier in this chapter (these results were generated

using Excel). See Chapter 16 for details on how to conduct regression analysis in Microsoft Excel.

As you can see, Figure 15–9 shows the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ under the *Coefficients* column; the values of the coefficient of determination (R^2) and the standard error of the estimate, under the *Regression Statistics* column; and the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ the t-statistics under the columns *Standard Error* and *t-Stat*.

Figure 15–9 provides one additional useful measure you can use to test hypotheses about the coefficients, called *p-values* (or *probability values*). The p-value represents the likelihood of finding the given t-statistic if the null hypothesis is true. An extremely low p-value indicates that the null hypothesis of a 0 coefficient should be rejected. More formally, when testing the hypothesis $H_0 : \beta_1 = 0$, if the p-value is less than the level of significance (α), the null hypothesis is rejected; otherwise, it isn't rejected.

In this example, the p-value for $\hat{\beta}_1$ is 0.002968105; the level of significance is 0.05; therefore, because the p-value is less than the level of significance, the null hypothesis is rejected, confirming the results found when testing the hypothesis with the t-statistic.



TIP

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.891132789
R Square	0.794117647
Adjusted R Square	0.759803922
Standard Error	0.152752523
Observations	8

ANOVA

	df	SS	MS	F
Regression	1	0.54	0.54	23.14285714
Residual	6	0.14	0.023333333	
Total	7	0.68		

	Coefficients	Standard Error	t Stat	P-value
Intercept	1.05	0.408928138	2.56768831	0.042466896
X (Monthly Hours)	0.15	0.031180478	4.810702354	0.002968105

FIGURE 15-9:
GPA regression
problem.

Assumptions of Simple Linear Regression

The simple regression model shown in this chapter is based on several extremely important assumptions. If any of these assumptions are violated, the reliability of the regression results is questionable. The most important assumptions include the following:

- » The expected value of each error term is 0; that is $E(\varepsilon_i) = 0$. So although some error terms are positive and some are negative, they add up to 0.
- » The variances of the error terms are finite and constant for all values of x_i ; this common variance is designated σ^2 .
- » The error terms are independent of each other (for example, they don't influence each other).
- » Each error term, ε_i , is independent of the corresponding value of X_i (the value of X_i doesn't influence the value of the error term and vice versa).
- » The error terms are *normally distributed*. Although this assumption isn't required for linear regression, it's often used and allows you to compute confidence intervals for the regression coefficients. It also allows you to test hypotheses about the coefficients.

With simple regression analysis, two of the most important violations of the assumptions include autocorrelation and heteroscedasticity:

- » **Autocorrelation** occurs when the error terms are correlated with each other (they are related to each other). It violates the assumption of independence. Two independent variables have a correlation of 0 between them. Autocorrelated error terms can cause the standard errors of the regression coefficients to be understated, thus increasing the risk that coefficients will be incorrectly found to be statistically significant (different from zero).
- » **Heteroscedasticity** occurs when the error terms don't have a constant variance. This problem can cause the standard errors of the regression coefficients to be understated, increasing the risk that coefficients will be incorrectly found to be statistically significant (different from zero).

Conducting Simple Regression Analysis with the TI-84 Plus Calculator

You can use the Texas Instruments TI-84 Plus and Plus CE calculators to conduct simple regression analysis. Based on the example of grade point averages and monthly hours of studying time in this chapter, the sample regression line can be estimated with the TI-84 by first entering the data for Y (GPA) into List 1 and the data for X (hours of studying) into List 2. (Details on entering numbers into lists are given at the end of Chapter 3.) Then, follow these steps:

1. Press the [STAT] button.
2. Use the right arrow key to select TESTS.
3. Choose F: LinRegTTest and then press [ENTER].

LinRegTTest requires the following information:

XList:

YList:

Freq:

β and p :

RegEQ:

4. For XList, enter 2^{nd} 2 (L2) and for YList enter 2^{nd} 1 (L1).

Freq will always be equal to 1 unless some or all data in the lists is repeated more than once.

5. For β and p , use the arrow keys to select ± 0 because the alternative hypothesis is that the slope coefficient is not equal to zero.

Skip over RegEQ (it is used to for previously saved regression equations).

6. Select Calculate and then press [ENTER].

This produces the following results:

$$y = a + bx$$

$$\beta \neq 0 \text{ and } p \neq 0$$

$$t = 4.810702354$$

$$p = 0.0029681051$$

$$df = 6$$

$$a = 1.05$$

$b = 0.15$

$s = 0.1527525232$

$r^2 = 0.7941176471$

$r = 0.8911327887$



REMEMBER

You have to scroll down with the down arrow key to see the entire list of outputs.

$y = a + bx$ shows that a represents the intercept of the estimated regression equation while b is the slope. $\beta \neq 0$ and $p \neq 0$ is a reminder that this is a two-tailed test.

t is the test statistic for testing the hypothesis that the slope coefficient of the regression line (β_1) equals zero. p is the p-value for this test; because it is less than the level of significance of 0.05, the null hypothesis can be rejected. This means that X does explain Y ; in other words, studying time does help explain a student's GPA. df refers to the degrees of freedom for this test; this equals $n - 2$ (in this case, $8 - 2 = 6$).

$a = 1.05$ and $b = 0.15$ are the intercept and slope coefficient, respectively, of the estimated sample regression line. The resulting sample regression line can be expressed as: $Y = 1.05 + 0.15X$.

s is the standard error of the estimate (SEE) for this regression equation. It can be thought of as the common standard deviation shared by all residuals (estimated errors) for this estimated regression equation.

r^2 is the R-squared measure; because it is close to 1, this indicates a good fit of the regression model to the sample data.

r is the correlation coefficient between X and Y . (Note that this equals the square root of r^2 .) In this case, the correlation is about 0.89. Because the correlation coefficient cannot exceed 1 (or fall below -1), a correlation of 0.89 indicates a very strong, positive relationship between X and Y . (Correlation is discussed in Chapter 5.) In other words, there is a strong relationship between the amount of time a student spends studying each month and the student's GPA.

Note that it is possible to compute the sample regression equation without any of the corresponding statistical measures. After entering the X (L2) and Y (L1) data into lists, follow these steps:

1. Press the [STAT] button.
2. Use the right arrow key to select CALC.

3. Choose 8: LinReg(a + bx) and then press [ENTER].

LinReg(a + bx) requires the following information:

XList:

YList:

FreqList:

Store RegEQ:

4. For XList, enter 2nd 2 (L2) and for YList enter 2nd 1 (L1) for the GPA example.

You can skip over FreqList and Store RegEQ.

5. Select Calculate and then press [ENTER].

This produces the following results:

$$y = a + bx$$

$$a = 1.05$$

$$b = 0.15$$

As before, a is the intercept of the estimated sample regression equation and b is the slope. The estimated sample regression equation is therefore: $y = 1.05 + 0.15x$.

IN THIS CHAPTER

- » Understanding how to implement functions in Excel
- » Discovering Excel's key statistical analysis functions
- » Performing complex statistical analyses with the Analysis ToolPak

Chapter **16**

Key Statistical Techniques in Excel

Microsoft Excel provides a large collection of tools that can help you with statistical calculations. You can use the built-in functions to calculate a single value or an array of values, or use one of Excel's add-in tools such as the Analysis ToolPak to calculate more advanced routines. In this chapter, you discover how to implement Excel functions, become familiar with the many statistical functions available in Excel, and see how Excel's add-in tools can help you generate results more quickly and easily than is possible with a calculator.

Implementing Excel Functions

You implement Excel functions by selecting a cell, typing an equals sign (=) followed by the name of the function, a left parenthesis ((), and then the values needed by the function to work, known as *arguments*. (If you need multiple values, you separate them with commas.) You then follow the input values with a right parenthesis)) and press the Enter key. For example, the square root function is written as =SQRT(9), where SQRT is the name of the function and 9 is the argument, as shown in Figure 16-1. Pressing the Enter key produces the result of 3.

	A	B
1	=SQRT(9)	
2		

FIGURE 16-1:
Excel's square root function.

Alternatively, you can provide the input values to the Excel function as a *cell reference*, or the location of the values. Excel cells are identified by row and column, with columns labeled with letters and rows labeled with numbers. For example, the cell shown in Figure 16-1 is referred to as A1 (column A, row 1).

In the example shown in Figure 16-2, the cell reference A2 is entered into the function instead of 9 because the value 9 is stored in cell A2. Using this approach makes it easier to change data used in a spreadsheet.

	A	B
1	Number	Square Root
2		=SQRT(A2)

FIGURE 16-2:
Using a cell reference.

Alternatively, you can use the Insert Function button (f_x) to choose the appropriate function and then enter the required information. The Insert Function button appears to the left of the formula bar. Clicking the Insert Function button opens the Insert Function dialog box, where you can search for the appropriate function or select a function from a list provided.

Checking Out Excel's Key Statistical Functions

Excel's statistical functions can be classified into one of several categories, including:

- » Measures of central tendency
- » Measures of dispersion
- » Measures of association

- » Discrete probability distributions
- » Continuous probability distributions
- » Confidence intervals
- » Regression analysis

Measures of central tendency

Excel provides several functions that you can use to measure the key properties of a data set (sample or population). A *population* is a collection of measurements that is being studied; a *sample* is a subset of the population. Often samples are chosen from a population in order to provide key insights into the population's properties without needing to analyze every element in the population.

Three of the most widely used measures of central tendency are mean, median, and mode. These measures are designed to show the “center” of a set of data. The *mean* is the average value within a data set; the *median* is the value that divides the lower half from the upper half of a data set; and the *mode* is the value that occurs most frequently in a data set.

Mean

The Excel function for computing the mean — the average value within a data set — is:

```
AVERAGE(number 1, number 2, ....)
```

For example, suppose that a corporation wants to compute its average sales over a five-year span. The necessary data is listed in Figure 16-3.

	A	B
1	Year	Sales (\$millions)
2	2018	100
3	2019	120
4	2020	95
5	2021	115
6	2022	133
7		
8	AVERAGE =AVERAGE(B2:B6)	

FIGURE 16-3:
Computing
average sales.

The sales are entered into the function as a range of values: B2:B6. This approach is usually used when data is found in consecutive rows or columns. You can enter this data by manually typing **B2:B6** or selecting the cells with a mouse. In this example, the average is:

$$(100 + 120 + 95 + 115 + 133)/5 = 112.6$$

Median

The median is the value that divides a data set in half; half of the members of the data set are less than or equal to the median, and half are greater than or equal to the median. The Excel function for computing the median is:

```
MEDIAN(number 1, number 2, ....)
```

For example, to compute the median sales over a five-year span for the same corporation in the previous example, the median function is implemented as shown in Figure 16-4.

	A	B
1	Year	Sales (\$millions)
2	2018	100
3	2019	120
4	2020	95
5	2021	115
6	2022	133
7		
8	MEDIAN =MEDIAN(B2:B6)	

FIGURE 16-4:
Computing
median sales.

If the data are sorted from low to high, they will be: 95, 100, 115, 120, 133. The value 115 is greater than the lowest two observations (95 and 100) and less than the highest two observations (120 and 133). Therefore, the median is 115.

Mode

The mode of a data set is the most commonly occurring value. Unlike the mean and median, the mode does not have to be unique; there can be two or more modes, and there is no mode if no values are repeated. Excel's function for the mode is:

```
MODE(number 1, number 2, ....)
```

Because the sales data set has no mode (no values are repeated), Excel expresses this result as #N/A (not applicable), as shown in Figure 16–5. For a data set with more than one mode, the function MODE.MULT is required to show all of the modes; MODE will only show one of them.

	A	B
1	Year	Sales (\$millions)
2	2018	100
3	2019	120
4	2020	95
5	2021	115
6	2022	133
7		
8	MODE	#N/A

FIGURE 16-5:
Computing the
mode of sales.

Measures of dispersion

Measures of dispersion show how “spread out” the members of a data set are around the mean. Some of the most important measures of dispersion are *variance* and *standard deviation*.

Variance

The two basic types of variance formulas are the *population variance* and the *sample variance*. Using the same sales example from the previous section, if the five years’ worth of data represent the entire history of the corporation, then the data represent a population; otherwise, they represent a sample.

The Excel function for computing the sample variance is:

```
VAR.S(number 1, number 2, ....)
```

where the S represents a sample. If this data represents a sample, the variance is 236.3.

The Excel function for computing the population variance is:

```
VAR.P(number 1, number 2, ....)
```

where the P represents a population. If this data represents a population, the variance is 189.04.

Standard deviation

One of the drawbacks of the variance measure is that it is measured in squared units. In the sales example, the variance is measured as “dollars squared.” Because this is a meaningless measure, the standard deviation is used to convert the units back to dollars. Algebraically, the standard deviation is the square root of the variance.

For a sample, the Excel function is:

```
STDEV.S(number 1, number 2, ....)
```

For a population, the Excel function is:

```
STDEV.P(number 1, number 2, ....)
```

Measures of association

Measures of association are used to measure the relationship between two data sets. The two standard measures of association are *covariance* and *correlation*.

Covariance

The covariance is a numerical value that can be interpreted as follows:

Covariance	Relationship between X and Y
Negative	Inverse (negative)
Zero	Independent (no relationship)
Positive	Direct (positive)



REMEMBER

An inverse (negative) relationship indicates that two variables tend to move in opposite directions. A direct (positive) relationship indicates that two variables tend to move in the same direction. An independent relationship means that two variables are unrelated to each other.

For a sample, the covariance formula is:

```
COVARIANCE.S(array 1, array 2)
```

For a population, the covariance formula is:

```
COVARIANCE.P(array 1, array 2)
```

An array is a row or column of data, expressed as a range of cells. Note that covariance can only be computed from two sets of data at a time, and that each data set must contain the same number of elements. For example, suppose that the corporation in the sales data example wants to analyze the relationship between sales and profits. If the sales and profits data are samples, then the covariance is computed as shown in Figure 16–6. Note that the first array (sales) is expressed as B2:B6, which means all the cells in column B from row 2 to row 6. The second array is expressed as C2:C6, which means all the cells in column C from row 2 to row 6.

FIGURE 16-6:
Computing the covariance between sales and profits.

	A	B	C	D
1	Year	Sales (\$millions)	Profits (\$millions)	
2	2018	100	20	
3	2019	120	21	
4	2020	95	19	
5	2021	115	23	
6	2022	133	24	
7		COVARIANCE	=COVARIANCE.S(B2:B6,C2:C6)	

The sample covariance equals 28.45. Because the covariance is positive, this indicates that there is a tendency for profits to increase in years when sales increase, and vice versa.

If the data represent two populations, then the covariance is computed as:

```
=COVARIANCE.P(B2:B6, C2:C6)
```

which equals 22.76.

Correlation

One of the drawbacks of the correlation measure is that it does not have a maximum or minimum value, which makes interpretation difficult. In the previous example, the sample covariance is 28.45. Although this is positive, it is hard to be sure if this is a strong positive relationship or a weak positive relationship. The correlation measure overcomes this difficulty; it is defined to equal a value

between -1 and 1 . The following table shows how the correlation measure can be interpreted:

Correlation	Interpretation
-1	Perfect negative relationship
Close to -1	Strong negative relationship
Negative and close to 0	Weak negative relationship
0	No relationship
Positive and close to 0	Weak positive relationship
Close to 1	Strong positive relationship
1	Perfect positive relationship



REMEMBER

Because of the way correlation is defined, its value will be the same whether the data consists of two samples or two populations.

In Excel, correlation is computed as:

```
=CORREL(array 1, array 2)
```

For the sales and profits example, the correlation is 0.89 , indicating a strong positive relationship between the corporation's sales and profits. In years with above-average sales, profits tend to be above average and vice versa.

Discrete probability distributions

A probability distribution assigns probabilities to the outcomes of a random experiment. For example, if a coin is flipped three times, a probability distribution can be defined to determine the likelihood of getting no heads, one head, two heads, or three heads.

When an experiment consists of only a finite number of possible outcomes, the corresponding probability distribution is said to be *discrete*. If the number of possible outcomes is infinite, the corresponding probability distribution is said to be *continuous*. Two of the most widely used discrete probability distributions are the *binomial distribution* and the *Poisson distribution*.

Binomial distribution

The binomial distribution is used for situations when a random experiment consists of a series of independent trials; each trial can only have two possible outcomes: success and failure.

A coin-flipping experiment can be described with the binomial distribution because each flip of the coin can be thought of as a trial, and each trial can only lead to two possible outcomes: heads or tails. The outcome of interest can be thought of as a “success” and the other a “failure.” For example, if a researcher is interested in the number of heads that turn up, then heads would be defined as success and tails would be defined as failure.

The binomial distribution is uniquely characterized by the number of trials and the probability of success on a single trial. The Excel formula for computing binomial probabilities is:

```
=BINOM.DIST(number_s, trials, probability_s, cumulative)
```

where:

- » `number_s` refers to the number of successes that occur during the experiment.
- » `trials` refers to the number of trials of the experiment that are conducted.
- » `probability_s` refers to the probability of obtaining a success on a single trial.
- » `cumulative` is equal to 1 if this is a “less than or equal to” probability or 0 if this is the probability of a single value.

Based on the coin-flipping experiment, there are three trials (flips). If heads is defined as a success, then the probability of a success is 0.5. To compute the probability of obtaining exactly one head when the coin is flipped three times, the Excel function `BINOM.DIST` is implemented as:

```
=BINOM.DIST(number_s, trials, probability_s, cumulative)  
=BINOM.DIST(1, 3, 0.5, 0)  
= 0.375
```

To compute the probability of obtaining one head or fewer (no heads or one head) when the coin is flipped three times, cumulative is set equal to 1:

```
=BINOM.DIST(number_s, trials, probability_s, cumulative)  
=BINOM.DIST(1, 3, 0.5, 1)  
= 0.500
```

Poisson distribution

The Poisson distribution is used for situations when events occur during a given interval of time. For example, the Poisson distribution can be used to determine the probability that five customers will enter a store over the next hour. The Poisson distribution is uniquely characterized by the average number of events that occur per unit of time. The Excel formula for computing Poisson probabilities is:

```
=POISSON.DIST(x, mean, cumulative)
```

where:

- » x refers to the number of events that occur per unit of time.
- » mean refers to the average number of events that occur per unit of time.
- » cumulative has the same meaning as it does with the `BINOM.DIST` formula.

For example, suppose that a retail store manager determines that on average five customers enter a store each hour, and wishes to determine the likelihood that during the upcoming hour seven customers will enter the store. This is computed as follows:

```
=POISSON.DIST(x, mean, cumulative)  
=POISSON.DIST(7, 5, 0)  
=0.1044
```

If the retail store manager wants to know the probability that during the upcoming hour seven customers or fewer will enter the store, cumulative is set equal to 1:

```
=POISSON.DIST(x, mean, cumulative)  
=POISSON.DIST(7, 5, 1)  
=0.8666
```

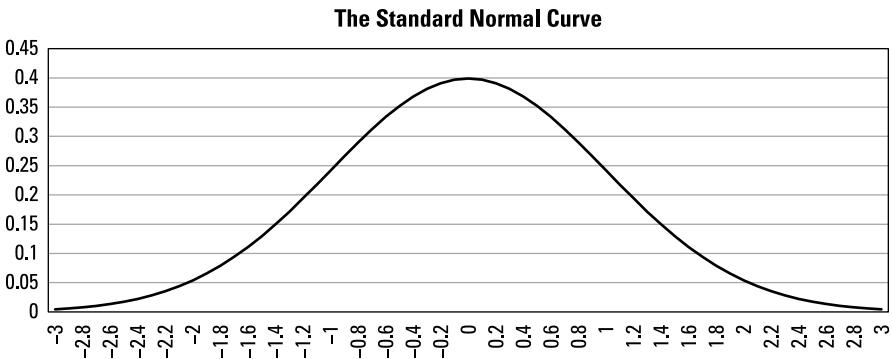
Continuous probability distributions

A continuous distribution is one in which the underlying experiment can have an infinite number of possible outcomes. For example, suppose that an experiment consists of observing the time that elapses until the next phone call. Because time can have a fractional value, such as 3.72 minutes, the number of possible times until the next phone call arrives is infinite. Therefore, the probability distribution corresponding to this experiment is continuous. Two of the most commonly used continuous probability distributions are the *normal distribution* and the *t-distribution*.

Normal distribution

The normal distribution is one of the most important distributions in statistics. It is used in many different disciplines, such as psychology, biology, economics, finance, and many others. The graph of the normal distribution can be described as a bell-shaped curve, which shows that there is a symmetry or balance between probabilities of values below the mean and above the mean. For example, if an individual's commuting time is normally distributed with a mean of 30 minutes, the probability that tomorrow's commute will be 25 minutes or less equals the probability that it will be 35 minutes or more. The graph shown in Figure 16-7 illustrates a special case of the normal distribution, known as the standard normal distribution.

FIGURE 16-7:
The standard
normal
distribution
curve.



The mean of this distribution is zero and the standard deviation is one. The area below the mean exactly matches the area above the mean, and the curve extends down to negative infinity and up to positive infinity.

There are infinitely many normal distributions, each uniquely characterized by the mean and standard deviation. The mean determines the location of the curve along the real number line, while the standard deviation determines how “spread out” the curve is from the mean. The Excel formula for computing normal probabilities is:

```
=NORM.DIST(x, mean, standard_dev, cumulative)
```

For example, suppose that it was determined that the returns to a stock are normally distributed with a mean of 8 percent per year and a standard deviation

of 12 percent per year. An investor wants to know the probability that next year the returns will be 10 percent or less. This would be computed as:

```
=NORM.DIST(x, mean, standard_dev, cumulative)  
=NORM.DIST(0.10, 0.08, 0.12, 1)  
=0.5662
```

In this case, $x = 0.10$ (10 percent) because this is the value of interest. The mean is 0.08 (8 percent) and the standard deviation is 0.12 (12 percent). Cumulative is set equal to 1, which indicates that the investor is looking for a “less than or equal to” probability.

Note that for “greater than or equal to” probabilities, an algebraic rearrangement of the NORM.DIST formula is needed. For example, suppose that the same investor wants to know the probability that next year the return to the stock will be 14 percent or more. The NORM.DIST function is only defined for “less than or equal to” or cumulative probabilities. The workaround for this problem is to compute:

```
1 - NORM.DIST(x, mean, standard_dev, cumulative)
```

This represents the probability that the variable of interest will be greater than or equal to x .

In this case, the probability that the return will be 14 percent or more is computed as:

```
1 - NORM.DIST(0.14, 0.08, 0.12, 1)  
= 1 - 0.6915 = 0.3085
```

The standard normal distribution

Because the standard normal distribution is widely used in practice, Excel includes a separate function for it:

```
= NORM.S.DIST(z, cumulative)
```

where S stands for *standard*.



TIP

Note that the function does not require the mean or standard deviation because these are defined as 0 and 1, respectively. Also, it is traditional to use the letter “ z ” to represent the standard normal distribution; x is used for all other normal distributions.

t-distribution

The t-distribution, also known as the *Student's t-distribution* is another heavily used continuous probability distribution. It is useful in many cases where the normal distribution is not appropriate, such as cases where small samples are used to draw conclusions about the mean of a population. The standard normal and t-distributions have many similarities and a few key differences.

The t-distribution has the following common features with the standard normal distribution:

- » A mean of zero
- » Characterized by a bell-shaped curve
- » Symmetrical about the mean
- » Continuous

The most important difference between the two distributions is that the t-distribution is characterized by a value known as the *degrees of freedom* instead of the mean and standard deviation. Degrees of freedom are closely related to the sample size used when choosing data from a population. Graphically, the t-distribution is more spread out than the standard normal distribution because it has a larger standard deviation. The graph of the t-distribution with 5 degrees of freedom is shown in Figure 16-8.

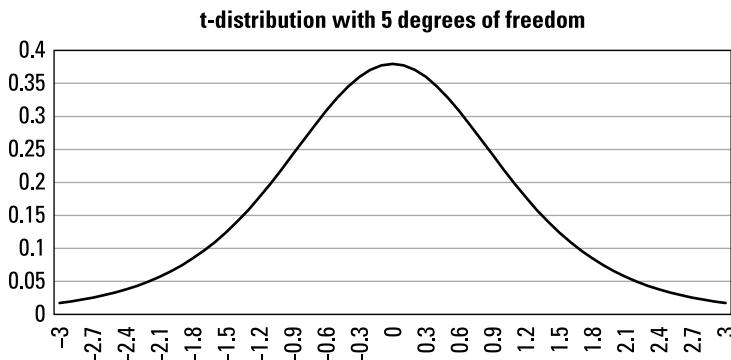


FIGURE 16-8:
The t-distribution
with 5 degrees
of freedom.

The properties of the t-distribution increasingly resemble those of the standard normal distribution as the degrees of freedom increase. The graphs shown in Figures 16-9 and 16-10 illustrate the relationship between the standard normal distribution and the t-distribution with 5 and 30 degrees of freedom, respectively.

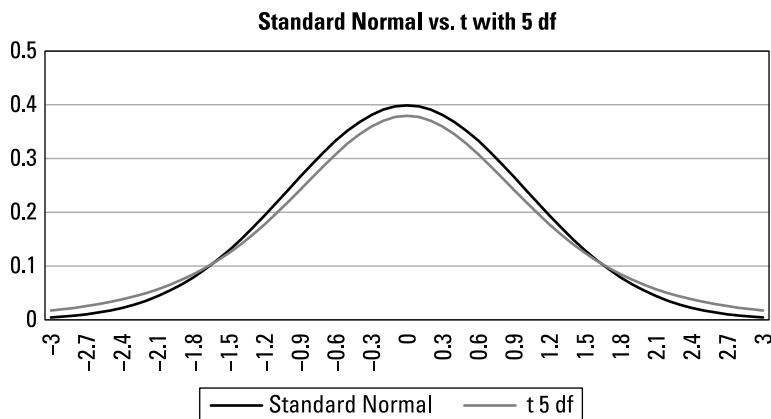


FIGURE 16-9:
Comparing the standard normal distribution with the t-distribution with 5 degrees of freedom.

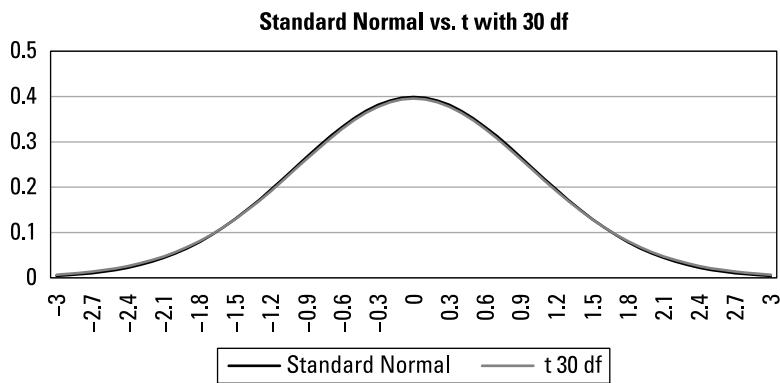


FIGURE 16-10:
Comparing the standard normal distribution with the t-distribution with 30 degrees of freedom.

The graphs show that with 5 degrees of freedom, the t-distribution has more area in the “tails” of the distribution and less in the center. With 30 degrees of freedom, it is difficult to see the difference between the standard normal distribution and the t-distribution.

Two key Excel functions for the t-distribution are T.DIST and T.INV.

You use the T.DIST function to find the probability that a t-distributed random variable has a value less than or equal to x . You use the T.INV function to find the location under the t-distribution that corresponds to a given “less than or equal to” probability.

For example, to find the probability that a t-distributed random variable with 5 degrees of freedom is less than or equal to 1, you use the Excel function T.DIST as follows:

```
=T.DIST(x, deg_freedom, cumulative)  
=T.DIST(1, 5, 1)  
= 0.8184
```

To find the location under the t-distribution with 5 degrees of freedom that corresponds to the lower 5 percent tail of the distribution, you use the Excel function T.INV as follows:

```
=T.INV(probability, deg_freedom)  
=T.INV(0.05, 5)  
=-2.02
```

This can be interpreted to mean that the t random variable has a 5 percent chance of being less than or equal to -2.02. Equivalently, it means that the t random variable has a 95 percent chance of being greater than or equal to -2.02.

Confidence intervals

A confidence interval shows the range of values within which a population measure such as the mean is believed to fall with a specified level of confidence. For example, suppose that a researcher determined that the mean rainfall in New York State each year has a 95 percent chance of being between 32 inches and 40 inches. This can be written as:

$$P(32 \leq \mu \leq 40) = 0.95$$

where μ is the population mean.

You obtain the confidence interval by computing the sample mean and then adding and subtracting a term known as the *margin of error* to create the lower limit and upper limit of the confidence interval, respectively. In this case, the sample mean is 36 and the margin of error is 4 so the lower limit equals $36 - 4 = 32$, and the upper limit equals $36 + 4 = 40$.



REMEMBER

The margin of error formula depends on whether the population standard deviation is known. If so, the margin of error is based on the standard normal distribution; otherwise, it is based on the t-distribution.



TIP

Excel provides a formula that you can use to determine the margin of error for a confidence interval, and you can then use this information to construct the appropriate confidence interval.

The Excel function for computing the margin of error of a confidence interval for the population mean when the population standard deviation is known is:

```
=CONFIDENCE.NORM(alpha, standard_dev, size)
```

where:

- » alpha is the “level of significance.” This equals one minus the confidence level. For example, with a 95 percent confidence interval, alpha equals 5 percent or 0.05.
- » standard_dev is the population standard deviation.
- » size is the sample size used to construct the confidence interval.

The Excel function for computing the margin of error of a confidence interval for the population mean when the population standard deviation is unknown is:

```
=CONFIDENCE.T(alpha, standard_dev, size)
```

As an example, suppose that a researcher is analyzing the salaries of major league baseball players. The researcher wants to construct a 95 percent confidence interval for the population mean (the mean salary of all major league baseball players). A sample of 22 players is chosen; the sample mean is \$4.8 million and the sample standard deviation is \$0.9 million. Because the confidence level is 95 percent, alpha is 5 percent (0.05). Because the problem does not specify the population standard deviation, it will be assumed that this value is unknown and the t-distribution will be used.

The margin of error is computed in Excel as:

```
=CONFIDENCE.T(alpha, standard_dev, size)  
=CONFIDENCE.T(0.05, 0.9, 22)  
= 0.40
```

This can be interpreted as \$0.40 million or \$400,000.

The lower limit of the confidence interval equals the sample mean minus the margin of error, or $4.8 - 0.40 = 4.40$. The upper limit of the confidence interval equals the sample mean plus the margin of error, or $4.8 + 0.40 = 5.20$. The 95 percent confidence interval for the mean salary of major league baseball players is therefore (4.40, 5.20). In other words, the confidence interval is the range between \$4.40 million and \$5.20 million.

Regression analysis

Regression analysis is used to estimate the relationship between a dependent variable and one or more independent variables. With simple regression analysis, there is a single independent variable; with multiple regression analysis, there are two or more independent variables.

With simple regression analysis, the sample regression line can be written as:

$$Y_i = \beta_0 + \beta_1 X_i$$

where:

\hat{Y}_i = the dependent variable

$\hat{\beta}_0$ = the intercept

$\hat{\beta}_1$ = the slope

X_i = the independent variable

Excel's built-in functions can be used to estimate the intercept and slope of the simple regression line.

The Excel function for the intercept is:

```
=INTERCEPT(Known_ys, Known_xs)
```

where:

» Known_ys is the data for the dependent variable (Y).

» Known_xs is the data for the independent variable (X).

For example, assume that a corporation wishes to analyze the relationship between its annual advertising expenditures and sales. In this case, sales are assumed to depend on advertising expenditures, so that sales are the dependent (Y) variable, and advertising expenditures are the independent variable (X). Figure 16-11 shows the layout of the data.

To estimate the intercept, the data are entered into the INTERCEPT function as follows:

```
=INTERCEPT(Known_ys, Known_xs)  
=INTERCEPT(B2:B6, C2:C6)  
= 98.55
```

FIGURE 16-11:
Computing the
coefficients of a
regression
equation.

	A	B	C
		Sales (Y) (\$millions)	Advertising Expenditures (X) (\$millions)
1	Year		
2	2018	120	1.2
3	2019	126	1.7
4	2020	133	2.1
5	2021	129	2
6	2022	140	2.3
7	INTERCEPT	=INTERCEPT(B2:B6,C2:C6)	
8	SLOPE	=SLOPE(B2:B6,C2:C6)	

The values are entered as a range of values; B2:B6 represents the sales (Y) for each year from 2018 to 2022 while C2:C6 represents the advertising expenditures (X) for the same period. The intercept of 98.55 indicates that in years when the corporation does not spend any money on advertising, its sales are expected to be \$98.55 million.

To estimate the slope, the Excel function is:

```
=SLOPE(Known_ys, Known_xs)
=SLOPE(B2:B6, C2:C6)
= 16.69
```

This indicates that each additional expenditure of \$1 million on advertising is expected to increase sales by \$16.69 million. The estimated sample regression equation is:

$$\hat{Y}_i = 98.55 + 16.69X_i$$

Going Deeper with the Analysis ToolPak

The Analysis ToolPak is a collection of advanced statistical routines. It contains a large number of important statistical procedures that would be difficult to implement with individual Excel functions. For example, it can perform hypothesis testing, regression analysis, and forecasting with exponential smoothing; it can also be used to compute the covariance and correlation of any number of variables. The ToolPak is an add-in package that is included with Excel, but it must be installed before it can be used.

Once installed, the Analysis ToolPak is found under the Data tab on the far right side of the ribbon, as shown in Figure 16-12. Clicking the Data Analysis button opens the dialog box shown in Figure 16-13.

FIGURE 16-12:
Excel's ribbon
showing the
Analysis ToolPak
button.

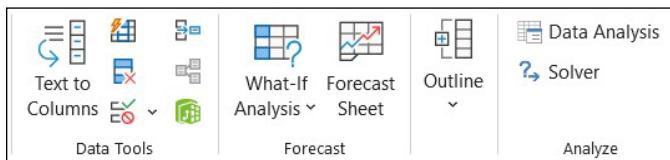
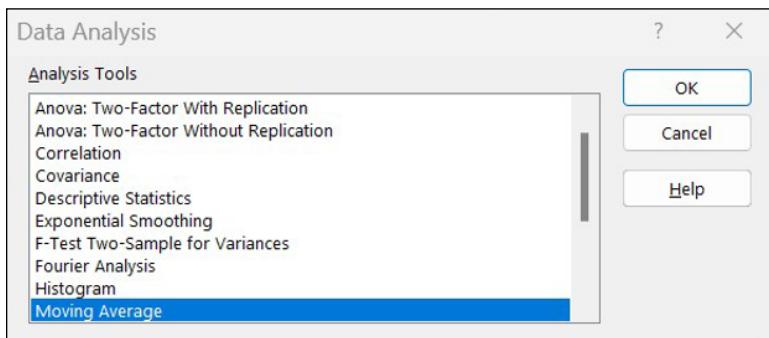


FIGURE 16-13:
The Analysis
ToolPak
dialog box.



Computing covariance and correlation

You can use the Analysis ToolPak to compute the covariance and correlation between two different data sets. For example, suppose that the data shown in Figure 16-14 represent the returns to two stocks from 2018–2022.

FIGURE 16-14:
Returns to Stocks
X and Y over a
five-year span.

	A	B	C
1	Year	Stock X	Stock Y
2	2018	3.60%	9.00%
3	2019	2.40%	-2.00%
4	2020	7.10%	11.00%
5	2021	8.40%	12.00%
6	2022	5.60%	5.00%

Choosing the Covariance tool from the Data Analysis dialog box opens the dialog box shown in Figure 16-15. Note that the data for the two variables (X and Y) must be stored in two consecutive columns or rows. If headers are included in the input range, the “Labels in first row” checkbox must be checked for columns of data or “Labels in first column” for rows of data. The data are entered as B1:C6, which means B1:B6 contains the data and header for the first variable (Stock X), and C1:C6 contains the data and header for the second variable (Stock Y).

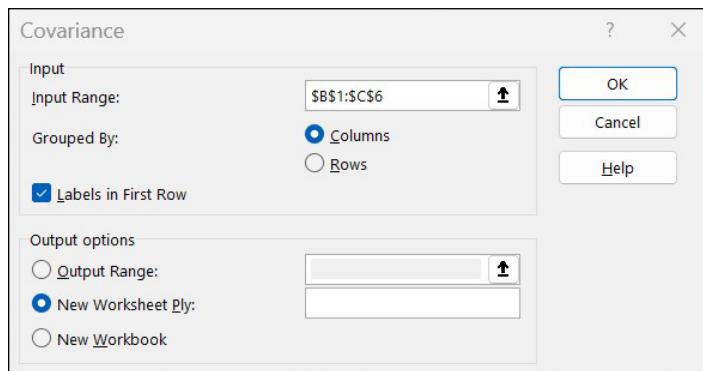


FIGURE 16-15:
The Analysis
ToolPak
Covariance
dialog box.

In the Output options section, the location of the output is specified. This can be a new tab, a new spreadsheet, or in a specified cell on the same spreadsheet. Here, the results start in Cell E1 on the same spreadsheet, as shown in Figure 16-16.

	E	F	G
	Stock X		Stock Y
Stock X	0.000483		
Stock Y	0.000896	0.0026	

FIGURE 16-16:
Variances and
covariances for
Stocks X and Y.

Note that the covariance between Stock X and itself equals the variance of Stock X, which is 0.000483. For Stock Y, the variance is 0.0026. The covariance between Stock X and Stock Y is 0.000896. The missing cell also shows the covariance between Stock X and Stock Y, which is left blank because this information is already shown in the table.

Choosing the Correlation tool produces the results shown in Figure 16-17. Note that the correlation between any variable and itself is equal to one, as shown in the table for Stocks X and Y. The correlation between Stock X and Y is 0.799256, which shows that there is a strong positive relationship between the two stocks. The missing cell also shows the correlation between Stock X and Stock Y, which is left blank.

FIGURE 16-17:
Correlations
between
Stocks X and Y.

	J	K
	Stock X	Stock Y
Stock X		1
Stock Y	0.799256	1



TIP

The Analysis ToolPak can also produce a table showing the covariances and correlations for a set of three or more variables at a time.

Computing descriptive statistics

You can use the Analysis ToolPak to compute many different descriptive statistics for a sample. For example, suppose that the data shown in Figure 16-18 represent the annual snowfall in Connecticut from 2013–2022.

FIGURE 16-18:
Snowfall
data over a
ten-year span.

	A	B
		Snowfall (Inches)
1	Year	
2	2013	16
3	2014	21
4	2015	18
5	2016	19
6	2017	13
7	2018	22
8	2019	24
9	2020	16
10	2021	17
11	2022	20

Choosing the Descriptive Statistics tool in the Analysis ToolPak opens up the dialog box shown in Figure 16-19. The input range refers to the cells that contain the data. Note that the data can be stored in rows or columns. When storing the data as a column, if a header is included in the Input Range, the “Labels in first row” checkbox must be checked. If the data are stored in a row and a header is used, the check box will change to “Labels in first column,” which must be checked. The summary statistics box must be checked to obtain the appropriate sample measures. Optionally, this procedure can also produce the margin of error for a confidence interval and the k th smallest and k th largest observations in the data set. The summary statistics for snowfall are shown in Figure 16-20.

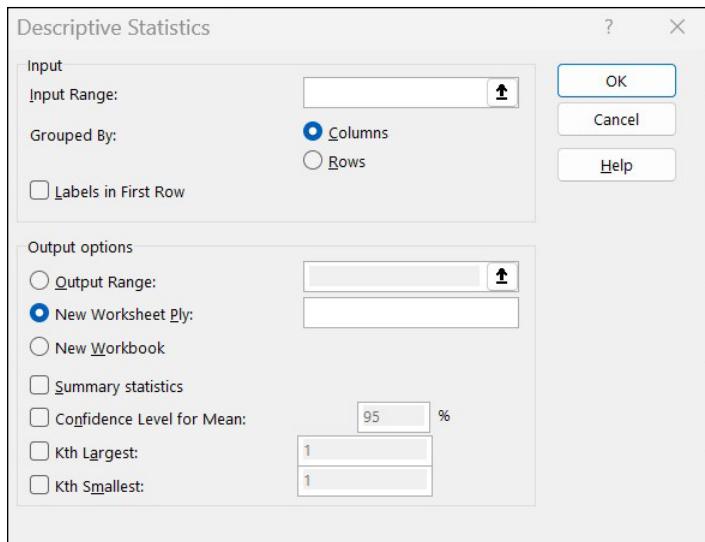


FIGURE 16-19:
The Analysis
ToolPak
Descriptive
Statistics
dialog box.



REMEMBER

E	F
<i>Snowfall (Inches)</i>	
Mean	18.6
Standard Error	1.03494498
Median	18.5
Mode	16
Standard Deviation	3.27278339
Sample Variance	10.71111111
Kurtosis	-0.3395474
Skewness	-0.0066562
Range	11
Minimum	13
Maximum	24
Sum	186
Count	10

FIGURE 16-20:
Descriptive
statistics for the
snowfall data.

Regression analysis

You can perform simple or multiple regression analysis with the Analysis ToolPak. Suppose that a corporation wants to analyze the relationship between its annual advertising expenditures and sales over the five-year span shown in Figure 16-21.

To analyze this data, choose the Regression tool to open the dialog box shown in Figure 16–22.

	A	B	C
		Sales (Y)	Advertising Expenditures
1	Year	(\$millions)	(X) (\$millions)
2	2018	120	1.2
3	2019	126	1.7
4	2020	133	2.1
5	2021	129	2.0
6	2022	140	2.3

FIGURE 16-21:
Advertising expenditures and sales over a five-year span.

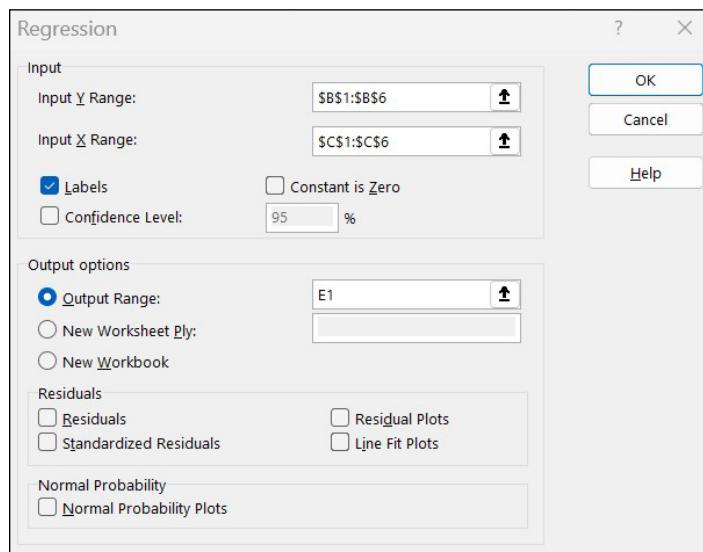


FIGURE 16-22:
The Analysis ToolPak Regression dialog box.

You must add values to the Input Y Range text field, which is the location of the data for the Y (dependent) variable, and the Input X Range text field, which is the location of the data for the X (independent) variable. In this case, Sales are the dependent variable and Advertising Expenditures are the independent variable. If labels are included, the Labels box must be checked. You can also specify where the output will be displayed. For this example, the information produced is shown in Figure 16–23.

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.951768818					
R Square	0.905863883					
Adjusted R Square	0.874485177					
Standard Error	2.658286011					
Observations	5					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	204.0005464	204.0005	28.86875	0.01262288	
Residual	3	21.19945355	7.066485			
Total	4	225.2				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.54918033	5.9000953	16.70298	0.000467	79.7724438	117.325917
Advertising Expenditures (X) (\$millions)	16.69398907	3.107035142	5.372964	0.012623	6.80601656	26.5819616

FIGURE 16-23:
Results of the
Analysis ToolPak
regression
routine.

This printout provides a great deal of information about the regression line and several key statistics for analyzing the results. The coefficients of the estimated regression line are found in the column labelled “Coefficients.” In this example, the regression line (with rounding) is:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$Y_i = 98.549 + 16.694 X_i$$

This indicates that the intercept is \$98.549 million and the slope is \$16.694 million, which matches the results from the Excel functions INTERCEPT and SLOPE.

Hypothesis testing

You can perform several different types of hypothesis tests using the Analysis ToolPak, such as testing hypotheses about the difference between two population means. For example, suppose that a researcher wishes to determine if the mean price of gas is the same on the East Coast and the West Coast of the United States. Samples are taken from gas stations in both regions (with prices expressed as dollars per gallon). Figure 16-24 lists the results.

In addition, it is assumed that the variances of the East Coast and West Coast gas prices are not equal. In this case, the appropriate hypothesis test is the two-sample t-test with unequal variances. Choosing the t-Test: Two-Sample Assuming Unequal Variances tool from the Data Analysis dialog box produces the dialog box shown in Figure 16-25.

	A	B	C
1	Station	East Coast	West Coast
2	1	3.88	3.89
3	2	3.99	3.93
4	3	3.94	3.70
5	4	4.05	3.61
6	5	4.03	4.00
7	6	3.81	4.01
8	7	3.95	4.29
9	8	3.96	4.08
10	9	3.82	3.96
11	10	3.83	4.02
12	11	3.57	
13	12	4.12	

FIGURE 16-24:
Gas prices at
12 gas stations on
the East Coast
and 10 gas
stations on the
West Coast.

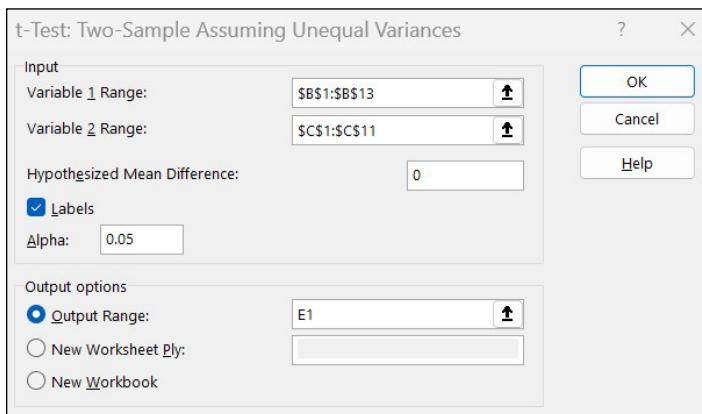


FIGURE 16-25:
The Analysis
ToolPak
two-sample t-test
dialog box.

Note that the hypothesized mean difference is set equal to zero, as the hypothesis being tested is:

$$H_0: \mu_1 = \mu_2$$

where:

μ_1 = the mean price of East Coast gas stations

μ_2 = the mean price of West Coast gas stations

East Coast gas prices are considered to be the first sample because the cells B1:B13 were entered into the “Variable 1 Range” box, and West Coast gas prices are considered to be the second sample because the cells C1:C11 were entered into the “Variable 2 Range” box. Because the researcher is trying to find out if the mean

price is the same on both coasts, this can be interpreted as a two-tailed test. The alternative hypothesis is that the means are not equal:

$$H_1: \mu_1 \neq \mu_2$$

Also note that alpha is set equal to 0.05, which is known as the *level of significance*. The probability of rejecting the null hypothesis when it is true is therefore 0.05 or 5 percent.

Clicking the OK button produces the output shown in Figure 16–26. The table shows the means and variances for the two samples as well as the number of observations within each sample. (For this test, the number of observations in each sample does not need to be equal.) The test statistic is approximately $t = -0.50$. This can be compared to the critical value for a two-tailed test, which is shown as ± 2.1098 . Because the t-statistic is not greater than 2.1098 or less than -2.1098, the null hypothesis is not rejected. In other words, there is not enough evidence to reject the null hypothesis that the mean gas prices are equal on both coasts.

The same test can be conducted with the p-value, which is shown as 0.6246 for a two-tailed test. Because this is greater than the assumed level of significance of 0.05, the null hypothesis is not rejected.

E	F	G
t-Test: Two-Sample Assuming Unequal Variances		
	<i>East Coast</i>	<i>West Coast</i>
Mean	3.9125	3.949
Variance	0.020947727	0.036187778
Observations	12	10
Hypothesized Mean Difference	0	
df	17	
t Stat	-0.498346495	
P(T<=t) one-tail	0.312312844	
t Critical one-tail	1.739606726	
P(T<=t) two-tail	0.624625687	
t Critical two-tail	2.109815578	

FIGURE 16-26:
The results of the Analysis ToolPak two-sample t-test routine.

If the researcher was interested in finding out if gas prices were higher on the East or West Coast, the one-tail p-value and critical t-values would have been used instead.

The Part of Tens

IN THIS PART . . .

See how statistical tests are based on the assumption of normality, and review several techniques available for testing whether a particular set of data is normally distributed.

Check out several types of problems that may arise when the assumptions of regression analysis are not met; two problems that can plague simple regression analysis are *autocorrelation* and *heteroscedasticity*.

IN THIS CHAPTER

- » Understanding logical fallacies that may arise in statistical analysis
- » Avoiding drawing incorrect conclusions from statistical results
- » Understanding the types of errors that can result in regression analysis
- » Understanding forecasting errors
- » Realizing how information may be presented incorrectly

Chapter 17

Ten Common Errors That Arise in Statistical Analysis

In the *For Dummies* Part of Tens fashion, this chapter discusses ten ways people may draw incorrect conclusions from statistical tests. These erroneous conclusions can result from several sources, including incorrect assumptions, misunderstanding the meaning of a statistical test, use of inappropriate data, and measurement error.

Any one of these mistakes can lead to erroneous conclusions being drawn, no matter how sophisticated the techniques being used. Part of the art of statistics is knowing which techniques to use under different circumstances and how to correctly interpret them. The following sections discuss different types of errors that may result from the incorrect application of statistical techniques.

Designing Misleading Graphs

Graphs may give a misleading picture of a sample or population if they're not well designed. For example, if you use scales on a graph that are substantially different from the values in the data you're analyzing, you may end up with a highly distorted view of the data.

Figures 17-1 and 17-2 represent the same data with two different histograms (see Chapter 2 for an overview of histograms). In this example, the data consist of the distribution of a bank's branches scattered throughout the four regions of the United States — North, South, East, and West.

Region	Branches
North	1,213
South	1,415
East	1,199
West	1,098

In Figure 17-1, the values on the vertical axis are separated by only 40 branches. With such closely spaced values on the vertical axis, the differences between the number of branches in each region appear to be very large. But, in fact, the difference between the largest number and the smallest number is only 317 (about 29 percent).

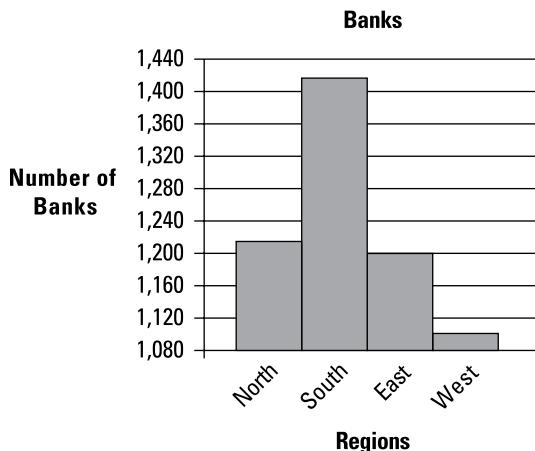


FIGURE 17-1:
Distribution of
bank branches by
geographical
region.

In Figure 17–2, the spacing of the values on the vertical axis is much wider, separated by 500 branches, making it appear that the differences between the numbers of branches are quite minimal. These figures show how easy it is to give a distorted view of data through poor design.

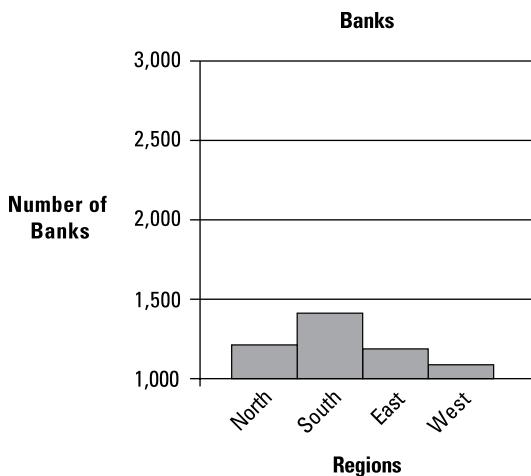


FIGURE 17-2:
Another look at
the distribution of
bank branches by
geographical
region.

Drawing the Wrong Conclusion from a Confidence Interval

When constructing a confidence interval, you can easily draw the wrong conclusion from the results. (Confidence intervals are covered in Chapter 11.) For example, suppose that a university constructs a 95 percent confidence interval for the mean GPA of its students. The sample mean is estimated to be 3.10; the 95 percent confidence interval is (2.95, 3.25).

It's tempting to conclude that the probability of the population mean being in the interval (2.95, 3.25) is 95 percent. Instead, this result indicates that for every confidence interval that's constructed from this population, in 95 cases out of 100, the confidence interval will contain the true population mean.

Misinterpreting the Results of a Hypothesis Test

One potential problem that may arise in hypothesis testing is being confused about what it means when the null hypothesis isn't rejected. It's important to distinguish between accepting the null hypothesis and failing to reject the null hypothesis. For example, suppose that a jury trial is in progress. For this hypothesis test, the following null and alternative hypotheses are used:

- » Null hypothesis (H_0): The defendant is not guilty.
- » Alternative hypothesis (H_1): The defendant is guilty.

If the null hypothesis is rejected, the defendant is guilty. If the null hypothesis isn't rejected, the defendant isn't necessarily innocent. There's simply insufficient evidence to show that he's guilty. There's a world of difference between being "innocent" and "not guilty!"



REMEMBER

The proper procedure in a hypothesis test is to conclude that a null hypothesis fails to be rejected unless strong contrary evidence exists against it. The conclusion should never be that the null hypothesis is accepted.

Placing Too Much Confidence in the Coefficient of Determination (R^2)

With regression analysis, researchers sometimes use the coefficient of determination to figure out whether one model "fits" the data better than another. The coefficient of determination assumes a value between 0 and 1; the closer it is to 1, presumably the better the regression model explains the relationship between X and Y. One of the drawbacks to the coefficient of determination is that it can be very close to 1 even for a model that makes no economic sense, such as a regression between two unrelated variables.

Another issue that arises with the coefficient of determination is that it automatically increases when new independent variables are added to a regression equation, even if the variables don't contribute any additional explanatory power to the regression. For this reason, the adjusted coefficient of determination is the preferred measure with multiple regression analysis because it increases only when newly added independent variables add at least some explanatory power.

Assuming Normality

Many statistical tests are based on the assumption of normality. For example, residuals are assumed to be normally distributed in regression analysis, enabling confidence intervals to be constructed for the slope coefficients.

For example, it's often assumed that the returns to stocks are normally distributed. In fact, although they're close to being normally distributed, they exhibit a property known as *fat tails*, where the actual probability of extreme outcomes (large positive returns and large negative returns) is greater than under the normal distribution. The assumption of normality causes investors to underestimate the true riskiness of their portfolios.



TIP

Several techniques are available for testing whether a particular set of data is normally distributed. For example, a Q-Q plot can be used to visually inspect data for normality. In addition, a formal hypothesis test of normality is available; it's known as the Jarque-Bera test. These types of techniques should be used before jumping to any conclusions about normality.

Thinking Correlation Implies Causality

One common error in statistical analysis is to assume that if two variables are correlated, one *causes* the other. Correlation simply indicates the tendency of two variables to move in the same or opposite directions. For example, new car sales tend to rise at the same time as new home sales, but no one would suggest that new home sales *cause* new car sales. (Equivalently, no one would suggest that new car sales are *caused by* new home sales.) These variables are positively correlated because they're both directly influenced by the economy. During an expansion, both new car sales and new home sales rise; during a recession, both fall.

One particularly well-known example of the dangers of assuming that correlation implies causality comes from the 19th century British economist William Stanley Jevons. Jevons was interested in applying statistical methods to the measurement of business cycles. He noticed that the business cycle had a tendency to follow changes in sunspot activity. Sunspots went through a cycle that lasted for about 11 years, while business cycles tended to last for just under 11 years. From his studies, Jevons concluded that the sunspots were actually responsible for the business cycle. (It's not as crazy as it sounds; sunspots can lead to changes in weather patterns, which would have a huge influence on the business cycles of a primarily agriculture-based economy. In spite of this, sunspots do *not* directly cause changes in the business cycle.)

Drawing Conclusions from a Regression Equation When the Data Do Not Follow the Assumptions

Several types of problems may arise when the assumptions of regression analysis are not met. (Simple regression analysis is covered in Chapter 15.) Two problems that can plague simple regression analysis are known as *autocorrelation* and *heteroscedasticity*.

- » **Autocorrelation** occurs when the error terms are correlated with each other (they are related to each other). It violates the assumption of independence. Two independent variables have a correlation of 0 between them. Autocorrelated error terms can cause the standard errors of the regression coefficients to be underestimated, thus increasing the risk that coefficients are incorrectly found to be statistically significant (different from zero).
- » **Heteroscedasticity** occurs when the error terms don't have a constant variance. This problem can cause the standard errors of the regression coefficients to be underestimated, increasing the risk that coefficients are incorrectly found to be statistically significant (different from zero).

When these problems are present, it is important to correct for them; otherwise, all results will be deceptive.

Using Regression Analysis to Make Predictions About Values Outside the Range of Sample Data

A regression model is estimated from sample data. The model only applies to the data found in this sample. For example, with the GPA case in Chapter 15, the independent variable (X) represents monthly hours of studying. In the data set used to estimate the sample regression model, the observations range from 11 hours to 16 hours per month. The estimated regression model can then be used to predict the GPA of a student whose studying time is between 11 and 16 hours per month. Using the model to predict the GPA of a student whose studying time is less than 11 hours or more than 16 hours is likely to lead to less accurate results.

One reason for this is that regression analysis is based on the assumption that the relationship between the independent (X) and dependent (Y) variables is linear. While this may hold for the data used to estimate a regression model, it is impossible to be sure if this is true for values outside the range of the sample data. It's also possible that other assumptions of regression analysis are violated outside the range of sample data.



REMEMBER

The results of any predictions for values of Y based on values of X that are not in the range of the sample data should be treated cautiously.

Placing Too Much Confidence in Forecasts

Many techniques are used to forecast future values of economic variables, such as stock prices, GDP growth, corporate sales, the demand for new products, and so on. Many of these techniques are highly sophisticated, which may give the false impression that they're extremely accurate. One major difficulty with forecasting techniques is that they're based on historical data that may not be repeated in the future. For example, if an economist is attempting to forecast future interest rates, his results don't capture any structural changes that occur in the economy during the forecast period, such as the selection of a new chairman of the Federal Reserve Board. In this case, future interest rates are unlikely to behave in exactly the same way that they have in the past, and the results of the forecast are likely to be inaccurate.

Two types of errors that may arise in forecasting are bias error and random error. *Bias error* occurs when a forecast is consistently greater than or less than actual values of a variable. *Random error* refers to unpredictable factors that can distort the results of a factor. These include earthquakes, strikes, sudden increases in oil prices, political turmoil, and so on.

With so much uncertainty surrounding forecasts, it would be a mistake to assume a high degree of accuracy.

Using the Wrong Distribution

In many situations, a variable is assumed to follow a specific probability distribution. For example, a computer chip manufacturer may assume that the number of defective chips produced by a specific process follows the binomial distribution. (The binomial distribution is covered in Chapter 8.) The binomial distribution is

based on several assumptions, one of which is that the trials are independent of each other. Suppose that in this process, one defective chip is highly likely to be followed by another defective chip (for example, repairs to the process are needed). In this case, the trials (chips) aren't truly independent of each other. As a result, any conclusions drawn about the distribution of defective chips are likely to be inaccurate. The manufacturer needs to find another distribution that more accurately reflects the distribution of the chips.

IN THIS CHAPTER

- » Keeping the most important statistical concepts fresh in your memory
- » Seeing how key statistical formulas are related

Chapter **18**

(Almost) Ten Key Categories of Formulas for Business Statistics

This chapter provides a brief overview of many key formulas discussed in the book. Think of this chapter as a handy reference guide so that you can quickly find the formulas that you need without having to search through the entire book.

Summary Measures of a Population or a Sample

Summary measures are used to describe key properties of a sample or a population. These measures can be classified as:

- » **Measures of central tendency** identify the *center* of a data set. Three of the most widely used measures of central tendency are the mean, median, and mode.

- The *mean* is another word for average.
- The *median* is a value that divides a sample or a population in half: Half of the elements in the data are less than or equal to the median, and half of the elements in the data are greater than or equal to the median.
- The *mode* is the most frequently occurring value in a sample or a population.

» **Measures of dispersion** are used to measure how spread out, or disperse, are the values of a sample or a population. Some of the most important measures of dispersion are the variance, standard deviation, percentiles, quartiles, and the interquartile range (IQR).

- **Variance:** The variance is calculated as the size of the average *squared* difference between the elements of a data set (a sample or a population) and the mean of the data set. The greater is the variance, the further the elements of the data set tend to be from the mean on average.
- **Standard deviation:** The square root of the variance. The standard deviation is more convenient to use than the variance due to the units in which these measures are calculated. As an example, if a sample consists of dollar prices, the sample standard deviation is measured in dollars, while the sample variance is measured in dollars *squared*, which is difficult to make sense of.
- **Percentiles:** Percentiles split a data set into 100 equal parts, each consisting of 1 percent of the values in the data set. For example, the 80th percentile represents the value in a sample or a population where 20 percent of the observations are greater than or equal to this value, and 80 percent are less than or equal to this value.
- **Quartiles:** Special types of percentiles, where the first quartile (Q1) is the 25th percentile, the second quartile (Q2) is the 50th percentile, and the third quartile (Q3) is the 75th percentile.
- **Interquartile range (IQR):** The difference between the third and first quartile.

» **Measures of association** provide a measure of how closely two samples or populations are related to each other. The two most important measures of association are:

- **Covariance:** A measure of the tendency for two variables to move in the same direction or in opposite directions. If two variables increase or decrease under the same circumstances, the covariance between them is positive. If two variables move in opposite directions, the covariance between them is negative. If two variables are unrelated to each other, the covariance between them is zero (or very close to zero).

- **Correlation:** Closely related to covariance; it has more convenient properties than covariance. Correlation can be thought of as the “standardized” version of covariance. For example, correlation always assumes a value between –1 and 1, whereas covariance has no lower or upper limits. As a result, it is easier to tell if the relationship between two variables is very strong or very weak with correlation than with covariance.

Probability

You use probability theory to model a large number of events in business applications. Probability theory is based on *set algebra*, and the important rules are

- » **Addition rule:** The formula for the Addition rule is:

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

The addition rule is designed to compute the probability of a *union* of two sets. In general, the union of sets A and B contains all the elements that are in set A, set B or both.

- » **Multiplication rule:** The Multiplication rule has two forms:

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

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

The multiplication rule is designed to compute the probability of the *intersection* of two sets. The intersection of sets A and B contains all the elements that are in *both* set A and set B.

- » **Complement rule:** The Complement rule has two forms:

$$P(A^C) = 1 - P(A)$$

$$P(A) = 1 - P(A^C)$$

The Complement rule tells you the probability of all elements that are *not* in a set. For example, suppose that set A contains all the black cards in a standard deck; the complement of A (written as A^C) is a set containing all the red cards. The probability of A^C can be computed with the Complement rule.

Discrete Probability Distributions

A discrete probability distribution is one where only a finite number of different outcomes may occur. The properties of a probability distribution may be summarized by a set of *moments*. Moments are numerical values that describe key properties of a probability distribution. Some of the most important are as follows:

- » The **expected value** is the first moment of a probability distribution. You compute it as

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

The expected value tells you the average value of X.

- » The **variance** is the second (central) moment of a probability distribution. You compute it as

$$\sigma^2 = \sum_{i=1}^n [X_i - E(X)]^2 P(X_i)$$

σ^2 represents the variance of X.

The variance tells you how much the different possible values of X are scattered around the expected value.

- » The **standard deviation** isn't a separate moment; it's the square root of the variance. The formula is

$$\sigma = \sqrt{\sum_{i=1}^n [X_i - E(X)]^2 P(X_i)}$$

The standard deviation is preferred to the variance since the variance is measured in squared units, which are difficult to interpret.

Following are two of the most widely used discrete probability distributions in business applications:

- » **Binomial distribution:** The binomial distribution is defined for a random process consisting of a series of independent trials in which only two different outcomes can occur on each trial. It enables you to determine the probability of a specified number of events occurring during a series of trials.

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

» **Poisson distribution:** The Poisson distribution is used to determine the probability that a specified number of events will occur during an interval of time.

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

Continuous Probability Distributions

A continuous probability distribution is defined for an infinite number of possible values. The normal distribution is one of the most widely used continuous probability distributions in business applications. You can get probabilities for the normal distribution from normal tables, specialized calculators, and spreadsheet programs.



REMEMBER

The normal distribution is important because many business situations may be accurately modeled with the normal distribution. For example, returns to stock prices are often assumed to follow the normal distribution.

Sampling Distributions

A sampling distribution is a special type of probability distribution defined for *sample statistics*. A sample statistic is a measure that describes the properties of a sample. Three of the most important sample statistics are the sample mean (\bar{X}), sample variance (s^2), and sample standard deviation (s). For more details about sampling distributions, see Chapter 10.

Based on a key result in statistics known as the *central limit theorem*, the sampling distribution of the sample mean is *normal* as long as the underlying population is normal or if you choose sample sizes of at least 30 from the population. To compute a probability for the sample mean, convert it into a standard normal random variable as follows:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

where:

\bar{X} = the sample mean

$\mu_{\bar{X}}$ = the mean of the sampling distribution of \bar{X}

$\sigma_{\bar{X}}$ = the standard deviation (also known as the *standard error*) of the sampling distribution of \bar{X}

Confidence Intervals for the Population Mean

A *confidence interval* is a set of numbers that is expected to contain the true value of the population mean with a specified probability. The formula you use to compute a confidence interval for the population mean depends on whether you know the population standard deviation (σ).

If you know the population standard deviation, the appropriate formula is

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where:

\bar{X} = the sample mean

$Z_{\alpha/2}$ = a quantile that represents the location of the right tail under the standard normal distribution with area $\alpha/2$

σ = the population standard deviation

n = the sample size

α = the level of significance

If you don't know the population standard deviation, you replace the population standard deviation with the sample standard deviation and the standard normal distribution with the t-distribution:

$$\bar{X} \pm t_{\alpha/2}^{n-1} \frac{s}{\sqrt{n}}$$

where:

$t_{\alpha/2}^{n-1}$ = a quantile (critical value) that represents the location of the right tail of the t-distribution with $n - 1$ degrees of freedom with an area of $\alpha/2$

s = the sample standard deviation

Testing Hypotheses about Population Means

Testing hypotheses about population means is a multi-step process, consisting of the null and alternative hypotheses, the level of significance, test statistic, critical value(s), and decision. (I walk you through all the steps of hypothesis testing in Chapter 12.)

You write the null hypothesis for testing the value of a single population mean as

$$H_0: \mu = \mu_0$$

where H_0 stands for the null hypotheses, μ is the true population mean and μ_0 is the hypothesized value of the population, or the value that you *think* is true.

The alternative hypothesis can assume one of three forms:

$$H_1: \mu > \mu_0 \text{ (known as a "right-tailed" test)}$$

$$H_1: \mu < \mu_0 \text{ (known as a "left-tailed" test)}$$

$$H_1: \mu \neq \mu_0 \text{ (known as a "two-tailed" test)}$$

To test a hypothesis, you must specify a level of significance — the probability of rejecting the null hypothesis when it's actually true.

When you're testing hypotheses about the population mean, the test statistic and the critical value (or values) depend on whether you know the population standard deviation.

- » When the population standard deviation is unknown, the appropriate test statistic is

$$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

where:

X = the sample mean

μ_0 = the hypothesized value of the population mean

s = the sample standard deviation

n = the sample size

- » When you know the population standard deviation (σ), the appropriate test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

To test hypotheses about the equality of two population means, the test statistic and critical values are different, but the basic process remains unchanged. In this case, though, you write the null hypothesis as $H_0: \mu_1 = \mu_2$, where μ_1 is the mean of population 1, and μ_2 is the mean of population 2.

- » For independent samples with equal population variances, the test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where s_p^2 is the estimated common “pooled” variance of the two populations — which you calculate with this formula:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The critical values of independent samples with equal population variances are based on the t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

- » If the independent samples are drawn from populations that don't have the same variance, the test statistic depends on the sizes of the two samples. If at least one sample is small (less than 30), the test statistic becomes

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}}$$

Here, the critical values are also drawn from the t-distribution, but the degrees of freedom calculation is much more complex:

$$df = \frac{\left[\left(s_1^2 / n_1 \right) + \left(s_2^2 / n_2 \right) \right]^2}{\frac{\left(s_1^2 / n_1 \right)^2}{(n_1 - 1)} + \frac{\left(s_2^2 / n_2 \right)^2}{(n_2 - 1)}}$$

- » If the independent samples are drawn from populations that don't have the same variance and both samples are large (at least 30), the test statistic becomes

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}}$$

In this case, the critical values are drawn from the standard normal distribution.

- » If the two samples aren't independent, they're known as *paired samples*. The test statistic is then based on the differences between the samples:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

where \bar{d} is the average difference between paired samples, and s_d is the standard deviation of the sample differences.

In this case, the critical values are taken from the t-distribution with $n - 1$ degree of freedom.

Testing Hypotheses about Population Variances

Testing hypotheses about population variances follows the same six-step procedure as testing hypotheses about population means (see previous section and Chapter 12 for details).

For testing hypotheses about the variance of a single population, the appropriate test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

where:

n = the sample size

s^2 = the sample variance

σ_0^2 = the hypothesized value of the population variance

The critical values are drawn from the chi-square distribution with $n - 1$ degrees of freedom.

For testing hypotheses about the equality of variances of two populations, the appropriate test statistic is

$$F = \frac{s_1^2}{s_2^2}$$

where:

s_1^2 = the variance of the sample drawn from population 1

s_2^2 = the variance of the sample drawn from population 2

The populations are assigned a number of 1 or 2 in such a way as to ensure that s_1^2 is greater than or equal to s_2^2 .

The critical values are drawn from the F-distribution, which has two different types of degrees of freedom: numerator and denominator. In this case, the numerator degrees of freedom equal $n_1 - 1$, and the denominator degrees of freedom equal $n_2 - 1$.

Using Regression Analysis

You use regression analysis to estimate the relationship between a dependent variable (Y) and one or more independent variables (X s).

- » Use **simple regression analysis** to estimate the relationship between a dependent variable (Y) and one independent variable (X).
- » Use **multiple regression analysis** to estimate the relationship between a dependent variable (Y) and two or more independent variables (X s).

Several tests allow you to validate the results of a regression equation. For example, if the coefficient of an independent variable equals 0, the variable doesn't belong in the regression. A hypothesis test helps you determine whether this coefficient equals 0.

It's also important to ensure that the underlying assumptions of regression analysis aren't being violated. Two potential problems can result with simple regression analysis if the assumptions aren't true:

- » **Autocorrelation** indicates that the error terms aren't independent of each other.
- » **Heteroscedasticity** indicates that the error terms don't have a common variance.

Index

Symbols

{ } (braces), 98
^ (caret), 297
| (conditional probabilities), 108–109
= (equals sign), 317
! (exclamation point), 130
! (factorial operator), 130
 \cap (intersections), 100–101
 λ (lambda), 138–139
 μ (mean, population), 38, 162
 ϵ (membership in set), 98
 $-\infty$ (negative infinity), 163–164
– (negative key), 163–164
() (parentheses), 317
 ∞ (positive infinity), 155, 164
 σ . See standard deviation
 Σ (summation operator), 36, 122
 \pm (two values), 195
 \cup (union), 99–100

Numbers

1-Var Stats function, 51, 70, 93
2-SampFTest, 284–285
2SampTTest, 241–243
2SampZTest, 239–241
2-Var Stats function, 92–93

A

absolute terms, 69
accuracy, samples and, 172
addition rule, 14, 110–112
alternative hypothesis, 18–19
 with dependent samples, 233
population variance, 251–253
 equality of two, 279
left-tailed test, 252
 two-tailed test, 252–253
for single population mean, 206–209

left-tailed test, 208
right-tailed test, 207–208
two-tailed test, 209
t-test, 306–307
 for two population means, 224
Analysis ToolPak, 334–342
 correlation, 335–337
 covariance, 335–337
 descriptive statistics, 337–338
 hypothesis testing, 340–342
 regression analysis, 338–340
arguments, 317
arithmetic mean, 36–38
 calculating population arithmetic mean, 38
 calculating sample arithmetic mean, 36–37
arrangements, 130, 263
assumption of normality, 349
autocorrelation, 313
average. See mean

B

bell-shaped curve
 normal distribution, 16, 149–150
t-distribution, 188
binomcdf function, 142
BINOM.DIST function, 134
binomial distribution, 16, 128–137
 computing binomial probabilities, 129–134
 binomial formula, 132–134
 combinations, 131–132
 factorial, 130
 computing with TI-84 Plus calculator, 141–142
graphing, 135–137
moments of, 134–135
 calculating expected value, 134–135
 computing variance and standard deviation, 135
overview, 128–129
statistical functions on Excel, 324–325
binomial formula, 132–134

binomial process, 128–129
binompdf function, 141–142
Blackstone, William, 211
braces ({}), 98
business statistics
 applications of, 7–8
 categories of formulas for, 353–362
 confidence intervals for population mean, 358
 continuous probability distributions, 357
 discrete probability distributions, 356–357
 probability theory, 355
 sampling distributions, 357–358
 summary measures, 353–355
 testing hypotheses about population means, 359–361
 testing hypotheses about population variances, 361–362
 using regression analysis, 362
data, properties of, 8–13
 analyzing with graphs, 8–11
 numerical measures, 11–13
probability, 13–17
 distributions, 15–17
 random variables, 14–15
sampling distributions, 17
statistical inference, 17–20
 confidence intervals, 18
 hypothesis testing, 18–19
 regression analysis, 19–20

with TI-84 Plus calculator, 270–272
staying positive with, 246–250
defining chi-square random variable, 248–249
moments of, 249–250
representing graphically, 247–248
testing hypotheses about population variance, 250–258
 alternative hypothesis, 251–253
 critical values, 254–258
 level of significance, 253
 null hypothesis, 250–251
 test statistic, 253–254
chi-square table, 254–255, 258
classes
 defined, 22
 figuring width, 24–25
 frequency of, 25
 on histograms, 175
CLT (central limit theorem), 17, 178–183
cluster samples, 171–172
coefficient of determination (R^2)
 computing, 305–306
 confiding too much in, 348
 overview, 303–305
coefficient of variation (CV), 67–68
coefficients, 295
COMBIN function, 132
combinations, 131–132
complement, mathematical operation, 102
complement rule, 14, 112–113, 182
conditional probabilities ($|$), 108–109
confidence coefficient, 195
confidence intervals
 computing with TI-84 Plus calculator, 201–203
 defined, 187
 drawing wrong conclusion from, 347
 estimating for population mean, 195–201
 known population standard deviation, 196–199
 unknown population standard deviation, 199–201
 formulas, 358
 normal distribution and, 151
 overview, 18
 point estimates vs. interval estimates, 194–195
 probabilities and t-table, 193–194

C

caret (^), 297
categories, 22
cell reference, 318
center of data. *See* measures of central tendency
central limit theorem (CLT), 17, 178–183
cheat sheet, 3
chi-square distribution, 245–272
 F-distribution defined by, 275–276
 goodness of fit tests, 258–272
 comparing population to normal distribution, 265–269
 comparing population to Poisson distribution, 259–265
 with TI-84 Plus calculator, 270–272
staying positive with, 246–250
defining chi-square random variable, 248–249
moments of, 249–250
representing graphically, 247–248
testing hypotheses about population variance, 250–258
 alternative hypothesis, 251–253
 critical values, 254–258
 level of significance, 253
 null hypothesis, 250–251
 test statistic, 253–254
chi-square table, 254–255, 258
classes
 defined, 22
 figuring width, 24–25
 frequency of, 25
 on histograms, 175
CLT (central limit theorem), 17, 178–183
cluster samples, 171–172
coefficient of determination (R^2)
 computing, 305–306
 confiding too much in, 348
 overview, 303–305
coefficient of variation (CV), 67–68
coefficients, 295
COMBIN function, 132
combinations, 131–132
complement, mathematical operation, 102
complement rule, 14, 112–113, 182
conditional probabilities ($|$), 108–109
confidence coefficient, 195
confidence intervals
 computing with TI-84 Plus calculator, 201–203
 defined, 187
 drawing wrong conclusion from, 347
 estimating for population mean, 195–201
 known population standard deviation, 196–199
 unknown population standard deviation, 199–201
 formulas, 358
 normal distribution and, 151
 overview, 18
 point estimates vs. interval estimates, 194–195
 probabilities and t-table, 193–194

- statistical functions on Excel, 331–332
- t-distribution, 188–192
- degrees of freedom, 189
 - graphing, 191–192
 - moments of, 189–190
 - properties of, 188–189
- confidence level, 195, 199
- constant (e), 138
- constant, slope as, 291
- continuous probability distributions
- chi-square distribution, 245–272
 - goodness of fit tests, 258–272
 - staying positive with, 246–250
 - testing hypotheses about population variance, 250–258
- comparing discrete and, 146–148
- F-distribution, 273–285
- computing with TI-84 Plus calculator, 283–285
 - defining F random variable, 275–276
 - moments of, 276–278
 - overview, 273–275
 - testing hypotheses about equality of two population variances, 278–283
- formulas, 357
- general discussion of, 16–17
- normal distribution, 16, 145–164
- comparing discrete and continuous distributions, 146–148
 - comparing population to, 265–269
 - computing standard normal probabilities, 152–159
 - computing with TI-84 Plus calculator, 162–164
 - graphing, 149–150
 - overview, 148–149
 - standard normal distribution, 151–152
- normal distribution as, 149
- statistical functions on Excel, 326–331
- normal distribution, 327–328
 - standard normal distribution, 328
- t-distribution, 329–331
- t-distribution
- degrees of freedom, 189
 - graphing, 191–192
 - left-tailed test with, 215
 - moments of, 189–190
- properties of, 188–189
- right-tailed test with, 214–215
- two-tailed test with, 216
- continuously compounded interest, 138
- convenience samples, 172
- correlation, 13, 72–86
- with Analysis ToolPak, 335–337
 - assuming causality with, 349
 - comparing covariance and, 83–86
 - interpreting correlation coefficient, 86–92
 - correlation and benefits of diversification, 89–92
 - showing relationship between two variables, 87–89
 - population covariance and correlation coefficient, 78–82
 - sample covariance and correlation coefficient, 73–78
- statistical functions on Excel, 323–324
- covariance, 13, 72–86
- with Analysis ToolPak, 335–337
 - comparing correlation and, 83–86
 - population covariance and correlation coefficient, 78–82
 - sample covariance and correlation coefficient, 73–78
- statistical functions on Excel, 322–323
- critical values
- defined, 18–19
 - equality of two population variances
 - left-tailed test, 281–282
 - right-tailed test, 280–281
 - two-tailed test, 282
 - known population standard deviation, 196–197, 216–219
 - left-tailed test with z-distribution, 218
 - right-tailed test with z-distribution, 217–218
 - two-tailed test with z-distribution, 218–219
 - overview, 212–213
 - population variance, 254–258
 - decision rule, 257–258
 - left-tailed test, 255–256
 - right-tailed test, 254–255
 - two-tailed test, 256–257
 - t-test, 309–311
 - unknown population standard deviation, 213–216

critical values (*continued*)

- left-tailed test with t-distribution, 215
 - right-tailed test with t-distribution, 214–215
 - two-tailed test with t-distribution, 216
- cumulative frequency distributions, 28–29
- cumulative probabilities, 152, 266
- CV (coefficient of variation), 67–68

D

data

- analyzing with graphs, 8–11
 - histograms, 8–9
 - line graphs, 9–10
 - pie charts, 10
 - scatter plots, 10–11
- defining properties and relationships with numerical measures
- determining relationship between two variables, 13
- measures of central tendency, 11–12
- measures of dispersion, 12–13

distribution of

- of, 22–29
 - cumulative frequency distributions, 28–29
 - frequency distribution for qualitative values, 27–28
 - frequency distributions for quantitative data, 23–27

relative position of

- 62–66
 - interquartile range, 66
 - percentiles, 63–64
 - quartiles, 64–66

decision rule

- equality of two population variances, 282–283
- population mean, 220–223
- population variance, 257–258

t-test

- degrees of freedom (df)**
 - chi-square distribution, 253–254
 - F-distribution, 274
 - hypothesis testing, 213
 - large number of, 229
 - overview, 189
 - t-table, 193
- denominator, 41
- dependent events, 110

dependent samples

- defined, 225
 - working with, 232–235
- descriptive statistics, 337–338
- df. *See degrees of freedom*
- discrete probability distributions
- comparing continuous and, 146–148
 - formulas, 356–357
- general discussion of, 16
- statistical functions on Excel, 324–326
 - binomial distribution, 324–325
 - Poisson distribution, 326

distribution of data

- cumulative frequency distributions, 28–29
- frequency distribution for qualitative values, 27–28
- frequency distributions for quantitative data, 23–27
 - figuring class width, 24–25
 - observing relative frequency distributions, 25–27

diversification

dollars squared

double-counting

E

- e (constant)**, 138
- elements**, 98
- empty sets**, 101
- equal population variances**, 226–229
- equal probability**, 167
- equality of two population variances**, 278–283
 - alternative hypothesis, 279
 - critical values, 280–282
 - decision about, 282–283
 - null hypothesis, 279
 - test statistic, 280
- equals sign (=)**, 317
- error term**, 296
- ESS (explained sum of squares)**, 304
- estimate**, 18
- estimated value**, 297
- events**, 103–105
 - computing probabilities of, 105–106
 - independent events, 104–105
 - mutually exclusive events, 104

Excel

Analysis ToolPak, 334–342
correlation, 335–337
covariance, 335–337
descriptive statistics, 337–338
hypothesis testing, 340–342
regression analysis, 338–340
BINOM.DIST function, 134
computing combinations, 132
computing factorial, 130
computing percentiles, 64
computing quartiles, 66
implementing functions, 317–318
POISSON.DIST function, 139
random number generator, 167–168
regression analysis, 311–312, 333–334
statistical functions, 318–334
 confidence intervals, 331–332
 continuous probability distributions, 326–331
 discrete probability distributions, 324–326
 measures of association, 322–324
 measures of central tendency, 319–321
 measures of dispersion, 321–322
exclamation point (!), 130
EXP function, 138
expected frequencies, 260–261, 266
expected value
 of binomial distribution, 134–135
F-distribution, 276–277
overview, 122–124
 of Poisson distribution, 139
explained sum of squares (ESS), 304

F

f_x (Function button), 318
FACT function, 130
factorial operator (!), 130
false negative (Type II error), 209–211, 274, 307
false positive (Type I error), 209–211, 253, 274, 307
F-distribution, 273–285
 computing with TI-84 Plus calculator, 283–285
 defining *F* random variable, 275–276
 moments of, 276–278

overview, 273–275

testing hypotheses about equality of two population variances, 278–283
 alternative hypothesis, 279
 critical values, 280–282
 decision about, 282–283
 null hypothesis, 279
 test statistic, 280
fifth root, 40
finite outcomes, 146
finite population correction factor, 178, 182
first moment, probability distribution
 binomial distribution, 134
 t-distribution, 189
Fisher, Ronald, 273
flat trend line, 34
forecasting, 20, 351
formulas
 alternative hypothesis, 251, 252
 binomial formula, 129, 132–134
 class width, 24
 coefficient of variation, 67
 combinations, 131
 confidence intervals
 known population standard deviation, 196
 population mean, 358
 unknown population standard deviation, 199
 continuous probability distributions, 357
 converting sample mean into standard normal random variable, 180
 cumulative rate of return, 39
 discrete probability distributions, 356–357
 expected value, 122
 Poisson distribution, 139
 probability distribution, 134
finite population correction factor, 178
interval estimates, 194–195
left-tailed test, 208
linear relationship, 72, 291
multiplication rule, 113
null hypothesis, 206, 224, 250
ordinary least squares estimators, 300
Poisson probabilities, 138, 262–263
population arithmetic mean, 38

- formulas (*continued*)
population correlation coefficient, 79, 81
population covariance, 78
population standard deviation, 60
population variance, 59
probabilities of events, 105
probability theory, 355
regression analysis, 362
relative frequency distributions, 25
right-tailed test, 208
sample arithmetic mean, 36
sample correlation coefficient, 76
sample covariance, 73
sample standard deviation, 55
sample variance, 54
sampling distribution of sample mean, 179
sampling distributions, 357–358
standard deviation
 binomial distribution, 135
 t-distribution, 190
standard error, 180
standard normal distribution, 159
standard normal random variable, 179
summary measures, 353–355
test statistic, 260
 equal population variances, 226
 sample variance, 253
 unequal population variances, 229, 231
testing hypotheses
 about population means, 359–361
 about population variances, 361–362
t-statistic, 307
two-tailed test, 209
variance, 125
variance of binomial distribution, 135
weighted arithmetic mean, 40
weighted mean, 261–262
fractions, rounding, 64
frequency distributions
 cumulative, 28–29
 for qualitative values, 27–28
 for quantitative data, 23–27
 figuring class width, 24–25
 observing relative frequency distributions, 25–27
F-table, 281, 283
function, random variable as, 116
Function button (f_x), 318
- ## G
- Gauss, Johann Carl Friedrich, 148
Gaussian distribution. *See* normal distribution
generalized least squares (GLS), 300
geometric mean, 38–40
goodness of fit tests, 258–272
 comparing population to normal distribution, 265–269
 comparing population to Poisson distribution, 259–265
 with TI-84 Plus calculator, 270–272
graphs, 8–11, 21–34
 analyzing distribution of data by class or category, 22–29
 cumulative frequency distributions, 28–29
 frequency distribution for qualitative values, 27–28
 frequency distributions for quantitative data, 23–27
 binomial distribution, 135–137
 chi-square distribution, 247–248
 designing misleading, 346–347
 histograms, 8–9, 29–30
 line graphs, 9–10, 31–32
 normal distribution, 149–150
 pie charts, 10, 32–33
 Poisson distribution, 140–141
 portraying sampling distributions, 175–177
 scatter plots, 10–11, 33–34, 292–295
 t-distribution, 191–192
greater than or equal to, 11, 42, 157, 181–182
Greek letters, 38
gross return, 38
- ## H
- heteroscedasticity, 313
histograms
 binomial distribution with, 135–137
 frequency distribution with, 29–30

overview, 8–9
Poisson distribution with, 140–141
portraying sampling distributions, 175–177
probability distribution with, 121
hypothesis testing
with Analysis ToolPak, 340–342
equality of two population variances, 278–283
goodness of fit tests, 258–272
misinterpreting results, 348
normal distribution and, 151
overview, 18–19, 205
population means formulas, 359–361
population regression equation, 303–311
population variance, 250–258
for single population mean, 206–223, 235–239
for two population means, 223–235, 239–243
hypothesized value, 250–251

I
“in between” standard normal probabilities, 158–159
inaccuracy, 128
independent chi-square random variables, 276
independent correlation, 78
independent events
determining, 109–110
multiplication rule with, 114
overview, 104–105
independent intervals, 137
independent samples, 225–232
equal population variances, 226–229
unequal population variances
both sample sizes are large, 231–232
at least one sample is small, 229–231
index, 36–37, 64
infinite outcomes, 146
intercept, 34
interest rates, 138
interquartile range (IQR), 13, 66
intersections (\cap), 100–101
interval estimates, 194–195
intervals. *See classes*

J
joint probabilities, 108
judgment samples, 173
K
known population standard deviation
computing with TI-84 Plus calculator, 202
critical values, 216–219
left-tailed test with z-distribution, 218
right-tailed test with z-distribution, 217–218
two-tailed test with z-distribution, 218–219
estimating confidence intervals for, 196–199

L
lambda (λ), 138–139
Latin letters, 38
left tails, 46
left-tailed test
alternative hypothesis, 208, 224, 233
defined, 207
for F-distribution, 281–282
population variance, 252, 255–256
rejecting null hypothesis, 220
with t-distribution, 215
with z-distribution, 218
less than or equal to, 11, 42, 152–155, 181
level of significance
confidence intervals, 195
hypothesis testing, 209–211
population variance, 253
t-test, 307
line graphs, 9–10, 31–32
linear relationship, 19–20, 72
applications of, 290
defining, 291–292
using scatter plots to identify, 292–295
LinReg(a + bx), 315–316
LinRegTTest, 314–315
long right tails, 47
lower limit, confidence interval, 196

M

- marginal probabilities, 106–107
- market capitalization, 289
- mathematical operations
 - complement, 102
 - intersections, 100–101
 - membership, 98
 - subsets, 98–99
 - union, 99–100
- mean, 36–42
 - arithmetic mean, 36–38
 - calculating population arithmetic mean, 38
 - calculating sample arithmetic mean, 36–37
 - computing with TI-84 Plus calculator, 50–52
 - defined, 11–12
 - determining relationship between median and, 44–47
 - negatively skewed, 45–46
 - positively skewed, 46–47
 - symmetrical about the mean, 45
 - geometric mean, 38–40
 - statistical functions on Excel, 319–320
 - weighted mean, 40–42
 - calculating weighted arithmetic mean, 40–42
- mean, population (μ), 38, 162
- measures of association, 13, 71–94
 - covariance and correlation, 72–86
 - comparing, 83–86
 - computing with TI-84 Plus calculator, 92–94
 - population covariance and correlation coefficient, 78–82
 - sample covariance and correlation coefficient, 73–78
 - interpreting correlation coefficient, 86–92
 - correlation and benefits of diversification, 89–92
 - showing relationship between two variables, 87–89
- statistical functions on Excel
 - correlation, 323–324
 - covariance, 322–323
- measures of central tendency, 11–12, 35–52
 - benefits of using histograms for, 175
 - computing with TI-84 Plus calculator, 50–52
- determining relationship between mean and median, 44–47
 - negatively skewed, 45–46
 - positively skewed, 46–47
 - symmetrical about the mean, 45
- mean, 36–42
 - arithmetic mean, 36–38
 - geometric mean, 38–40
 - weighted mean, 40–42
- median, 42–44
- mode, 48–49
- overview, 11–12
- statistical functions on Excel
 - mean, 319–320
 - median, 320
 - mode, 320–321
- measures of dispersion, 12–13, 62–66
 - benefits of using histograms for, 175
- interquartile range, 66
- measuring, 12–13
- percentiles, 63–64
- quartiles, 64–66
- statistical functions on Excel
 - standard deviation, 322
 - variance, 321
- median
 - computing with TI-84 Plus calculator, 50–52
 - defined, 11–12
 - determining relationship between mean and, 44–47
 - negatively skewed, 45–46
 - positively skewed, 46–47
 - symmetrical about the mean, 45
 - overview, 42–44
 - statistical functions on Excel, 320
- membership, 98
- membership in set (ϵ), 98
- Microsoft Excel. *See* Excel
- mode
 - computing with TI-84 Plus calculator, 50–52
 - defined, 11–12
 - discovering, 48–49
 - statistical functions on Excel, 320–321
- moments, 121–125
 - of binomial distribution

calculating expected value, 134–135
computing variance and standard deviation, 135
of chi-square distribution, 249–250
expected value, 122–124
of F-distribution, 276–278
of sampling distribution, 177–178
summation operator (Σ), 122
of t-distribution, 189–190
variance and standard deviation, 124–125
multiple regression, 289
multiplication rule, 14, 113–114
mutually exclusive events, 104, 112
mutually exclusive sets, 101

N

nCr function, 132
negative infinity ($-\infty$), 163–164
negative key (-), 163–164
negative probabilities, table, 153–154, 181
negative relationship, 72–73
negatively related variables, 33–34
negatively skewed distribution, 45–46, 136, 175
nonlinear relationship, 292
nonprobability sampling
 convenience samples, 172
 judgment samples, 173
 purposive samples, 173
 quota samples, 172–173
normal distribution, 16, 145–164
 comparing discrete and continuous distributions, 146–148
 comparing population to, 265–269
 computing standard normal probabilities, 152–159
 in between, 158–159
 greater than or equal to, 157
 less than or equal to, 152–155
 properties, 155–156
 computing with TI-84 Plus calculator, 162–164
graphing, 149–150
overview, 148–149
standard normal distribution, 151–152
statistical functions on Excel, 327–328
normalcdf function, 162–164

normality, assumption of, 349
null hypothesis, 18–19
 equality of two population variances, 279
population variance, 250–251
for single population mean, 206
t-test, 306–307
for two population means, 224
numerator, 41
numerical measures
 determining relationship between two variables, 13
measures of central tendency, 11–12
measures of dispersion, 12–13
moments, 121–125
expected value, 122–124
summation operator (Σ), 122
variance and standard deviation, 124–125

O

observed frequencies, 260–261, 265
online references
 binomial table with 19 values, 134
 cheat sheet for this book, 3
 continuous distribution online calculator, 146
ordinary least squares (OLS) estimators, 300
outliers
 defined, 36
 interquartile range and, 66
overestimation, 297
overrepresentation, 170

P

paired samples. *See* dependent samples
parameters, 17, 174
parentheses (()), 317
PERCENTILE function, 64
percentiles, 12–13, 63–64
perfect negative correlation, 78
permutations, 132
pie charts, 10, 32–33
point estimates, 194–195
point estimators, 194
Poisson, Siméon Denis, 138

- Poisson distribution, 16
calculating expected value, 139
comparing population to, 259–265
computing variance and standard deviation, 140
computing with TI-84 Plus calculator, 142–143
graphing, 140–141
overview, 137–139
sampling distribution and, 175–177
statistical functions on Excel, 326
- Poisson process, 137
- poissoncdf function, 142–143
- POISSON.DIST function, 139
- poissonpdf function, 142–143
- population
cluster samples, 171–172
comparing
to normal distribution, 265–269
to Poisson distribution, 259–265
convenience samples, 172
defined, 17, 36, 165
judgment samples, 173
purposive samples, 173
quota samples, 172–173
simple random samples, 167–168
stratified samples, 169–171
systematic samples, 168–169
population arithmetic mean, 38
population covariance, 78–82
population mean, 174
estimating confidence intervals for, 195–201
known population standard deviation, 196–199
unknown population standard deviation, 199–201
hypothesis testing for single
alternative hypothesis, 206–209
comparing critical values, 212–219
computing test statistic, 211–212
level of significance, 209–211
null hypothesis, 206
with TI-84 Plus calculator, 235–239
using decision rule, 220–223
hypothesis testing for two means, 223–235
alternative hypotheses, 224
determining test statistics for, 225–235
- null hypothesis, 224
with TI-84 Plus calculator, 239–243
- population parameter, 174
- population regression equation
defining, 295–296
estimating, 297–303
testing
coefficient of determination (R^2), 303–305
computing coefficient of determination, 305–306
t-test, 306–311
- population standard deviation, 60–62
- population variance, 250–258
alternative hypothesis, 251–253
left-tailed test, 252
right-tailed test, 251–252
two-tailed test, 252–253
- critical values, 254–258
decision rule, 257–258
left-tailed test, 255–256
right-tailed test, 254–255
two-tailed test, 256–257
- equality of two, 278–283
alternative hypothesis, 279
critical values, 280–282
decision about equality of two population variances, 282–283
null hypothesis, 279
test statistic, 280
- formula, 59–60
- level of significance, 253
- null hypothesis, 250–251
test statistic, 253–254
- positive infinity (∞), 155, 164
- positive probabilities, table, 152–153
- positive relationship, 72
- positively related variables, 33–34
- positively skewed distribution
binomial distribution, 136
chi-square distribution, 246–250
defining chi-square random variable, 248–249
moments of, 249–250
representing graphically, 247–248
- F-distribution
computing with TI-84 Plus calculator, 283–285

- defining F random variable, 275–276
moments of, 276–278
overview, 273–275
testing hypotheses about equality of two population variances, 278–283
histograms showing, 175
overview, 46–47
- probabilities, 13–17
characterizing probability distribution with moments, 121–125
expected value, 122–124
summation operator, 122
variance and standard deviation, 124–125
- distributions, 15–17
continuous probability distributions, 16–17
discrete probability distributions, 16
- random variables, 14–15
assigning probabilities to, 119–121
defining role of, 116–119
- t-table, 193–194
- types of
conditional probabilities, 108–109
determining independence of events, 109–110
joint probabilities, 108
unconditional probabilities, 106–107
- probability distributions, 15–17, 127–143
binomial distribution, 128–137
binomial formula, 132–134
combinations, 131–132
computing with TI-84 Plus calculator, 141–142
factorial, 130
graphing, 135–137
moments of, 134–135
overview, 128–129
calculating, 119–120
continuous probability distributions, 16–17, 146–148
discrete probability distributions, 16, 146–148
- normal distribution, 145–164
computing standard normal probabilities, 152–159
computing with TI-84 Plus calculator, 162–164
graphing, 149–150
overview, 148–149
standard normal distribution, 151–152
- Poisson distribution
calculating expected value, 139
computing variance and standard deviation, 140
computing with TI-84 Plus calculator, 142–143
graphing, 140–141
overview, 137–139
- sampling distributions
central limit theorem, 178–183
moments of, 177–178
overview, 174
portraying graphically, 175–177
using wrong distribution, 351–352
visualizing with histogram, 121
- probability sampling
cluster samples, 171–172
simple random samples, 167–168
stratified samples, 169–171
systematic samples, 168–169
- probability theory, 13–14, 97–114
bettting on uncertain outcomes
computing probabilities of events, 105–106
events, 103–105
sample space, 103
- formulas, 355
- rules
addition rule, 110–112
complement rule, 112–113
multiplication rule, 113–114
- types of probabilities
conditional probabilities, 108–109
determining independence of events, 109–110
joint probabilities, 108
unconditional probabilities, 106–107
- working with sets
complement, 102
intersection, 100–101
membership, 98
subset, 98–99
union, 99–100
- properties, standard normal distribution, 155–156
- purposive samples, 173
- p-values (probability values), 312

Q

qualitative data

defined, 8

frequency distribution for, 27–28

quantile, 196

quantitative data

defined, 8

frequency distributions for, 23–27

figuring class width, 24–25

observing relative frequency distributions, 25–27

QUARTILE function, 66

quartiles, 13, 64–66

questionnaires, 166

quota samples, 172–173

R

R^2 . *See* coefficient of determination

RANDBETWEEN function

for cluster samples, 171

for simple random samples, 167–168

for systematic samples, 169

random experiment, 14

computing probabilities of events, 105–106

events

independent events, 104–105

mutually exclusive events, 104

sample space, 103

random number generator, 167–168

random variables, 14–15

assigning probabilities to, 119–121

calculating probability distribution, 119–120

visualizing probability distribution with histogram, 121

chi-square, 248–249

defining role of, 116–119

F-distribution, 275–276

t-distributed, 193–194

thinking of statistics as, 174

regression analysis

with Analysis ToolPak, 338–340

errors

from drawing conclusions, 350

in making predictions, 350–351

formulas, 362

normal distribution and, 151

overview, 19–20

simple regression analysis, 289–316

assumptions of simple linear regression, 313

conducting with TI-84 Plus calculator, 314–316

linear relationship, 290–295

population regression equation, 295–311

using statistical software, 311–312

statistical functions on Excel, 333–334

rejecting null hypothesis, 311

decision rule for, 220–223

left-tailed test, 215

overview, 206

right-tailed test, 214

two-tailed test, 216

relative frequency distributions, 25–27

relative ranking, 63

relative terms, 69

relative variation, 67–69

coefficient of variation, 67–68

comparing relative risks of two portfolios, 68–69

replacement, sampling with, 168

residual sum of squares (RSS), 304

residuals, 298

returns to stocks, 151

right-tailed test

alternative hypothesis, 207–208, 224, 233

for F-distribution, 280–281

population variance, 254–255

rejecting null hypothesis, 220

with t-distribution, 214–215

with z-distribution, 217–218

risk

comparing two portfolios, 68–69

correlation between stocks, 89

rounding numbers, 25, 64

RSS (residual sum of squares), 304

rules, probability theory

addition rule, 110–112

complement rule, 112–113

multiplication rule, 113–114

S

sample, defined, 36, 165
sample arithmetic mean, 36–37
sample covariance, 73–78
sample mean, 174
 converting to standard normal random variable, 179–183
 moments of sampling distribution, 177–178
 portraying sampling distributions graphically, 175–177
sample regression line, 20, 297–303
sample space, 103
sample standard deviation, 55–59
sample statistics, 166
sample variance, 54–55
sampling distributions, 17
 central limit theorem, 178–183
 formulas, 357–358
 moments of, 177–178
 overview, 174
 portraying graphically, 175–177
sampling techniques
 nonprobability sampling
 convenience samples, 172
 judgment samples, 173
 purposive samples, 173
 quota samples, 172–173
 overview, 165–167
 probability sampling
 cluster samples, 171–172
 simple random samples, 167–168
 stratified samples, 169–171
 systematic samples, 168–169
scatter plots, 10–11, 33–34, 71
 identifying linear relationship, 292–295
 showing strong negative relationship, 88
 showing strong positive relationship, 87
 showing unrelated variables, 89
second central moment, probability distribution
 binomial distribution, 134
 t-distribution, 189–190
sequence of numbers, 168–169
SER (standard error of the regression), 308

sets, 14
complement, 102
intersection, 100–101
membership, 98
subset, 98–99
union, 99–100
simple random samples, 167–168
simple regression analysis, 19–20, 289–316
 assumptions of, 313
 conducting with TI-84 Plus calculator, 314–316
linear relationship, 290–295
 defining, 291–292
 using scatter plots to identify, 292–295
population regression equation
 defining, 295–296
 estimating, 297–303
 testing, 303–311
 using statistical software, 311–312
slope, 34, 291–292
specialized statistical packages, 311–312
spread of data. *See* measures of dispersion
squared percentage, 190
squared units, 54, 58, 190
squared values, 298
standard deviation (σ), 12, 162
 binomial distribution, 135
 F-distribution, 276, 278
 Poisson distribution, 140
population, 60–62
sample, 55–59
statistical functions on Excel, 322
variance vs., 55–59, 124–125
standard error, 177, 180
standard error of the regression (SER), 308
standard normal probability tables
 converting sample mean to standard normal random variable, 179–183
 known population standard deviation, 197
for negative values, 153–154
for positive values, 152–153
standard normal random variable
 converting normal random variable to, 266
 converting sample mean to, 179–183

- standard normal (z) distribution
- in between, 158–159. *See also* normal distribution
 - greater than or equal to, 157
 - left-tailed test with, 218
 - less than or equal to, 152–155
 - for negative values, 267
 - overview, 151–152
 - for positive values, 268
 - properties, 155–156
 - right-tailed test with, 217–218
 - statistical functions on Excel, 328
 - t-distribution vs., 191–192
 - two-tailed test with, 218–219
- STAT main menu, 50
- statistic, defined, 174
- statistical analysis
- errors in, 345–352
 - assuming normality, 349
 - confiding in coefficient of determination, 348
 - confiding in forecasts, 351
 - designing misleading graphs, 346–347
 - drawing conclusions from regression equation, 350
 - drawing wrong conclusion from confidence interval, 347
 - misinterpreting hypothesis test results, 348
 - thinking correlation implies causality, 349
 - using regression analysis to make predictions, 350–351
 - using wrong distribution, 351–352
- normal distribution in, 151
- statistical functions, Excel, 318–334
- confidence intervals, 331–332
 - continuous probability distributions, 326–331
 - normal distribution, 327–328
 - standard normal distribution, 328
 - t-distribution, 329–331
- discrete probability distributions
- binomial distribution, 324–325
 - Poisson distribution, 326
- measures of association
- correlation, 323–324
 - covariance, 322–323
- measures of central tendency
- mean, 319–320
 - median, 320
 - mode, 320–321
- measures of dispersion
- standard deviation, 322
 - variance, 321
- regression analysis, 333–334
- statistical inference, 17–20, 166
- confidence intervals, 18
 - hypothesis testing, 18–19
 - regression analysis, 19–20
- strata, 169–170
- stratified samples, 169–171
- strong negative relationship, 86
- strong positive relationship, 86
- Student's t-distribution. *See* t-distribution
- subjective judgment, 172, 173
- subsets, 98–99
- subtracting, 110–111
- summary measures, 353–355
- summation operator (Σ), 36, 122
- symmetric distribution, 136, 148
- symmetrical about the mean distribution, 175
- normal distribution as, 149
- overview, 45
- standard normal curve as, 155–156
- t-distribution, 188
- symmetry, 17
- systematic samples, 168–169

T

- tables
- chi-square
 - with larger numbers of degrees of freedom, 258
 - overview, 254–255
 - for negative values, 153–154
 - for positive values, 152–153
- tails, 45
- t-distribution
- degrees of freedom, 189
 - graphing, 191–192
 - left-tailed test with, 215
 - moments of, 189–190

properties of, 188–189
purpose of, 310
right-tailed test with, 214–215
statistical functions on Excel, 329–331
two-tailed test with, 216
test statistics, 18–19
equality of two population variances, 280
population variance, 253–254
for single population mean, 211–212
t-test, 307–309
for two population means
with dependent samples, 232–235
using independent samples, 225–232
TI-84 Plus and Plus CE calculators
binomial distribution, 141–142
confidence intervals, 201–203
known population standard deviation, 202
unknown population standard deviation, 203
covariance and correlation with, 92–94
F-distribution, 283–285
goodness of fit tests with, 270–272
hypothesis testing with
single population mean, 235–239
two population means, 239–243
largest value expressed, 164
mean, median, and mode with, 50–52
measures of dispersion with, 69–70
normal distribution, 162–164
Poisson distribution, 142–143
smallest value expressed, 164
TInterval, 203
total sum of squares (TSS), 304
trend line, 34, 88, 293
trials, 14
t-statistic, 307
t-table, 193–194
t-test, 306–311
alternative hypothesis, 306–307
critical values, 309–311
decision rule, 311
level of significance, 307
null hypothesis, 306–307
test statistic, 307–309

TTest, 237–239
two values (\pm), 195
2SampTTest, 241–243
2SampZTest, 239–241
two-tailed test
alternative hypothesis, 209, 224, 233
defined, 207
for F-distribution, 282
population variance, 252–253, 256–257
rejecting null hypothesis, 220, 222
with t-distribution, 216
with z-distribution, 218–219
Type I error (false positive), 209–211, 253, 274, 307
Type II error (false negative), 209–211, 274, 307

U

unconditional probabilities, 106–107
uncountable outcomes, 146
underestimation, 297
underrepresentation, 170
unequal population variances
both sample sizes are large, 231–232
at least one sample is small, 229–231
unexplained variation, 304
union (\cup), 99–100
unique values, 100
universal set, 103
unknown population standard deviation
computing with TI-84 Plus calculator, 203
critical values, 213–216
left-tailed test with t-distribution, 215
right-tailed test with t-distribution, 214–215
two-tailed test with t-distribution, 216
estimating confidence intervals for, 199–201
unrelated variables, 33–34
upper limit, confidence interval, 196

V

variables, linear relationship, 290–295
defining, 291–292
using scatter plots to identify, 292–295

variance, 12, 53–70
of binomial distribution, 135
computing measures of dispersion with TI-84 Plus calculator, 69–70
F-distribution, 276–278
finding sample standard deviation, 55–59
measuring relative variation, 67–69
coefficient of variation, 67–68
comparing relative risks of two portfolios, 68–69
of Poisson distribution, 140
population standard deviation, finding, 60–62
population variance, 250–258
alternative hypothesis, 251–253
critical values, 254–258
finding, 59–60
level of significance, 253
null hypothesis, 250–251
test statistic, 253–254
relative position of data, 62–66
interquartile range, 66
percentiles, 63–64
quartiles, 64–66
of sample, 54–55
standard deviation vs., 54, 124–125
statistical functions on Excel, 321

variation, 303
Venn diagrams, 98–99
volatility, 257

W
weak negative relationship, 86
weak positive relationship, 86
weighted arithmetic mean, 40–42
weighted least squares (WLS), 300
weighted mean, 40–42, 122, 261–262
width, interval, 199
widths, class, 24–25

X
 χ^2 GOF-Test, 271

Z
z distribution. *See* standard normal distribution
zero correlation, 73
zero covariance, 73
zero probability, 147
ZInterval, 202
ZTest, 236–237

About the Author

Alan Anderson currently teaches finance, economics, statistics, and math at several different schools, including Fordham University, New York University, Manhattanville College, Purchase College, and Fairfield University. He has also spent many years in the “real world” as an economist, risk manager, and fixed income analyst. (He prefers academia!)

Alan received his PhD in economics from Fordham University, and also holds an M.S. in financial engineering from New York University.

Author's Acknowledgments

I'd like to acknowledge several people who helped get this book put together. First, thanks goes to Lindsay Berg, my acquisitions editor. Thanks for giving me a chance to write this book. Thanks also goes to Katharine Dvorak, my development editor who kept me focused. And thanks to M. Higuera for providing a creative and sound technical review. You all were a great team . . . thank you.

Publisher's Acknowledgments

Executive Editor: Lindsay Berg

Senior Managing Editor: Kristie Pyles

Project Editor: Katharine Dvorak

Technical Editor: M. Higuera, Ed.D.

Production Editor: Saikarthick Kumarasamy

Cover Image: © kertlis/Getty Images

Take dummies with you everywhere you go!

Whether you are excited about e-books, want more from the web, must have your mobile apps, or are swept up in social media, dummies makes everything easier.



Find us online!



dummies.com

dummies®
A Wiley Brand

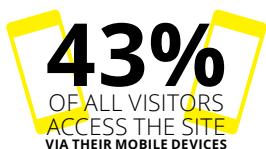
Leverage the power

Dummies is the global leader in the reference category and one of the most trusted and highly regarded brands in the world. No longer just focused on books, customers now have access to the dummies content they need in the format they want. Together we'll craft a solution that engages your customers, stands out from the competition, and helps you meet your goals.

Advertising & Sponsorships

Connect with an engaged audience on a powerful multimedia site, and position your message alongside expert how-to content. Dummies.com is a one-stop shop for free, online information and know-how curated by a team of experts.

- Targeted ads
- Video
- Email Marketing
- Microsites
- Sweepstakes
- sponsorship



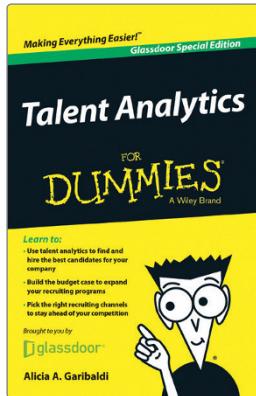
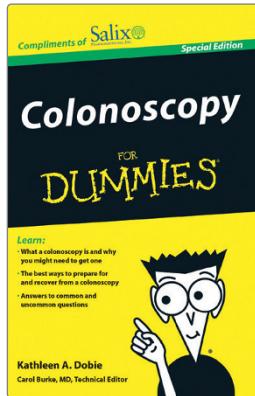
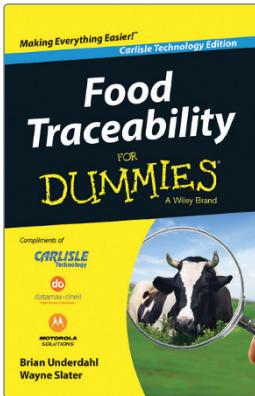
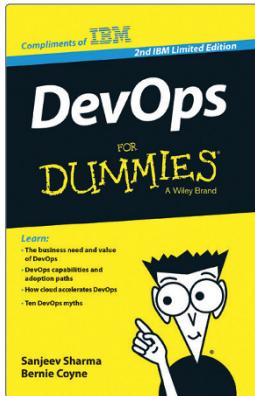
of dummies



Custom Publishing

Reach a global audience in any language by creating a solution that will differentiate you from competitors, amplify your message, and encourage customers to make a buying decision.

- Apps
- eBooks
- Audio
- Books
- Video
- Webinars



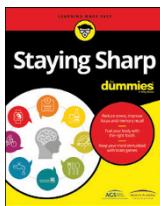
Brand Licensing & Content

Leverage the strength of the world's most popular reference brand to reach new audiences and channels of distribution.

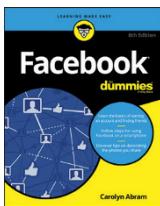
For more information, visit dummies.com/biz



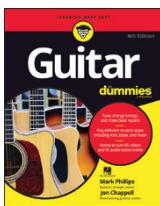
PERSONAL ENRICHMENT



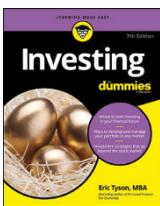
9781119187790
USA \$26.00
CAN \$31.99
UK £19.99



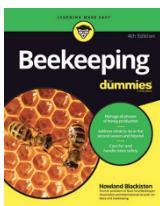
9781119179030
USA \$21.99
CAN \$25.99
UK £16.99



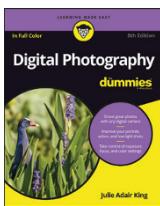
9781119293354
USA \$24.99
CAN \$29.99
UK £17.99



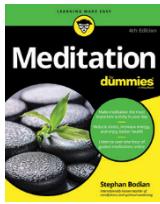
9781119293347
USA \$22.99
CAN \$27.99
UK £16.99



9781119310068
USA \$22.99
CAN \$27.99
UK £16.99



9781119235606
USA \$24.99
CAN \$29.99
UK £17.99



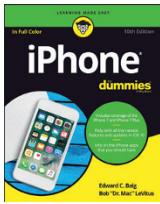
9781119251163
USA \$24.99
CAN \$29.99
UK £17.99



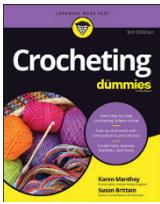
9781119235491
USA \$26.99
CAN \$31.99
UK £19.99



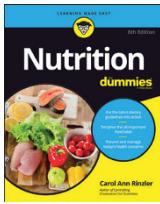
9781119279952
USA \$24.99
CAN \$29.99
UK £17.99



9781119283133
USA \$24.99
CAN \$29.99
UK £17.99

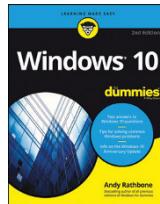


9781119287117
USA \$24.99
CAN \$29.99
UK £16.99

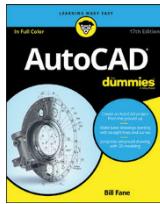


9781119130246
USA \$22.99
CAN \$27.99
UK £16.99

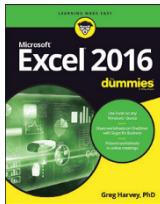
PROFESSIONAL DEVELOPMENT



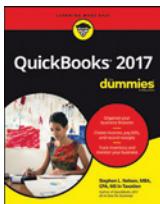
978111911041
USA \$24.99
CAN \$29.99
UK £17.99



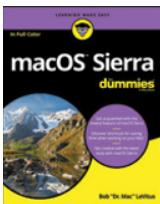
9781119255796
USA \$39.99
CAN \$47.99
UK £27.99



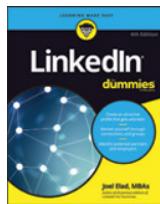
9781119293439
USA \$26.99
CAN \$31.99
UK £19.99



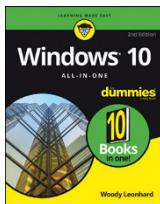
9781119281467
USA \$26.99
CAN \$31.99
UK £19.99



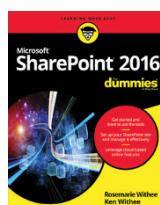
9781119280651
USA \$29.99
CAN \$35.99
UK £21.99



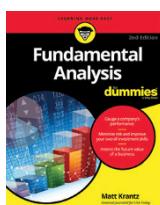
9781119251132
USA \$24.99
CAN \$29.99
UK £17.99



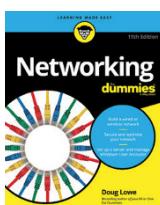
9781119130563
USA \$34.00
CAN \$41.99
UK £24.99



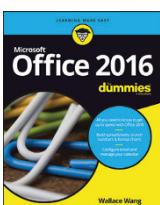
9781119181705
USA \$29.99
CAN \$35.99
UK £21.99



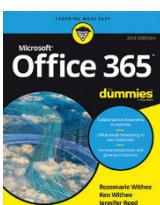
9781119263593
USA \$26.99
CAN \$31.99
UK £19.99



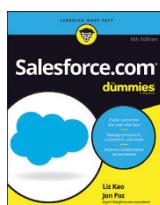
9781119257769
USA \$29.99
CAN \$35.99
UK £21.99



9781119293477
USA \$26.99
CAN \$31.99
UK £19.99



9781119265313
USA \$24.99
CAN \$29.99
UK £21.99



9781119239314
USA \$29.99
CAN \$35.99
UK £21.99

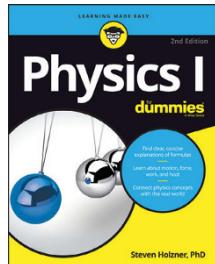
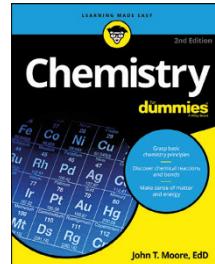
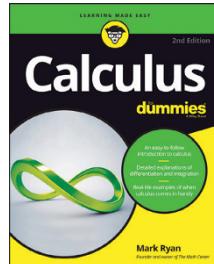
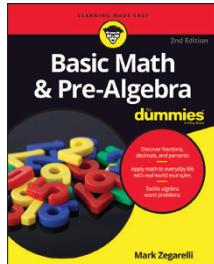
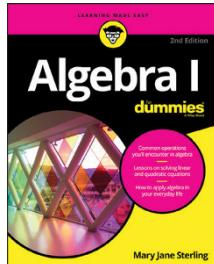


9781119293323
USA \$29.99
CAN \$35.99
UK £21.99

Learning Made Easy



ACADEMIC



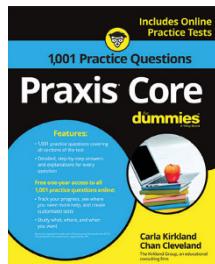
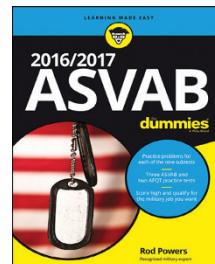
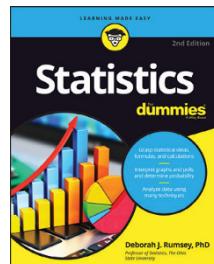
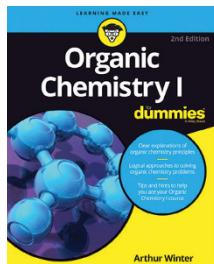
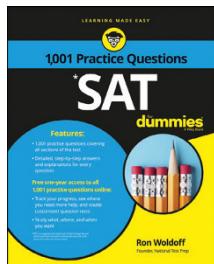
9781119293576
USA \$19.99
CAN \$23.99
UK £15.99

9781119293637
USA \$19.99
CAN \$23.99
UK £15.99

9781119293491
USA \$19.99
CAN \$23.99
UK £15.99

9781119293460
USA \$19.99
CAN \$23.99
UK £15.99

9781119293590
USA \$19.99
CAN \$23.99
UK £15.99



9781119215844
USA \$26.99
CAN \$31.99
UK £19.99

9781119293378
USA \$22.99
CAN \$27.99
UK £16.99

9781119293521
USA \$19.99
CAN \$23.99
UK £15.99

9781119239178
USA \$18.99
CAN \$22.99
UK £14.99

9781119263883
USA \$26.99
CAN \$31.99
UK £19.99

Available Everywhere Books Are Sold

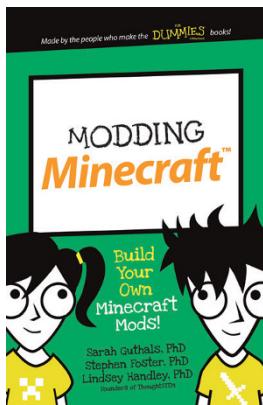
dummies.com

dummies®
A Wiley Brand

Small books for big imaginations



9781119177173
USA \$9.99
CAN \$9.99
UK £8.99



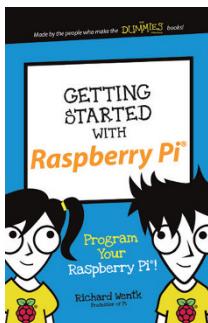
9781119177272
USA \$9.99
CAN \$9.99
UK £8.99



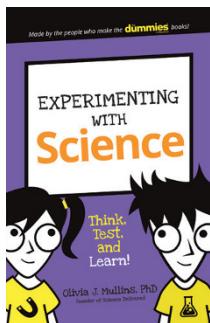
9781119177241
USA \$9.99
CAN \$9.99
UK £8.99



9781119177210
USA \$9.99
CAN \$9.99
UK £8.99



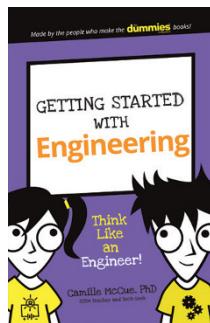
9781119262657
USA \$9.99
CAN \$9.99
UK £6.99



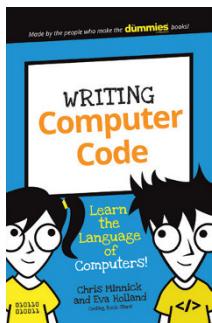
9781119291336
USA \$9.99
CAN \$9.99
UK £6.99



9781119233527
USA \$9.99
CAN \$9.99
UK £6.99



9781119291220
USA \$9.99
CAN \$9.99
UK £6.99



9781119177302
USA \$9.99
CAN \$9.99
UK £8.99

Unleash Their Creativity

dummies.com

dummies[®]
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.