LEARNING MADE EASY

2nd Edition

Statistics II

for dummies®

A Wiley Brand

Learn to analyze big data sets

Explore important intermediate statistical techniques

Use statistical software for real-world applications

Deborah J. Rumsey, PhD

Associated Professor of Statistics
The Ohio State University

Statistics II

for dummies®
A Wiley Brand

Statistics II

2nd Edition

by Deborah J. Rumsey, PhD

for dummies®
A Wiley Brand

Statistics II For Dummies®, 2nd Edition

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

Copyright © 2022 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. Wiley Publishing, 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: 2021946535

ISBN 978-1-119-82739-9 (pbk); ISBN 978-1-119-82740-5 (ebk); ISBN 978-1-119-82741-2 (ebk)

Contents at a Glance

Table of Contents

Introduction

So you've gone through some of the basics of statistics. Means, medians, and standard deviations all ring a bell. You know about surveys and experiments and the basic ideas of correlation and simple regression. You've studied probability, margin of error, and a few hypothesis tests and confidence intervals. Are you ready to load your statistical toolbox with a new level of tools? *Statistics II For Dummies,* 2nd Edition, picks up right where *Statistics For Dummies,* 2nd Edition, (John Wiley & Sons) leaves off and keeps you moving along the road of statistical ideas and techniques in a positive, step-by-step way.

The focus of *Statistics II For Dummies,* 2nd Edition, is on finding more ways of analyzing data. I provide step-by-step instructions for using techniques such as multiple regression, nonlinear regression, one-way and two-way analysis of variance (ANOVA), and Chi-square tests, and I give you some practice with big data sets, which are all the rage right now. Using these new techniques, you estimate, investigate, correlate, and congregate even more variables based on the information at hand, and you see how to put the tools together to create a great story about your data (nonfiction, I hope!).

About This Book

This book is designed for those who have completed the basic concepts of statistics through confidence intervals and hypothesis testing (found in *Statistics For Dummies,* 2nd Edition) and are ready to plow ahead to get through the final part of Stats I, or to tackle Stats II. However, I do pepper in some brief overviews of Stats I as needed, just to remind you of what was covered and to make sure you're up to speed. For each new technique, you get an overview of when and why it's used, how to know when you need it, step-by-step directions on how to apply it, and tips and tricks from a seasoned data analyst (yours truly). Because it's very important to be able to know which method to use when, I emphasize what makes each technique distinct and what the results tell you. You will also see many applications of the techniques used in real life.

I also include interpretation of computer output for data analysis purposes. I show you how to use the software to get the results, but I focus more on how to interpret the results found in the output, because you're more likely to be interpreting this kind of information than doing the programming specifically. Because the equations and calculations can get too involved if you are solving them by hand, you often use a computer to get your results. I include instructions for using Minitab to conduct many of the calculations in this book. Most statistics teachers who cover these topics use this approach as well. (What a relief!)

This book is different from the other Stats II books in many ways. Notably, this book features the following:

- **Full explanations of Stats II concepts.** Many statistics textbooks squeeze all the Stats II topics at the very end of their Stats I coverage; as a result, these topics tend to get condensed and presented as if they're optional. But no worries; I take the time to clearly and fully explain all the information you need to survive and thrive.
- **Dissection of computer output.** Throughout the book, I present many examples that use statistical software to analyze the data. In each case, I present the computer output and explain how I got it and what it means.
- **An extensive number of examples.** I include plenty of examples to cover the many different types of problems you'll face. Some examples are short, and some are quite extensive and include multiple variables.
- **Lots of tips, strategies, and warnings.** I share with you some trade secrets, based on my experience teaching and supporting students and grading their papers.
- **Understandable language.** I try to keep things conversational to help you understand, remember, and put into practice statistical definitions, techniques, and processes.
- **Clear and concise, step-by-step procedures.** In most chapters, you can find steps that intuitively explain how to work through Stats II problems — and remember how to do it on your own later on.

Throughout this book, I've used several conventions that I want you to be aware of:

- I indicate multiplication by using a times sign, indicated by a lowered asterisk *.
- I indicate the null and alternative hypotheses as H_o (for the null hypothesis) and H_a (for the alternative hypothesis).
- The statistical software package I use and display throughout the book is Minitab 18, but I simply refer to it as Minitab.
- Whenever I introduce a new term, I *italicize* it.
- Keywords and numbered steps appear in **boldface.**

At times I get into some of the more technical details of formulas and procedures for those individuals who may need to know about them — or just really want to get the full story. These minutiae are marked with a Technical Stuff icon. I also include sidebars along with the essential text, usually in the form of a real-life statistics example or some bonus information you may find interesting. You can feel free to skip those icons and sidebars because you won't miss any of the main information you need (but by reading them, you may just be able to impress your stats professor with your above-and-beyond knowledge of Stats II!).

Foolish Assumptions

Because this book deals with Stats II, I assume you have one previous course in introductory statistics under your belt (or at least have read *Statistics For Dummies,* 2nd Edition), with topics taking you up through the Central Limit Theorem and perhaps an introduction to confidence intervals and hypothesis tests (although I review these concepts briefly in Chapter 4). Prior experience with simple linear regression isn't necessary. Only college algebra is needed for the math details. Some experience using statistical software is also a plus, but not required.

As a student, you may be covering these topics in one of two ways: either at the tail end of your Stats I course (perhaps in a hurried way, but in some way nonetheless); or through a two-course sequence in statistics in which the topics in this book are the focus of the second course. If so, this book provides you the information you need to do well in those courses.

You may simply be interested in Stats II from an everyday point of view, or perhaps you want to add to your understanding of studies and statistical results presented in the media. If this sounds like you, you can find plenty of real-world examples and applications of these statistical techniques in action, as well as cautions for interpreting them.

Icons Used in This Book

I use icons in this book to draw your attention to certain text features that occur on a regular basis. Think of the icons as road signs that you encounter on a trip. Some signs tell you about shortcuts, and others offer more information that you may need; some signs alert you to possible warnings, while others leave you with something to remember.

COMPUTER OUTPUT

When you see this icon, it means I'm explaining how to carry out that particular data analysis using Minitab. I also explain the information you get in the computer output so you can interpret your results.

REMEMBER

I use this icon to reinforce certain ideas that are critical for success in Stats II, such as things I think are important to review as you prepare for an exam.

TECHNICAL STUFF

When you see this icon, you can skip over the information if you don't want to get into the nitty-gritty details. They exist mainly for people who have a special interest or obligation to know more about the technical aspects of certain statistical issues.

TIP

This icon points to helpful hints, ideas, or shortcuts that you can use to save time; it also includes alternative ways to think about a particular concept.

WARNING

I use warning icons to help you stay away from common misconceptions and pitfalls you may face when dealing with ideas and techniques related to Stats II.

Beyond the Book

In addition to all the great content included in the book itself, you can find even more content online. Check out this book's online Cheat Sheet on dummies.com. It covers the major formulas needed for Statistics II. You can access it by going to `www.dummies.com` and then typing "Statistics II For Dummies Cheat Sheet" into the search bar.

I've also included two major data sets that are analyzed in Chapters 20 and 21, so you can follow along with me or do your own analysis (not required!). Go to `www.dummies.com/go/statisticsIIfd2e` to access these files.

Where to Go from Here

This book is written in a nonlinear way, so you can start anywhere and still understand what's happening. However, I can make some recommendations if you want some direction on where to start.

If you're thoroughly familiar with the ideas of hypothesis testing and simple linear regression, start with Chapter 5 (multiple regression). Use Chapter 1 if you need a reference for the jargon that statisticians use in Stats II.

If you've covered all topics up through the various types of regression (simple, multiple, nonlinear, and logistic) or a subset of those as your professor deemed important, proceed to Chapter 10, the basics of analysis of variance (ANOVA).

Chapter 15 is the place to begin if you want to tackle categorical (qualitative) variables before hitting the quantitative stuff. You can work with the Chi-square test there.

Nonparametric statistics are presented starting in Chapter 17. Start there if you want the full details on the most common nonparametric procedures, used when you do not necessarily have an assumed distribution (for example, a normal).

If you want to see a bunch of Stats II ideas put into practice right off the bat, head to Chapter 19 where I discuss a multi-stage approach to analyzing a big data set, or Chapter 21, where you look into a big data set on refrigerators and see how it's analyzed in a multi-stage approach.

1

Tackling Data Analysis and Model-Building Basics

IN THIS PART . . .

Understand why data analysis is both a science and an art.

Make sure you use the right type of analysis for the job.

Work with the normal and binomial distribtions.

Reaquaint yourself with confidence intervals and hypothesis tests.

IN THIS CHAPTER

- » **Realizing your role as a data analyst**
- » **Avoiding statistical faux pas**
- » **Delving into the jargon of Stats II**

Chapter 1

Beyond Number Crunching: The Art and Science of Data Analysis

Because you're reading this book, you're likely familiar with the basics of statistics and you're ready to take it up a notch. That next level involves using what you know, picking up a few more tools and techniques, and finally putting it all to use to help you answer more realistic questions by using real data. In statistical terms, you're ready to enter the world of the *data analyst.*

In this chapter, you review the terms involved in statistics as they pertain to data analysis at the Stats II level. You get a glimpse of the impact that your results can have by seeing what these analysis techniques can do. You also gain insight into some of the common misuses of data analysis and their effects.

Data Analysis: Looking before You Crunch

It used to be that statisticians were the only ones who really analyzed data because the only computer programs available were very complicated to use, requiring a great deal of knowledge about statistics to set up and carry out analyses.

The calculations were tedious and at times unpredictable, and they required a thorough understanding of the theories and methods behind the calculations to get correct and reliable answers.

Today, anyone who wants to analyze data can do it easily. Many user-friendly statistical software packages are made expressly for that purpose — Microsoft Excel, Minitab, and SAS are just a few. Free online programs are available, too, such as R, which helps you do just what it says — crunch your numbers and get an answer.

Each software package has its own pros and cons (and its own users and protesters). My software of choice and the one I reference throughout this book is Minitab, because it's very easy to use, the results are precise, and the software's loaded with all the data-analysis techniques used in Stats II. Although a site license for Minitab isn't cheap, the student version is available for rent for only a few bucks a semester.

REMEMBER

The most important idea when applying statistical techniques to analyze data is to know what's going on behind the number crunching so you (not the computer) are in control of the analysis. That's why knowledge of Stats II is so critical.

WARNING

Many people don't realize that statistical software can't tell you when and when not to use a certain statistical technique. You have to determine that on your own. As a result, people think they're doing their analyses correctly, but they can end up making all kinds of mistakes. In the following sections, I give examples of some situations in which innocent data analyses can go wrong and why it's important to spot and avoid these mistakes before you start crunching numbers.

Bottom line: Today's software packages really are too good to be true if you don't have a clear and thorough understanding of the Stats II that's beneath the surface.

Nothing (not even a straight line) lasts forever

Bill Prediction is a statistics student who is studying the effect of study time on a student's exam score. Bill collects data on statistics students and uses his trusty software package to predict exam scores based on study time. His computer comes up with the equation $y = 10x + 30$, where y represents the test score you get if you study for a certain number of hours (x). Notice that this model is the equation of a straight line with a y-intercept of 30 and a slope of 10.

So using this model, Bill predicts that if you don't study at all, you'll get a 30 on the exam (plugging $x = 0$ into the equation and solving for y; this point represents the *y*-intercept of the line). He also predicts, using this model, that if you study for 5 hours, you'll get an exam score of $y = (10 * 5) + 30 = 80$ So, the point (5,80) is also on this line.

But then Bill goes a little crazy and wonders what would happen if you studied for 40 hours (because it always seems that long when he's studying). The computer tells him that if he studies for 40 hours, his test score is predicted to be $(10 * 40) + 30 = 430$ points. Wow, that's a lot of points! Problem is, the exam only goes up to a total of 100 points. Bill wonders where his computer went wrong.

But Bill puts the blame in the wrong place. He needs to remember that there are limits on the values of *x* that make sense in this equation. For example, because *x* is the amount of study time, *x* can never be a number less than zero. If you plug a negative number in for *x*, say $x = -10$, you get $y = (10 * -10) + 30 = -70$, which makes no sense. However, the equation itself doesn't know that, nor does the computer that found it. The computer simply graphs the line you give it, assuming it'll go on forever in both the positive and negative directions.

WARNING

After you get a statistical equation or model, you need to specify for what values the equation applies. Equations don't know when they work and when they don't; it's up to the data analyst to determine that. This idea is the same for applying the results of any data analysis that you do.

Data snooping isn't cool

WARNING

Statisticians have come up with a saying that you may have heard: "Figures don't lie. Liars figure." Make sure that you find out about all the analyses that were performed on a data set, not just the ones reported as being statistically significant.

Suppose Bill Prediction (from the previous section) decides to try to predict scores on a biology exam based on study time, but this time his model doesn't fit. Not one to give in, Bill insists there must be some other factors that predict biology exam scores besides study time, and he sets out to find them.

Bill measures everything from soup to nuts. His set of 20 possible variables includes study time, GPA, previous experience in statistics, math grades in high school, and whether you chew gum during the exam. After his multitude of various correlation analyses, the variables that Bill finds to be related to exam score are study time, math grades in high school, GPA, and gum chewing during the exam. It turns out that this particular model fits pretty well (by criteria I discuss in Chapter 6 on multiple linear regression models).

But here's the problem: By looking at all possible correlations between his 20 variables and the exam score, Bill is actually doing 20 separate statistical analyses. Under typical conditions that I describe in Chapter 4, each statistical analysis has a 5 percent chance of being wrong just by chance. I bet you can guess which one of Bill's correlations likely came out wrong in this case. And hopefully, he didn't rely on a stick of gum to boost his grade in biology.

WARNING

Looking at data until you find something in it is called *data snooping.* Data snooping results in giving the researcher his five minutes of fame but then leads him to lose all credibility because no one can repeat his results.

No (data) fishing allowed

Some folks just don't take no for an answer, and when it comes to analyzing data, that can lead to trouble.

Sue Gonnafindit is a determined researcher. She believes that her horse can count by stomping his foot. (For example, she says "2" and her horse stomps twice.) Sue collects data on her horse for four weeks, recording the percentage of time the horse gets the counting right. She runs the appropriate statistical analysis on her data and is shocked to find no significant difference between her horse's results and those you would get simply by guessing.

Determined to prove her results are real, Sue looks for other types of analyses that exist and plugs her data into anything and everything she can find (never mind that those analyses are inappropriate to use in her situation). Using the famous hunt-and-peck method, at some point she eventually stumbles upon a significant result. However, the result is bogus because she tried so many analyses that weren't appropriate and ignored the results of the appropriate analysis because it didn't tell her what she wanted to hear.

Funny thing, too. When Sue went on a late-night TV program to show the world her incredible horse, someone in the audience noticed that whenever the horse got to the correct number of stomps, Sue would interrupt him and say "Good job!" and the horse quit stomping. He didn't know how to count; all he knew to do was to quit stomping when she said, "Good job!"

WARNING

Redoing analyses in different ways in order to try to get the results you want is called *data fishing,* and folks in the stats biz consider it to be a major no-no. (However, people unfortunately do it all too often to verify their strongly held beliefs.) By using the wrong data analysis for the sake of getting the results you desire, you mislead your audience into thinking that your hypothesis is actually correct when it may not be.

Getting the Big Picture: An Overview of Stats II

Stats II is an extension of Stats I (introductory statistics), so the jargon follows suit and the techniques build on what you already know. In this section, you get an introduction to the terminology you use in Stats II along with a broad overview of the techniques that statisticians use to analyze data and find the story behind it. (If you're still unsure about some of the terms from Stats I, you can consult your Stats I textbook or see my other book, *Statistics For Dummies, 2nd Edition* [Wiley], for a complete rundown.)

Population parameter

REMEMBER

A *parameter* is a number that summarizes the *population*, which is the entire group you're interested in investigating. Examples of parameters include the mean of a population, the median of a population, or the proportion of the population that falls into a certain category.

Suppose you want to determine the average length of a cellphone call among teenagers (ages 13–18). You're not interested in making any comparisons; you just want to make a good guesstimate of the average time. So you want to estimate a population parameter (such as the mean or average). The population is all cellphone users between the ages of 13 and 18 years old. The parameter is the average length of a phone call this population makes.

Sample statistic

Typically you can't determine population parameters exactly; you can only estimate them. But all is not lost; by taking a representative *sample* (a well-chosen subset of individuals) from the population and studying it, you can come up with a good estimate of the population parameter. A *sample statistic* is a single number that summarizes that subset.

For example, in the cellphone scenario from the previous section, you select a sample of teenagers and measure the duration of their cellphone calls over a period of time (or look at their cellphone records if you can gain access legally). You take the average of the cellphone call duration. For example, the average duration of 100 cellphone calls may be 12.2 minutes — this average is a statistic. This particular statistic is called the *sample mean* because it's the average value from your sample data.

Many different statistics are available to study different characteristics of a sample, such as the proportion, the median, and standard deviation.

Confidence interval

A *confidence interval* is a range of likely values for a population parameter. A confidence interval is based on a sample and the statistics that come from that sample. The main reason you want to provide a range of likely values rather than a single number is that sample results vary.

For example, suppose you want to estimate the percentage of people who eat chocolate. According to the Simmons Market Research Bureau, 78 percent of adults reported eating chocolate, and of those, 18 percent admitted eating sweets frequently. What's missing in these results? These numbers are only from a single sample of people, and those sample results are guaranteed to vary from sample to sample. You need some measure of how much you can expect those results to move if you were to repeat the study.

This expected variation in your statistic from sample to sample is measured by the *margin of error*, which reflects a certain number of standard deviations of your statistic that you add and subtract to have a certain confidence in your results (see Chapter 4 for more on margin of error). If the chocolate-eater results were based on 1,000 people, the margin of error would be approximately 3 percent. This means the actual percentage of people who eat chocolate in the entire population is expected to be 78 percent, ± 3 percent (that is, between 75 percent and 81 percent).

Hypothesis test

A *hypothesis test* is a statistical procedure that you use to test an existing claim about the population, using your data. The claim is noted by H_o (the null hypothesis). If your data support the claim, you fail to reject H_o. If your data don't support the claim, you reject H_o and conclude an alternative hypothesis, H_a. The reason most people conduct a hypothesis test is not to merely show that their data support an existing claim, but rather to show that the existing claim is false, in favor of the alternative hypothesis.

The Pew Research Center studied the percentage of people who turn to ESPN for their sports news. Its statistics, based on a survey of about 1,000 people, found that in 2000, 23 percent of people said they went to ESPN; in 2020, only 20.9 percent reported going to ESPN. The question is this: Does this 2.1 percent reduction in viewers represent a significant trend that ESPN should worry about?

To test these differences formally, you can set up a hypothesis test. You set up your null hypothesis as the result you have to believe without your study, H_o = No difference exists between 2000 and 2020 data for ESPN viewership. Your alternative hypothesis (H_a) is that a difference is there. To run a hypothesis test, you look at the difference between your statistic from your data and the claim that has been already made about the population (in H_o), and you measure how far apart they are in units of standard deviations.

With respect to the example, using the techniques from Chapter 4, the hypothesis test shows that 23 percent and 20.9 percent aren't far enough apart in terms of standard deviations to dispute the claim (H_o). You can't say the percentage of viewers of ESPN in the entire population changed from 2000 to 2020.

REMEMBER

As with any statistical analysis, your conclusions can be wrong just by chance, because your results are based on sample data, and sample results vary. In Chapter 4, I discuss the types of errors that can be made in conclusions from a hypothesis test.

Analysis of variance (ANOVA)

ANOVA is the acronym for *analysis of variance.* You use ANOVA in situations where you want to compare the means of more than two populations. For example, say you want to compare the lifetimes of four brands of tires in number of miles. You take a random sample of 50 tires from each group, for a total of 200 tires, and set up an experiment to compare the lifetime of each tire, and record it. You now have four means and four standard deviations, one for each data set.

Then, to test for differences in average lifetime for the four brands of tires, you basically compare the variability between the four data sets to the variability within the entire data set, using a ratio. This ratio is called the *F-statistic.* If this ratio is large, the variability between the brands is more than the variability within the brands, giving evidence that not all the means are the same for the different tire brands. If the *F*-statistic is small, not enough difference exists between the treatment means compared to the general variability within the treatments (here the brands) themselves. In this case, you can't say that the means are different for the groups. (I give you the full scoop on ANOVA plus all the jargon, formulas, and computer output in Chapters 10 and 11.)

Multiple comparisons

Suppose you conduct ANOVA, and you find a difference in the average lifetimes of the four brands of tire (see the preceding section). Your next questions would probably be, "Which brands are different?" and "How different are they?" To answer these questions, you use multiple-comparison procedures.

A *multiple-comparison procedure* is a statistical technique that compares means to each other and finds out which ones are different and which ones aren't. With this information, you're able to put the groups in order from those with the largest mean to those with the smallest mean, realizing that sometimes two or more groups were too close to tell and are placed together in a group.

Many different multiple-comparison procedures exist to compare individual means and come up with an ordering in the event that your *F*-statistic does find that some difference exists. Some of the multiple-comparison procedures include Tukey's test, LSD (least significant difference), and pairwise *t*-tests. Some procedures are better than others, depending on the conditions and your goal as a data analyst. I discuss multiple-comparison procedures in detail in Chapter 11.

WARNING

Never take that second step to compare the means of the groups if the ANOVA procedure doesn't find any significant results during the first step. Computer software will never stop you from doing a follow-up analysis, even if it's wrong to do so.

Interaction effects

An *interaction effect* in statistics operates the same way that it does in the world of medicine. Sometimes if you take two different medicines at the same time, the combined effect is much different than if you were to take the two individual medications separately.

REMEMBER

Interaction effects can come up in statistical models that use two or more variables to explain or compare outcomes. In this case you can't automatically study the effect of each variable separately; you have to first examine whether or not an interaction effect is present.

For example, suppose medical researchers are studying a new drug for depression and want to know how this drug affects the change in blood pressure for a low dose versus a high dose. They also compare the effects for children versus adults. It could also be that dosage level affects the blood pressure of adults differently than the blood pressure of children. This type of model is called a *two-way ANOVA model,* with a possible interaction effect between the two factors (age group and dosage level). Chapter 12 covers this subject in depth.

Correlation

The term *correlation* is often misused. Statistically speaking, the correlation measures the strength and direction of the linear relationship between two *quantitative variables* (variables that represent counts or measurements only).

REMEMBER

You aren't supposed to use correlation to talk about relationships unless the variables are quantitative. For example, it's wrong to say that a correlation exists between eye color and hair color. (In Chapter 14, you explore associations between two categorical variables.)

Correlation is a number between −1.0 and +1.0. A correlation of +1.0 indicates a perfect positive relationship; as you increase one variable, the other one increases in perfect sync. A correlation of −1.0 indicates a perfect negative relationship between the variables; as one variable increases, the other one decreases in perfect sync. A correlation of zero means you found no linear relationship at all between the variables. Most correlations in the real world fall somewhere in between −1.0 and +1.0; the closer to −1.0 or +1.0, the stronger the relationship is; the closer to 0, the weaker the relationship is.

Figure 1-1 shows a plot of the number of coffees sold at football games in Buffalo, New York, as well as the air temperature (in degrees Fahrenheit) at each game. This data set seems to follow a downhill straight line fairly well, indicating a negative correlation. The correlation turns out to be −0.741; the number of coffees sold has a fairly strong negative relationship with the temperature of the football game. This makes sense because on days when the temperature is low, people get cold and want more coffee. I discuss correlation further, as it applies to model building, in Chapter 5.

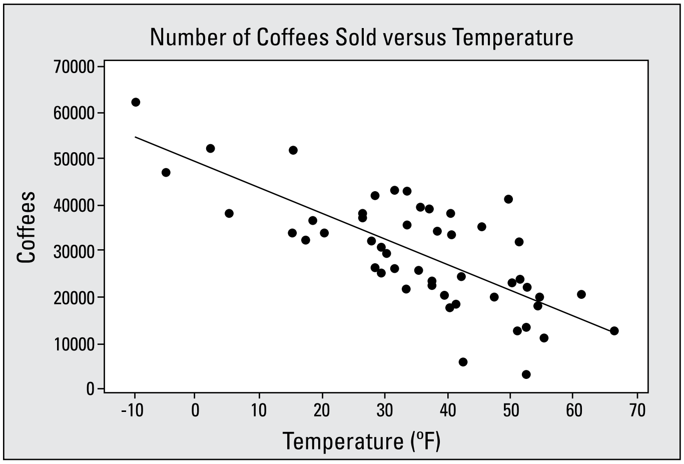


FIGURE 1-1: Coffees sold at various air temperatures on football game day.

Linear regression

After you've found a correlation and determined that two variables have a fairly strong linear relationship, you may want to try to make predictions for one variable based on the value of the other variable. For example, if you know that a fairly strong negative linear relationship exists between coffees sold and the air temperature at a football game (see the previous section), you may want to use this information to predict how much coffee is needed for a game, based on the temperature. This method of finding the best-fitting line is called *linear regression.*

Many different types of regression analyses exist, depending on your situation. When you use only one variable to predict the response, the method of regression is called *simple linear regression* (see Chapter 5). Simple linear regression is the best known of all the regression analyses and is a staple in the Stats I course sequence.

However, you use other flavors of regression for other situations.

- If you want to use more than one variable to predict a response, you use *multiple linear regression* (see Chapter 6).
- If you want to make predictions about a variable that has only two outcomes, yes or no, you use *logistic regression* (see Chapter 9).
- For relationships that don't follow a straight line, you have a technique called (no surprise) *nonlinear regression* (see Chapter 8).

Chi-square tests

Correlation and regression techniques all assume that the variable being studied in most detail (the response variable) is quantitative — that is, the variable measures or counts something. You can also run into situations where the data being studied isn't quantitative, but rather categorical — that is, the data represents categories, not measurements or counts. To study relationships in categorical data, you use a Chi-square test for independence. If the variables are found to be unrelated, they're declared independent. If they're found to be related, they're declared dependent.

Suppose you want to explore the relationship between age group and eating breakfast. Because each of these variables is categorical, or qualitative, you use a Chi-square test for independence. You survey 70 adults and 70 children and find that 25 adults eat breakfast and 45 do not; for the children, 35 do eat breakfast and 35 do not. Table 1-1 organizes this data and sets you up for the Chi-square test for this scenario.

TABLE 1-1 Table Setup for the Breakfast and Age Group Question

	Do Eat Breakfast	Don't Eat Breakfast	Total
Adult	25	45	70
Child	35	35	70

REMEMBER

A Chi-square test first calculates what you expect to see in each cell of the table if the variables are independent (these values are brilliantly called the *expected cell counts*). The Chi-square test then compares these expected cell counts to what you observed in the data (called the *observed cell counts*) and compares them using a Chi-square statistic.

In the breakfast age-group comparison, fewer adults than children eat breakfast ($25/70 = 35.7$ percent compared to $35/70 = 50$ percent). Even though you know results will vary from sample to sample, this difference turns out to be enough to declare a relationship between age group and eating breakfast, according to the Chi-square test of independence. Chapter 15 reveals all the details of doing a Chi-square test.

You can also use the Chi-square test to see whether your theory about what percent of each group falls into a certain category is true or not. For example, can you guess what percentage of M&M'S fall into each color category? You can find more on these Chi-square variations, as well as the M&M'S question, in Chapter 16.

IN THIS CHAPTER

- Deciphering the difference between categorical and quantitative variables
- Choosing appropriate statistical techniques for the task at hand
- Evaluating bias and precision levels
- Interpreting the results properly

Chapter 2

Finding the Right Analysis for the Job

One of the most critical elements of statistics and data analysis is the ability to choose the right statistical technique for each job. Carpenters and mechanics know the importance of having the right tool when they need it and the problems that can occur if they use the wrong tool. They also know that the right tool helps to increase their odds of getting the results they want the first time around, using the "work smarter, not harder" approach.

In this chapter, you look at some of the major statistical analysis techniques from the point of view of the carpenters and mechanics — knowing what each statistical tool is meant to do, how to use it, and when to use it. You also zoom in on mistakes some number crunchers make in applying the wrong analysis or doing too many analyses.

REMEMBER

Knowing how to spot these problems can help you avoid making the same mistakes, but it also helps you to steer through the ocean of statistics that may await you in your job and in everyday life.

If many of the ideas you find in this chapter seem like a foreign language to you and you need more background information, don't fret. Before continuing on in this chapter, head to your nearest Stats I book or check out another one of my books, *Statistics For Dummies, 2nd Edition* (Wiley).

Categorical versus Quantitative Variables

After you've collected all the data you need from your sample, you want to organize it, summarize it, and analyze it. Before plunging right into all the number crunching, though, you need to first identify the type of data you're dealing with. The type of data you have points you to the proper types of graphs, statistics, and analyses you're able to use.

TECHNICAL STUFF

Before I begin, here's an important piece of jargon: Statisticians call any quantity or characteristic you measure on an individual a *variable*; the data collected on a variable is expected to vary from person to person (hence the creative name).

There are two major types of variables:

- **Categorical.** A *categorical variable,* also known as a *qualitative variable,* classifies the individual based on categories. For example, political affiliation may be classified into four categories: Democrat, Republican, Independent, and Other. Similarly, type of pet can take on three categories: Cat, Dog, and Other. Categorical variables can take on numerical values only as placeholders; the numbers themselves don't mean anything special.
- **Quantitative.** A *quantitative variable* measures or counts a quantifiable characteristic, such as height, weight, number of children you have, your GPA in college, or the number of hours of sleep you got last night. The quantitative variable value represents a quantity (count) or a measurement and has numerical meaning. That is, you can add, subtract, multiply, or divide the values of a quantitative variable, and the results make sense as numbers.

Because the two types of variables represent such different types of data, it makes sense that each type has its own set of statistics. Categorical variables, such as political affiliation, are somewhat limited in terms of the statistics that can be performed on them.

For example, suppose you have a sample of 500 classmates classified by dominant hand — 80 are left-handed and 420 are right-handed. How can you summarize this information? You already have the total number in each category (this statistic is called the *frequency*). You're off to a good start, but frequencies are hard to interpret because you find yourself trying to compare them to a total in your mind in order to get a proper comparison. For example, in this case you may be thinking, "Eighty left-handers out of what? Let's see, it's out of 500. Hmmm . . . what percentage is that?"

The next step is to find a means to relate these numbers to each other in an easy way. You can do this by using the *relative frequency*, which is the percentage of data that falls into a specific category of a categorical variable. You can find a category's relative frequency by dividing the frequency by the sample total and then multiplying by 100. In this case, you have $\frac{80}{500} = 0.16 * 100 = 16$ percent left-handers and $\frac{420}{500} = 0.84 * 100 = 84$ percent right-handers in the class.

You can also express the relative frequency as a proportion in each group by leaving the result in decimal form and not multiplying by 100. This statistic is called the *sample proportion.* In this example, the sample proportion of left-handed students is 0.16, and the sample proportion of right-handed students is 0.84.

REMEMBER

You mainly summarize categorical variables by using two statistics: the number in each category (frequency) and the percentage in each category (relative frequency).

Statistics for Categorical Variables

The types of statistics done on categorical data may seem limited; however, the wide variety of analyses you can perform using frequencies and relative frequencies offers answers to an extensive range of possible questions you may want to explore.

In this section, you see that the proportion in each group is the number-one statistic for summarizing categorical data. Beyond that, you see how you can use proportions to estimate, compare, and look for relationships between the groups that comprise the categorical data.

Estimating a proportion

You can use relative frequencies to make estimates about a single population proportion. (Refer to the earlier section, "Categorical versus Quantitative Variables," for an explanation of relative frequencies.)

Suppose you want to know what proportion of registered voters in the United States identify as Democrats, Republicans, and Independents. According to a random sample of 12,000 registered voters in the U.S. conducted by the Pew Research Center in 2019, the percentage of Democrat, Republican, and Independent registered voters was 33 percent, 29 percent, and 34 percent, respectively. Now, because the Pew researchers based these results on only a sample of the population and not on the entire population, their results will vary if they take another random sample of 12,000 people. This variation in sample results is cleverly called — you guessed it — *sampling variability*.

The sampling variability is measured by the *margin of error* (the amount that you add and subtract from your sample statistic), which for this sample is only about 0.9 percent. (To find out how to calculate margin of error, turn to Chapter 4.) That means, for example, that the estimated percentage of all registered voters in the U.S. identifying as Democrat is somewhere between $33 - .9 = 32.1$ percent and $33 + .9 = 33.9$ percent.

REMEMBER

The margin of error, combined with the sample proportion, forms what statisticians call a confidence interval for the population proportion. Recall from Stats I that a *confidence interval* is a range of likely values for a population parameter, formed by taking the sample statistic plus or minus the margin of error. (For more on confidence intervals, see Chapter 4.)

Comparing proportions

Researchers, the media, and even everyday folks like you and me love to compare groups (whether you like to admit it or not). For example, what proportion of Democrats support oil drilling in Alaska, compared to Republicans? What percentage of women watch college football, compared to men? What proportion of readers of *Statistics II For Dummies, 2nd Edition* pass their stats exams with flying colors, compared to nonreaders?

To answer these questions, you need to compare the sample proportions using a hypothesis test for two proportions (see Chapter 4 or your Stats I textbook).

Suppose you've collected data on a random sample of 1,000 voters in the U.S. (who identify as male or female), and you want to compare the proportion of female voters to the proportion of male voters and find out how they compare. Suppose that in your sample, you find that the proportion of females is 0.53, and the proportion of males is 0.47. So for this sample of 1,000 people, you have a higher proportion of females than males.

But here's the big question: Are these sample proportions different enough to say that the entire population of American voters has more females in it than males? After all, sample results vary from sample to sample. The answer to this question requires comparing the sample proportions by using a hypothesis test for two proportions. I demonstrate and expand on this technique in Chapter 4.

Looking for relationships between categorical variables

Suppose you want to know whether two categorical variables are related; for example, is gender related to political affiliation? Answering this question requires putting the sample data into a two-way table (using rows and columns to represent the two variables) and analyzing the data by using a Chi-square test (see Chapter 15).

By following this process, you can determine if two categorical variables are independent (unrelated) or if a relationship exists between them. If you find a relationship, you can use percentages to describe it.

Table 2-1 shows an example of data organized in a two-way table. The data was collected by the Pew Research Center.

TABLE 2-1 Gender and Political Affiliation of 56,735 U.S. Voters

Gender	Republican	Democrat	Other
Males	32%	27%	41%
Females	29%	36%	35%

Notice that the percentage of male Republicans in the sample is 32 and the percentage of female Republicans in the sample is 29. These percentages are quite close in relative terms. However, the percentage of female Democrats seems much higher than the percentage of male Democrats (36 percent versus 27 percent); also, the percentage of males in the "Other" category is quite a bit higher than the percentage of females in the same category (41 percent versus 35 percent).

These large differences in the percentages indicate that gender and political affiliation are related in the sample. But do these trends carry over to the population of all American voters? This question requires a hypothesis test to answer. Because gender and political affiliation are both categorical variables, the particular hypothesis test you need in this situation is a Chi-square test. (I discuss Chi-square tests in detail in Chapter 15.)

COMPUTER OUTPUT

To make a two-way table from a data set by using Minitab, first enter the data in two columns, where column one is the row variable (in this case, gender) and column two is the column variable (in this case, political affiliation). For example, suppose the first person is a male Democrat. In row one of Minitab, enter *M* (for male) in column one and *D* (Democrat) in column two. Then go to Stat>Tables>Cross Tabulation and Chi-square. Click in the ROWS box, then move to the large box on the left where the variables are listed and double-click your row variable name to select it, click in the COLUMNS box, move to the variable list, and double-click your column variable name to select it. Click OK.

WARNING

People often use the word *correlation* to discuss relationships between variables, but in the world of statistics, correlation only relates to the relationship between two quantitative (numerical) variables, not two categorical variables. *Correlation* measures how closely the relationship between two quantitative variables, such as height and weight, follows a straight line and tells you the direction of that line as well. In total, for any two quantitative variables, x and y, the correlation measures the strength and direction of their linear relationship. As one increases, what does the other one do?

Because categorical variables don't have a numerical order to them, they don't increase or decrease in value. For example, just because $\text{cat} = 1$ and $\text{dog} = 2$ doesn't mean that a dog is worth twice as much as a cat (although some dog owners want to disagree). Therefore, you can't use the word *correlation* to describe the relationship between, say, gender and political affiliation. (Chapter 5 covers correlation.)

The appropriate term to describe the relationships of categorical variables is *association.* You can say that political affiliation is associated with gender and then explain how. (For full details on association, see Chapter 14.)

Building models to make predictions

You can build models to predict the value of a categorical variable based on other related information. In this case, building models involves more than working with a lot of little plastic pieces and some irritatingly sticky glue.

When you build a statistical model, you look for variables that help explain, estimate, or predict some response you're interested in; the variables that do this are called *explanatory variables.* You sort through the explanatory variables and figure out which ones do the best job of predicting the response. Then you put them together into a type of equation like $y = 2x + 4$ where $x = \text{shoe size}$ and $y = \text{estimated calf length}$. That equation is a *model.*

For example, suppose you want to know which factors or variables can help you predict someone's political affiliation. Is someone without children more likely to be a Republican or a Democrat? What about a middle-aged person who proclaims Hinduism as their religion?

In order for you to compare these complex relationships, you must build a model to evaluate each group's impact on political affiliation (or some other categorical variable). This kind of model-building is explored in depth in Chapter 9, where I discuss the topic of logistic regression.

TIP

Logistic regression builds models to predict the outcome of a categorical variable, such as political affiliation. If you want to make predictions about a quantitative variable, such as income, you need to use the standard type of regression (check out Chapters 5 and 6).

Statistics for Quantitative Variables

Quantitative variables, unlike categorical variables, have a wider range of statistical functions that you can perform, depending on what questions you want to ask. The main reason for this wider range is that *quantitative data* are numbers that represent measurements or counts, so it makes sense that you can order, add or subtract, and multiply or divide them — and the results all have numerical meaning. In this section, I present the major data-analysis techniques for quantitative data. I expand on each technique in later chapters of this book.

Making estimates

Quantitative variables take on numerical values that involve counts or measurements, so they have means, medians, standard deviations, and all those good things that categorical variables don't have. Researchers often want to know what the average or median value is for a population (these are called *parameters*). To do this requires taking a sample and making a good guess, also known as an *estimate*, of that parameter.

To find an estimate for any population parameter requires a confidence interval. For quantitative variables, you would find a confidence interval to estimate the population mean, median, or standard deviation, but by far the most common parameter of interest is the population mean.

A confidence interval for the population mean is the sample mean plus or minus a margin of error. (To calculate the margin of error in this case, see Chapter 4.) The result will be a range of likely values you have produced for the real population mean. Because the variable is quantitative, the confidence interval will take on the same units as the variable does. For example, household incomes will be in thousands of dollars.

There is no rule of thumb regarding how large or small the margin of error should be for a quantitative variable; it depends on what the variable is counting or measuring. For example, if you want average household income for the state of New York, a margin of error of plus or minus $5,000 is not unreasonable. If the variable is the average number of steps from the first floor to the second floor of a two-story home in the U.S., the margin of error will be much smaller. Estimates of categorical variables, on the other hand, are percentages; most people want those confidence intervals to be within plus or minus 2 to 3 percent.

Making comparisons

Suppose you want to look at income (a quantitative variable) and how it relates to a categorical variable, such as identifying gender or region of the country. One question may be: Do people that live on the coasts make more money than the people who don't? In this case, you can compare the mean incomes of two populations — those that live on the coasts and those that don't. This assessment requires a hypothesis test of two means (often called a *t*-test for independent samples). I present more information on this technique in Chapter 4.

WARNING

When comparing the means of *more* than two groups, don't simply look at all the possible *t*-tests that you can do on the pairs of means, because you have to control for an overall error rate in your analysis. Too many analyses can result in errors — adding up to disaster. For example, if you conduct 100 hypothesis tests, each one with a 5 percent error rate, then 5 of those 100 tests will come out statistically significant on average, just by chance, even if no real relationship exists.

If you want to compare the average wage in different regions of the country (the East, the Midwest, the South, and the West, for example), this comparison requires a more sophisticated analysis because you're looking at four groups rather than just two. The procedure for comparing more than two means is called *analysis of variance* (ANOVA, for short), and I discuss this method in detail in Chapters 10 and 11.

Exploring relationships

One of the most common reasons why data is collected is to look for relationships between variables. With quantitative variables, the most common type of relationship people look for is a linear relationship; that is, as one variable increases, does the other increase or decrease along with it in a similar way? Relationships between any variables are examined using specialized plots and statistics. Because a linear relationship is so common, it has its own special statistic called *correlation*. You find out how statisticians make graphs and statistics to explore relationships in this section, paying particular attention to linear relationships.

Suppose you're an avid golfer and you want to figure out how much time you should spend on your putting game. The question is this: Is the number of putts related to your total score? If the answer is yes, then spending time on your putting game makes sense. If not, then you can slack off on it. These are both quantitative variables, and you're looking for a connection between them. You collect data on 100 rounds of golf played by golfers at your favorite course over a weekend. Following are the first few lines of your data set.

Round	Number of Putts	Total Score
1	23	76
2	27	80
3	28	80
4	29	80
5	30	80
6	29	82
7	30	83
8	31	83
9	33	83
10	26	84

The first step in looking for a connection between putts and total scores (or any other quantitative variables) is to make a scatterplot of the data. A *scatterplot* graphs your data set in two-dimensional space by using an X,Y plane. You can take a look at the scatterplot of the golf data in Figure 2-1. Here, *x* represents the number of putts, and *y* represents the total score. For example, the point in the lower-left corner of the graph represents someone who had only 23 putts and a total score of 75. (For instructions on making a scatterplot using Minitab, see Chapter 5.)

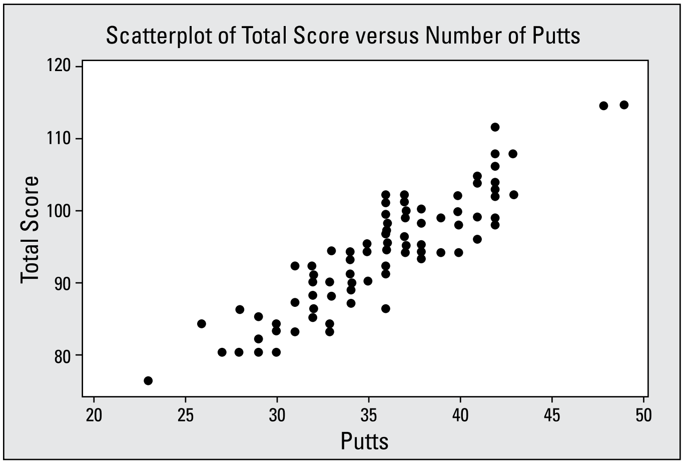


FIGURE 2-1: The two-dimensional scatterplot helps you look for relationships in data.

According to Figure 2-1, it appears that as the number of putts increases, so does the golfer's total score. It also shows that the variables increase in a linear way; that is, the data form a pattern that resembles a straight line. The relationship seems pretty strong — the number of putts plays a big part in determining the total score.

Now you need a measure of how strong the relationship is between *x* and *y* and whether it goes uphill or downhill. Different measures are used for different types of patterns seen in a scatterplot. Because the relationship you see in this case resembles a straight line, the correlation is the measure that you use to quantify the relationship. *Correlation* is the number that measures how close the points follow a straight line. Correlation is always between −1.0 and +1.0, and the more closely the points follow a straight line, the closer the correlation is to −1.0 or +1.0.

- **A positive correlation means that as *x* increases on the *x*-axis, *y* also increases on the *y*-axis.** Statisticians call this type of relationship an *uphill relationship.*
- **A negative correlation means that as *x* increases on the *x*-axis, *y* goes down.** Statisticians call this type of relationship — you guessed it — a *downhill relationship.*

For the golf data set, the correlation is $0.896 = 0.90$, which is extremely high as correlations go. The sign of the correlation is positive, so as you increase the number of putts, your total score increases (an uphill relationship). For instructions on calculating a correlation in Minitab, see Chapter 5.

Predicting y using x

If you want to predict some response variable (*y*) using one explanatory variable (*x*) and you want to use a straight line to do it, you can use *simple linear regression* (see Chapter 5 for all the fine points on this topic). Linear regression finds the best-fitting line — called the *regression line* — that cuts through the data set. After you get the regression line, you can plug in a value of *x* and get your prediction for *y*. (For instructions on using Minitab to find the best-fitting line for your data, see Chapter 5.)

To use the golf example from the previous section, suppose you want to predict the total score you can get for a certain number of putts. In this case, you want to calculate the linear regression line. By running a regression analysis on the data set, the computer tells you that the best line to use to predict total score using number of putts is the following:

$$\text{Total score} = 39.6 + 1.52 * \text{Number of putts}$$

So if you have 35 putts in an 18-hole golf course, your total score is predicted to be about $39.6 + 1.52 * 35 = 92.8$, or 93. (Not bad for 18 holes!)

WARNING

Don't try to predict y for x-values that fall outside the range of where the data was collected; you have no guarantee that the line still works outside of that range or that it will even make sense. For the golf example, you can't say that if $x\ (\text{the number of putts}) = 1$, the total score would be $39.6 + 1.52 * 1 = 41.12$ (unless you just call it good after your ball hits the green). This mistake is called *extrapolation.*

You can discover more about simple linear regression, and expansions on it, in Chapters 5 and 6.

Avoiding Bias

Bias is the bane of a statistician's existence; it's easy to create and very hard (if not impossible) to deal with in most situations. The statistical definition of *bias* is the systematic overestimation or underestimation of the actual value. In language the rest of us can understand, it means that the results are always off by a certain amount in a certain direction.

For example, a bathroom scale may always report a weight that's five pounds more than it should be (I'm convinced this is true of the scale at my doctor's office).

Bias can show up in a data set in a variety of different ways. Here are some of the most common ways bias can creep into your data.

- **Selecting the sample from the population:** Bias occurs when you either leave some groups out of the process that should have been included, or give certain groups too much weight.

  For example, TV surveys that ask viewers to phone in their opinion are biased because no one has selected a prior sample of people to represent the population — viewers who want to be involved select themselves to participate by calling in on their own. Statisticians have found that folks who decide to participate in "call-in" or website polls are very likely to have stronger opinions than those who have been randomly selected but choose not to get involved in such polls. Such samples are called *self-selected samples* and are typically very biased.

- **Designing the data-collection instrument:** Poorly designed instruments, including surveys and their questions, can result in inconsistent or even incorrect data. A survey question's wording plays a large role in whether or not results are biased. A leading question can make people feel like they should answer a certain way. For example, "Don't you think that the president should be allowed to have a line-item veto to prevent government spending waste?" Who would feel they should say *no* to that?

- **Collecting the data:** In this case, bias can infiltrate the results if someone makes errors in recording the data or if interviewers deviate from the script.
- **Deciding how and when the data is collected:** The time and place you collect data can affect whether your results are biased. For example, if you conduct a telephone survey during the middle of the day, people who work from 9 to 5 aren't able to participate. Depending on the issue, the timing of this survey could lead to biased results.

The best way to deal with bias is to avoid it in the first place, but you also can try to minimize it by doing the following:

- **Using a random process to select the sample from the population.** The only way a sample is truly random is if every single member of the population has an equal chance of being selected. Self-selected samples aren't random.
- **Making sure the data is collected in a fair and consistent way.** Be sure to use neutral wording in the question, and time the survey properly.

DON'T PUT ALL YOUR DATA IN ONE BASKET!

An animal science researcher once came to me with a data set he was very proud of. He was studying cows and the variables involved in helping determine their longevity. His super-mega data set contained over 100,000 observations. He was thinking, "Wow, this is gonna be great! I've been collecting this data for years and years, and I can finally have it analyzed. There's got to be loads of information I can get out of this. The papers I'll write, the talks I'll be invited to give . . . the raise I'll get!" He turned his precious data over to me with an expectant smile and sparkling eyes.

But after looking at his data for a few minutes, I had a terrible realization: All his data came from exactly one cow. With no other cows to compare with and a sample size of just one, he had no way to even measure how much those results would vary if he wanted to apply them to another cow. His results were so biased toward that one animal that I couldn't do anything with the data. After I summoned the courage to tell him so, it took a while to peel him off the floor. The moral of the story, I suppose, is to run your big plans by a statistician before you go down a cow path like this guy did.

Measuring Precision with Margin of Error

Precision is the amount of movement you expect to have in your sample results if you repeat your entire study again with a new sample. Precision comes in two forms:

- **Low precision** means that you expect your sample results to move a lot (not a good thing).
- **High precision** means that you expect your sample results to remain fairly close in the repeated samples (a good thing).

In this section, you find out what precision does and doesn't measure, and you see how to measure the precision of a statistic in general terms.

UP CLOSE AND PERSONAL: SURVEY RESULTS

The Gallup organization states its survey results in a universal, statistically correct format. Using a specific example from a recent survey it conducted, here's the language it uses to report its results:

> "These results are based on telephone interviews with a randomly selected national sample of 1,002 adults, aged 18 years and older, conducted June 9–11. For results based on this sample, one can say with 95 percent confidence that the maximum error attributable to sampling and other random effects is ± 3 percentage points. In addition to sampling error, question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of public opinion polls."
>
> The first sentence of the quote refers to how the Gallup organization collected the data, as well as the size of the sample. As you can guess, precision is related to the sample size, as seen in the section, "Measuring Precision with Margin of Error."

The second sentence of the quote refers to the precision measurement: How much did Gallup expect these sample results to vary? The fact that Gallup is 95 percent confident means that if this process were repeated a large number of times, in 5 percent of the cases the results would be wrong, just by chance. This inconsistency occurs if the sample selected for the analysis doesn't represent the population — not due to biased reasons, but due to chance alone. Check out the section, "Avoiding Bias," to get the information on why the third sentence is included in this quote.

Before you report or try to interpret any statistical results, you need to have some measurement of how much those results are expected to vary from sample to sample. This measurement is called the *margin of error.* You always hope, and may even assume, that statistical results shouldn't change much with another sample, but that's not always the case.

The margin of error is affected by two elements:

- The sample size
- The amount of diversity in the population (also known as the *population standard deviation*)

You can read more about these elements in Chapter 4, but here's the big picture: As your sample size increases, you have more data to work with, and your results become more precise. As a result, the margin of error goes down.

On the other hand, a high amount of diversity in your population reduces your level of precision because the diversity makes it harder to get a handle on what's going on. As a result, the margin of error increases. (To offset this problem, just increase the sample size to get your precision back.)

TIP

To interpret the margin of error, just think of it as the amount of play you allow in your results to cover most of the other samples you could have taken.

Suppose you're trying to estimate the proportion of people in the population who support a certain issue, and you want to be 95 percent confident in your results. You sample 1,002 individuals and find that 65 percent support the issue. The margin of error for this survey turns out to be ±3 percentage points (you can find the details of this calculation in Chapter 4). That result means that you could expect the sample proportion of 65 percent to change by as much as 3 percentage points either way if you were to take a different sample of 1,002 individuals. In other words, you believe the actual population proportion is somewhere between $65-3=62$ percent and $65+3=68$ percent. That's the best you can say.

WARNING

Any reported margin of error is calculated on the basis of having zero bias in the data. However, this assumption is rarely true. Before interpreting any margin of error, check first to be sure that the sampling process and the data-collection process don't contain any obvious sources of bias. Ignore results that are based on biased data, or at least take them with a great deal of skepticism.

For more details on how to calculate margin of error in various statistical techniques, turn to Chapter 4.

Knowing Your Limitations

The most important goal of any data analyst is to remain focused on the big picture — the question that you or someone else is asking — and make sure that the data analysis used is appropriate and comprehensive enough to answer that question correctly and fairly.

TIP

Here are some tips for analyzing data and interpreting the results, in terms of the statistical procedures and techniques that you may use — at school, in your job, and in everyday life. These tips are implemented and reinforced throughout this book:

- **Be sure that the research question being asked is clear and definitive.** Some researchers don't want to be pinned down on any particular set of questions because they have the intent of mining the data — looking for any relationship they can find and then stating their results after the fact. This practice can lead to overanalyzing the data, making the results subject to skepticism by statisticians.
- **Double-check that you clearly understand the type of data being collected.** Is the data categorical or quantitative? The type of data used drives the approach that you take in the analysis.
- **Make sure that the statistical technique you use is designed to answer the research question.** If you want to make comparisons between two groups and your data is quantitative, use a hypothesis test for two means. If you want to compare five groups, use analysis of variance (ANOVA). Use this book as a resource to help you determine the technique you need.
- **Look for the limitations of the data analysis.** For example, if the researcher wants to know whether negative political ads affect the population of voters, and they base their study on a group of college students, you can find severe limitations here. For starters, student reactions to negative ads don't necessarily carry over to all voters in the population. In this case, it's best to limit the conclusions to college students in that class (which no researcher would ever want to do). Better to take a sample that represents the intended population of all voters in the first place (a much more difficult task, but well worth it).

IN THIS CHAPTER

- Characterizing normal and sampling distributions
- Making friends with the standard normal (*Z*-) table
- Finding probabilities for the normal distribution
- Feeling comfortable with the sampling distributions of $\bar{X}$ and $\hat{p}$
- Building confidence intervals and hypothesis tests

Chapter 3

Having the Normal and Sampling Distributions in Your Back Pocket

You have most likely heard of the bell-shaped curve, and likely have some experience with it from Stats I. And you've probably run into (or had a run-in with) the term *sampling distributions*, one of the more difficult statistical concepts to understand. The normal distribution and sampling distributions are very important, probably the most important distributions in Stats I and II. Many techniques require the data to have come from a population with a normal distribution, and many techniques use sampling distributions as key components, so you want to know them very well as you proceed through the rest of this book. No sweat, this chapter's got you covered.

In this chapter, I give you an overview of the normal and sampling distributions from Stats I and their connection to each other. Throughout this book, I discuss their importance as prerequisites for many of the other statistical techniques, including confidence intervals and hypothesis tests (Chapter 4), regression (Chapters 5 to 9), and ANOVA (Chapters 10 to 13), among others.

Recognizing the VIP Distribution — the Normal

In this section, I go over the main points of the normal distribution as they pertain to many other chapters in this book. If a statistical technique has a condition that the data come from a normal distribution, then you need to know what that means and how to work with it. First, you look at the characteristics of the normal, then you standardize to the standard normal distribution, then you walk through the steps of the standard normal table that I use in this book, and finally you walk through some examples of finding probabilities for a normal distribution.

Characterizing the normal

The normal distribution has a distinct, bell-shaped curve. The mean is in the middle, and the distribution is symmetric, meaning it looks the same on each side if you cut it down the middle. Each different normal distribution has its own mean, μ, and standard deviation, σ. The mean can be any number, and the standard deviation can be any non-negative number. The 68-95-99.7% Rule for normal distributions says the following: About 68% of its values lie within one standard deviation of the mean (middle), about 95% of its values lie within two standard deviations of the mean, and about 99.7% (virtually all) of its values lie within three standard deviations of the mean. Figure 3-1 shows some different normal distributions. The tick marks on the horizontal axis denote the standard deviations.

Standardizing to the standard normal (Z-) distribution

In Figure 3-1c, you see a normal distribution with a mean of 0 and a standard deviation of 1. This is a special normal distribution — a VIP of VIPs — and is called the *standard normal distribution*. This distribution is so special that it was given its own letter, *Z*, to describe it. The *Z*-distribution is special because whatever value you have on the *Z*-distribution, it represents the number of standard deviations you are above (if positive) or below (if negative) the mean of 0. For example, if $z = 2$, that means you are 2 standard deviations above the mean, because the mean is 0 and the standard deviation is 1. If $z = -1$, that means you are 1 standard deviation below the mean.

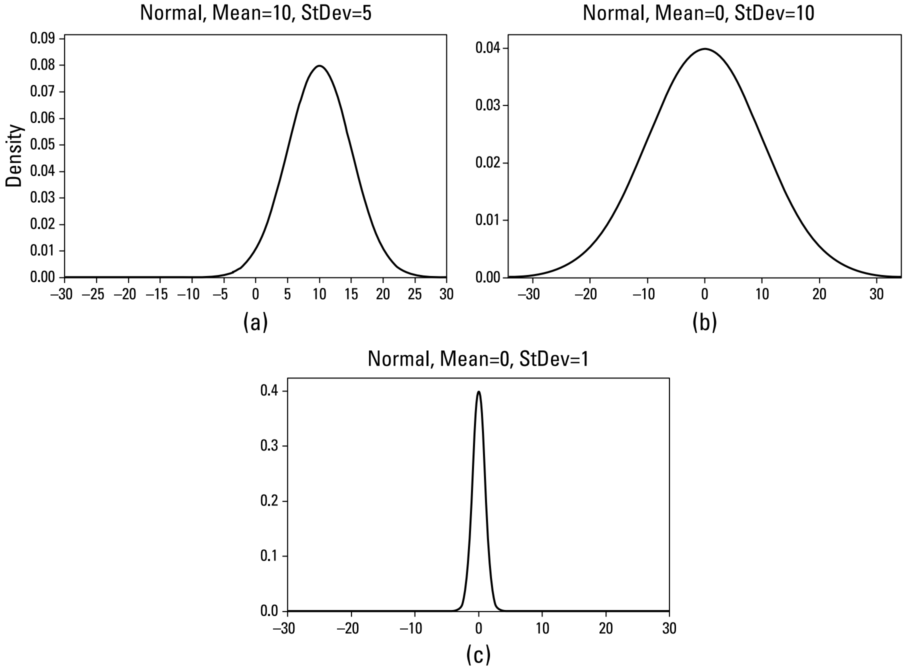


FIGURE 3-1: Three normal distributions, with means and standard deviations of a) 10 and 5; b) 0 and 10; and c) 0 and 1, respectively.

The Z-distribution sets a standard by which other distributions are compared. This comparison is made in the following way. Suppose X is normal with mean 80 and standard deviation 5. If you get a score of 90 on an exam, what is your Z-value? In other words, how many standard deviations are you above or below the mean? Well, your score is 90, and $90-80$ represents a difference of 10. In terms of number of standard deviations, you divide by 5 (the standard deviation) to get 2. And this tells you that on this exam, a 90 is 2 standard deviations above the mean, and so $z=2$. Similarly, a 70 equals $(70-80)/5=-2$, or 2 standard deviations below the mean, and so $z=-2$.

The notation is important here. Note that

- A capital italicized letter represents a random variable or a name of a distribution in the general sense. For example, "Test scores have a Z-distribution."
- A lowercase italicized letter is a specific value of a random variable, or a specific value on a distribution. For example, "Bob's test score has a value of $z=2$ standard deviations above the mean."

In general, to get from X to Z, you use what statistics students around the world call the Z-formula: $Z=\frac{X-\mu}{\sigma}$. It's a formula that you'll soon have memorized (if you haven't already), and it'll be like a song you can't get out of your head.

Why is this formula so important? It represents a process called standardizing a score; it takes away *X*'s original units, and for a normal it puts its values on a scale from about −3 to 3 (3 standard deviations below or above the mean.) Any value on any normal distribution can be standardized this way, so to find probabilities for any normal distribution, you only need to standardize the values and provide a single table to look at — cleverly called the standard normal table, or *Z*-table. This table can be found in Table A-5 in the Appendix.

TIP

A great use for the *Z*-formula is to compare two groups that have different means and different standard deviations. For example, suppose you got a 90 on an exam with mean 80 and standard deviation 5, and you got a 90 on an exam with mean 75 and standard deviation 10. On which exam did you do better, relative to the other students? If you standardize both exam scores, you can easily find out. Your first exam score is $(90-80)/5=2$ standard deviations above the mean, and so $Z=2$. Your second exam score is $(90-75)/10=1.5$ standard deviations above the mean, and so $Z=1.5$. So you did better on the first exam than the second exam, relative to the rest of the class — even though you got the same score of 90 each time.

Using the normal table

Because the *Z*-distribution is standardized, and you can change any normal distribution *X* to the *Z*-distribution through the *Z*-formula, statisticians created a single table to find probabilities for any normal distribution — and it is based on the standard normal distribution.

The *Z*-table in this book shows $P(Z<z)$ for basically any *Z* from −3 to +3. Here are the steps for finding the probability that *Z* is less than *z* using the *Z*-table in this book:

1. **Write the *Z*-value with 2 digits after the decimal point.**

   For example, if you want $P(Z<1)$, you write $P(Z<1.00)$.

2. **Look in the row for the leading digit and the first digit after the decimal point.**

   In this case, look in the row that says 1.0.

3. **Look in the column that has the second digit after the decimal point.**

   In this case, go to the first column, .00.

4. **Intersect the row and column to find $P(Z<z)$.**

   When you do that here, you get 0.8413. So, the probability of scoring less than 90 on this exam is $P(X<90)=P(Z<1.00)=0.8413$.

5. **If you want a greater-than probability, take 1 minus your answer from Step 4.**

   That's because the total probability under the entire curve equals 1, and you want the upper part when given the lower part. For example, $\text{P}(Z > 1) = 1 - \text{P}(Z < 1) = 1 - 0.8413 = 0.1587$.

6. **If you want an in-between probability, such as $\text{P}(a < Z < b)$, change both values to *Z*, look them both up, and subtract (largest value minus smallest) to get $\text{P}(Z < b) - \text{P}(Z < a)$.**

   For example, $\text{P}(-2 < Z < 1) = \text{P}(Z < 1.00) - \text{P}(Z < -2.00) = 0.8413 - 0.1587 = 0.6826$.

TIP

Different textbooks may have different formats of the *Z*-table. Make sure you understand how your table works before proceeding to try to find probabilities.

Finding probabilities for the normal distribution

Other normal distributions, of course, have other means besides 0 and other standard deviations besides 1. To find a probability for any normal distribution *X* with mean μ and standard deviation σ, just standardize it and look it up. That is, use the *Z*-formula to change *X* to *Z*, and look it up on the *Z*-table to find the probabilities.

For example, say you want to find the probability that someone scored less than 90 on a statistics exam that had mean 80 and standard deviation 10. So you want to find $\text{P}(X < 90)$. Change *X* to *Z* using the *Z*-formula and you get $\text{P}(X < 90) = \text{P}\left(\frac{X - 80}{10} < \frac{90 - 80}{10}\right) = \text{P}(Z < 1)$. From the previous section, you know this probability from the *Z*-table is 0.8413.

As another example, suppose you want $\text{P}(X < 92.5)$. (Yes, some of us professors actually do give half points!) This equals $\text{P}\left(Z < \frac{92.5 - 80}{10}\right) = \text{P}(Z < 1.25)$. Look up the row for 1.2, and the column for .05, and intersect them to get 0.8944. This is the probability that someone scored less than 92.5 on the exam.

REMEMBER

If you want a greater-than probability using the *Z*-table, you need to take 1 minus your answer from Step 4. So, for example, $\text{P}(X > 92.5) = 1 - \text{P}(X < 92.5) = 1 - 0.8944 = 0.1056$.

If you want an in-between probability, find the probability of being less than each of the numbers in the inequality, and subtract the larger one minus the smaller one to get a non-negative answer. For example, if you want $\text{P}(90 < X < 92.5)$, you find $P(X < 92.5)$ and subtract $P(X < 90)$ to get the part in between. In this case, from the previous example, $P(X < 92.5) = 0.8944$ and $P(X < 90) = 0.8413$. Subtract them to get $0.8944 - 0.8413 = 0.0531$.

Finally Getting Comfortable with Sampling Distributions

In Stats I, one of the most difficult and often dreaded topics is sampling distributions. However, sampling distributions, once broken down, really aren't that bad, plus they possess some neat qualities and have a strong connection to the normal distribution. In this section, you gain more experience and find a higher comfort level with sampling distributions as they pertain to other statistical techniques in this book.

A sampling distribution is a population of values that comes from taking random samples over and over and collecting the resulting statistics into one big pot. This pot is called a *sampling distribution* for that particular statistic. For example, you can start with a class of 1,000 students and take random samples of size 100 from this population, over and over and over, and each time you can find the mean age of the sample. When you are done taking as many different random samples of 100 as you possibly can, and you put all the mean ages together into one big pot, this pot is called the *sampling distribution of the sample mean*.

As another example, you could take all possible random samples of size 2,000 from all registered voters in Columbus, Ohio, and find the proportion in each sample who voted for the elected president in the last election. This pot contains many sample proportions, and is called the *sampling distribution of the sample proportion*.

REMEMBER

You can have a lot of different types of sampling distributions for a lot of different statistics from any population you can dream up. You have to make sure that the samples are random, that you have all possible samples, and you have the same sample size each time.

The mean and standard error of a sampling distribution

A sampling distribution is a population of sample statistics. The big pot of all possible sample means from samples of size *n* is called the *sampling distribution of the sample mean*, or the sampling distribution of $\bar{X}$. The big pot of all possible sample proportions from samples of size n is called the *sampling distribution of the sample proportion*, or the sampling distribution of $\hat{p}$.

Each sampling distribution is its own population. That means it has its own population mean, and its own population standard error. (It's the same as a standard deviation, but because it applies to sampling distributions, it is called a *standard error*.)

Sampling distribution of $\bar{X}$

If X is the original population that the samples came from, and it has mean μ_X, then the mean of the sampling distribution of $\bar{X}$ is denoted by, and is the same as, the mean of X. So the mean of $\bar{X}$ is equal to μ_X.

Also, if the standard deviation of the population X is equal to σ_x, then the standard error of the sampling distribution of $\bar{X}$ is denoted by $\sigma_{\bar{x}}$ and is equal to the population standard deviation of X, divided by the square root of n, written as $\frac{\sigma_x}{\sqrt{n}}$. Here you are finding the average when you are finding $\bar{X}$, so the standard error contains an n in the denominator like the average does. That means the standard error of $\bar{X}$ gets smaller and smaller as you take larger and larger sample sizes. Big samples don't change as much as small samples. $\bar{X}$ can be a pretty precise statistic!

Now the shape of the population of all possible sample means (for example, the shape of the sampling distribution of $\bar{X}$) is either normal or approximately normal (if the samples are large enough). If X is a distribution from a population that has a normal distribution, then you know that $\bar{X}$ also has a normal distribution. But the crazy and wonderful news is, and you saw this in Stats I, if X is not a normal distribution, but your samples are large enough (at least > 30 for most folks), then $\bar{X}$ has an approximate normal distribution. This is due to what you know as the Central Limit Theorem. See how it all fits together? And this is big news because whether X is a normal distribution or not, you can use Z to solve problems if the right conditions are met.

The formula for using Z to solve probability problems for $\bar{X}$ is $Z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$, and when you plug in what you know for $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$, you get $Z = \frac{\bar{X} - \mu_x}{\sigma_x / \sqrt{n}}$.

REMEMBER

Notice the difference between $\mu_{\bar{x}}$ and μ_x and the difference between $\sigma_{\bar{x}}$ and σ_x. It's a very important difference. Don't be afraid to read the previous sections again if you are still confused; sometimes it takes a couple of whacks to get it.

For example, suppose X = distribution of exam scores and X has a normal distribution with mean 80 and standard deviation 10. What's the chance that the average of 36 exams is less than 75? To answer this question, you know $\bar{X}$ is the average of 36 exams, and that it has a normal distribution (because the exams do — you do not need the Central Limit Theorem here). You know the mean is $\mu_{\bar{x}} = \mu_x = 80$ and the standard error is $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{10}{\sqrt{36}} = 1.67$. So you find $P(\bar{X} < 75) = P\left(Z < \frac{75-80}{10/\sqrt{36}}\right) = P(Z < -3.00) = 0.0013$. (Highly unlikely.)

Sampling distribution of $\hat{p}$

Suppose X has a binomial distribution with n trials with p = the proportion of successes/yeses in the population and X in general represents the number of successes/yeses in a sample of size n from the population. Let $\hat{p}$ represent the proportion of successes/yeses in the sample. You can form a sampling distribution for $\hat{p}$ by taking all possible random samples of size n from the binomial population and finding all possible proportions of yeses from those samples.

The mean of the sampling distribution of $\hat{p}$ is denoted by $\mu_{\hat{p}}$. Its mean is equal to this same p from the binomial distribution. Note that both are proportions; one is from the sample, $\hat{p}$, and one is from the population, p.

Also, because the standard deviation of the binomial distribution is equal to $\sigma_x = np(1-p)$, the standard error of the sampling distribution of $\hat{p}$ is denoted by $\sigma_{\hat{p}}$ and is equal to $\sqrt{\frac{p(1-p)}{n}}$.

TECHNICAL STUFF

From Stats I, you know the mean of the binomial itself is np, but that is the case where you are counting X = number of successes in n trials. If you are measuring the proportion of successes in n trials, you divide by n to get $np/n = p$. That's the mean of $\hat{p}$. The standard error of $\hat{p}$ is found by looking at how much each observation can vary (between 1 = success with probability p and 0 = failure with probability $1-p$), and dividing by n. (The square root is there because you are taking the square root of the variance when you find the standard deviation.)

Again, because you have an n in the denominator of the standard error for $\hat{p}$, that means the standard error of $\hat{p}$ gets smaller and smaller as you take larger and larger sample sizes. Big samples don't change as much as small samples; $\hat{p}$ can also be a pretty precise statistic!

The shape of the distribution of all possible sample proportions (for example, the shape of the sampling distribution of $\hat{p}$) is approximately normal, again by another version of the Central Limit Theorem, if the sample sizes are large enough. In the binomial's case, you need np and $n(1-p)$ to both be at least 10. That means you can use Z to solve problems regarding $\hat{p}$ as long as the conditions are met.

The formula for using Z to solve probability problems for $\hat{p}$ is $Z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}$, and when you plug in what you know for $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$, you get $Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$.

For example, suppose X = number of exams that are passed out to a large class, and suppose the pass rate is reported to be 90%. What is the chance that out of a sample of 100 exams, less than 85% of them pass? To answer this question, you know $\hat{p}$ is the proportion of 100 exams that had a "success" (passed), and that it has an approximate normal distribution. Remember, this is because the conditions are met: $np = 100(.90) = 90$ and $n(1-p) = 100(1-.90) = 10$ are both at least 10. You know the mean of $\hat{p}$ is $p = .90$ and the standard error is $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.90(1-.90)}{100}} = .03$. So you find $\text{P}(\hat{p} < 0.85) = \text{P}\left(Z < \frac{.85 - .90}{.03}\right) = \text{P}(Z < -1.67) = 0.0475$.

Heads Up! Building Confidence Intervals and Hypothesis Tests

The information from the previous sections is very important for later chapters, where you use properties of $\bar{X}$ and $\hat{p}$ to build formulas for other statistical techniques, such as confidence intervals and hypothesis tests. In this section, I provide an overview of how the main formulas for confidence intervals and hypothesis tests for μ_x and p come about, as they are based in large part on $\bar{X}$ and $\hat{p}$ and their means and standard errors. Keeping these techniques in your mind is much easier if you understand where they came from.

Confidence interval for the population mean

Say your goal is to estimate the mean of a population, μ_x. Your technique is to make a confidence interval for it. That is, you start with the sample mean, $\bar{X}$, and you add and subtract a margin of error (as you learned in Stats I). The margin of error is made up of a certain number of standard errors of $\bar{X}$. Do you remember the 68-95-99.7% Rule (from the earlier section, "Characterizing the normal"), where 95% of the values lie within about 2 standard deviations of the mean in a normal distribution? Because the sampling distribution of $\bar{X}$ has an exact or approximate (if $n > 30$) normal distribution, you can just take $\bar{X}$, your sample mean, and add and subtract about 2 standard errors, and you'll have a 95% confidence interval. That's what's going on. The general formula looks like this: $\bar{X} \pm Z\frac{\sigma_x}{\sqrt{n}}$, where Z is the number of standard errors you add and subtract to get the confidence level you want — a common value of Z is 1.96, about 2! See the standard error of $\bar{X}$ doing its job? Now that you see how the formula comes to be, you can understand a great deal more about confidence intervals in Chapter 4 and beyond!

Confidence interval for the population proportion

The same idea applies to get the formula for a confidence interval for the population proportion, p. If the conditions are met (np and $n[1-p]$ are both at least 10), then $\hat{p}$ has an approximate normal distribution, and you can again make a 95% confidence interval by taking the sample proportion, $\hat{p}$, plus or minus about 2 standard errors (1.96 of them, to be exact). The standard error of $\hat{p}$ as found previously in this chapter, is $\sqrt{\frac{p(1-p)}{n}}$, which contains the very thing you are trying to estimate: p. You can't have that because you don't know p, so you substitute $\hat{p}$ for p, and there you go! So the final formula for a confidence interval for the population proportion is $\hat{p} \pm Z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. (See Chapter 4 for more on building formulas for confidence intervals.)

Hypothesis test for population mean

A hypothesis test for the population mean μ_x is used when you want to challenge some reported value for the mean, based on your sample data. The ideas are all fleshed out in Chapter 4, but briefly, you have a null hypothesis H_o: $\mu_x = \mu_0$, and an alternative hypothesis, H_a, which says μ_x is >, <, or $\neq \mu_0$, and H_o is on trial. Starting with your sample mean, you standardize it using the ideas presented earlier in this chapter, and you use a table to see where it falls. If the population has a normal distribution (or if the sample size is > 30), then you can use the standard normal (Z-) distribution to compare your test statistic to H_o, and gauge how likely or unlikely it is that your data matches the value in H_o.

That is, you find the p-value for your test statistic, the probability of being where you were with your data, or further out if H_o were the truth. If the p-value is small, you reject the reported value for the population mean presented in H_o. If the p-value is large, you fail to reject. Those are the nuts and bolts of doing a hypothesis test for the population mean. All these ideas are filled out in Chapter 4.

Now let's focus on the test statistic, the centerpiece of the hypothesis test. You take your sample mean, $\bar{X}$, and standardize it, which means you subtract the mean and divide by the standard error, as you saw previously in this chapter: $Z = \frac{\bar{X} - \mu_x}{\sigma_x / \sqrt{n}}$.

But, because μ_x is not known, you use its reported value, and assume that is the true value until proven otherwise. So you substitute μ_0 for μ_x and you get the test statistic $Z = \frac{\bar{X} - \mu_0}{\sigma_x / \sqrt{n}}$. Again, you can see that standard error of $\bar{X}$ at work in the denominator of the test statistic.

Hypothesis test for the population proportion

Similarly, for a hypothesis test for p, the population proportion, you start with the sample proportion, $\hat{p}$, and standardize it to get the test statistic. That is, you subtract the mean of $\hat{p}$, and divide by the standard error of $\hat{p}$, both of which you found earlier in this chapter, to get $Z = \frac{\hat{p} - p}{\sqrt{\frac{p\ (1-p\ )}{n}}}$. Because you don't know the actual value of p in this situation (you are hypothesizing about its value), you assume H_0 is true until proven otherwise, and you use p_0 for p in this formula. In the end, your test statistic looks like this:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

If the conditions are met where np and $n(1-p)$ are both at least 10, you can use a Z-table to look up your test statistic and find the p-value.

Knowing how test statistics for the population mean and standard error are built sets you up well for the information covered in Chapter 4 and other techniques used in this book. You can also see how incredibly important the normal distribution is, and how much it is used in statistics.

IN THIS CHAPTER

- Utilizing confidence intervals to estimate parameters
- Testing models by using hypothesis tests
- Finding the probability of getting it right and getting it wrong
- Discovering power in a large sample size

Chapter 4

Reviewing Confidence Intervals and Hypothesis Tests

One of the major goals in statistics is to use the information you collect from a sample to get a better idea of what's going on in the entire population you're studying (because populations are generally large and exact information is often unknown). Unknown values that summarize the population are called *population parameters.* Researchers typically want to either get a handle on what those parameters are or test a hypothesis about the population parameters.

In Stats I, you probably went over confidence intervals and hypothesis tests for one and two population means and one and two population proportions. Your instructor hopefully emphasized that no matter which parameters you're trying to estimate or test, the general process is the same. If not, don't worry; this chapter drives that point home.

This chapter reviews the basic concepts of confidence intervals and hypothesis tests, including the probabilities of making errors by chance. I also discuss how statisticians measure the ability of a statistical procedure to do a good job — of detecting a real difference in the populations, for example.

Estimating Parameters by Using Confidence Intervals

Confidence intervals are a statistician's way of covering their you-know-what when it comes to estimating a population parameter. For example, instead of just giving a one-number guess as to what the average household income is in the United States, a statistician gives a range of likely values for this number. They do this because

- All good statisticians know sample results vary from sample to sample, so a one-number estimate isn't any good.
- Statisticians have developed some awfully nice formulas to give a range of likely values, so why not use them?

In this section, you get the general formula for a confidence interval, including the margin of error, and a good look at the common approach to building confidence intervals. I also discuss interpretation and the chance of making an error.

Getting the basics: The general form of a confidence interval

The big idea of a confidence interval is coming up with a range of likely values for a population parameter. The *confidence level* represents the chance that if you were to repeat your sample-taking over and over, you'd get a range of likely values that actually contains the actual population parameter. In other words, the confidence level is the long-term chance of being correct.

REMEMBER

The general formula for a confidence interval is

$$\text{Confidence interval} = \text{Sample statistic} \pm \text{Margin of error}$$

The confidence interval has a certain level of precision (measured by the margin of error). Precision measures how close you expect your results to be to the truth.

For example, suppose you want to know the average amount of time a student at Ohio State University spends listening to music on their phone per day. The average time for the entire population of OSU students who are phone users is the parameter you're looking for. You take a random sample of 1,000 students and find that the average time a student uses their phone per day to listen to music is 2.5 hours, and the standard deviation is 0.5 hour. Is it right to say that the population of all OSU students use their phones an average of 2.5 hours per day for music listening? You hope and may want to assume that the average for the whole population is close to 2.5, but it probably isn't exact.

What's the solution to this problem? The solution is to not only report the average from your sample, but also to report some measure of how much you expect that sample average to vary from one sample to the next, with a certain level of confidence. The number that you use to represent this level of precision in your results is called the *margin of error.*

Finding the confidence interval for a population mean

The sample-statistic part of the confidence-interval formula is fairly straightforward.

- **To estimate the population mean,** you use the sample mean plus or minus a margin of error, which is based on standard error. The sample mean has a standard error of $\frac{\sigma}{\sqrt{n}}$. In this formula, you can see the population standard deviation (σ) and the sample size (n).
- **To estimate the population proportion,** you use the sample proportion plus or minus a margin of error.

In many cases, the standard deviation of the population, σ, is not known. To estimate the population mean by using a confidence interval when σ is unknown, you use the formula $\bar{x} \pm t_{n-1}\left(\frac{s}{\sqrt{n}}\right)$. This formula contains the sample standard deviation (s), the sample size (n), and a t-value representing how many standard errors you want to add and subtract to get the confidence you need. To get the margin of error for the mean, you see the standard error, $\frac{s}{\sqrt{n}}$, is being multiplied by a factor of t. Notice that t has $n-1$ as a subscript to indicate which of the myriad t-distributions you use for your confidence interval. The $n-1$ is called *degrees of freedom.*

The value of t in this case represents the number of standard errors you add and subtract to or from the sample mean to get the confidence you want. If you want to be 95 percent confident, for example, you add and subtract 1.96 of those standard errors. If you want to be 99.7 percent confident, you add or subtract about three of them. (See Table A-1 in the Appendix to find t-values for various confidence levels; use $\left(\frac{1-\text{confidence level}}{2}\right)$ for the area to the right and find the t-value that goes with it.)

TIP

If you know the population standard deviation, you should certainly use it. In that case, you use the corresponding number from the Z-distribution (standard normal distribution) in the confidence interval formula. (The Z-distribution in the Appendix can give you the numbers you need.) But I would be remiss in saying that while textbooks and teachers always include problems where σ is known, rarely is σ known in the real world. Why teach it this way? This issue is up for debate; for now, just go with it, and I can keep you posted.

For the phone example from the preceding section, a random sample of 1,000 of all OSU students spend an average of 2.5 hours using their phones to listen to music. The standard deviation is 0.5 hour. Plugging this information into the formula for a confidence interval, you get $2.5 \pm 1.96\left(\frac{0.5}{\sqrt{1{,}000}}\right)$. You conclude that *all* OSU-student phone-users spend an average of between 2.47 and 2.53 hours listening to music on their phones.

What changes the margin of error?

What do you need to know in order to come up with a margin of error? Margin of error, in general, depends on three elements:

- The standard deviation of the population, σ (or an estimate of it, denoted by s, the sample standard deviation)
- The sample size, n
- The level of confidence you need

You can see these elements in action in the formula for margin of error of the sample mean: $\pm t_{n-1} * \frac{s}{\sqrt{n}}$. Here I assume that σ isn't known; t_{n-1} represents the value on the t-distribution table (see Table A-1 in the Appendix) with $n-1$ degrees of freedom.

Each of these three elements has a major role in determining how large the margin of error will be when you estimate the mean of a population. In the following sections, I show how each of the elements of the margin of error formula work separately and together to affect the size of the margin of error.

Population standard deviation

The standard deviation of the population is typically combined with the sample size in the margin of error formula, with the population standard deviation on top of the fraction and n on the bottom. (In this case, the standard error of the population, σ, is estimated by the standard deviation of the sample, s, because σ is typically unknown.)

This combination of standard deviation of the population and sample size is known as the *standard error* of your statistic. It measures how much the sample statistic deviates from its mean in the long term.

REMEMBER

How does the standard deviation of the population (σ) affect margin of error? As it gets larger, the margin of error increases, so your range of likely values is wider.

Suppose you have two gas stations, one on a busy corner (gas station #1) and one farther off the main drag (gas station #2). You want to estimate the average time between customers at each station. At the busy gas station #1, customers are constantly using the gas pumps, so you basically have no downtime between customers. At gas station #2, customers sometimes come all at once, and sometimes you don't see a single person for an hour or more. So the time between customers varies quite a bit.

For which gas station would it be easier to estimate the overall average time between customers as a whole? Gas station #1 has much more consistency, which represents a smaller standard deviation of time between customers. Gas station #2 has much more variability in time between customers. That means σ for gas station #1 is smaller than σ for gas station #2. So the average time between customers is easier to estimate at gas station #1.

Sample size

Sample size affects margin of error in a very intuitive way. Suppose you're trying to estimate the average number of pets per household in your city. Which sample size would give you better information: 10 homes or 100 homes? I hope you'd agree that 100 homes would give more precise information (as long as the data on those 100 homes was collected properly).

If you have more data to base your conclusions on and that data is collected properly, your results will be more precise. Precision is measured by margin of error, so as the sample size increases, the margin of error of your estimate goes down.

WARNING

Bigger is only better in terms of sample size if the data is collected properly — that is, with minimal bias. If the quality of the data can't be maintained with a larger sample size, it does no good to have it.

Confidence level

For each problem at hand, you have to address how confident you need to be in your results over the long term, and, of course, more confidence comes with a price in the margin of error formula. This level of confidence in your results over the long term is reflected in a number called the *confidence level,* which you report as a percentage. In general, more confidence requires a wider range of likely values. So, as the confidence level increases, so does the margin of error.

REMEMBER

Every margin of error is interpreted as plus or minus a certain number of standard errors. The number of standard errors added and subtracted is determined by the confidence level. If you need more confidence, you add and subtract more standard errors. If you need less confidence, you add and subtract fewer standard errors. The number that represents how many standard errors to add and subtract is different from situation to situation. For one population mean, you use a value on the t-distribution, represented by t_{n-1}, where n is the sample size (see Table A-1 in the Appendix).

Suppose you have a sample size of 20, and you want to estimate the mean of a population with 90 percent confidence. The number of standard errors you add and subtract is represented by t_{n-1}, which in this case is $t_{19} = 1.73$. (To find these values of t, see Table A-1 in the Appendix, with $n-1$ degrees of freedom for the row, and $\frac{(1-\text{confidence level})}{2}$ for the column.)

Now suppose you want to be 95 percent confident in your results, with the same sample size of $n = 20$. The degrees of freedom are $20-1=19$ (row) and the column is for $\frac{(1-.95)}{2} = .025$. The t-table gives you the value of $t_{19} = 2.09$.

Notice that this value of t is larger than the value of t for 90 percent confidence, because in order to be more confident, you need to go out more standard deviations on the t-distribution table to cover more possible results.

Large confidence, narrow intervals — just the right size

A narrow confidence interval is much more desirable than a wide one. For example, claiming that the average cost of a new home is $150,000 plus or minus $100,000 isn't helpful at all because your estimate is anywhere between $50,000 and $250,000. (Who has an extra $100,000 to throw around?) But you *do* want a high confidence level, so your statistician has to add and subtract more standard errors to get there, which makes the interval that much wider (a downer).

Wait, don't panic — you can have your cake and eat it too! If you know you want to have a high level of confidence but you don't want a wide confidence interval, just increase your sample size to meet that level of confidence.

Suppose the standard deviation of the house prices from a previous study is $s = \$15{,}000$, and you want to be 95 percent confident in your estimate of average house price. Using a large sample size, your value of t (from the last row of Table A-1 in the Appendix) is 1.96.

With a sample of 100 homes, your margin of error is $\pm 1.96 * \frac{15{,}000}{\sqrt{100}} = \$2{,}940$.

If this is too large for you but you still want 95 percent confidence, crank up your value of n. If you sample 500 homes, the margin of error decreases to $\pm 1.96 * \frac{15{,}000}{\sqrt{500}}$, which brings you down to $1,314.81.

TECHNICAL STUFF

You can use a formula to find the sample size you need to meet a desired margin of error. That formula is $n = \left(\frac{t_{n-1} s}{MOE} \right)^2$, where MOE is the desired margin of error (as a proportion), s is the sample standard deviation, and t is the value on the t-distribution that corresponds with the confidence level you want. (For large sample sizes, the t-distribution is approximately equal to the Z-distribution.)

Interpreting a confidence interval

Interpreting a confidence interval involves a couple of subtle but important issues. The big idea is that a *confidence interval* presents a range of likely values for the population parameter, based on your sample. However, you interpret it not in terms of your own sample, but in terms of an infinite number of other samples out there that could have been selected, yours just being one of them. For example, suppose 1,000 people each took a sample and they each formed a 95 percent confidence interval for the mean. The "95 percent confidence" part means that of those 1,000 confidence intervals, about 950 of them can be expected to be correct on average. (Correct means the confidence interval actually contains the true value of the parameter.)

WARNING

A 95 percent confidence interval doesn't mean that your particular confidence interval has a 95 percent chance of capturing the actual value of the parameter; after the sample has been taken, the parameter is either in the interval or it isn't. A confidence interval represents the chances of capturing the actual value of the population parameter over many different samples.

Suppose a polling organization wants to estimate the percentage of people in the United States who drive a car with more than 100,000 miles on it, and it wants to be 95 percent confident in its results. The organization takes a random sample of

1,200 people and finds that 420 of them (35 percent) drive a car with that minimum mileage; the margin of error turns out to be plus or minus 3 percent. (See your Stats I text for determining margin of error for percentages.)

The meaty part of the interpretation lies in the confidence level — in this case, the 95 percent. Because the organization took a sample of 1,200 people in the U.S., asked each of them whether their car had more than 100,000 miles on it, and made a confidence interval out of the results, the polling organization was, in essence, accounting for all the other samples out there that it could have gotten by building in the margin of error (±3 percent). The organization wanted to cover its bases on 95 percent of those other situations, and ±3 percent satisfies that.

Another way of thinking about the confidence interval is to say that if the organization sampled 1,200 people over and over again and made a confidence interval from its results each time, 95 percent of those confidence intervals would be right. (You just have to hope that yours is one of those right results.)

TIP

Using stat notation, you can write confidence levels as $(1-\alpha)\%$. So if you want 95 percent confidence, you write it as $1-0.05$. Here, α represents the chance that your confidence interval is one of the wrong ones. This number, α, is also related to the random chance of making a certain kind of error with a hypothesis test, which I explain in the later section, "False alarms and missed opportunities: Type I and II errors."

What's the Hype about Hypothesis Tests?

Suppose a shipping company claims that its packages are on time 92 percent of the time, or a campus official claims that 75 percent of students live off campus. If you're questioning these claims, how can you use statistics to investigate?

In this section, you see the big ideas of hypothesis testing that are the basis for the data-analysis techniques in this book. You review and expand on the concepts involved in a hypothesis test, including the hypotheses, the test statistic, and the *p*-value.

What H_o and H_a really represent

You use a hypothesis test in situations where you have a certain model in mind and want to see whether that model fits your data. Your model may be one that just revolves around the population mean (testing whether that mean is equal to ten, for example). Your model may be testing the slope of a regression line (whether or not it's zero, for example, with zero meaning you find no relationship between *x* and *y*). You may be trying to use several different variables to predict the marketability of a product, and you believe a model using customer age, price, and shelf location can help predict it, so you need to run one or more hypothesis tests to see whether that model works. (This particular process is called *multiple regression*, and you can find more information on it in Chapter 6.)

A hypothesis test is made up of two hypotheses:

- **The null hypothesis, H_o:** H_o symbolizes the current situation — the one that everyone assumed was true until you got involved.
- **The alternative hypothesis, H_a:** H_a represents the alternative model that you want to consider. It stands for the researcher's hypothesis, and the burden of proof lies on the researcher.

REMEMBER

H_o is the model that's on trial. If you get enough evidence against it, you conclude H_a, which is the model you're claiming is the right one. If you don't get enough evidence against H_o, then you can't say that your model (H_a) is the right one.

Gathering your evidence into a test statistic

A *test statistic* is the statistic from your sample, standardized so you can look it up on a table, basically. Although each hypothesis test is a little different, the main thought is the same. Take your statistic and standardize it in the appropriate way so you can use the corresponding table for it. Then look up your test statistic on the table to see where it stands. That table may be the *t*-table, the Chi-square table, or a different table. The type of test you need to use for your data dictates which table you use.

In the case of testing a hypothesis for a population mean, μ, you use the sample mean, $\bar{x}$, as your statistic. To standardize it, you take $\bar{x}$ and convert it to a value of *t* by using the formula $t_{n-1} = \dfrac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$, where μ_o is the value in H_o. This value is your test statistic, which you compare to the *t*-distribution.

Determining strength of evidence with a p-value

If you want to know whether your data has the brawn to stand up against H_o, you need to figure out the *p*-value and compare it to a predetermined cutoff, α

(typically 0.05). The *p-value* is a measure of the strength of your evidence against H_o. You calculate the *p*-value through these steps:

1. **Calculate the test statistic (refer to the preceding section for more information on this).**
2. **Look up the test statistic on the appropriate table (such as the *t*-table, found in the Appendix).**
3. **Find the percentage of values on the table that fall beyond your test statistic.** This percentage is the *p*-value.
4. **If your H_a is "not equal to," double the percentage that you got in Step 3 because your test statistic could have gone either way before the data was collected.** (See your Stats I textbook or *Statistics For Dummies*, 2nd Edition for full details on obtaining *p*-values for hypothesis tests.)

Your friend α is the cutoff for your *p*-value. (α is typically set at 0.05, but sometimes it's another number, like 0.10 for example.) If your *p*-value is less than your predetermined value of α, reject H_o because you have sufficient evidence against it. If your *p*-value is greater than or equal to α, you can't reject H_o.

For example, if your *p*-value is 0.002, your test statistic is so far away from H_o that the chance of getting this result by chance if the null hypothesis were true is only 2 out of 1,000. So, you conclude that H_o is very likely to be false. If your *p*-value turns out to be 0.30, this same result is expected to happen 30 percent of the time anyway, so you see no red flags there, and you can't reject H_o; you don't have enough evidence against it. If your *p*-value is close to the cutoff line, say $p = 0.049$ or 0.51, you say the result is marginal and let the reader make their own conclusions. That's the main advantage of the *p*-value: It lets other folks determine whether your evidence is strong enough to reject H_o in their minds.

False alarms and missed opportunities: Type I and II errors

Any technique you use in statistics to make a conclusion about a population based on a sample of data has the chance of making an error. The errors I am talking about, Type I and Type II errors, are due to random chance (assuming the data were collected properly).

REMEMBER

The way you set up your test can help to reduce these kinds of errors, but they're always out there. As a data analyst, you need to know how to measure and understand the impact of the errors that can occur with a hypothesis test and what you can do to possibly make those errors smaller. In the following sections, I show you how you can do just that.

Making false alarms with Type I errors

A *Type I error* is the conditional probability of rejecting H_o, given that H_o is true. I think of a Type I error as a false alarm: You blew the whistle when you shouldn't have.

The chance of making a Type I error is equal to α, which is predetermined before you begin collecting your data. This α is the same α that represents the chance of missing the boat in a confidence interval. It makes some sense that these two probabilities are both equal because the probability of rejecting H_o when you shouldn't (a Type I error) is the same as the chance that the true population parameter falls out of the range of likely values when it shouldn't. That chance is α.

Suppose someone claims that the mean time to deliver packages for a company is 3.0 days on average (so H_o is $\mu = 3.0$), but you believe it's not equal to that (so H_a is $\mu \neq 3.0$). Your α level is 0.05, and because you have a two-sided test, you have 0.025 on each side. Your sample of 100 packages has a mean of 3.5 days with a standard deviation of 1.5 days. The test statistic equals $\frac{3.5-3.0}{\frac{1.5}{\sqrt{100}}} = 3.33$, which is greater than 1.96 (the value in the last row and the 0.025 column of the *t*-distribution table — see the Appendix). So 3.0 is not a likely value for the mean time of delivery for all packages, and you reject H_o.

But suppose that just by chance, your sample contained some longer-than-normal delivery times and that, in reality, the company's claim is right. You just made a Type I error; that is, you made a false alarm about the company's claim.

TIP

To reduce the chance of a Type I error, reduce your value of α. However, I don't recommend reducing it too far. On the positive side, this reduction makes it harder to reject H_o because you need more evidence in your data to do so. On the negative side, by reducing your chance of a false alarm (Type I error), you increase the chance of a missed opportunity (Type II error).

Missing an opportunity with a Type II error

A *Type II error* is the conditional probability of not rejecting H_o, given that H_o is false. I call it a missed opportunity because you were supposed to be able to find a problem with H_o and reject it, but you didn't. You didn't blow the whistle when you should have.

The chance of making a Type II error depends on a couple of things.

- **Sample size:** If you have more data, you're less likely to miss something that's going on. For example, if a coin actually is unfair, flipping the coin only ten times may not reveal the problem. But if you flip the coin 1,000 times, you have a good chance of seeing a pattern that favors heads over tails, or vice versa.
- **Actual value of the parameter:** A Type II error is also related to how big the problem is that you're trying to uncover. For example, suppose a company claims that the average delivery time for packages is 3.5 days. If the actual average delivery time is 5.0 days, you won't have a very hard time detecting that with your sample (even a small sample). But if the actual average delivery time is 4.0 days, you have to do more work to actually detect the problem.

TIP

To reduce the chance of a Type II error, take a larger sample size. A greater sample size makes it easier to reject H_0 (although it increases the chance of a Type I error).

REMEMBER

Type I and Type II errors sit on opposite ends of a seesaw — as one goes up, the other goes down. Try to meet in the middle by choosing a large sample size (the bigger, the better; see Figures 4-1 and 4-2) and a small α level (0.05 or less) for your hypothesis test.

The power of a hypothesis test

Type II errors, which I explain in the preceding section, show the downside of a hypothesis test. But statisticians, despite what many may think, actually try to look on the bright side once in a while; so, instead of looking at the chance of *missing* a difference from H_0 that is actually there, they look at the chance of *detecting* a difference that is really there. This detection is called the *power of a hypothesis test.*

REMEMBER

The power of a hypothesis test is 1 minus the probability of making a Type II error. So *power* is a number between 0 and 1 that represents the chance that you rejected H_0 when H_0 was false. (You can even sing about it: "If H_0 is false and you know it, clap your hands. . ..") Remember that power (just like Type II errors) depends on two elements: the sample size and the actual value of the parameter (see the preceding section for a description of these elements).

In the following sections, you discover what power means in statistics (not being one of the bigwigs, mind you); you also find out how to quantify power by using a power curve.

Throwing a power curve

The specific calculations for the power of a hypothesis test are beyond the scope of this book (so you can breathe a sigh of relief), but computer programs and graphs are available online to show you what the power is for different hypothesis tests and various sample sizes (just type "power curve for the [blah blah blah] test" into an Internet search engine).

These graphs are called *power curves* for a hypothesis test. A power curve is a special kind of graph that gives you an idea of how much of a difference from H_o you can detect with the sample size that you have. Because the precision of your test statistic increases as your sample size increases, sample size is directly related to power. But it also depends on how much of a difference from H_o you're trying to detect. For example, if a package delivery company claims that its packages arrive in 2 days or less, do you want to blow the whistle if it's actually 2.1 days? Or wait until it's 3 days? You need a much larger sample size to detect the 2.1-days situation versus the 3-days situation just because of the precision level needed.

In Figure 4-1, you can see the power curve for a particular test of H_o: $\mu = 0$ versus H_a: $\mu > 0$. You can assume that σ (the standard deviation of the population) is equal to 2 (I give you this value in each problem) and doesn't change. I set the sample size at 10 throughout.

The horizontal (x) axis on the power curve shows a range of actual values of μ. For example, you hypothesize that μ is equal to 0, but it may actually be 0.5, 1.0, 2.0, 3.0, or any other possible value. If μ equals 0, then H_o is true, and the chance of detecting this (and therefore rejecting H_o) is equal to 0.05, the set value of α. You work from that baseline. (Notice the low power in this situation makes sense because there's nothing to detect for values of μ that are close to 0.) So, on the graph in Figure 4-1, when $x = 0$, you get a y-value of 0.05.

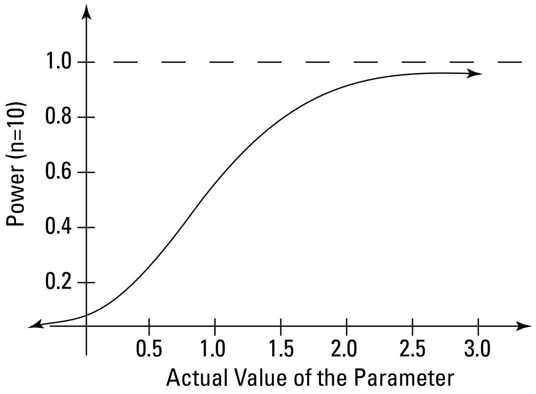


FIGURE 4-1: Power curve for H_o: $\mu = 0$ versus H_a: $\mu > 0$, for $n = 10$ and $\sigma = 2$.

Suppose that μ is actually 0.5, not 0, as you hypothesized. A computer tells you that the chance of rejecting H_o (what you're supposed to do here) is $0.197 = 0.20$, which is the power. So, you have about a 20 percent chance of detecting this difference with a sample size of 10. As you move to the right, away from 0 on the horizontal (*x*) axis, you can see that the power goes up and the *y*-values get closer and closer to 1.0.

For example, if the actual value of μ is 1.0, the difference from 0 is easier to detect than if it's 0.50. In fact, the power at 1.0 is equal to $0.475 = 0.48$, so you have almost a 50 percent chance of catching the difference from H_o in this case. And as the values of the mean increase, the power gets closer and closer to 1.0. Power never reaches 1.0 because statistics can never prove anything with 100 percent accuracy, but you can get close to 1.0 if the actual value is far enough from your hypothesis.

Controlling the sample size

How can you increase the power of your hypothesis test? You don't have any control over the actual value of the parameter, because that number is unknown. So what do you have control over? The sample size. As the sample size increases, it becomes easier to detect a real difference from H_o.

Figure 4-2 shows the power curve with the same numbers as Figure 4-1, except for the sample size (*n*), which is 100 instead of 10. Notice that the curve increases much more quickly and approaches 1.0 when the actual mean is 1.0, compared to your hypothesis of 0. You want to see this kind of curve that moves up quickly toward the value of 1.0, while the actual values of the parameter increase on the *x*-axis.

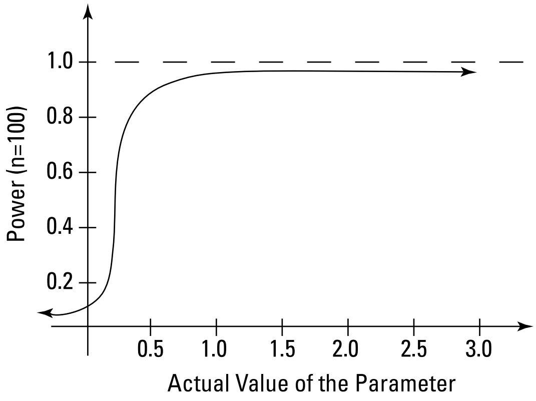


FIGURE 4-2: Power curve for $H_o: \mu = 0$ versus $H_a: \mu > 0$, for $n = 100$ and $\sigma = 2$.

If you compare the power of your test when μ is 1.0 for the $n = 10$ situation (in Figure 4-1) versus the $n = 100$ situation (in Figure 4-2), you see that the power increases from 0.475 to more than 0.999. Table 4-1 shows the different values of power for the $n = 10$ case versus the $n = 100$ case, when you test H_o: $\mu = 0$ versus H_a: $\mu > 0$, assuming a value of $\sigma = 2$.

TABLE 4-1

Comparing the Values of Power for $n = 10$ versus $n = 100$ (H_o is $\mu = 0$)

Actual Value of μ	*Power When* $n = 10$	Power When $n = 100$
0.00	$0.050 = 0.05$	$0.050 = 0.05$
0.50	$0.197 = 0.20$	$0.804 = 0.81$
1.00	$0.475 = 0.48$	approx. 1.0
1.50	$0.766 = 0.77$	approx. 1.0
2.00	$0.935 = 0.94$	approx. 1.0
3.00	$0.999 =$ approx. 1.0	approx. 1.0

REMEMBER

You can find power curves for a variety of hypothesis tests under many different scenarios. Each has the same general look and feel to it: starting at the value of α when H_o is true, increasing in an *S*-shape as you move from left to right on the *x*-axis, and finally approaching the value of 1.0 at some point. Power curves with large sample sizes approach 1.0 faster than power curves with small sample sizes.

WARNING

It's possible to have too much power. For example, if you make the power curve for $n = 10{,}000$ and compare it to Figures 4-1 and 4-2, you find that it's practically at 1.0 already for any number other than 0.0 for the mean. In other words, the actual mean could be 0.05 and with your hypothesis H_o: $\mu = 0.00$, you would reject H_o because of your huge sample size. Unless a researcher really wants to detect very small differences from H_o (such as in medical studies or quality control situations), inflated values of *n* are usually suspect. People sometimes increase *n* just to be able to say they've found a difference, no matter how small, so watch for that. If you zoom in enough, you can always detect something, even if that something makes no practical difference. Beware of surveys and experiments with an excessive sample size, such as one in the tens of thousands. Their results are guaranteed to be inflated.

POWER IN MANUFACTURING AND MEDICINE

The power of a test plays a role in the manufacturing process. Manufacturers often have very strict specifications regarding the size, weight, and/or quality of their products. During the manufacturing process, manufacturers want to be able to detect deviations from these specifications, even small ones, so they must determine how much of a difference from H_o they want to detect, and then figure out the sample size needed in order to detect that difference when it appears. For example, if a candy bar is supposed to weigh 2.0 ounces, the manufacturer may want to blow the whistle if the actual average weight shifts to 2.2 ounces. Statisticians can work backward in calculating the power and find the sample size they need to know to stop the process.

Medical scientists also think about power when they set up their studies (called *clinical trials*). Suppose they're checking to see whether an antidepressant adversely affects blood pressure (as a side effect of taking the drug). Scientists need to be able to detect small differences in blood pressure, because for some patients, any change in blood pressure is important to note and treat.

2

Using Different Types of Regression to Make Predictions

IN THIS PART . . .

Get up to speed on linear regression.

Work with more than one *x* variable in multiple regression.

Select the best model with model building.

Work with nonlinear regression.

Model yes/no data with logistic regression.

IN THIS CHAPTER

- **Using scatterplots and correlation coefficients to examine relationships**
- **Building a simple linear regression model to estimate *y* from *x***
- **Testing how well the model fits**
- **Interpreting the results and making good predictions**

Chapter **5**

Getting in Line with Simple Linear Regression

Looking for relationships and making predictions is one of the staples of data analysis. Everyone wants to answer questions like, “Can I predict how many units I’ll sell if I spend *x* amount of advertising dollars?” or “Does drinking more diet cola really relate to more weight gain?” or “Do children’s backpacks seem to get heavier with each year of school, or is it just me?”

Linear regression tries to find relationships between two or more variables and comes up with a model that tries to describe those relationships, much like how the line $y = 2x + 3$ explains the relationship between *x* and *y*. But unlike in math, where functions like $y = 2x + 3$ tell the entire story about the two variables, in statistics things don’t come out that perfectly; some variability and error is involved (that’s what makes it fun!).

This chapter is partly a review of the concepts of simple linear regression presented in a typical Stats I textbook. But the fun doesn’t stop there. I expand on the ideas about regression that you picked up in your Stats I course and set you up for some of the other types of regression models you see in Chapters 6 through 9.

In this chapter, you see how to build a simple linear regression model that examines the relationship between two variables. You also see how simple linear regression works from a model-building standpoint.

Exploring Relationships with Scatterplots and Correlations

Before looking ahead to predicting a value of *y* from *x* using a line, you need to

- Establish that you have a legitimate reason to do so by using a straight line.
- Feel confident that using a line to make that prediction will actually work well.

In order to accomplish both of these important steps, you need to first plot the data in a pairwise fashion so you can look for a visual relationship; then you need to somehow quantify that relationship in terms of how well those points follow a line. In this section, you do just that, using scatterplots and correlations.

Here's a perfect example of a situation where simple linear regression is useful: In 2018, a paper appeared in the *International Journal of Environmental Research and Public Health* called, "The Impact of Backpack Loads on School Children." This paper discussed the great concern over the weight of the textbooks in students' backpacks and the problems it presents for students. The paper referenced a study where researchers weighed a variety of textbooks from each of four core areas studied in grades 1 to 12 (reading, math, science, and history — where's statistics?) over a range of textbook brands and found the average total weight for all four books for each grade.

The study consulted pediatricians and chiropractors, who recommended that the weight of a student's backpack should not exceed 15 percent of their body weight. From there, the board hypothesized that the total weight of the textbooks in these four areas increased for each grade level and wanted to see whether it could find a relationship between the average child's weight in each grade and the average weight of their books. So, along with the average weight of the four core-area textbooks for each grade, researchers also recorded the average weight for the students in that grade. The results are shown in Table 5-1.

In this section, you begin exploring whether or not a relationship exists between these two quantitative variables. You start by displaying the pairs of data using a two-dimensional scatterplot to look for a possible pattern, and you quantify the strength and direction of that pattern using the correlation coefficient.

TABLE 5-1 Average Textbook Weight and Student Weight (Grades 1–12)

Grade	Average Student Weight (In Pounds)	Average Textbook Weight (In Pounds)
1	48.50	8.00
2	54.50	9.44
3	61.25	10.08
4	69.00	11.81
5	74.50	12.28
6	85.00	13.61
7	89.00	15.13
8	99.00	15.47
9	112.00	17.36
10	123.00	18.07
11	134.00	20.79
12	142.00	16.06

Using scatterplots to explore relationships

In order to explore a possible relationship between two variables, such as textbook weight and student weight, you first plot the data in a special graph called a *scatterplot.* A scatterplot is a two-dimensional graph that displays pairs of data, one pair per observation in the (x, y) format. Figure 5-1 shows a scatterplot of the textbook-weight data from Table 5-1.

You can see that the relationship appears to follow the straight line that's included on the graph, except possibly for the last point, where textbook weight is 16.06 pounds and student weight is 142 pounds (for grade 12). This point appears to be an *outlier* — it's the only point that doesn't fall into the pattern. So overall, an uphill, or *positive* linear relationship appears to exist between textbook weight and student weight; as student weight increases, so does textbook weight.

COMPUTER OUTPUT

To make a scatterplot in Minitab, enter the data in columns one and two of the spreadsheet. Go to Graph>Scatterplot. Click Simple and then click OK. Click on the response variable (y) in the left-hand box, and click Select. This variable shows up as the y variable in the scatterplot. Click on the explanatory (x) variable in the left-hand box, and click Select. It shows up in the x variable box. Click OK, and you get the scatterplot.

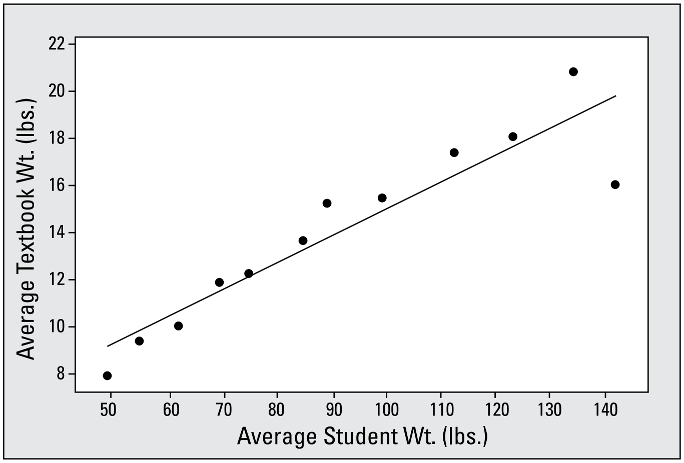


FIGURE 5-1: Scatterplot of average student weight versus average textbook weight in grades 1–12.

Collating the information by using the correlation coefficient

After you've displayed the data using a scatterplot (see the preceding section), the next step is to find a statistic that quantifies the relationship somehow. The *correlation coefficient* (also known as *Pearson's correlation coefficient,* especially in statistical software packages) measures the strength and direction of the linear relationship between two quantitative variables *x* and *y*. It's a number between –1 and +1 that's *unit-free,* which means that if you change from pounds to ounces, the correlation coefficient doesn't change. (What a messed-up world it would be if this wasn't the case!)

If the relationship between *x* and *y* is uphill, or positive (as *x* increases, so does *y*), the correlation is a positive number. If the relationship is downhill, or negative (as *x* increases, *y* gets smaller), then the correlation is negative. The following list translates different correlation values:

- **A correlation value of zero means that you can find no linear relationship between *x* and *y*.** (It may be that a different relationship exists, such as a curve; see Chapter 8 for more on this.)
- **A correlation value of +1 or –1 indicates that the points fall in a perfect, straight line.** (Negative values indicate a downhill relationship; positive values indicate an uphill relationship.)
- **A correlation value close to +1 or –1 signifies a strong linear relationship.** A general rule of thumb is that correlations close to or beyond 0.7 or –0.7 are considered to be strong.
- **A correlation closer to +0.5 or –0.5 shows a moderate linear relationship.**

- **A correlation closer to +0.3 or –0.3 shows a weak linear relationship.** In general, beware of correlations this low.
- **A correlation closer to 0 shows no linear relationship.**

You can calculate the correlation coefficient by using a formula involving the standard deviation of *x*, the standard deviation of *y*, and the covariance of *x* and *y*, which measures how *x* and *y* move together in relation to their means. However, the formula isn't the focus here (you can find it in your Stats I textbook or in my other book, *Statistics For Dummies*, 2nd Edition, published by John Wiley& Sons); it's the concept that's important. Any computer package can calculate the correlation coefficient for you with a simple click of the mouse.

COMPUTER OUTPUT

To have Minitab calculate a correlation for you, go to Stat>Basic Statistics>Correlation. Click on the name of the first of your two variables, and then click Select. Click on the name of the second variable, and then click Select. Then click OK.

The correlation for the textbook-weight example is (can you guess before looking at it?) 0.926, which is very close to 1.0. This correlation means that a very strong linear relationship is present between average textbook weight and average student weight for grades 1 to 12, and that relationship is positive and linear (it follows a straight line). This correlation is confirmed by the scatterplot shown in Figure 5-1.

WARNING

Data analysts should never make any conclusions about a relationship between *x* and *y* based solely on either the correlation or the scatterplot; the two elements need to be examined together. It's possible (but, of course, not a good idea) to manipulate graphs to look better or worse than they really are just by changing the scales on the axes. Because of this, statisticians never go with the scatterplot alone to determine whether or not a linear relationship exists between *x* and *y*. A correlation without a scatterplot is dangerous, too, because the relationship between *x* and *y* may be very strong but just not linear.

Building a Simple Linear Regression Model

After you have a handle on which *x* variables may be related to *y* in a linear way, you go about the business of finding that straight line that best fits the data. You find the slope and *y*-intercept, put them together to make a line, and use the equation of that line to make predictions for *y*. All this is part of building a simple linear regression model.

In this section, you set the foundation for regression models in general (including those you can find in Chapters 6 through 9). You plot the data, come up with a model that you think makes sense, assess how well it fits, and use it to guesstimate the value of y given another variable, x.

Finding the best-fitting line to model your data

After you've established that x and y have a strong linear relationship, as evidenced by both the scatterplot and the correlation coefficient (close to or beyond 0.7 and −0.7; see the previous sections), you're ready to build a model that estimates y using x. In the textbook-weight case, you want to estimate average textbook weight using average student weight.

The most basic of all the regression models is the *simple linear regression model* that comes in the general form of $y = \alpha + \beta x + \varepsilon$. Here, α represents the y-intercept of the line, β represents the slope, and ε represents the error in the model due to chance.

REMEMBER

A straight line that's used in simple linear regression is just one of an entire family of models (or functions) that statisticians use to express relationships between variables. A *model* is just a general name for a function that you can use to describe what outcome will occur based on some given information about one or more related variables.

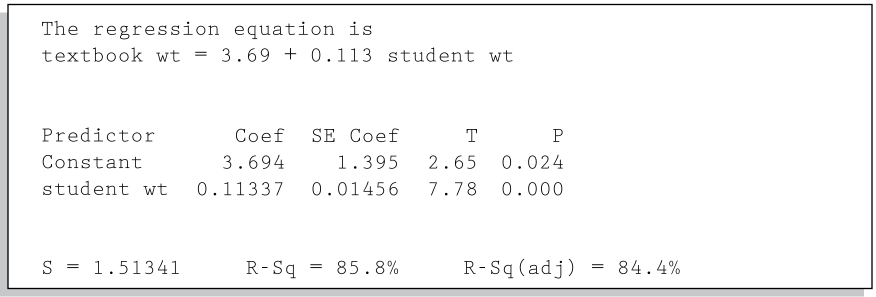
Note that you will never know the true model that describes the relationship perfectly. The best you can do is estimate it based on data.

To find the right model for your data, the idea is to scour all possible lines and choose the one that fits the data best. Thankfully, you have an algorithm that does this for you (computers use it in their calculations). Formulas also exist for finding the slope and y-intercept of the best-fitting line by hand. The best-fitting line based on your data is $y = a + bx$, where a estimates α and b estimates β from the true model. (You can find those formulas in your Stats I text or in *Statistics For Dummies*, 2nd Edition.)

COMPUTER OUTPUT

To run a linear regression analysis in Minitab, go to Stat>Regression>Regression> Fit Regression Model. Highlight the response (y) variable in the left-hand box, and click Select. The variable shows up in the Response Variable box. Then highlight your explanatory (x) variable, and click Select. This variable shows up in the (Continuous) Predictor Variable box. Click OK. (Simple linear regression as described in this chapter assumes the predictor variable (x) is quantitative, not categorical.)

The equation of the line that best describes the relationship between average textbook weight and average student weight is $y = 3.69 + 0.113x$, where *x* is the average student weight for that grade, and *y* is the average textbook weight. Figure 5-2 shows the Minitab output of this analysis.



```
The regression equation is
textbook wt = 3.69 + 0.113 student wt


Predictor        Coef  SE Coef     T      P
Constant        3.694    1.395  2.65  0.024
student wt    0.11337  0.01456  7.78  0.000


S = 1.51341     R-Sq = 85.8%     R-Sq(adj) = 84.4%
```

FIGURE 5-2: Simple linear regression analysis for the textbook-weight example.

TECHNICAL STUFF

By writing $y = 3.69 + 0.113x$, you mean that this equation represents your estimated value of *y*, given the value of *x* that you observe with your data. Statisticians technically write this equation by using a caret (or *hat*, as statisticians call it), like $\hat{y}$, so everyone can know it's an estimate, not the actual value of *y*. This *y*-hat is your estimate of the average value of *y* over the long term, based on the observed values of *x*. However, in many Stats I texts, the hat is left off because statisticians have an unwritten understanding as to what *y* represents. This issue comes up again in Chapters 6 through 9. (By the way, if you think *y*-hat is a funny term here, it's even funnier in Mexico, where statisticians call it *y-sombrero* — no kidding!)

The y-intercept of the regression line

Selected parts of that Minitab output shown in Figure 5-2 are of importance to you at this point. First, you can see that under the Coef column you have the numerical values on the right side of the equation of the line — in other words, the slope and *y*-intercept. The number 3.69 represents the coefficient of "Constant," which is a fancy way of saying that's the *y*-intercept (because the *y*-intercept is just a constant — it never changes). The *y*-intercept is the point where the line crosses the *y*-axis; in other words, it's the value of *y* when *x* equals zero.

TIP

The y-intercept of a regression line may or may not have a practical meaning, depending on the situation. To determine whether the y-intercept of a regression line has practical meaning, look at the following:

- **Does the *y*-intercept fall within the actual values in the data set?** If yes, it has practical meaning.
- **Does the *y*-intercept fall into negative territory where negative *y*-values aren't possible?** For example, if the *y*-values are weights, they can't be negative. If this is the case, then the *y*-intercept has no practical meaning. The *y*-intercept is still needed in the equation, though, because it just happens to be the place where the line, if extended to the *y*-axis, crosses the *y*-axis.
- **Does the value $x = 0$ have practical meaning?** For example, if *x* is temperature at a football game in Green Bay, then $x = 0$ is a value that's relevant to examine. If $x = 0$ has practical meaning, then the *y*-intercept does too, because it represents the value of *y* when $x = 0$. If the value of $x = 0$ doesn't have practical meaning in its own right (such as when *x* represents height of a toddler), then the *y*-intercept doesn't either.

In the textbook example, the y-intercept doesn't really have a practical meaning because students don't weigh zero pounds, so you don't really care what the estimated textbook weight is for that situation. But you do need to find a line that fits the data you do have (where average student weights go from 48.5 to 142 pounds). That best-fitting line must include a y-intercept, and for this problem, that y-intercept happens to be 3.69 pounds.

The slope of the regression line

The value 0.113 from Figure 5-2 indicates the coefficient (or number in front) of the student-weight variable. This number is also known as the *slope.* It shows that the change in y (textbook weight) is associated with a one-unit increase in x (student weight). As student weight increases by 1 pound, textbook weight increases by about 0.113 pound, on average. To make this relationship more meaningful, you can multiply both quantities by 10 to say that as student weight increases by 10 pounds, the textbook weight goes up by about 1.13 pounds on average.

TIP

Whenever you get a number for the slope, take that number and put it over 1 to help you get started on a proper interpretation of slope. For example, a slope of 0.113 is rewritten as $\frac{0.113}{1}$. Using the idea that slope equals rise over run, or change in y over change in x, you can interpret the value of 0.113 in the following way: As x increases on average by 1 pound, y increases by 0.113 pound.

Making point estimates by using the regression line

When you have a line that estimates *y* given *x*, you can use it to give a one-number estimate for the (average) value of *y* for a given value of *x*. This is called making a *point estimate.* The basic idea is to take a reasonable value of *x*, plug it into the equation of the regression line, and see what you get for the value of *y*.

In the textbook-weight example, the best-fitting line (or model) is the line $y = 3.69 + 0.113x$. For an average student who weighs 60 pounds, for example, a one-number point estimate of the average textbook weight is $3.69 + (0.113 * 60) = 10.47$ pounds (those poor little kids!). If the average student weighs 100 pounds, the estimated average textbook weight is $3.69 + (0.113 * 100) = 14.99$, or nearly 15 pounds, plus or minus something. (You find out what that something is in the following section.)

No Conclusion Left Behind: Tests and Confidence Intervals for Regression

After you have the slope of the best-fitting regression line for your data (see the previous sections), you need to step back and take into account the fact that sample results will vary. You shouldn't just say, "Okay, the slope of this line is 2. I'm done!" It won't be exactly 2 the next time. This variability is why statistics professors harp on adding a margin of error to your sample results; you want to be sure to cover yourself by adding that plus or minus.

In hypothesis testing, you don't just compare your sample mean to the population mean and say, "Yep, they're different alright!" You have to standardize your sample result using the standard error so that you can put your results in the proper perspective (see Chapter 4 for a review of confidence intervals and hypothesis tests).

The same idea applies here with regression. The data were used to figure out the best-fitting line, and you know it fits well for those data. That's not to say that the best-fitting line will work perfectly well for a new data set taken from the same population. So, in regression, all your results should involve the standard error with them in order to allow for the fact that sample results vary. That also goes for estimating and testing for the slope and *y*-intercept and for any predictions that you make.

In Stats I courses, the concept of margin of error is often skipped over after the best-fitting regression line is found, but this is a very important idea and should always be included. (Okay, enough of the soap box for now. Let's get out there and do it!)

Scrutinizing the slope

Recall the *slope* of the regression line is the amount by which you expect the *y* variable to change on average as the *x* variable increases by 1 unit — the old rise-over-run idea (see the section, "The slope of the regression line," earlier in this chapter). Now, how do you deal with knowing that the best-fitting line will change with a new data set? You just apply the basic ideas of confidence intervals and hypothesis tests (see Chapter 4).

A confidence interval for slope

A *confidence interval* in general has this form: your statistic plus or minus a margin of error. The margin of error includes a certain number of standard deviations (or standard errors) from your statistic. How many standard errors you add and subtract depends on what confidence level, $1-\alpha$, you want. The size of the standard error depends on the sample size and other factors.

The equation of the best-fitting simple linear regression line, $y = a + bx$, includes a slope (b) and a y-intercept (a). Because these were found using the data, they're only estimates of what's going on in the population, and therefore they need to be accompanied by a margin of error.

The formula for a $1-\alpha$ level confidence interval for the slope of a regression line is $b \pm t^*_{n-2} * SE_b$, where the standard error is denoted $SE_b = \dfrac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}}$, where $s = \sqrt{\dfrac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2}$. The value of t^* comes from the t-distribution with $n-2$ degrees of freedom, and the area to its right is equal to $\alpha \div 2$. (See Chapter 4 regarding the concept of α.)

TECHNICAL STUFF

In case you wonder why you see $n-2$ degrees of freedom here, as opposed to $n-1$ degrees of freedom used in t-tests for the population mean in Stats I, here's the scoop. From Stats I you know that a *parameter* is a number that describes the population; it's usually unknown, and it can change from scenario to scenario. For each parameter in a model, you lose 1 degree of freedom. The regression line contains two parameters — the slope and the y-intercept — and you lose 1 degree of freedom for each one. With the t-test from Stats I, you only have one parameter, the population mean, to worry about; hence, you use $n-1$ degrees of freedom.

REMEMBER

You can find the value of *t** in any *t*-distribution table (check your textbook for one, or check the Appendix of this book). Or you can always have Minitab calculate it for you. For example, suppose you want to find a 95 percent confidence interval based on sample size $n = 10$. The value of *t** is found in Table A-1 in the Appendix in the row marked $10 - 2 = 8$ degrees of freedom, and the column marked 0.025 (because $\alpha \div 2 = 0.05 \div 2 = 0.025$). This value of *t** is 2.306. (*Statistics For Dummies*, 2nd Edition can tell you a lot more about the *t*-distribution and the *t*-table.)

To put together a 95 percent confidence interval for the slope using computer output, you pull off the pieces that you need. For the textbook-weight example, in Figure 5-2 you see that the slope is equal to 0.11337. (Recall that slope is the coefficient of the *x* variable in the equation, which is why you see the abbreviation *Coef* in the output.)

Because the slope changes from sample to sample, it's a random variable with its own distribution, its own mean, and its own standard error. (Recall from Stats I that the standard error of a statistic is likened to the standard deviation of a random variable.) If you look just to the right of the slope in Figure 5-2, you see SE Coef; this stands for the standard error of the slope (which is 0.01456 in this case).

Now all you need is the value of *t** from the *t*-table (Table A-1 in the Appendix). Because $n = 12$, you look in the row where degrees of freedom is $12 - 2 = 10$. You want a 95 percent confidence interval, so you look in the column for $(1 - 0.95) \div 2 = 0.25$. The *t** value you get is 2.228.

Putting these pieces together, a 95 percent confidence interval for the slope of the best-fitting regression line for the textbook-weight example is $0.11337 \pm 2.228 * 0.01456$, which goes from 0.0809 to 0.1458. The units are in pounds (textbook weight) per pounds (child weight). Note that this interval is large due to the small sample size, which increases the standard error.

A hypothesis test for slope

You may be interested in conducting a hypothesis test for the slope of a regression line as another way to assess how well the line fits. If the slope is zero or close to it, the regression line is basically flat, signifying that no matter the value of *x*, you'll always estimate *y* by using its mean. This means that *x* and *y* aren't related at all, so a specific value of *x* doesn't help you predict a specific value for *y*. You can also test to see if the slope is some value other than zero, but that's atypical. So for all intents and purposes, I use the hypotheses $H_o : \beta = 0$ versus $H_a : \beta \neq 0$, where β is the slope of the true model.

To conduct a hypothesis test for the slope of a simple linear regression line, you follow the basic steps of any hypothesis test. You take the statistic (*b*) from your data, subtract the value in H_0 (in this case it's zero), and standardize it by dividing by the standard error (see Chapter 4 for more on this process).

Using the formula for standard error for b, the test statistic for the hypothesis test of whether or not the slope equals zero is $\frac{b-0}{SE_b}$, where $SE_b = \frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}}$, and $s = \sqrt{\frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2}$. On the Minitab output from Figure 5-2, the test statistic is located right next to the SE Coef column; it's cleverly marked T. In this case, $T = 7.78$. Compare this value to $t^* = 2.228$ from the t-table. Because $T > t^*$, you have strong evidence to reject H_0 and conclude that the slope of the regression line for the textbook-weight data is not zero. (In fact, it has to be greater than that, according to your data.)

You can also just find the exact p-value on the coefficients part of the Minitab output in Figure 5-2, right next to the T column, in the column marked P (for p-value). In this case, the p-value for the test for slope is 0.000, which means it's less than 0.0005. You conclude that the slope of this line is not zero, so textbook weight is significantly related to student weight. (See Chapter 4 to brush up on p-values.)

TECHNICAL STUFF

To test to see whether the slope is some value other than zero, just plug that value in for b_0 in the formula for the test statistic. You may also want to conduct one-sided hypothesis tests to see whether the slope is strictly greater than zero or strictly less than zero. In those cases, you find the same test statistic but compare it to the value t^* where the area to the right (or left, respectively) is α.

Inspecting the y-intercept

The y-intercept is the place where the regression line $y = a + bx$ crosses the y-axis and is denoted by a (see the earlier section, "The y-intercept of the regression line"). Sometimes the y-intercept can be interpreted in a meaningful way, and sometimes not. This differs from slope, which is always interpretable. In fact, between the two elements of slope and intercept, the slope is the star of the show, with the y-intercept serving as the somewhat less famous but still noticeable sidekick.

There are times when the y-intercept makes no sense. For example, suppose you use rain to predict bushels per acre of corn; if you have zero rain, you have zero corn, but if the regression line crosses the y-axis somewhere else besides zero (and it most likely will), the y-intercept will make no sense. Another situation is where no data were collected near the value of $x = 0$; interpreting the y-intercept at that point is not appropriate. For example, using a student's score on midterm

1 to predict their score on midterm 2, unless the student didn't take the exam at all (in which case it doesn't count), they'll get at least some points.

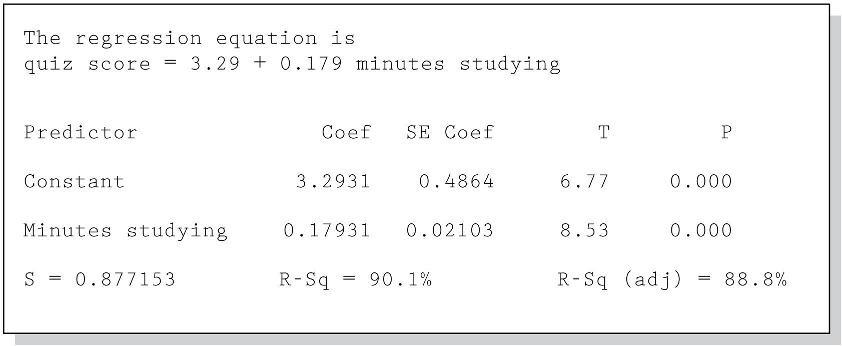
Many times, however, the *y*-intercept is of interest to you and has a value that you can interpret, such as when you're talking about predicting coffee sales using temperature for football games. Some games get cold enough to have zero and subzero temperatures (like Packers games, for example — Go, Pack, Go!).

Suppose I collect data on ten of my students who recorded their study time (in minutes) for a 10-point quiz, along with their quiz scores. The data have a strong linear relationship by all the methods used in this chapter (for example, refer to the earlier section, "Exploring Relationships with Scatterplots and Correlations"). I went ahead and conducted a regression analysis, and the results are shown in Figure 5-3.

Because there are students who (heaven forbid!) didn't study at all for the quiz, the *y*-intercept of 3.29 points (where study time $x = 0$) can be interpreted safely. Its value is shown in the Coef column in the row marked Constant (see the section, "The *y*-intercept of the regression line," for more information). The next step is to give a confidence interval for the *y*-intercept of the regression line, where you can take conclusions beyond just this sample of ten students.

The formula for a $1-\alpha$ level confidence interval for the *y*-intercept (*a*) of a simple linear regression line is $a \pm t^*_{n-2} SE_a$. The standard error, $\mathrm{SE_a}$, is equal to $SE_a = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}$, where $s = \sqrt{\frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2}$, where again, the value of *t** comes from the *t*-distribution with $n-2$ degrees of freedom whose area to the right is equal to $\alpha \div 2$. Using the output from Figure 5-3 and the *t*-table, I'm 95 percent confident that the quiz score (*y*) for someone with a study time of $x = 0$ minutes is $3.29 \pm 2.306 * 0.4864$, which is anywhere from 2.17 to 4.41, on average. Note that 2.306 comes from the *t*-table with $10 - 2 = 8$ degrees of freedom and 0.4864 is the SE for the *y*-intercept from Figure 5-3. (So studying for zero minutes for my quiz is not something to aspire to.)

By the way, to find out how much time studying affected the quiz score for these students, you can get an estimate of the slope on the output from Figure 5-3 that the coefficient for slope is 0.1793, which says each minute of studying is related to an increase in score of 0.1793 of a point, plus or minus the margin of error, of course. Or, 10 more minutes relates to 1.793 more points. On a 10-point quiz, it all adds up!



```
The regression equation is
quiz score = 3.29 + 0.179 minutes studying


Predictor               Coef   SE Coef        T         P

Constant              3.2931    0.4864     6.77     0.000

Minutes studying     0.17931   0.02103     8.53     0.000

S = 0.877153         R-Sq = 90.1%         R-Sq (adj) = 88.8%
```

FIGURE 5-3: Regression analysis for study time and quiz score data.

TECHNICAL STUFF

Testing a hypothesis about the y-intercept isn't really something you'll find yourself doing much because most of the time you don't have a preconceived notion about what the y-intercept would be (nor do you really care ahead of time). The confidence interval is much more useful. However, if you do need to conduct a hypothesis test for the y-intercept, you take your y-intercept, subtract the value in H_0, and divide by the standard error, found on the computer output in the row for Constant and the column for SE Coef. (The default value is to test to see whether the y-intercept is zero.) The test is in the T (T-value) column of the output, and its p-value is shown in the P (p-value) column. In the study time and quiz score example, the p-value is 0.000, so the y-intercept is significantly different from zero. All this means is that the line crosses the y-axis somewhere else.

Building confidence intervals for the average response

When you have the slope and y-intercept for the best-fitting regression line, you put them together to get the line $y = a + bx$. The value of y here really represents the average value of y for a particular value of x. For example, in the textbook-weight data, Figure 5-2 shows the regression line $y = 3.69 + 0.11337x$, where $x =$ average student weight and y = average textbook weight. If you put in 100 pounds for x, you get $y = 3.69 + 0.1137 * 100 = 15.02$ pounds of textbook weight for the group averaging 100 pounds. This number, 15.02, is an estimate of the average weight of textbooks for children of this weight.

But you can't stop there. Because you're getting an estimate of the average textbook weight using y, you also need a margin of error for y to go with it, to create a confidence interval for the average y at a given x that generalizes to the population.

Take your estimate, y, which you get by plugging your given x value into the regression line, and then add and subtract the margin of error for y. The formula for a $1-\alpha$ confidence interval for the mean of y for a given value of x (call it x^*) is equal to $y \pm t^*_{n-2} SE_{\hat{\mu}}$, where y is the value of the equation of the line when you plug in x^* for x. The standard error for y is equal to $SE_y = s\sqrt{1+\frac{1}{n}+\frac{\left(x^*-\bar{x}\right)^2}{\sum_i \left(x_i-\bar{x}\right)^2}}$, where $s = \sqrt{\frac{1}{n-2}\sum_i \left(y_i-\hat{y}_i\right)^2}$. Luckily, Minitab does these calculations for you and reports a confidence interval for the mean of y for a given x^*.

COMPUTER OUTPUT

To find a confidence interval for the mean value of y using Minitab, you ask for a regression analysis (see instructions in the earlier section, “Finding the best-fitting line to model your data”) and click on Predict. That is, you go to Stat> Regression>Regression>Predict. In the Response box, you see your y variable. You also see a box that contains both your x value and an empty column for numbers. Enter the value of x^* that you want to find a prediction for, and click OK. (If you want to change the confidence level from 95%, click Options and change the level before clicking OK.) On the computer output, the confidence interval is labeled 95% CI. You also see an interval labeled PI, which is a prediction interval. (Prediction intervals are different from confidence intervals, and I discuss them in the next section.)

Returning to the textbook-weight example, the computer output for finding a 95 percent confidence interval for the average textbook weight for 100-pound children is shown in Figure 5-4. The result is (14.015, 16.048) pounds. I’m 95 percent confident that the average textbook weight for the group of children averaging 100 pounds is between 14.015 and 16.048 pounds. (Get out those rolling backpacks, kids!)

WARNING

You should only make predictions for the average value of y for x values that are within the range of where the data was collected. Failure to do so will result in the statistical no-no called *extrapolation* (see the later section, “Knowing the Limitations of Your Regression Analysis”).

Making the band with prediction intervals

Suppose that instead of the mean value of y, you want to take a guess at what y would be for some future value of x. Because you’re looking into the future, you have to make a prediction, and to do that, you need a range of likely values of y for a given x^*. This is what statisticians call a *prediction interval.*

The formula for a 1 – α level prediction interval for *y* at a given value *x** is $SE_y = s\sqrt{1+\frac{1}{n}+\frac{\left(x^*-\bar{x}\right)^2}{\sum_i \left(x_i-\bar{x}\right)^2}}$, where $s=\sqrt{\frac{1}{n-2}\sum_i \left(y_i-\hat{y}_i\right)^2}$. Again, Minitab easily makes these calculations for you.

COMPUTER OUTPUT

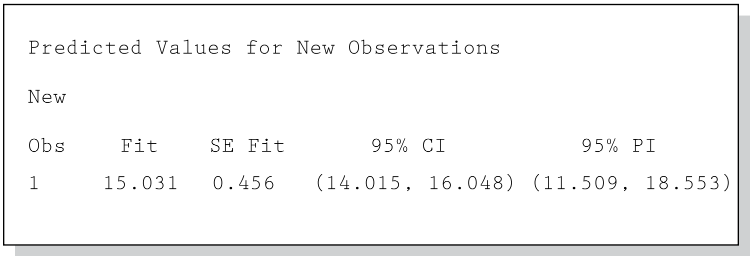
To find a $1-\alpha$ level prediction interval for the value of *y* for a given *x** using Minitab, you ask for a regression analysis (see instructions in the earlier section, "Finding the best-fitting line to model your data") and click Predict. In the column that is labeled with your *x*-variable, enter the value of *x* that you want, and click OK. (The default confidence level is 95 percent. If you want to change it, click Options and then change the level before clicking OK.) On the computer output, the prediction interval is labeled 95% PI, and it appears right next to the confidence interval for the mean of *y* for that same *x**.

Predicting textbook weight using student weight

For the textbook-weight data, suppose you've already made your regression line and now a new student comes on the scene. You want to predict this student's textbook weight. This means you want a prediction interval rather than a confidence interval, because you want to predict the textbook weight for one person, not the average weight for a group.

Suppose this new student weighs 100 pounds. To find the prediction interval for the textbook weight for this student, you use $x^*=100$ pounds and let Minitab do its thing.

The computer output in Figure 5-4 shows that the 95 percent prediction interval for textbook weight for a single 100-pound child is (11.509, 18.533) pounds. Note that this is wider than the confidence interval of (14.015, 16.048) for the mean textbook weight for 100-pound children found in the earlier section, "Building confidence intervals for the average response." This difference is due to the increased variability in looking at one child and predicting one textbook weight.



```
Predicted Values for New Observations

New

Obs    Fit    SE Fit       95% CI           95% PI

1     15.031   0.456   (14.015, 16.048) (11.509, 18.553)
```

FIGURE 5-4: Prediction interval of textbook weight for a 100-pound child.

Comparing prediction and confidence intervals

Note that the formulas for prediction intervals and confidence intervals are very similar. In fact, the prediction interval formula is exactly the same as the confidence interval formula, except that it adds a 1 under the square root. Because of this difference in the formulas, the margin of error for a prediction interval is larger than for a confidence interval.

This difference also makes sense from a statistical standpoint. A prediction interval has more variability than a confidence interval because it's harder to make a prediction about *y* for a single value of *x** than it is to estimate the average value of *y* for a given *x**. (For example, individual test scores vary more than average test scores do.) A prediction interval will be wider than a confidence interval; it will have a larger margin of error.

A similarity between prediction intervals and confidence intervals is that their margin of error formulas both contain *x**, which means the margin of error in either case depends on which value of *x** you use. It turns out in both cases that if you use the mean value of *x* as your *x**, the margin of error for each interval is at its smallest because there's more data around the mean of *x* than at any other value. As you move away from the mean of *x*, the margin of error increases for each interval.

Checking the Model's Fit (The Data, Not the Clothes!)

After you've established a relationship between *x* and *y* and have come up with an equation of a line that represents that relationship, you may think your job is done. (Many researchers erringly stop here, so I'm depending on you to break the cycle!) The most-important job remains to be completed: checking to be sure that the conditions of the model are truly met and that the model fits well in more specific ways than the scatterplot and correlation measure (which I cover in the earlier section, "Exploring Relationships with Scatterplots and Correlations").

This section presents methods for defining and assessing the fit of a simple linear regression model.

Defining the conditions

Two major conditions must be met before you apply a simple linear regression model to a data set:

- The *y*'s must have an approximately normal distribution for each value of *x*.
- The *y*'s must have a constant amount of spread (standard deviation) for each value of *x*.

Normal y's for every x

For any value of *x*, the population of possible *y*-values must have a normal distribution. The mean of this distribution is the value for *y* that's on the best-fitting line for that *x*-value. That is, some of your data fall above the best-fitting line, some data fall below the best-fitting line, and a few may actually land right on the line.

TIP

If the regression model is fitting well, the data values should be scattered around the best-fitting line in such a way that about 68 percent of the values lie within one standard deviation of the line, about 95 percent of the values lie within two standard deviations of the line, and about 99.7 percent of the values lie within three standard deviations of the line. This specification, as you may recall from your Stats I course, is called the *68-95-99.7% Rule,* and it applies to all bell-shaped data (for which the normal distribution applies).

You can see in Figure 5-5 how for each *x*-value, the *y*-values you may observe tend to be located near the best-fitting line in greater numbers, and as you move away from the line, you see fewer and fewer *y*-values, both above and below the line. More than that, they're scattered around the line in a way that reflects a bell-shaped curve, the normal distribution. This indicates a good fit.

Why does this condition make sense? The data you collect on *y* for any particular *x*-value vary from individual to individual; for example, not all students' textbooks weigh the same, even for students who weigh the exact same amount. But those values aren't allowed to vary any way they want to. To fit the conditions of a linear regression model, for each given value of *x*, the data should be scattered around the line according to a normal distribution. Most of the points should be close to the line, and as you get farther from the line, you can expect fewer data points to occur. So condition number one is that the data have a normal distribution for each value of *x*.

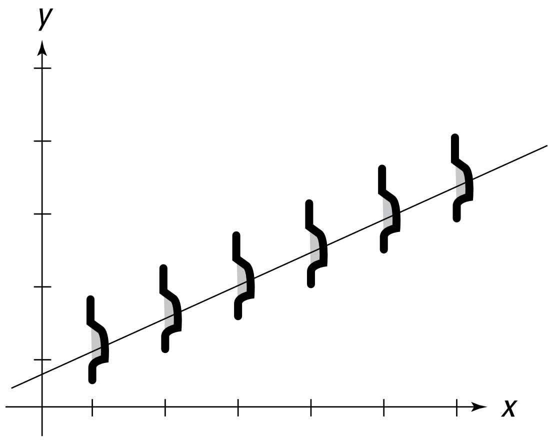


FIGURE 5-5: Conditions of a simple linear regression model.

Same spread for every x

In order to use the simple linear regression model, as you move from left to right on the *x*-axis, the spread in the *y*-values around the line should be the same, no matter which value of *x* you're looking at. This requirement is called the *homoscedasticity condition.* (How they came up with that mouthful of a word just for describing the fact that the standard deviations stay the same across the *x*-values, I'll never know.) This condition ensures that the best-fitting line works well for all relevant values of *x*, not just in certain areas.

You can see in Figure 5-5 that no matter what the value of *x* is, the spread in the *y*-values stays the same throughout. If the spread got bigger and bigger as *x* got larger and larger, for example, the line would lose its ability to fit well for those large values of *x*.

Finding and exploring the residuals

To check to see whether the *y*-values come from a normal distribution, you need to measure how far off your predictions were from the actual data that came in. These differences are called *errors,* or *residuals.* To evaluate whether a model fits well, you need to check those errors and see how they stack up.

TIP

In a model-fitting context, the word *error* doesn't mean "mistake." It just means a difference between the data and the prediction based on the model. The word I like best to describe this difference is *residual,* however. It sounds more upbeat.

The following sections focus on finding a way to measure these residuals that the model makes. You also explore the residuals to identify particular problems that occurred in the process of trying to fit a straight line to the data. In other words, you can discover that looking at residuals helps you assess the fit of the model and diagnose problems that caused a bad fit, if that was the case.

Finding the residuals

A *residual* is the difference between the observed value $\hat{y}$ of y (from the best-fitting line) and the predicted value of y, also known as y (from the data set). Its notation is $(y - \hat{y})$. Specifically, for any data point, you take its observed y-value (from the data) and subtract its expected y-value (from the line). If the residual is large, the line doesn't fit well in that spot. If the residual is small, the line fits well in that spot.

For example, suppose you have a point in your data set (2,4) and the equation of the best-fitting line is $y = 2x + 1$. The expected value of y in this case is $(2 * 2) + 1 = 5$. The observed value of y from the data set is 4. Taking the observed value minus the estimated value, you get $4 - 5 = -1$. The residual for that particular data point (2,4) is −1. If you observe a y-value of 6 and use the same straight line to estimate y, then the residual is $6 - 5 = +1$.

REMEMBER

In general, a positive residual means you underestimated y at that point; the line is below the data. A negative residual means you overestimated y at that point; the line is above the data.

Standardizing the residuals

Residuals in their raw form are in the same units as the original data, making them hard to judge out of context. To make interpreting the residuals easier, statisticians typically *standardize* them — that is, subtract the mean of the residuals (zero) and divide by the standard deviation of all the residuals. The residuals are a data set just like any other data set, so you can find their mean and standard deviation like you always do. Standardizing just means converting to a Z-score so that you see where it falls on the standard normal distribution. (See your Stats I text or *Statistics For Dummies,* 2nd Edition for information on Z-scores.)

Making residual plots

You can plot the residuals on a graph called a *residual plot.* (If you've standardized the residuals, you call it a *standardized residual plot.*) Figure 5-6 shows the Minitab output for a variety of standardized residual plots, all getting at the same idea: checking to be sure the conditions of the simple linear regression model are met.

COMPUTER OUTPUT

To make a residual plot in Minitab, click Stat>Regression>Regression>Fit Regression Model, and then click Graphs. Under Residual Plots, select Four in One. You will get the four plots shown in Figure 5-6.

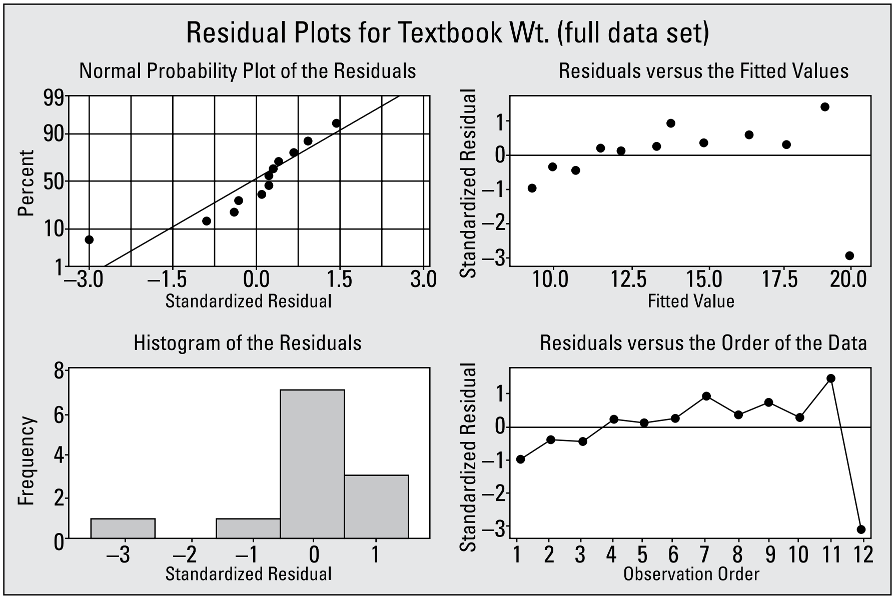


FIGURE 5-6: Standardized residual plots for textbook-weight data.

Checking normality

If the condition of normality is met, you can see on the residual plot lots of (standardized) residuals close to zero; as you move farther away from zero, you can see fewer residuals. ***Note:*** You shouldn't expect to see a standardized residual at or beyond +3 or –3. If this occurs, you can consider that point an outlier, which warrants further investigation. (For more on outliers, see the section, "Scoping for outliers," later in this chapter.)

REMEMBER

The residuals should also occur at random — some above the line, and some below the line. If a pattern occurs in the residuals, the line may not be fitting right.

You can also look at the upper-left plot in Figure 5-6, the normal probability plot of the residuals. This plot examines what you got for your residuals versus what you would expect to get if the residuals had a normal distribution. If the normal probability plot shows a straight line, you are on target for having normality. Strong departures from a straight line show the residuals may not be normally distributed.

The plots in Figure 5-6 seem to have an issue with the very last observation, the one for 12th graders. In this observation, the average student weight (142) seems to follow the pattern of increasing with each grade level, but the textbook weight (16.06) is less than for 11th graders (20.79) and is the first point to break the pattern.

You can also see in the plot in the upper-right corner of Figure 5-6 that the very last data value has a standardized residual that sticks out from the others and has a value of -3 (something that should be a very rare occurrence). So the value you expected for y based on your line was off by a factor of 3 standard deviations. And because this residual is negative, what you observed for y was much lower than you may have expected it to be using the regression line.

The other residuals seem to fall in line with a normal distribution, as you can see in the upper-right plot of Figure 5-6. The residuals concentrate around zero, with fewer appearing as you move farther away from zero. You can also see this pattern in the upper-left plot of Figure 5-6, which shows how close to normal the residuals are. The line in this graph represents the equal-to-normal line. If the residuals follow close to the line, then normality is okay. If not, you have problems (in a statistical sense, of course). You can see the residual with the highest magnitude is -3, and that number falls outside the line quite a bit.

The lower-left plot in Figure 5-6 makes a histogram of the standardized residuals, and you can see it doesn't look much like a bell-shaped distribution. It doesn't even look *symmetric* (the same on each side when you cut it down the middle). The problem again seems to be the residual of -3, which skews the histogram to the left.

The lower-right plot of Figure 5-6 plots the residuals in the order presented in the data set in Table 5-1. Because the data was ordered already, the lower-right residual plot looks like the upper-right residual plot in Figure 5-6, except the dots are connected. This lower-right residual plot makes the residual of -3 stand out even more.

Checking the spread of the y's for each x

The graph in the upper-right corner of Figure 5-6 also addresses the homoscedasticity condition. If the condition is met, then the residuals for every x-value have about the same spread. If you cut a vertical line down through each x-value, the residuals have about the same spread (standard deviation) each time, except for the last x-value, which again represents grade 12. That means the condition of equal spread in the y-values is met for the textbook-weight example.

REMEMBER

If you look at only one residual plot, choose the one in the upper-right corner of Figure 5-6, the plot of the fitted values (the values of y on the line) versus the standardized residuals. Most problems with model fit will show up on that plot because a residual is defined as the difference between the observed value of y and the fitted value of y. In a perfect world, all the fitted values have no residual at all; a large residual (such as the one where the estimated textbook weight is 20 pounds for students averaging 142 pounds; see Figure 5-1) is indicated by a point far off from zero. This graph also shows you deviations from the overall pattern of the line; for example, if large residuals are on the extremes of this graph (very low or very high fitted values), the line isn't fitting in those areas. On balance, you can say this line fits well at least for grades 1 through 11.

Using r² to measure model fit

One important way to assess how well the model fits is to use a statistic called the *coefficient of determination,* or r^2. This statistic takes the value of the correlation, r, and squares it to give you a percentage. You interpret r^2 as the percentage of variability in the y variable that's explained by, or due to, its relationship with the x variable. (Note this doesn't mean X necessarily causes Y to be what it is, it's just helping to explain why it varies as much as it does.)

The y-values of the data you collect have a great deal of variability in and of themselves. You look for another variable (x) that helps you explain that variability in the y-values. After you put that x variable into the model and find that it's highly correlated with y, you want to find out how well this model did at explaining why the values of y are different.

REMEMBER

Note that you have to interpret r^2 using different standards than those for interpreting r. Because squaring a number between -1 and $+1$ results in a smaller number (except for $+1$, -1, and 0, which stay the same or switch signs), an r^2 of 0.49 isn't too bad, because it's the square of $r = 0.7$, which is a strong correlation.

TIP

The following are some general guidelines for interpreting the value of r^2:

- **If the model containing *x* explains a lot of the variability in the *y*-values, then r^2 is high** (in the 80 to 90 percent range is considered to be extremely high). Values like 0.70 are still considered fairly high. A high percentage of variability means that the line fits well because there's not much left to explain about the value of y other than using x and its relationship to y. So a larger value of r^2 is a good thing.
- **If the model containing *x* doesn't help much in explaining the difference in the *y*-values, then the value of r^2 is small** (closer to zero; between, say, 0.00 and 0.30, roughly). The model, in this case, wouldn't fit well. You need another variable to explain y other than the one you already tried.

- **Values of r^2 that fall in the middle** (between, say, 0.30 and 0.70) mean that *x* does help somewhat in explaining *y*, but it doesn't do the job well enough on its own. In this case, statisticians would try to add one or more variables to the model to help explain *y* more fully as a group (read more about this in Chapter 6).

For the textbook-weight example, the value of *r* (the correlation coefficient) is 0.93. Squaring this result, you get $r^2 = 0.8649$. That number means approximately 86 percent of the variability you find in average textbook weights for all students (*y*-values) is explained by the average student weight (*x*-values). This percentage tells you that the model of using year in school to estimate backpack weight is a good bet.

In the case of simple linear regression, you have only one *x* variable, but in Chapter 6, you can see models that contain more than one *x* variable. In that situation, you use r^2 to help sort out the contribution that those *x* variables as a group bring to the model.

Scoping for outliers

Sometimes life isn't perfect (oh really?), and you may find a residual in your otherwise tidy data set that totally sticks out. It's called an *outlier*, and it has a standardized value at or beyond +3 or −3. It threatens to blow the conditions of your regression model and send you crying to your professor.

Before you panic, the best thing to do is to examine that outlier more closely. First, can you find an error in that data value? Did someone report her age as 642, for instance? (After all, mistakes do happen.) If you do find a certifiable error in your data set, you remove that data point (or fix it if possible) and analyze the data without it. However, if you can't explain away the problem by finding a mistake, you must think of another approach.

If you can't find a mistake that caused the outlier, you don't necessarily have to trash your model; after all, it's only one data point. Analyze the data with that data point, and analyze the data again without it. Then report and compare both analyses. This comparison gives you a sense of how influential that one data point is, and it may lead other researchers to conduct more research to zoom in on the issue you brought to the surface.

In Figure 5-1, you can see the scatterplot of the full data set for the textbook-weight example. Figure 5-7 shows the scatterplot for the data set minus the outlier. The scatterplot fits the data better without the outlier. The correlation increases to 0.993, and the value of r^2 increases to 0.986. The equation for the regression line for this data set is $y = 1.78 + 0.139x$.

The slope of the regression line doesn't change much by removing the outlier (compare it to Figure 5-2, where the slope is 0.113). However, the *y*-intercept changes: It's now 1.78 without the outlier compared to 3.69 with the outlier. The slopes of the lines are about the same, but the lines cross the *y*-axis in different places. It appears that the outlier (the last point in the data set) has quite an effect on the best-fitting line.

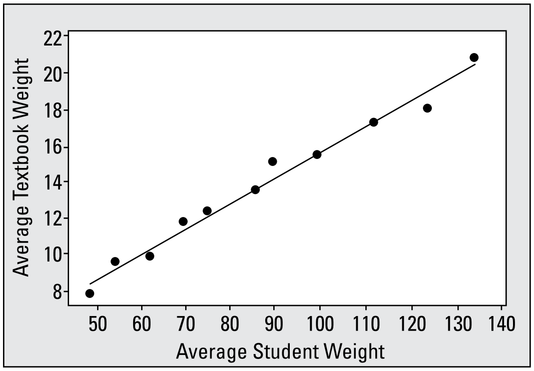


FIGURE 5-7: Scatterplot of textbook-weight data minus the outlier.

Figure 5-8 shows the residual plots for the regression line for the data set without the outlier. Each of these plots shows a much better fit of the data to the model compared to Figure 5-6. This result tells you that the data for grade 12 is influential in this data set and that the outlier needs to be noted and perhaps explored further. Do students peak when they're juniors in high school? Or do they just decide when they're seniors that it isn't cool to carry books around? (A statistician's job isn't to wonder why, but to do and analyze, and bring the why questions to their collaborators.)

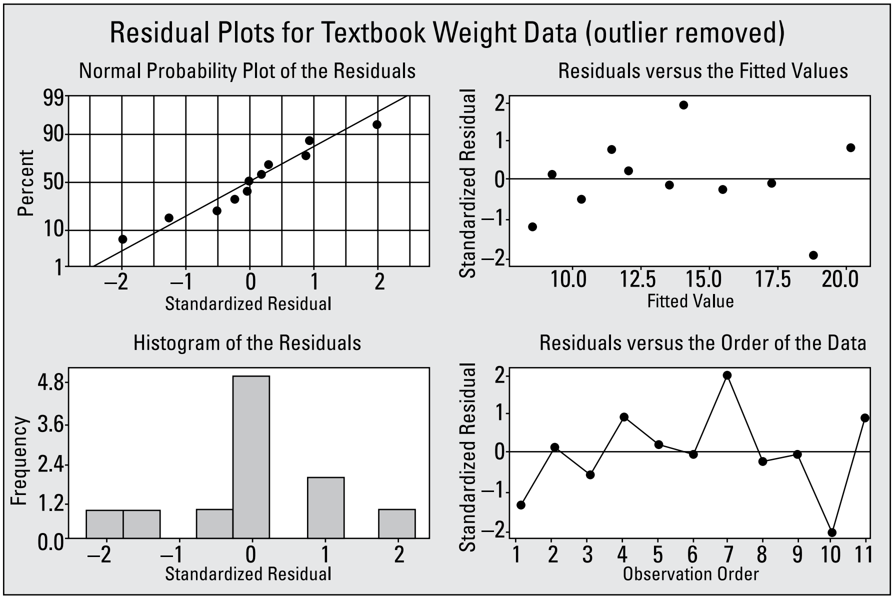


FIGURE 5-8: Residual plots for textbook-weight data minus the outlier.

Knowing the Limitations of Your Regression Analysis

The bottom line of any data analysis is to make the correct conclusions given your results. When you're working with a simple linear regression model, there's the potential to make three major errors. This section shows you those errors and tells you how to avoid them.

Avoiding slipping into cause-and-effect mode

In a simple linear regression, you investigate whether *x* is related to *y*, and if you get a strong correlation and a scatterplot that shows a linear trend, then you find the best-fitting line and use it to estimate the value of *y* for reasonable values of *x*.

WARNING

There's a fine line, however (no pun intended), that you don't want to cross with your interpretation of regression results. Be careful to not automatically interpret slope in a cause-and-effect mode when you're using the regression line to estimate the value of *y* using *x*. Doing so can result in a leap of faith that can send you into the frying pan. Unless you have used a controlled experiment to get the data, you can only assume that the variables are correlated; you can't really give a stone-cold guarantee about why they're related.

In the textbook-weight example, you estimate the average weight of the students' textbooks by using the students' average weight, but that doesn't mean increasing a particular child's weight causes their textbook weight to increase. For example, because of the strong positive correlation, you do know that students with lower weights are associated with lower total textbook weights, and students with higher weights tend to have higher textbook weights. But you can't take one particular third-grade student, increase their weight, and presto — suddenly their textbooks weigh more.

The variable underlying the relationship between a child's weight and the weight of their backpack is the grade level of the student from an academic standpoint; as grade level increases, so might the size and number of their books, as well as the homework coming home. Student grade level drives both student weight and textbook weight. In this situation, student grade level is what statisticians call a *lurking variable;* it's a variable that wasn't included in the model but is related to both the outcome and the response. A lurking variable confuses the issue of what's causing what to happen.

REMEMBER

If the collected data was the result of a well-designed experiment that controls for possible confounding variables, you can establish a cause-and-effect relationship between x and y if they're strongly correlated. Otherwise, you can't establish such a relationship. (See your Stats I text or *Statistics For Dummies,* 2nd Edition for information regarding experiments.)

Extrapolation: The ultimate no-no

Plugging values of x into the model that fall outside of the reasonable boundaries of x is called *extrapolation.* And one of my colleagues sums up this idea very well: "Friends don't let friends extrapolate."

When you determine a best-fitting line for your data, you come up with an equation that allows you to plug in a value for x and get a predicted value for y. In algebra, if you find the equation of a line and graph it, the line typically has an arrow on each end, indicating it goes on forever in either direction. But that doesn't work for statistical problems (because statistics represents the *real* world). When you're dealing with real-world units like height, weight, IQ, GPA, house prices, and the weight of your statistics textbook, only certain numbers make sense.

So the first point is, don't plug in values for x that don't make any sense. For example, if you're estimating the price of a house (y) by using its square footage (x), you wouldn't think of plugging in a value of x like 10 square feet or 100 square feet, because houses simply aren't that small.

You also wouldn't think about plugging in values like 1,000,000 square feet for x (unless your "house" is the Ohio State football stadium or something). It wouldn't make sense. Likewise, if you're estimating tomorrow's temperature using today's temperature, negative numbers for *x* could possibly make sense, but if you're estimating the amount of precipitation tomorrow given the amount of precipitation today, negative numbers for *x* (or *y* for that matter) don't make sense.

REMEMBER

Choose only reasonable values of *x* for which you try to make estimates about *y* — that is, look at the values of *x* for which your data was collected, and stay within those bounds when making predictions. In the textbook-weight example, the smallest average student weight is 48.5 pounds, and the largest average student weight is 142 pounds. Choosing student weights between 48.5 and 142 to plug in for *x* in the equation is okay, but choosing values less than 48.5 or more than 142 isn't a good idea. You can't guarantee that the same linear relationship (or any linear relationship for that matter) continues outside the given boundaries.

Think about it: If the relationship you found actually continued for any value of *x*, no matter how large, then a 250-pound lineman from OSU would have to carry $3.69 + 0.113 * 250 = 31.94$ pounds of books around in their backpack. Of course, this would be easy for them, but what about the rest of us?

Sometimes you need more than one variable

A simple linear regression model is just what it says it is: simple. I don't mean easy to work with, necessarily, but simple in the uncluttered sense. The model tries to estimate the value of *y* by only using one variable, *x*. However, the number of real-world situations that can be explained by using a simple, one-variable linear regression is small. Often one variable just can't do all the predicting.

If one variable alone doesn't result in a model that fits well enough, you can try to add more variables. It may take many variables to make a good estimate for *y*, and you have to be careful in how you choose them. In the case of stock market prices, for example, they're still looking for that ultimate prediction model.

As another example, health insurance companies try to estimate how long you'll live by asking you a series of questions (each of which represents a variable in the regression model). You can't find one single variable that estimates how long you'll live; you must consider many factors: your health, your weight, whether or not you smoke, genetic factors, how much exercise you do each week, and the list goes on and on and on.

The point is that regression models don't always use just one variable, *x*, to estimate *y*. Some models use two, three, or even more variables to estimate *y*. Those models aren't called simple linear regression models; they're called *multiple linear regression models* because of their employment of multiple variables to make an estimate. (You explore multiple linear regression models in Chapter 6.)

IN THIS CHAPTER

- Getting the basic ideas behind a multiple regression model
- Finding, interpreting, and testing coefficients
- Checking model fit

Chapter 6

Multiple Regression with Two X Variables

The idea of regression is to build a model that estimates or predicts one quantitative variable (*y*) by using at least one other quantitative variable (*x*). Simple linear regression uses exactly one *x* variable to estimate the *y* variable. (See Chapter 5 for all the information you need on simple linear regression.) *Multiple linear regression,* on the other hand, uses more than one *x* variable to estimate the value of *y*.

In this chapter, you see how multiple regression works and how to apply it to build a model for *y*. You see all the steps necessary for the process, including determining which *x* variables to include, estimating their contributions to the model, finding the best model, using the model for estimating *y*, and assessing the fit of the model. It may seem like a mountain of information, but you won't regress on the topic of regression if you take this chapter one step at a time.

Getting to Know the Multiple Regression Model

Before you jump right into using the multiple regression model, get a feel for what it's all about. In this section, you see the usefulness of multiple regression as well as the basic elements of the multiple regression model. Some of the ideas are just an extension of the simple linear regression model (see Chapter 5). Some of the concepts are a little more complex, as you might guess, because the model is more complex. But the concepts and the results should make intuitive sense, which is always good news.

Discovering the uses of multiple regression

One situation in which multiple regression is useful is when the *y* variable is hard to track down — that is, its value can't be measured straight up, and you need more than one other piece of information to help get a handle on what its value will be. For example, you may want to estimate the price of gold today. It would be hard to imagine being able to do that with only one other variable. You may base your estimate on recent gold prices, the price of other commodities on the market that move with or against gold, and a host of other possible economic conditions associated with the price of gold.

Another case for using multiple regression is when you want to figure out what factors play a role in determining the value of *y*. For example, you may want to find out what information is important to real estate agents in setting a price for a house going on the market.

Looking at the general form of the multiple regression model

The general idea of simple linear regression is to fit the best straight line through a body of data that you possibly can and use that line to make estimates for *y* based on certain *x*-values. The equation of the best-fitting line in simple linear regression is $y = b_0 + b_1x_1$, where b_0 is the *y*-intercept and b_1 is the slope. (The equation also has the form $y = a + bx$; see Chapter 5.)

In the multiple regression setting, you have more than one *x* variable that's related to *y*. Call these *x* variables $x_1, x_2, \ldots x_k$. In the most basic multiple regression model, you use some or all of these *x* variables to estimate *y*, where each *x* variable is taken to the first power. This process is called finding the best-fitting linear function for the data. This linear function looks like the following:

$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$, and you can call it the *multiple (linear) regression model.* You use this model to make estimates about *y* based on given values of the *x* variables.

TECHNICAL STUFF

A *linear* function is an equation whose *x* terms are taken to the first power only. For example, $y = 2x_1 + 3x_2 + 24x_3$ is a linear equation using three *x* variables. If any of the *x* terms are squared, the function becomes a *quadratic* one; if an *x* term is taken to the third power, the function becomes a *cubic* function, and so on. In this chapter, I consider only linear functions.

Stepping through the analysis

Your job in conducting a multiple regression analysis is to do the following (the computer can help you do Steps 3 through 6):

1. **Come up with a list of possible *x* variables that may be helpful in estimating *y*.**
2. **Collect data on the *y* variable and your *x* variables from Step 1.**
3. **Check the relationships between each *x* variable and *y* (using scatterplots and correlations), and use the results to eliminate those *x* variables that aren't strongly related to *y*.**
4. **Look at possible relationships between the *x* variables to make sure you aren't being redundant (in statistical terms, you're trying to avoid the problem of multicolinearity).**

   If two *x* variables relate to *y*in the same way, you don't need both in the model.
5. **Use those *x* variables (from Step 4) in a multiple regression analysis to find the best-fitting model for your data.**
6. **Use the best-fitting model (from Step 5) to predict *y* for given *x*-values by plugging those *x*-values into the model.**

I outline each of these steps in the following sections.

Looking at x's and y's

The first step of a multiple regression analysis comes way before the number crunching on the computer; it occurs even before the data is collected. Step 1 is where you sit down and think about what variables may be useful in predicting your response variable *y*. This step will likely take more time than any other step, except maybe the data-collection process. Deciding which *x* variables may be candidates for consideration in your model is a deal-breaking step, because you can't go back and collect more data after the analysis is over.

REMEMBER

Always check to be sure that your response variable, *y*, and at least one of the *x* variables are quantitative. For example, if *y* isn't quantitative but at least one *x* is, a logistic regression model may be in order (see Chapter 9).

Suppose you're in the marketing department for a major national company that sells plasma TVs. You want to sell as many TVs as you can, so you want to figure out which factors play a role in plasma TV sales. In talking with your advertising people and remembering what you learned in those business classes in college, you know that one powerful way to get sales is through advertising. You think of the types of advertising that may be related to sales of plasma TVs, and your team comes up with two ideas.

- **TV ads:** Of course, how better to sell a TV than through a TV ad?
- **Internet ads:** Hit 'em with an ad while they are on the social sites if, for example, they have been checking TV sets on Amazon.

By coming up with a list of possible *x* variables to predict *y*, you have just completed Step 1 of a multiple regression analysis, according to the list in the previous section. Note that all three variables that I use in the TV example are quantitative (the TV ad and Internet ad variables and the TV sales response variable), which means you can go ahead and think about a multiple regression model by using the two types of ads to predict TV sales.

Collecting the Data

Step 2 in the multiple regression analysis process is to collect the data for your *x* and *y* variables. To do this, make sure that for each individual in the data set, you collect all the data for that individual at the same time (including the *y*-value and all *x*-values) and keep the data all together for each individual, preserving any relationships that may exist between the variables. You must then enter the data into a table format by using Minitab or any other software package (with each column representing a variable and each row representing all the data from a single individual) to get a glimpse of the data and to organize it for later analyses.

To continue with the TV sales example from the preceding section, suppose that you start thinking about all the reams of data you have available to you regarding the plasma TV industry. You remember working with the advertising department before to do a media blitz by using, among other things, TV and local Internet ads. So you have data on these variables from a variety of store locations. You take a sample of 22 store locations in different parts of the country and put together the data on how much money was spent on each type of advertising, along with the plasma TV sales for that location. You can see the data in Table 6-1.

TABLE 6-1 Advertising Dollars and Sales of Plasma TVs

Location	Sales (In Millions of Dollars)	TV Ads (In Thousands of Dollars)	Local Internet Ads (In Thousands of Dollars)
1	9.73	0	20
2	11.19	0	20
3	8.75	5	5
4	6.25	5	5
5	9.10	10	10
6	9.71	10	10
7	9.31	15	15
8	11.77	15	15
9	8.82	20	5
10	9.82	20	5
11	16.28	25	25
12	15.77	25	25
13	10.44	30	0
14	9.14	30	0
15	13.29	35	5
16	13.30	35	5
17	14.05	40	10
18	14.36	40	10
19	15.21	45	15
20	17.41	45	15
21	18.66	50	20
22	17.17	50	20

In reviewing this data, the question is whether the amount of money spent on these two forms of advertising can do a good job of estimating sales (in other words, are the ads worth the money?). And if so, do you need to include spending for both types of ads to estimate sales, or is one of them enough? Looking at the numbers in Table 6-1, you can see that higher sales may be related at least to higher amounts spent on TV advertising; the situation with Internet advertising may not be so clear. So will the final multiple regression model contain both *x* variables or only one? In the following sections, you can find out.

Pinpointing Possible Relationships

The third step in doing a multiple regression analysis (see the list in the "Stepping through the analysis" section) is to find out which (if any) of your possible *x* variables are actually related to *y*. If an *x* variable has no relationship with *y*, including it in the model is pointless. Data analysts use a combination of scatterplots and correlations to examine relationships between pairs of variables (as you can see in Chapter 5). Although you can view these two techniques under the heading of looking for relationships, I walk you through each one separately in the following sections to discuss their nuances.

Making scatterplots

You make scatterplots in multiple linear regression to get a handle on whether your possible *x* variables are even related to the *y* variable you're studying. To investigate these possible relationships, you make one scatterplot of each *x* variable with the response variable *y*. If you have *k* different *x* variables being considered for the final model, you make *k* different scatterplots.

COMPUTER OUTPUT

To make a scatterplot in Minitab, enter your data in columns, where each column represents a variable and each row represents all the data from one individual. Go to Graph>Scatterplots>Simple. Select your *y* variable on the left-hand side, and click Select. That variable appears in the *y*-variable box on the right-hand side. Then select your *x* variable on the left-hand side, and click Select. That variable appears in the *x*-variable box on the right-hand side. Click OK.

Scatterplots of both TV ad spending versus plasma TV sales and Internet ad spending versus plasma TV sales are shown in Figure 6-1.

You can see from Figure 6-1a that TV spending does appear to have a fairly strong linear relationship with sales. This observation provides evidence that TV ad spending may be useful in estimating plasma TV sales. Figure 6-1b shows a linear relationship between Internet ad spending and sales, but the relationship isn't as strong as the one between TV ads and sales. However, it still may be somewhat helpful in estimating sales.

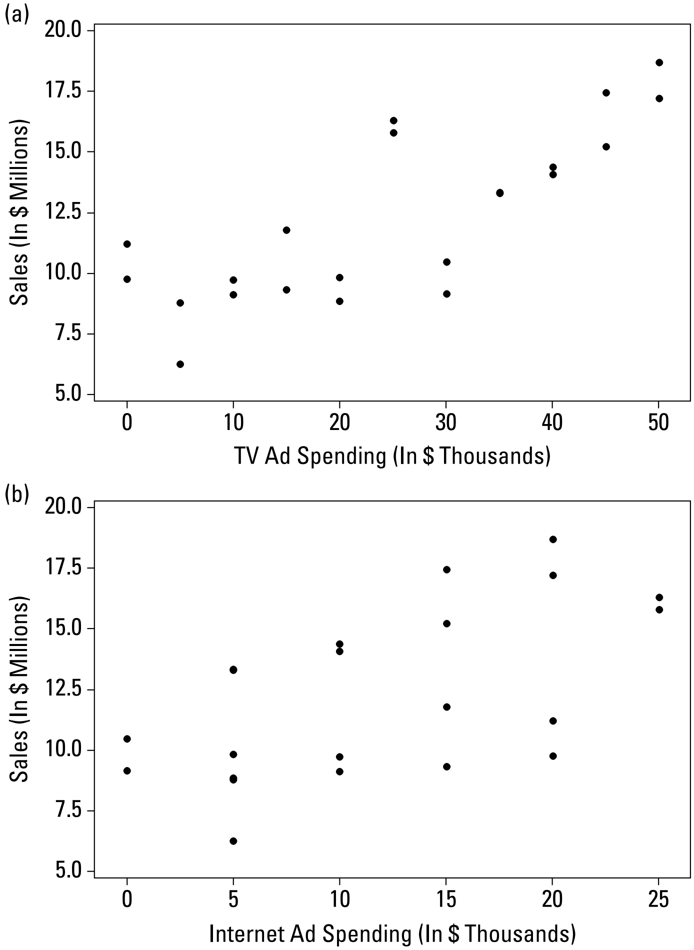


FIGURE 6-1: Scatterplots of a) TV ad spending and b) Internet ad spending versus plasma TV sales.

Correlations: Examining the bond

The second part of Step 3 involves calculating and examining the correlations between the *x* variables and the *y* variable. (Of course, if a scatterplot of an *x* variable and the *y* variable fails to come up with a pattern, then you drop that *x* variable altogether and don't proceed to find the correlation.)

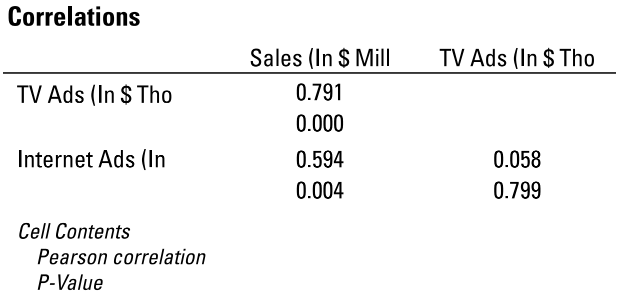
Whenever you employ scatterplots to explore possible linear relationships, correlations are typically not far behind. The *correlation coefficient* is a number that measures the strength and direction of the linear relationship between two variables, *x* and *y*. (See Chapter 5 for the lowdown on correlation.)

This step involves two parts:

- Finding and interpreting the correlations.
- Testing the correlations to see which ones are statistically significant (thereby determining which *x* variables are significantly related to *y*).

Finding and interpreting correlations

You can calculate a set of all possible correlations between all pairs of variables — which is called a *correlation matrix* — in Minitab. You can see the correlation matrix Minitab output for the TV data from Table 6-1 in Figure 6-2. Note the correlations between the *y* variable (sales) and each *x* variable, as well as the correlation between TV ads and Internet ads.



```
Correlations

                    Sales (In $ Mill   TV Ads (In $ Tho
TV Ads (In $ Tho         0.791
                         0.000

Internet Ads (In         0.594              0.058
                         0.004              0.799

Cell Contents
  Pearson correlation
  P-Value
```

FIGURE 6-2: Correlation values and *p*-values for the TV sales example.

COMPUTER OUTPUT

Minitab can find a correlation matrix between any pairs of variables in the model, including the *y* variable and all the *x* variables. To calculate a correlation matrix for a group of variables in Minitab, first enter your data in columns (one for each variable). Then go to Stat>Basic Statistics> Correlation. Highlight the variables from the left-hand side for which you want correlations, and click Select.

REMEMBER

To find the values of the correlation matrix from the computer output, intersect the row and column variables for which you want to find the correlation; the top number in that intersection is the correlation of those two variables. For example, the correlation between TV ads and TV sales is 0.791, because it intersects the TV row with the Sales column in the correlation matrix in Figure 6-2.

Testing correlations for significance

Using the rule-of-thumb approach from Stats I (also reviewed in Chapter 5), a correlation that's close to 1 or −1 (starting around ±0.75) is strong; a correlation close to 0 is very weak or nonexistent; and at around ±0.6 to 0.7, the relationships become moderately strong. The correlation between TV ads and TV sales of 0.791 indicates a fairly strong positive linear relationship between these two variables, based on the rule-of-thumb approach. The correlation between Internet ads and TV sales shown in Figure 6-2 is 0.594, which is moderate by my rule-of-thumb.

Many times in statistics, a rule-of-thumb approach to interpreting a correlation coefficient is sufficient. However, you're in the big leagues now, so a next step is conducting a null hypothesis to decide whether or not your model fits, based on the data you have. The model you are checking is the one that says the correlation is 0 (this is H_o). If you reject H_o, the correlation is statistically significant, based on the data; you have evidence that the *x* variable is helping to predict the *y* variable.

Now, that phrase, *statistically significant*, should ring a bell. It's your old friend the hypothesis test calling out to you (see Chapter 3 for a brush-up on hypothesis testing). Just like a hypothesis test for the mean of a population or the difference in the means of two populations, you also have a test for the correlation between two variables within a population.

REMEMBER

The null hypothesis to test a correlation is $H_o: \rho = 0$ (no relationship between the variables) versus $H_a: \rho \neq 0$ (a relationship exists between the variables). The letter ρ is the Greek version of *r* and represents the true correlation of *x* and *y* in the entire population; *r* is the correlation coefficient of the sample.

- **If you can't reject H_o based on your data**, you can't reject the model that the correlation is 0. That means you don't have evidence that the x variable is really helping the y variable out in your regression model, and it shouldn't be included. Note that a relationship still may exist in the population but you didn't see one in your data.
- **If you can reject H_o based on your data,** you conclude that the correlation isn't equal to zero, and there is evidence that the *x* variable is helping out the *y* variable in the regression model. You can say the relationship is statistically significant — that is, a relationship like this would occur very rarely in your sample just by chance if the truth was that there was no relationship. That's what a *p*-value really means.

WARNING

When determining whether to include your *x* variable in a model to predict *y*, *p*-values are just one of the issues to look at. Also think about things like how much it would cost to use the x variable, and whether that *x* variable makes sense when predicting *y*. For example, ice cream sales are related to murder rate (yes!), but that doesn't mean you should stop ice cream sales to prevent murders.

Any statistical software package can calculate a hypothesis test of a correlation for you. The actual formulas used in that process are beyond the scope of this book. However, the interpretation is the same as for any test: If the *p*-value is smaller than your predetermined value of α (typically 0.05), reject H_0 and conclude you have evidence that *x* and *y* are related. Otherwise, you can't reject H_0, and you conclude that you don't have enough evidence to indicate that the variables are related.

COMPUTER OUTPUT

In Minitab, you can conduct a hypothesis test for a correlation by clicking on Stat>Basic Statistics>Correlation, and checking the Display *p*-values box. Choose the variables you want to find correlations for, and click Select. You'll get output in the form of a little table that shows the correlations between the variables for each pair, with the respective *p*-values under each one. You can see the correlation output for the ads and sales example in Figure 6-2.

Looking at Figure 6-2, the correlation of 0.791 between TV ads and sales has a *p*-value of 0.000, which means it's actually less than 0.0005. That's a highly significant result, much less than 0.05 (your predetermined α level). So TV ad spending is strongly related to sales. The correlation between Internet ad spending and sales was 0.594, which is also found to be statistically significant with a *p*-value of 0.004.

Checking for Multicolinearity

You have one more very important step to complete in the relationship-exploration process before going on to using the multiple regression model. You need to complete Step 4: looking at the relationship between the *x* variables themselves and checking for redundancy. Failure to do so can lead to problems during the model-fitting process.

WARNING

Multicolinearity is a term you use if two *x* variables are highly correlated. Not only is it redundant to include both related variables in the multiple regression model, but it's also problematic. The bottom line is this: If two *x* variables are significantly correlated, only include one of them in the regression model, not both. If you include both, the computer won't know what numbers to give as coefficients for each of the two variables because they share their contribution to determining the value of y. Multicolinearity can really mess up the model-fitting process and give answers that are inconsistent and often not repeatable in subsequent studies.

To head off the problem of multicolinearity, along with the correlations you examine regarding each x variable and the response variable y, you also need to find the correlations between all pairs of x variables. If two x variables are highly correlated, don't leave them both in the model, or multicolinearity will result. To see the correlations between all the x variables, have Minitab calculate a correlation matrix of all the variables (see the section, "Finding and interpreting correlations"). You can ignore the correlations between the y variable and the x variables and only choose the correlations between the x variables shown in the correlation matrix. Find those correlations by intersecting the rows and columns of the x variables for which you want correlations.

REMEMBER

If two x variables x_1 and x_2 are strongly correlated (that is, their correlation is beyond +0.7 or −0.7), then one of them would do just about as good a job of estimating y as the other, so you don't need to include them both in the model. If x_1 and x_2 aren't strongly correlated, then both of them working together would do a better job of estimating sales than either variable alone.

For the ad-spending example, you have to examine the correlation between the two x variables — TV ad spending and Internet ad spending — to be sure no multicolinearity is present. The correlation between these two variables (as you can see in Figure 6-2) is only 0.058. You don't even need a hypothesis test to tell you whether or not these two variables are related; they're clearly not.

The p-value for the correlation between the spending for the two ad types is 0.799 (see Figure 6-2), which is much, much larger than 0.05 ever thought of being and therefore isn't statistically significant. The large p-value for the correlation between spending for the two ad types confirms your thoughts that both variables together may be helpful in estimating y because each makes its own contribution. It also tells you that keeping them both in the model won't create any multicolinearity problems. (This completes Step 4 of the multiple regression analysis, as listed in the "Stepping through the analysis" section.)

Finding the Best-Fitting Model for Two x Variables

After you have a group of x variables that are all related to y and not related to each other (refer to previous sections), you're ready to perform Step 5 of the multiple regression analysis (as listed in the "Stepping through the analysis" section). You're ready to find the best-fitting model for the data.

In the multiple regression model with two x variables, you have the general equation $y = b_0 + b_1x_1 + b_2x_2$, and you already know which x variables to include in the model (by doing Step 4 in the previous section); the task now is to figure out which coefficients (numbers) to put in for b_0, b_1, and b_2, so you can use the resulting equation to estimate y. This specific model is the *best-fitting multiple linear regression model.* This section tells you how to get, interpret, and test those coefficients in order to complete Step 5 in the multiple regression analysis.

TECHNICAL STUFF

Finding the best-fitting linear equation is like finding the best-fitting line in simple linear regression, except that you're not finding a line. When you have two x variables in multiple regression, for example, you're estimating a best-fitting plane for the three-dimensional data.

Getting the multiple regression coefficients

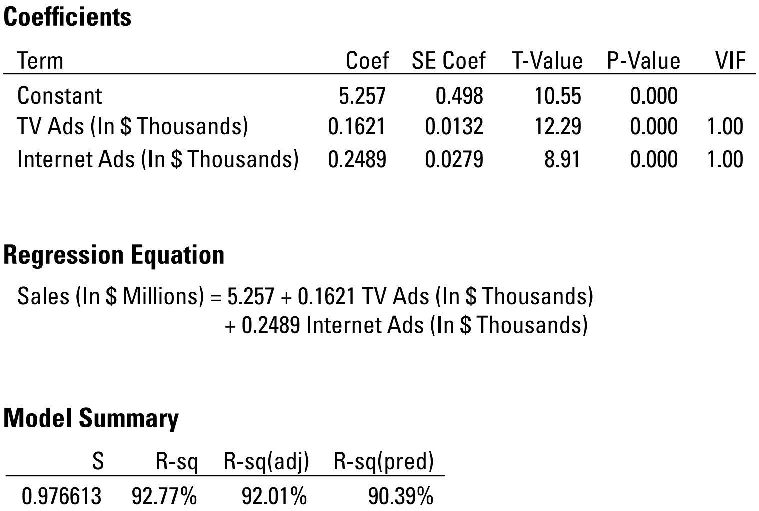
In the simple linear regression model, you have the straight line $y = b_0 + b_1x$; the coefficient of x is the slope, and it represents the change in y per unit change in x. In a multiple linear regression model, the coefficients b_1, b_2, and so on quantify in a similar matter the sole contribution that each corresponding x variable (x_1, x_2) makes in predicting y. The coefficient b_0 indicates the amount by which to adjust all these values in order to provide a final fit to the data (like the y-intercept does in simple linear regression).

Computer software does all the nitty-gritty work for you to find the proper coefficients (b_0, b_1, and so on) that fit the data best. The coefficients that Minitab settles on to create the best-fitting model are the ones that, as a group, minimize the sum of the squared residuals (sort of like the variance in the data around the selected model). The equations for finding these coefficients by hand are too unwieldy to include in this book; a computer can do all the work for you. The results appear in the regression output in Minitab. You can find the multiple regression coefficients $(b_0, b_1, b_2, \ldots, b_k)$ on the computer output under the column labeled COEF.

COMPUTER OUTPUT

To run a multiple regression analysis in Minitab, click on Stat>Regression>Regression>Fit Regression Model. Then choose the response variable (y) and click on Select. Then choose your predictor variables (x variables) one by one, and click Select. Click on OK, and the computer carries out the analysis.

For the plasma TV sales example from the previous sections, Figure 6-3 shows the multiple regression Minitab output for the ads and plasma TV sales. You can see the coefficients in the COEF column for the multiple regression model. The first coefficient (5.257) is just the constant term (or b_0 term) in the model and isn't affiliated with any *x* variable. This constant just sort of goes along for the ride in the analysis; it's the number that you tack on the end to make the numbers work out right. The second coefficient in the COEF column is 0.162; this value is the coefficient of the x_1 (TV ad amount) term, also known as b_1. The third coefficient in the COEF column is 0.249, which is the value for b_2 in the multiple regression model and is the coefficient that goes with x_2 (Internet ad amount).



Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5.257	0.498	10.55	0.000	
TV Ads (In $ Thousands)	0.1621	0.0132	12.29	0.000	1.00
Internet Ads (In $ Thousands)	0.2489	0.0279	8.91	0.000	1.00

Regression Equation

Sales (In $ Millions) = 5.257 + 0.1621 TV Ads (In $ Thousands)
+ 0.2489 Internet Ads (In $ Thousands)

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.976613	92.77%	92.01%	90.39%

FIGURE 6-3: (Selected) Regression output for the ads and plasma TV sales example.

Putting these coefficients into the multiple regression equation, you see the regression equation is $\text{Sales} = 5.267 + 0.162\ (\text{TV ads}) + 0.249$ (Internet ads), where sales are in millions of dollars and ad spending is in thousands of dollars.

So you have your coefficients (no sweat, right?), but where do you go from here? What does it all mean? The next section guides you through interpretation.

Interpreting the coefficients

In simple linear regression (covered in Chapter 5), the coefficients represent the slope and *y*-intercept of the best-fitting line and are straightforward to interpret. The slope, in particular, represents the change in *y* due to a 1-unit increase in *x* because you can write any slope as a number over 1 (and *slope* is rise over run).

In the multiple regression model, the interpretation's a little more complicated. Due to all the mathematical underpinnings of the model and how it's finalized (believe me, you don't want to go there unless you're looking for a PhD in statistics), the coefficients have a different meaning.

REMEMBER

The coefficient of an x variable in a multiple regression model is the amount by which y changes if that x variable increases by one unit and the values of all other x variables in the model *don't change.* So basically, you're looking at the marginal contribution of each x variable when you hold the other variables in the model constant.

In the ads and sales regression analysis (see Figure 6-3), the coefficient of x_1 (TV ad spending) equals 0.1621. Sales (plasma TV sales) increases by 0.1621 million dollars when TV ad spending increases by 1.0 thousand dollars and spending on Internet ads doesn't change. (Note that keeping more digits after the decimal point reduces rounding error when in units of millions.)

TIP

You can more easily interpret the number "0.16211 million dollars" by converting it to a dollar amount without the decimal point: $0.16211 million is equal to $162,110. (To get this value, I just multiplied $0.16211 by 1,000,000.) So plasma TV sales increased by $162,110 for each $1,000 increase in TV ad spending, while Internet ad spending remained the same. Similarly, the coefficient of x_2 (Internet ad spending) was equal to 0.24887. So plasma TV sales increased by 0.24887 million dollars (or $248,870), when Internet ad spending increased by $1,000 and TV ad spending remained the same.

WARNING

Don't forget the units of each variable in a multiple regression analysis. This mistake is one of the most common in Stats II. If you were to forget about units in the ads and sales example, you would think that sales increased by 0.24887 dollars with $1 in Internet ad spending!

Knowing the multiple regression coefficients (b_1 and b_2, in this case) and their interpretation, you can now answer the original question: Is the money spent on TV or Internet ads worth it? The answer is a resounding *yes!* Not only that, but you also can say how much you expect sales to increase per $1,000 you spend on TV or Internet advertising. Note that this conclusion assumes the model fits the data well. You have some evidence of that through the scatterplots and correlation tests, but more checking needs to be done before you can run to your manager and tell them the good news. The next section tells you what to do next.

Testing the coefficients

To officially determine whether you have the right x variables in your multiple regression model, you do a formal hypothesis test. You test whether there is evidence in the data of a regression coefficient being zero. Note that if the coefficient of an *x* variable is zero, when you put that coefficient into the model, you get zero times that *x* variable, which equals zero. This result is essentially saying that if an *x* variable's coefficient has evidence in the data of being equal to zero, you don't need that *x* variable in the model.

REMEMBER

With any regression analysis, the computer automatically performs all the necessary hypothesis tests for the regression coefficients. Along with the regression coefficients you can find on the computer output, you see the test statistics and *p*-values for a test of each of those coefficients in the same row for each coefficient. Each one is testing H_0: Coefficient $= 0$ versus H_a: Coefficient $\neq 0$.

REMEMBER

The general format for finding a test statistic in most any situation is to take the statistic (in this case, the coefficient), subtract the value in H_0 (zero), and divide by the standard error of that statistic (for this example, the standard error of the coefficient). (For more information on the general format of hypothesis tests, see Chapter 4.)

To test a regression coefficient, the test statistic (using the labels from Figure 6-3) is $\left(\text{Coef} - 0\right) / \text{SE Coef}$. In noncomputer language, that means you take the coefficient, subtract zero, and divide by the standard error (SE) of the coefficient. The standard error of a coefficient here is a measure of how much the coefficient is expected to vary when you take a new sample. (Refer to Chapter 4 for more on standard error.)

The test statistic has a *t*-distribution with $n - k - 1$ degrees of freedom, where n equals the sample size and k is the number of predictors (*x* variables) in the model. This number of degrees of freedom works for any coefficient in the model (except you don't bother with a test for the constant, because it has no *x* variable associated with it).

The test statistic for testing each coefficient is listed in the column marked T (because it has a *t*-distribution) on the Minitab output. You compare the value of the test statistic to the *t*-distribution with $n - k - 1$ degrees of freedom (using Table A-1 in the Appendix) and come up with your *p*-value. If the *p*-value is less than your predetermined α (usually 0.05), then you reject $H_0$and conclude that the coefficient of that *x* variable isn't zero and that variable makes a significant contribution toward estimating *y* (given the other variables are also included in the model). If the *p*-value is larger than 0.05, you can't reject H_0, so that *x* variable makes no significant contribution toward estimating *y* (when the other variables are included in the model).

In the case of the ads and plasma TV sales example, Figure 6-3 shows that the coefficient for the TV ads is 0.1621 (the second number in column two). The standard error is listed as being 0.0132 (the second number in column three). To find the test statistic for TV ads, take 0.1621 minus zero and divide by the standard error, 0.0132. You get a value of $t = 12.29$, which is the second number in column four. Comparing this value of t to a t-distribution with $n - k - 1 = 22 - 2 - 1 = 19$ degrees of freedom (Table A-1 in the Appendix), you see the value of t is way off the scale. That means the p-value is smaller than can be measured on the t-table. Minitab lists the p-value in column five of Figure 6-3 as 0.000 (meaning it's less than 0.001). This result leads you to conclude that the coefficient for TV ads is statistically significant, and TV ads should be included in the model for predicting TV sales.

The Internet ads coefficient is also significant with a p-value of 0.000 by the same reasoning; you find these results by looking across the Internet ads row of Figure 6-3. Based on your coefficient tests and the lack of multicolinearity between TV and Internet ads (see the earlier section, "Checking for Multicolinearity"), you should include both the TV ads variable and the Internet ads variable in the model for estimating TV sales.

Predicting y by Using the x Variables

When you have your multiple regression model, you're finally ready to complete Step 6 of the multiple regression analysis: to predict the value of y given a set of values for the x variables. To make this prediction, you take those x values for which you want to predict y, plug them into the multiple regression model, and simplify.

In the ads and plasma TV sales example (see the analysis in Figure 6-3), the best-fitting model is $y = 5.26 + 0.162x_1 + 0.249x_2$. In the context of the problem, the model is $\text{Sales} = 5.26 + 0.162$ TV ad spending $(x_1) + 0.249$ Internet ad spending (x_2).

WARNING

Remember that the units for plasma TV sales is in millions of dollars and the units for ad spending for both TV and Internet ads is in thousands of dollars. That is, \$20,000 spent on TV ads means $x_1 = 20$ in the model. Similarly, \$10,000 spent on Internet ads means $x_2 = 10$ in the model. Forgetting the units involved can lead to serious miscalculations.

Suppose you want to estimate plasma TV sales if you spend \$20,000 on TV ads and \$10,000 on Internet ads. Plug $x_1 = 20$ and $x_2 = 10$ into the multiple regression model, and you get $y = 5.26 + 0.162(20) + 0.249(10) = 10.99$. In other words, if you spend \$20,000 on TV advertising and \$10,000 on Internet advertising, you estimate that sales will be \$10.99 million.

This estimate at least makes sense in terms of the data from the 22 store locations shown in Table 6-1. Location 10 spent $20,000 on TV ads and $5,000 on Internet ads (short of what you had) and got sales of $9.82 million. Location 11 spent a little more on TV ads and a lot more on Internet ads than what you had and got sales of $16.28 million. Your estimates of sales for Store Locations 10 and 11 are $5.26+0.162*20+0.249*5=\$9.745$ million, and $5.26+0.162*25+0.249*25=\$15.535$ million, respectively. These estimates turned out to be pretty close to the actual sales at those two locations ($9.82 million and $16.28 million, respectively, as shown in Table 6-1), giving at least some confidence that your estimates will be close for the other store locations not chosen for the study.

WARNING

Be careful to put in only values for the *x* variables that fall in the range of where the data lies. In other words, Table 6-1 shows data for TV ad spending between $0 and $50,000; Internet ad spending goes from $0 to $25,000. It wouldn't be appropriate to try to estimate sales for spending amounts of $75,000 for TV ads and $50,000 for Internet ads, respectively, because the regression model you came up with only fits the data that you collected. You have no way of knowing whether that same relationship continues outside that area. This no-no of estimating *y* for values of the *x* variables outside their range is called *extrapolation.* As one of my colleagues says, "Friends don't let friends extrapolate."

Checking the Fit of the Multiple Regression Model

Before you run to your boss in triumph saying you've slam-dunked the question of how to estimate plasma TV sales, you first have to make sure all your i's are dotted and all your t's are crossed, as you do with any other statistical procedure. In this case, you have to check the conditions of the multiple regression model. These conditions mainly focus on the *residuals* (the difference between the estimated values for *y* and the observed values of *y* from your data). If the model is close to the actual data you collected, you can feel somewhat confident that if you were to collect more data, it would fall in line with the model as well, and your predictions should be good.

In this section, you see what the conditions are for multiple regression, as well as specific techniques that statisticians use to check each of those conditions. The main character in all this condition-checking is the residual.

Noting the conditions

The conditions for multiple regression concentrate on the error terms, or residuals. The residuals are the amount that's left over after the model has been fit. They represent the difference between the actual value of *y* observed in the data set and the estimated value of *y* based on the model. Following are the conditions for the residuals of the multiple regression model; note that all conditions need to be met in order to give the go-ahead for a multiple regression model:

- They have a normal distribution with a mean of zero.
- They have the same variance for each fitted (predicted) value of y.
- They're independent (meaning they don't affect each other).

Plotting a plan to check the conditions

It may sound like you have a ton of things to check here and there, but luckily, Minitab gives you all the information you need to know in a series of four graphs, all presented at one time. These plots are called the *residual plots,* and they graph the residuals so that you can check to see whether the conditions from the previous section are met.

You can get the set of residual plots in two flavors.

- **Regular residuals:** The regular residual plots (the vanilla-flavored ones) show you exactly what the residuals are for each value of *y*. Their units depend on the variables in the model; use them *only* if you want to mainly look for patterns in the data. Figure 6-4 shows the plots of the regular residuals for the TV sales example. These residuals are in units of millions of dollars.
- **Standardized residuals:** The standardized residual plots (the strawberry-flavored kind) take each residual and convert it to a *Z*-score by subtracting the mean and dividing by the standard deviation of all the residuals. Figure 6-5 shows the plots of the standardized residuals for the TV sales example. Use these plots if you want to not only look for patterns in the data but also assess the standardized values of the residuals in terms of values on a *Z*-distribution to check for outliers. (Most statisticians use standardized residual plots.)

  Note that the plots in Figure 6-5 look almost exactly the same as those in Figure 6-4. It's not surprising that the shapes of all graphs are the same for both types of residuals. Note, however, that the values of the regular residuals in Figure 6-4 are in millions of dollars and the standardized residuals in Figure 6-5 are from the standard normal distribution, which has no units.

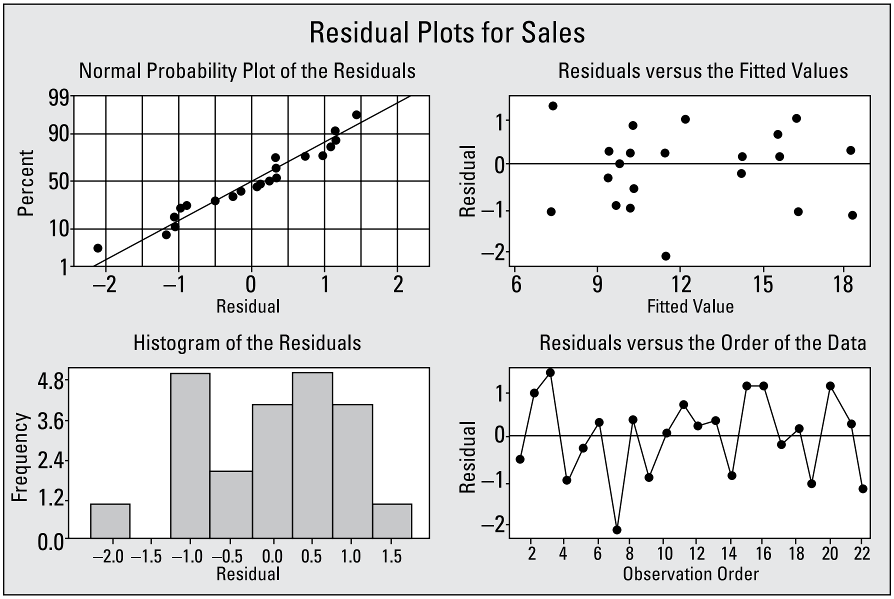


FIGURE 6-4: Residual plots for the ads and plasma TV sales example.

COMPUTER OUTPUT

To make residual plots in Minitab, go to Stat>Regression>Regression>Fit Regression Model. Select your response (y) variable and your predictor (x) variables. Click on Graphs, and choose either Regular or Standardized for the residuals, depending on which one you want. Then click on Four-in-one, which indicates you want to get all four residual plots shown in Figure 6-4 (using regular residuals) and Figure 6-5 (using standardized residuals).

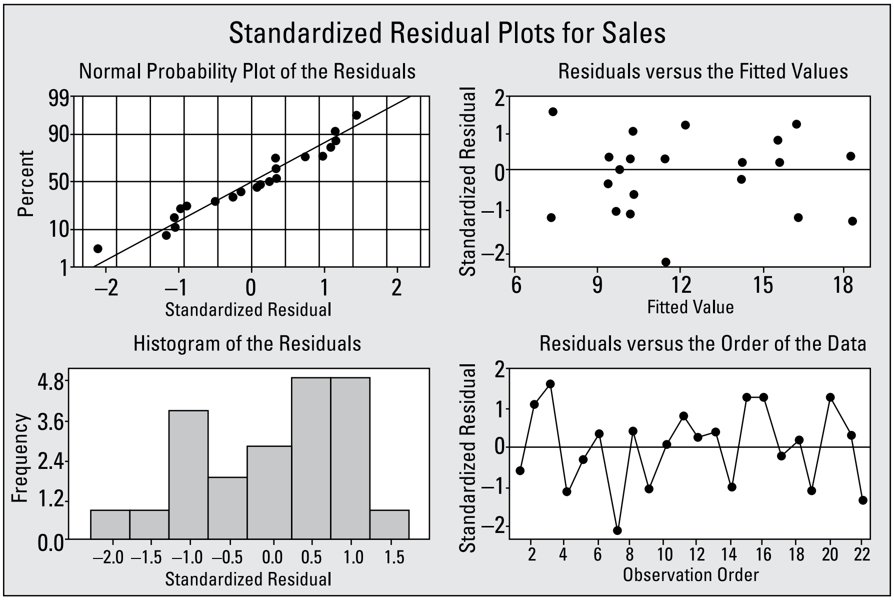


FIGURE 6-5: Standardized residual plots for the ads and plasma TV example.

Checking the three conditions

The following sections show you how to check the residuals to see whether your data set meets the three conditions of the multiple regression model.

Meeting the first condition: Normal distribution with mean zero

The first condition to meet is that the residuals must have a normal distribution with mean zero. The upper-left plot of Figure 6-4 shows how well the residuals match a normal distribution. Residuals falling in a straight line means the normality condition is met. By the looks of this plot, I'd say that condition is met for the ad and sales example.

The upper-right plot of Figure 6-4 shows what the residuals look like for the various estimated *y* values. Look at the horizontal line going across that plot: It's at zero as a marker. The residuals should average out to be at that line (zero). This Residuals versus Fitted Values plot checks the mean-of-zero condition and holds for the ads and sales example looking at Figure 6-4.

TIP

As an alternative check for normality apart from using the regular residuals, you can look at the standardized residuals plot (see Figure 6-5) and check out the upper-right plot. It shows how the residuals are distributed across the various estimated (fitted) values of *y*. Standardized residuals are supposed to follow a standard normal distribution — that is, they should have mean of zero and standard deviation of one. So when you look at the standardized residuals, they should be centered around zero in a way that has no predictable pattern, with the same amount of variability around the horizontal line that crosses at zero as you move from left to right.

In looking at the upper-right plot of Figure 6-5, you should also find that most (95 percent) of the standardized residuals fall within two standard deviations of the mean, which in this case is −2 to +2 (via the 68-95-99.7% Rule — remember that from Stats I?). You should see more residuals hovering around zero (where the middle lump would be on a standard normal distribution), and you should have fewer and fewer of the residuals as you go away from zero. The upper-right plot in Figure 6-5 confirms a normal distribution for the ads and sales example on all the counts mentioned here.

The lower-left plots of Figures 6-4 and 6-5 show histograms of the regular and standardized residuals, respectively. These histograms should reflect a normal distribution; the shape of the histograms should be approximately symmetric and look like a bell-shaped curve. If the data set is small (as is the case here with only 22 observations), the histogram may not be as close to normal as you would like;

in that case, consider it part of the body of evidence that all four residual plots show you. The histograms shown in the lower-left plots of Figure 6-4 and 6-5 aren't terribly normal looking; however, because you can't see any glaring problems with the upper-right plots, don't be worried.

Satisfying the second condition: Variance

The second condition in checking the multiple regression model is that the residuals have the same variance for each fitted (predicted) value of *y*. Look again at the upper-right plot of Figure 6-4 (or Figure 6-5). You shouldn't see any change in the amount of spread (variability) in the residuals around that horizontal line as you move from left to right. Looking at the upper-right graph of Figure 6-4, there's no reason to say condition number two hasn't been met.

WARNING

One particular problem that raises a red flag with the second condition is if the residuals fan out, or increase in spread, as you move from left to right on the upper-right plot. This fanning out means that the variability increases more and more for higher and higher predicted values of *y*, so the condition of equal variability around the fitted line isn't met, and the regression model wouldn't fit well in that case.

Checking the third condition: Independence

The third condition is that the residuals are independent; in other words, they don't affect each other. Looking at the lower-right plot on either Figure 6-4 or 6-5, you can see the residuals plotted by *observation number*, which is the order in which the data came in the sample. If you see a pattern, you have trouble; for example, if you were to connect the dots, so to speak, you might see a pattern of a straight line, a curve, or any kind of predictable up or down trend. You can see no patterns in the lower-right plots, so the independence condition is met for the ads and plasma TV sales example.

WARNING

If the data must be collected over time, such as stock prices over a ten-year period, the independence condition may be a big problem because the data from the previous time period may be related to the data from the next time period. This kind of data requires time series analysis and is beyond the scope of this book.

IN THIS CHAPTER

- Evaluating different methods for choosing a multiple regression model
- Understanding how forward selection and backward selection work
- Using the best subsets methods to find a good model

Chapter 7

How Can I Miss You If You Won't Leave? Regression Model Selection

Suppose you're trying to estimate some quantitative variable, *y*, and you have many *x* variables available at your disposal. You have so many variables related to *y*, in fact, that you feel like I do in my job every day — overwhelmed with opportunity. Where do you go? What do you do? Never fear, this chapter is for you.

In this chapter, you uncover criteria for determining when a model fits well. I discuss different model selection procedures and all the details of the most statistician-approved method for selecting the best model. Plus, you get to find out what factors come into play when a punter kicks a football. (You can think about that while you're reading.)

REMEMBER

Note that the term *best* has many connotations here. You can't find one end-all-be-all model that everyone comes up with in the end. That's to say that each data analyst can come up with a different model, and each model can still do a good job of predicting *y*.

Getting a Kick out of Estimating Punt Distance

Before you jump into a model selection procedure to predict *y* by using a set of *x* variables, you have to do some legwork. The variable of interest is *y*, and that's a given. But where do the *x* variables come from? How do you choose which ones to investigate as being possible candidates for predicting *y*? And how do those possible *x* variables interact with each other toward making that prediction?

You must answer all these questions before using any model selection procedure. However, this part is the most challenging and the most fun; a computer can't think up *x* variables for you!

Suppose you're at a football game and the opposing team has to punt the ball. You see the punter line up and get ready to kick the ball, and some questions come to you: "Gee, I wonder how far this punt will go? I wonder what factors influence the distance of a punt? Can I use those factors in a multiple regression model to try to estimate punt distance? Hmm, I think I'll consult my *Statistics II For Dummies*, 2nd Edition book on this and analyze some data during halftime. . . ."

Well, maybe that's pushing it, but it's still an interesting line of questioning for football players, golfers, soccer players, and even baseball players. Everyone's looking for more distance and a way to get it.

In the following sections, you can see how to identify and assess different *x* variables in terms of their potential contribution to predicting *y*.

Brainstorming variables and collecting data

Starting with a blank slate and trying to think of a set of *x* variables that may be related to *y* may sound like a daunting task, but in reality, it's probably not as bad as you think. Most researchers who are interested in predicting some variable *y* in the first place have some ideas about which variables may be related to it. After you come up with a set of logical possibilities for *x*, you collect data on those variables, as well as on *y*, to see what their actual relationship with *y* may be.

The Virginia Polytechnic Institute did a study to try to estimate the distance of a punt in football (something Ohio State fans aren't familiar with). Possible variables they thought may be related to the distance of a punt included the following:

- Hang time (time in the air, in seconds)
- Right leg strength (measured in pounds of force)
- Left leg strength (in pounds of force)
- Right leg flexibility (in degrees)
- Left leg flexibility (in degrees)
- Overall leg strength (in pounds)

The data collected on a sample of 13 punts (by right-footed punters) is shown in Table 7-1.

TABLE 7-1 Data Collected for Punt Distance Study

Distance (In Feet)	Hang Time	Right Leg Strength	Left Leg Strength	Right Leg Flexibility	Left Leg Flexibility	Overall Leg Strength
162.50	4.75	170	170	106	106	240.57
144.00	4.07	140	130	92	93	195.49
147.50	4.04	180	170	93	78	152.99
163.50	4.18	160	160	103	93	197.09
192.00	4.35	170	150	104	93	266.56
171.75	4.16	150	150	101	87	260.56
162.00	4.43	170	180	108	106	219.25
104.93	3.20	110	110	86	92	132.68
105.67	3.02	120	110	90	86	130.24
117.59	3.64	130	120	85	80	205.88
140.25	3.68	120	140	89	83	153.92
150.17	3.60	140	130	92	94	154.64
165.17	3.85	160	150	95	95	240.57

Other variables you may think of that are related to punt distance may include the direction and speed of the wind at the time of the punt, the angle at which the ball was snapped, the average distance of punts made in the past by a particular punter, whether the game is at home or away in a hostile environment, and so on. However, these researchers seem to have enough information on their hands to build a model to estimate punt distance.

For the sake of simplicity, you can assume the kicker is right-footed, which isn't always the case, but it represents the overwhelming majority of kickers.

Looking just at this raw data set in Table 7-1, you can't figure out which variables, if any, are related to distance of the punt or how those variables may be related to punt distance. You need more analyses to get a handle on this.

Examining scatterplots and correlations

REMEMBER

After you've identified a set of possible *x* variables, the next step is to find out which of these variables are highly related to *y* in order to start trimming down the set of possible candidates for the final model. In the punt distance example, the goal is to see which of the six variables in Table 7-1 are strongly related to punt distance. There are two ways to look at these relationships.

- **Scatterplot:** A graphical technique
- **Correlation:** A one-number measure of the linear relationship between two variables

Seeing relationships through scatterplots

To begin examining the relationships between the *x* variables and *y*, you use a series of scatterplots. Figure 7-1 shows all the scatterplots — not only of each *x* variable with *y* but also of each *x* variable with the other *x* variables. The scatterplots are in the form of a *matrix*, which is a table made of rows and columns. For example, the first scatterplot in row two of Figure 7-1 looks at the variables of distance (which appears in column one) and hang time (which appears in row two). This scatterplot shows a possible positive (uphill) linear relationship between distance and hang time.

Note that Figure 7-1 is essentially a symmetric matrix across the diagonal line. The scatterplot for distance and hang time is the same as the scatterplot for hang time and distance; the *x* and *y* axes are just switched. The essential relationship shows up either way. So you only have to look at all the scatterplots below the diagonal (where the variable names appear) or above the diagonal. You don't need to examine both.

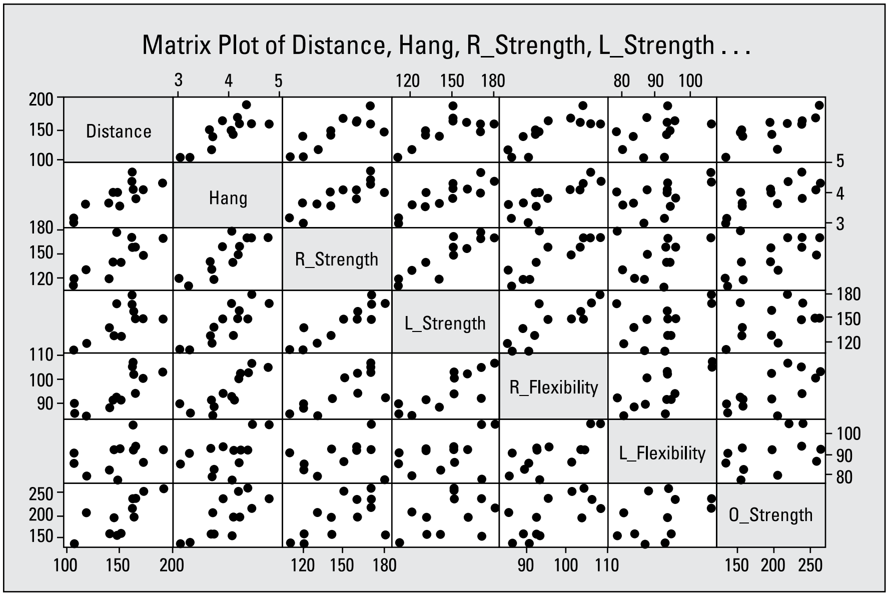


FIGURE 7-1: A matrix of all scatterplots between pairs of variables in the punting distance example.

COMPUTER OUTPUT

To get a matrix of all scatterplots between a set of variables in Minitab, go to Graph>Matrix Plot and choose Matrix of Plots>Simple. Highlight all the variables in the left-hand box for which you want scatterplots by clicking on them; click Select, and then click OK. You'll see the matrix of scatterplots with a format similar to Figure 7-1.

Looking across row one of Figure 7-1, you can see that all the variables seem to have a positive linear relationship with punt distance except left leg flexibility. Perhaps the reason left leg flexibility isn't much related to punt distance is because the left foot is planted into the ground when the kick is made — for a right-footed kicker, the left leg doesn't have to be nearly as flexible as the right leg, which does the kicking. So it doesn't appear that left leg flexibility contributes a great deal to the estimation of punt distance on its own.

You can also see in Figure 7-1 that the scatterplots showing relationships between pairs of *x* variables are to the right of column one and below row one. (Remember, you need to look on only the bottom part of the matrix or the top part of the matrix to see the relevant scatterplots.) It appears that hang time is somewhat related to each of the other variables (except left leg flexibility, which doesn't contribute to estimating *y*). So hang time could possibly be the most important single variable in estimating the distance of a punt.

Looking for connections by using correlations

Scatterplots can give you some general ideas as to whether two variables are related in a linear way. However, pinpointing that relationship requires a numerical value to tell you how strongly the variables are related (in a linear fashion) as well as the direction of that relationship. That numerical value is the *correlation* (also known as *Pearson's correlation*; see Chapter 5). So the next step toward trimming down the possible candidates for *x* variables is to calculate the correlation between each *x* variable and *y*.

COMPUTER OUTPUT

To get a set of all the correlations between any set of variables in your model by using Minitab, go to Stat>Basic Statistics>Correlation. Then highlight all the variables you want correlations for, and click Select. (To include the *p*-values for each correlation, click the Display *p*-values box.) Then click OK. You'll see a listing of all the variables' names across the top row and down the first column. Intersect the row depicting the first variable with the column depicting the second variable in order to find the correlation for that pair.

Table 7-2 shows the correlations you can calculate between $y = punt$ distance and each of the *x* variables. These results confirm what the scatterplots were telling you. Distance seems to be related to all the variables except left leg flexibility because that's the only variable that didn't have a statistically significant correlation with distance using the α level 0.05. (For more on the test for correlation, see Chapter 6.)

TABLE 7-2 Correlations between Distance of a Punt and Other Variables

x Variable	Correlation with Punt Distance	p-value
Hang time	0.819	0.001*
Right leg strength	0.791	0.001*
Left leg strength	0.744	0.004*
Right leg flexibility	0.806	0.001*
Left leg flexibility	0.408	0.167
Overall leg strength	0.796	0.001*

* *Statistically significant at level* $\alpha = 0.05$

If you take a look at Figure 7-1, you can see that hang time is related to other *x* variables such as right foot and left foot strength, right leg flexibility, and so on. This is where things start to get sticky. You have hang time related to distance, and lots of other variables related to hang time. Although hang time is clearly the most related to distance, the final multiple regression model may not include hang time.

Here's one possible scenario: You find a combination of other *x* variables that can do a good job estimating *y* together. And all those other variables are strongly related to hang time. This result may mean that in the end, you don't need to include hang time in the model. Strange things happen when you have many different *x* variables to choose from.

After you narrow down the set of possible *x* variables for inclusion in the model to predict punt distance, the next step is to put those variables through a selection procedure to trim down the list to a set of essential variables for predicting *y*.

WARNING

A note on multiple hypothesis tests. In this case, you are looking at six *p*-values from six different hypothesis tests. If a big group of hypothesis tests were conducted at the 0.05 level (where $\alpha = .05$), you would expect about 5% of the tests to be rejected (come up significant) just by chance. (That's what the α-level means; it's also known as a Type I error — see Chapter 4.) The more tests you do, the more you need to worry about making this error. Oftentimes people use an adjusted α-level of α/n, where n is the number of tests (called the *Bonferroni adjustment*), and this is their cutoff for each test. This reduces the overall chance of making a Type I error. In this example, $0.05/6 = .008$ and all the tests are still significant as before.

Just Like Buying Shoes: The Model Looks Nice, But Does It Fit?

When you get into model selection procedures, you find that many different methods exist for selecting the best model, according to a wide range of criteria. Each one can result in models that differ from each other, but that's something I love about statistics: Sometimes there's no one single best answer.

The three model selection procedures covered in this section are

- Best subsets procedure
- Forward selection
- Backward selection

Of all the model selection procedures out there, the one that gets the most votes with statisticians is the *best subsets procedure,* which examines every single possible model and determines which one fits best, using certain criteria.

In this section, you see different methods statisticians use to assess and compare the fit of different models. I show you how the best subsets procedure works for model selection in a step-by-step manner. Then I show you how to take all the information given to you and wade through it to make your way to the answer — the best-fitting model based on a subset of the available *x* variables. Finally, you see how this procedure is applied to find a model to predict punt distance.

Assessing the fit of multiple regression models

For any model selection procedure, assessing the fit of each model being considered is built into the process. In other words, as you go through all the possible models, you're always keeping an eye on how well each model fits. So before you get into a discussion of how to do the best subsets procedure, you need criteria to assess how well a particular model fits a data set.

Although there are tons of different statistics for assessing the fit of regression models, I discuss the most popular ones: R^2 (simple linear regression only), R^2 adjusted, and Mallow's C-p. All three models appear on the bottom line of the Minitab output when you do any sort of model selection procedure. Here's a breakdown of the assessment techniques.

- **R^2:** R^2 is the percentage of the variability in the *y* values that's explained by the model. It falls between 0 and 100 percent (0 and 1.0). In simple linear regression (see Chapter 5), a high value of R^2 means the line fits well, and a low value of R^2 means the line doesn't fit well.

  When you have multiple regression, however, there's a bit of a catch here. As you add more and more variables (no matter how significant), the value of R^2 increases or stays the same — it never goes down. This can result in an inflated measure of how well the model fits. Of course, statisticians have a fix for the problem, which leads me to the next item on this list.

- **R^2 adjusted:** R^2 adjusted takes the value of R^2 and adjusts it downward according to the number of variables in the model. The higher the number of variables in the model, the lower the value of R^2 adjusted will be, compared to the original R^2.

  A high value of R^2 adjusted means the model you have is fitting the data very well (the closer to 1, the better). I typically find a value of 0.70 to be considered okay for R^2 adjusted, and the higher the better.

  Always use R^2 adjusted rather than the regular R^2 to assess the fit of a multiple regression model. With every addition of a new variable into a multiple regression model, the value of R^2 stays the same or increases. It will never go down because a new variable will either help explain some of the variability in the *y*'s (thereby increasing R^2 by definition), or it will do nothing (leaving R^2 exactly where it was before). So theoretically, you could just keep adding more and more variables into the model just for the sake of getting a larger value of R^2.

R^2 adjusted is important because it keeps you from adding more and more variables by taking into account how many variables there already are in the model. The value of R^2 adjusted can actually decrease if the added value of the additional variable is outweighed by the number of variables in the model. This gives you an idea of how much or how little added value you get from a bigger model (bigger isn't always better).

- **Mallow's C-p:** Mallow's C-p takes the amount of error left unexplained by a model of p as the number of x variables, divides that number by the average amount of error left over from the full model (with all the x variables), and adjusts that result for the number of observations (n) and the number of x variables used (p). In general, the smaller Mallow's C-p is, the better, because when it comes to the amount of error in your model, less is more. A C-p value close to p (the number of x variables in the model) reflects a model that fits well.

Model selection procedures

The process of finding the "best" model is not cut and dried. (Heck, even the definition of "best" here isn't cut and dried.) Many different procedures exist for going through different models in a systematic way, evaluating each one, and stopping at the right model. Three of the more common model selection procedures are forward selection, backward selection, and the best subsets model. In this section you get a very brief overview of the forward and backward selection procedures, and then you get into the details of the best subsets model, which is the one statisticians use most.

Going with the forward selection procedure

The *forward selection procedure* starts with a model with no variables in it and adds variables one at a time according to the amount of contribution they can make to the model.

Start with an entry-level value of α. Then run hypothesis tests (see Chapter 4 for instructions) for each x variable to see how it's related to y. The x variable with the smallest p-value wins and is added to the model, as long as its p-value is smaller than the entry level. You keep doing this with the remaining variables until the one with the smallest p-value doesn't make the entry level. Then you stop.

REMEMBER

The drawback of the forward selection procedure is that it starts with nothing and adds variables one at a time as you go along; after a variable is added, it's never removed. The best model might not even get tested.

Opting for the backward selection procedure

The *backward selection procedure* does the opposite of the forward selection method. It starts with a model with all the *x* variables in it and removes variables one at a time. Those that make the least amount of contribution to the model are removed first. You choose a removal level to begin; then you test all the *x* variables and find the one with the largest *p*-value. If the *p*-value of this *x* variable is higher than the removal level, that variable is taken out of the model.

You continue removing variables from the model until the one with the largest *p*-value doesn't exceed the removal level. Then you stop.

REMEMBER

The drawback of the backward selection procedure is that it starts with everything and removes variables one at a time as you go along; after a variable is removed, it never comes back. Again, the best model might not even be tested.

Using the best subsets procedure

The best subsets procedure has fewer steps than the forward or backward selection model because the computer formulates and analyzes all possible models in a single step. In this section, you see how to get the results and then use them to come up with a best multiple regression model for predicting *y*.

Here are the steps for conducting the best subsets model selection procedure to select a multiple regression model; note that Minitab does all the work for you to crunch the numbers:

1. **Conduct the best subsets procedure in Minitab, using all possible subsets of the *x* variables being considered for inclusion in the final model.**

   COMPUTER OUTPUT

   To carry out the best subsets selection procedure in Minitab, go to Stat>Regression>Regression>Best Subsets. Highlight the response variable (*y*), and click Select. Highlight all the (free) predictor (*x*) variables, click Select, and then click OK.

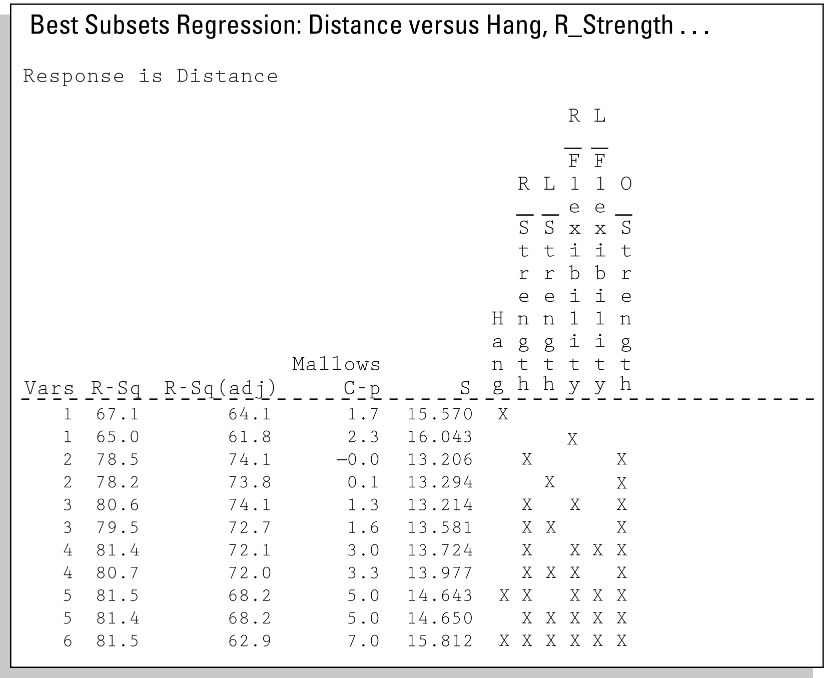
   The output contains a listing of all models that contain one *x* variable, all models that contain two *x* variables, all models that contain three *x* variables, and so on, all the way up to the full model (containing all the *x* variables). Each model is presented in one row of the output.

2. **Choose the best of all the models shown in the best subsets Minitab output by finding the model with the largest value of R^2 adjusted and the smallest value of Mallow's C-p; if two competing models are about equal, choose the model with the fewer number of variables.**

If the model fits well, R^2 adjusted is high. So you also want to look for the smallest possible model that has a high value of R^2 adjusted and a small value of Mallow's C-p compared to its competitors. And if it comes down to two similar models, always make your final model as easy to interpret as possible by selecting the model with fewer variables.

The secret to a punter's success: An example

Returning to the punt distance example from earlier in this chapter, suppose that you analyzed the punt distance data by using the best subsets model selection procedure. Your results are shown in Figure 7-2. This section follows Minitab's footsteps in getting these results and provides you with a guide for interpreting the results.



```
Best Subsets Regression: Distance versus Hang, R_Strength . . .

Response is Distance

                                           R L
                                           _ _
                                           F F
                                       R L l l O
                                       _ _ e e _
                                       S S x x S
                                       t t i i t
                                       r r b b r
                                       e e i i e
                                     H n n l l n
                                     a g g i i g
                     Mallows         n t t t t t
Vars R-Sq  R-Sq(adj)     C-p       S g h h y y h
   1  67.1        64.1     1.7  15.570 X
   1  65.0        61.8     2.3  16.043       X
   2  78.5        74.1    -0.0  13.206   X       X
   2  78.2        73.8     0.1  13.294     X     X
   3  80.6        74.1     1.3  13.214   X   X   X
   3  79.5        72.7     1.6  13.581   X X     X
   4  81.4        72.1     3.0  13.724   X   X X X
   4  80.7        72.0     3.3  13.977   X X X   X
   5  81.5        68.2     5.0  14.643 X X   X X X
   5  81.4        68.2     5.0  14.650   X X X X X
   6  81.5        62.9     7.0  15.812 X X X X X X
```

FIGURE 7-2: Best subsets procedure results for the punt distance example.

Assuming that you already used Minitab to carry out the best subsets selection procedure on the punt distance data, you can now analyze the output from Figure 7-2. Each variable shows up as a column on the right side of the output. Each row represents the results from a model containing the number of variables shown in column one. The *X*'s at the end of each row tell you which variables were included in that model. The number of variables in the model starts at 1 and increases to 6 because six *x* variables are available in the data set.

The models with the same number of variables are ordered by their values of R^2 adjusted and Mallows C-p, from best to worst. The top-two models (for each number of variables) are included in the computer output.

For example, rows one and two of Figure 7-2 (both marked 1 in the Vars column) show the top-two models containing one *x* variable; rows three and four show the top-two models containing two *x* variables; and so on. Finally, the last row shows the results of the full model containing all six variables. (Only one model contains all six variables, so you don't have a second-best model in this case.)

Looking at the first two rows of Figure 7-2, the top one-variable model is the one including hang time only. The second-best one-variable model includes only right foot flexibility. The right foot flexibility model has a lower value of R^2 and a higher Mallow's C-p than the hang time model, which is why it's the second best.

Row three shows that the best two-variable model for estimating punt distance is the model containing right leg strength and overall leg strength. The best three-variable model is in row five; it shows that the best three-variable model includes right foot strength, right foot flexibility, and overall leg strength. The best four-variable model is found in row seven and includes right foot strength, right and left foot flexibility, and overall foot strength. The best five-variable model is found in row nine and includes every variable except left foot strength. The only six-variable model with all variables included is listed in the last row.

Among the best one-variable, two-variable, three-variable, four-variable, and five-variable models, which one should you choose for your final multiple regression model? Which model is the best of the best? With all these results, it would be easy to have a major freakout over which one to pick, but never fear — Mallow's is here (along with his friendly sidekick, the R^2 adjusted).

Looking at Figure 7-2 column three, you see that as the number of variables in the model increases, R^2 adjusted peaks out and then drops way off. That's because R^2 adjusted takes into account the number of variables in the model and reduces R^2 accordingly. You can see that R^2 adjusted peaks out at a level of 74.1 percent for two models. The corresponding models are the top two-variable model (right leg strength and overall leg strength) and the best three-variable model (right foot strength, right foot flexibility, and overall leg strength).

Now look at Mallow's C-p for these two models. Notice that Mallow's C-p is zero for the best two-variable model and 1.3 for the best three-variable model. Both values are small compared to others in Figure 7-2, but because Mallow's C-p is smaller for the two-variable model, and because it has one less variable in it, you should choose the two-variable model (right leg strength and overall leg strength) as the final model, using the best subsets procedure.

IN THIS CHAPTER

- » Getting a feel for nonlinear regression
- » Making use of scatterplots
- » Fitting a polynomial to your data set
- » Exploring exponential models to fit your data

Chapter 8

Getting Ahead of the Learning Curve with Nonlinear Regression

In Stats I, you concentrate on the *simple linear regression model,* where you look for one quantitative variable, *x*, that you can use to make a good estimate of another quantitative variable, *y*, using a straight line. The examples you look at in Stats I fall right in line with this kind of model, such as using height to estimate weight or using study time to estimate an exam score. (For more information and examples for using simple linear regression models, see Chapter 5.)

But not all situations fall into the straight line category. Take gas mileage and speed, for example. At low speeds, gas mileage is lower, and at high speeds, gas mileage is lower; but at medium speeds, gas mileage is higher. This low-high-low relationship between speed and gas mileage represents a curved relationship. Relationships that don't resemble straight lines are called *nonlinear relationships* (clever, huh?). Looked at simply, nonlinear regression takes the stage when you want to predict some quantitative variable (*y*) by using another quantitative variable (*x*) but the pattern you see in the data collected resembles a curve, not a straight line.

In this chapter, you see how to make your way around the curved road of data that leads to nonlinear regression models. The good news is twofold: You can use many of the same techniques you use for regular regression, and in the end, Minitab does the analysis for you.

Anticipating Nonlinear Regression

Nonlinear regression comes into play in situations where you have graphed your data on a *scatterplot* (a two-dimensional graph showing the *x* variable on the *x*-axis and the *y* variable on the *y*-axis; see the next section, "Starting Out with Scatterplots"), and you see a pattern emerging that looks like some type of curve. Examples of data that follow a curve include changes in population size over time, demand for a product as a function of supply, and the length of time that a battery lasts. When a data set follows a curved pattern, the time has come to move away from the linear regression models (covered in Chapters 5 and 6) and move on to a nonlinear regression model.

Suppose a manager is considering the purchase of new office management software but is hesitating. They want to know how long it typically takes someone to get up to speed using the software. So she takes a random sample of workers and looks at their times to complete the task and the number of tries.

What's the statistical question here? The manager wants a model that shows what the learning curve looks like (on average). (A *learning curve* shows the decrease in time to do a task with more and more practice.) In this scenario, you have two variables: time to complete the task and trial number (for example, the first try is designated by 1, the second try by 2, and so on). Both variables are *quantitative* (numerical) and you want to find a connection between two quantitative variables. At this point, you can start thinking regression.

A *regression model* produces a function (be it a line or otherwise) that describes a pattern or relationship. The relationship here is task time versus number of times the task is practiced. But what type of regression model do you use? After all, four types are described in this book: simple linear regression, multiple regression, nonlinear regression, and logistic regression. You need more clues.

The word "curve" in learning curve is a clue that the relationship being modeled here may not be linear. That word signals that you're talking about a nonlinear regression model. If you think about what a possible learning curve may look like, you can imagine task time on the *y*-axis and the number of the trial on the *x*-axis.

You may guess that the *y*-values will be high at first, because the first couple of times you try a new task, it takes longer to perform. Then, as the task is repeated, the task time decreases, but at some point more practice doesn't reduce task time much. So the relationship may be represented by some sort of curve, like the one I simulate in Figure 8-1 (which can be fit by using an exponential function).

This example illustrates the basics of nonlinear regression; the rest of the chapter shows you how the model breaks down.

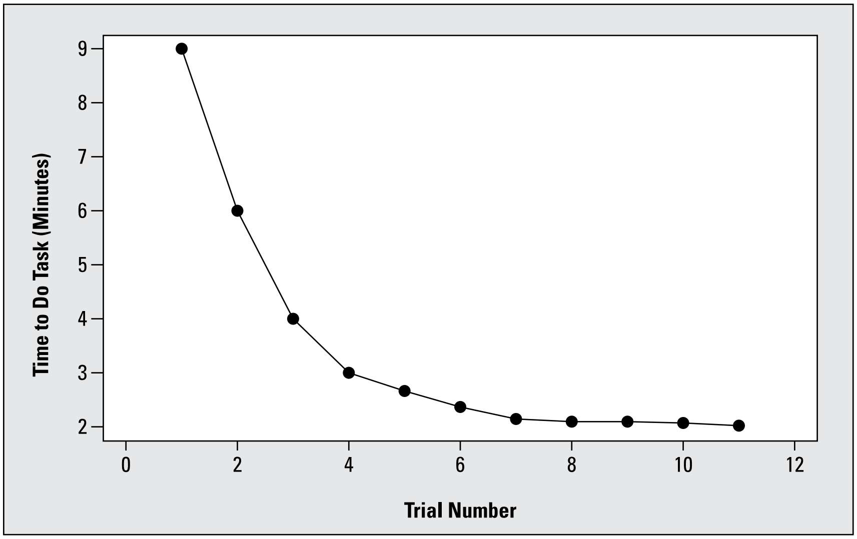


FIGURE 8-1: Learning curve for time performing a new task.

Starting Out with Scatterplots

As with any type of data analysis, before you dive in and select a model that you think fits the data (or that's supposed to fit the data), you have to step back and take a look at the data to see whether any patterns emerge. To do this, look at a scatterplot of the data, and see whether or not you can draw a smooth curve through the data and find that most of the points follow along that curve.

Suppose you're interested in modeling how quickly a rumor spreads. One person knows a secret and tells it to another person, and now two know the secret; each of them tells a person, and now four know the secret; some of those people may pass it on, and so it goes on down the line. Pretty soon, a large number of people know the secret, which is a secret no longer.

To collect your data, you count the number of people who know a secret by tracking how many people know the secret over a six-day period from your population. The data are shown in Table 8-1. Note that the spread of the secret catches fire on day 5 — this is how an exponential model works. You can see a scatterplot of the data in Figure 8-2.

TABLE 8-1

Number of People Knowing a Secret over a 6-Day Period

x (Day)	y (Number of People)
1	1
2	2
3	5
4	7
5	17
6	30

In this situation, the explanatory variable, *x*, is day, and the response variable, *y*, is the number of people who know the secret. Looking at Figure 8-2, you can see a pattern between the values of *x* and *y*. But this pattern isn't linear. It curves upward. If you tried to fit a line to this data set, how well would it fit?

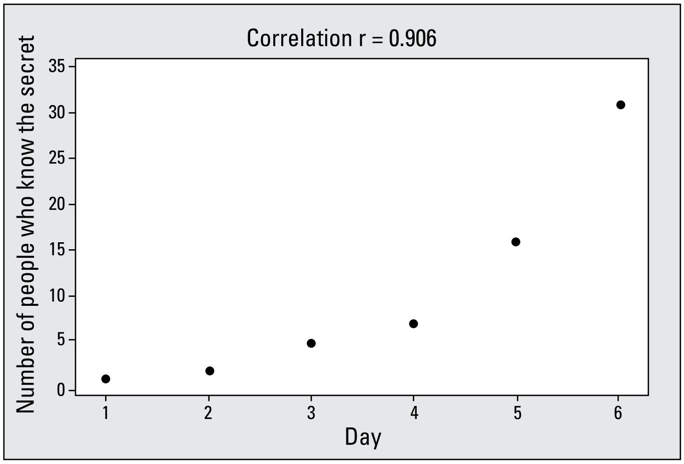


FIGURE 8-2: A scatterplot showing the spread of a secret over a six-day period.

To figure this out, look at the correlation coefficient between *x* and *y*, which is found in Figure 8-2 to be 0.906 (see Chapter 5 for more on correlation). You can interpret this correlation as a strong, positive (uphill) linear relationship between *x* and *y*. However, in this case, the correlation is misleading because the scatterplot appears to be curved.

REMEMBER

If the correlation looks good (close to +1 or −1), don't stop there. As with any regression analysis, it's very important to take into account both the scatterplot and the correlation when making a decision about how well the model being considered would fit the data. The contradiction in this example between the scatterplot and the correlation is a red flag that a straight-line model isn't the best idea.

WARNING

The correlation coefficient measures only the strength and direction of the linear relationship between *x* and *y* (see Chapter 5). However, you may run into situations (like the one shown in Figure 8-2) where a correlation is strong, yet the scatterplot shows a curve would fit better. Don't rely solely on either the scatterplot or the correlation coefficient alone to make your decision about whether to go ahead and fit a straight line to your data.

The bottom line here is that fitting a line to data that appear to have a curved pattern isn't the way to go. Instead, explore models that have curved patterns themselves.

The following sections address two major types of nonlinear (or curved) models that are used to model curved data: polynomials (that are not straight lines — that is, curves like quadratics or cubics), and exponential models (that start out small and quickly increase, or the other way around). Because the pattern of the data in Figure 8-2 starts low and bends upward, the correct model to fit this data is an exponential regression model. (This model is also appropriate for data that start out high and bend down low.)

Handling Curves in the Road with Polynomials

One major family of nonlinear models is the *polynomial* family. You use these models when a polynomial function (beyond a straight line) best describes the curve in the data. (For example, the data may follow the shape of a parabola, which is a second-degree polynomial.) You typically use polynomial models when the data follow a pattern of curves going up and down a certain number of times.

For example, suppose a doctor examines the occurrence of heart problems in patients as it relates to their blood pressure. The doctor finds that patients with very low or very high blood pressure had a higher occurrence of problems, while patients whose blood pressure fell in the middle, constituting the normal range, had fewer problems. This pattern of data has a U-shape, and a parabola would fit this data well.

In this section, you see what a polynomial regression model is, how you can search for a good-fitting polynomial for your data, and how you can assess polynomial models.

Bringing back polynomials

You may recall from algebra that a *polynomial* is a sum of *x* terms raised to a variety of powers, and each *x* is preceded by a constant called the *coefficient* of that term. For example, the model $y = 2x + 3x^2 + 6x^3$ is a polynomial. The general form for a polynomial regression model is $y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_k x^k + \varepsilon$. Here, *k* represents the total number of terms in the model. The ε represents the error that occurs simply due to chance. (Not a bad kind of error, just random fluctuations from a perfect model.)

Here are a few of the more common polynomials you run across when analyzing data and fitting models. Remember, the simplest model that fits is the one you use (don't try to be a hero in statistics — save that for Batman and Robin). The models I discuss in this book are some of your old favorites from algebra: second-, third-, and fourth-degree polynomials.

- **Second-degree (or quadratic) polynomial:** This model is called a *second-degree* (or *quadratic*) *polynomial,* because the largest exponent is 2. An example model is $y = 2x + 3x^2$. A second-degree polynomial forms a parabola shape — either an upside-down or right-side-up bowl; it changes direction one time (see Figure 8-3).
- **Third-degree polynomial:** This model has 3 as the highest power of *x*. It typically has a sideways S-shape, changing directions two times (see Figure 8-4).
- **Fourth-degree polynomial:** Fourth-degree polynomials involve x^4. They typically change directions in curvature three times to look like the letter W or the letter M, depending on whether they're upside down or right-side up (see Figure 8-5).

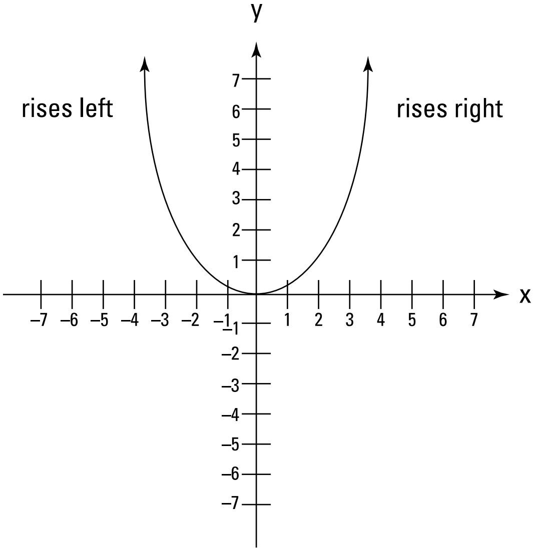


FIGURE 8-3: Example of a second-degree polynomial.

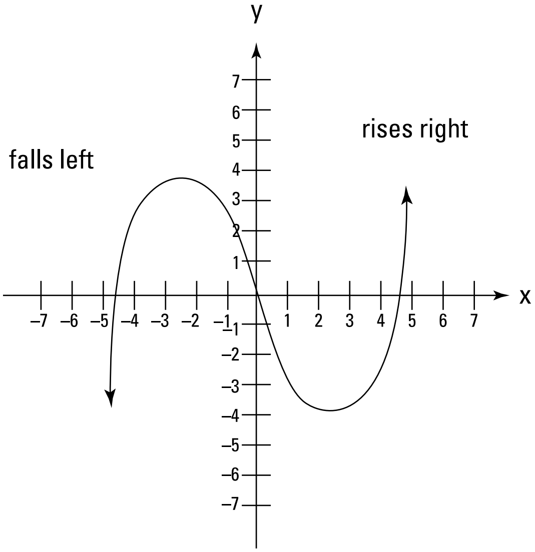


FIGURE 8-4: Example of a third-degree polynomial.

In general, if the largest exponent on the polynomial is n, the number of curve changes in the graph is typically $n - 1$. For more information on graphs of polynomials, refer to your algebra textbook or *Algebra For Dummies*, by Mary Jane Sterling (Wiley).

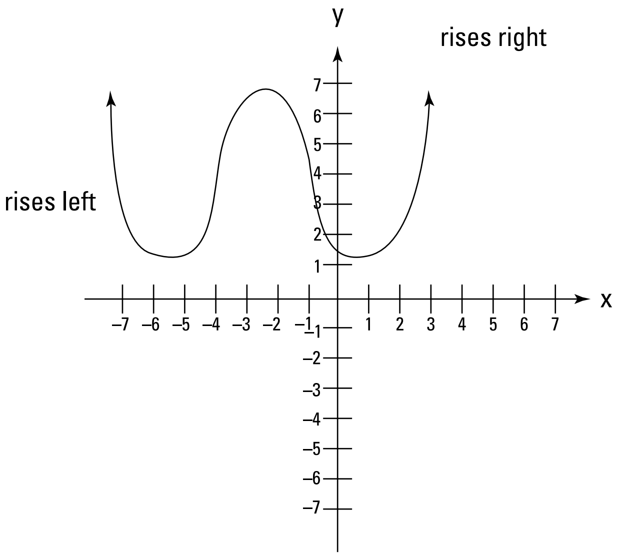


FIGURE 8-5: Example of a fourth-degree polynomial.

TECHNICAL STUFF

The nonlinear models in this chapter involve only one explanatory variable, *x*. You can include more explanatory variables in a nonlinear regression, raising each separate variable to a power, but those models are beyond the scope of this book. I give you information on basic multiple regression models in Chapter 6.

Searching for the best polynomial model

REMEMBER

When fitting a polynomial regression model to your data, you always start with a scatterplot so you can look for patterns; the scatterplot will give you some idea of the type of model that may work. Always start with the simplest model possible and work your way up as needed. Don't plunge in with a high-order polynomial regression model right off the bat. Here are a couple of reasons why:

- **High-order polynomials are hard to interpret, and their models are complex.** For example, with a straight line, you can interpret the values of the *y*-intercept and slope easily, but interpreting a tenth-degree polynomial is difficult (and that's putting it mildly).
- **High-order polynomials tend to cause overfitting.** If you're fitting the model as close as you can to every single point in a data set, your model may not hold for a new data set, meaning that your estimates for *y* could be way off.

COMPUTER OUTPUT

To fit a polynomial to a data set in Minitab, go to Stat>Regression>Fitted Line Plot and click on the type of regression model you want: linear, quadratic, or cubic. (It doesn't go beyond a third-degree polynomial, but these options should cover 90 percent of the cases where a polynomial is appropriate.) Click on the Response (*y*) box, then click on the *y* variable from the left-hand box, and click Select. This variable will appear in the Response (*y*) box. Click on the *x* variable from the left-hand box, and click Select; it will appear in the Predictor (*x*) box. Click OK. (***Note:*** If you want to do a more complicated or higher-order nonlinear equation, you can click on Regression>Nonlinear regression and type your equation into the box. For more details, look in Minitab Help.)

Following are steps that you can use to see if a polynomial fits your data. (Statistical software can jump in and fit the models for you after you tell it which ones to fit.)

1. **Make a scatterplot of your data, and look for any patterns, such as a straight line or a curve.**
2. **If the data resemble a straight line, try to fit a first-degree polynomial (straight line) to the data first: $y = b_0 + b_1x$.**

   If the scatterplot doesn't show a linear pattern, or if the correlation isn't close to +1 or –1, move to Step 3.
3. **If the data resemble the shape of a parabola, try to fit a second-degree polynomial: $y = b_0 + b_1x + b_2x^2$.**

   If the data fit the model well, stop here and refer to the later section, "Assessing the fit of a polynomial model." If the model still doesn't fit well, move to Step 4.
4. **If you see curvature that's more complex than a parabola, try to fit a third-degree polynomial: $y = b_0 + b_1x + b_2x^2 + b_3x^3$.**

   If the data fit the model well, stop here and refer to the later section, "Assessing the fit of a polynomial model." If the model still doesn't fit well, move to Step 5.
5. **Continue trying to fit higher-order polynomials until you find one that fits or until the order of the polynomial (largest exponent) simply gets too large to find a reliable pattern.**

   TIP

   How large is too large? Typically, if you can't fit the data by the time the degree of the polynomial reaches three, then perhaps a different type of model would work better. Or you may determine that you observe too much scatter and haphazard behavior in the data to try to fit any model.

Minitab can do each of these steps for you up to degree three (that's step 4); from there, you need a more sophisticated statistical software program, such as SAS or SPSS, or you can use the Regression>Nonlinear regression feature in Minitab.

However, most of the models you need to fit go up to the second- or third-degree polynomials.

Using a second-degree polynomial to pass the quiz

The first step in fitting a polynomial model is to graph the data in a scatterplot and see whether the data fall into a particular pattern. Many different types of polynomials exist to fit data that have a curved type of pattern. One of the most common patterns found in curved data is the quadratic pattern, or second-degree polynomial, which goes up and comes back down, or goes down and comes back up, as the *x* values move from left to right (see Figure 8-3). The second-degree polynomial is the simplest and most commonly used polynomial beyond the straight line, so it deserves special consideration. (After you master the fundamentals of second-degree polynomials, you can apply them to polynomials with higher powers.)

Suppose 20 students take a statistics quiz. You record the quiz scores (which have a maximum score of ten) and the number of hours students report studying for the quiz. You can see the results in Figure 8-6.

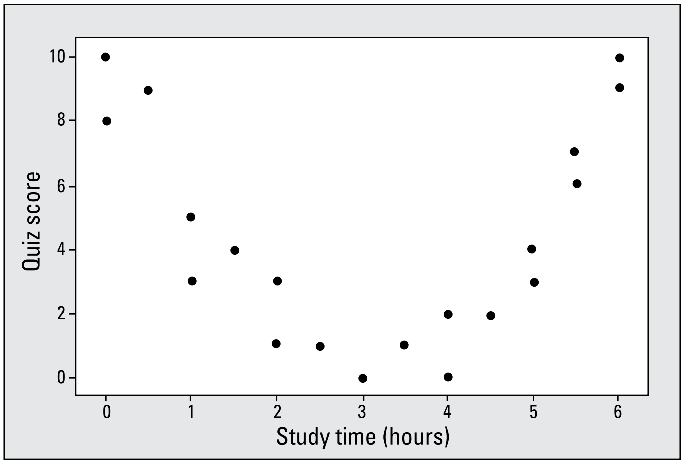


FIGURE 8-6: Scatterplot showing study time and quiz scores.

Looking at Figure 8-6, it appears that three camps of students are in this class. Camp 1, on the left end of the *x*-axis, understood the stuff (as reflected in their higher scores) and hardly had to study at all (you can see that their study time on the *x*-axis is low). Camp 3 also did very well on the quiz (as indicated by their high quiz scores) but had to study a great deal to get those grades (as shown on the far-right end of the *x*-axis). The students in the middle, Camp 2, didn't seem to fare well.

All in all, based on the scatterplot, it does appear that study time may explain quiz scores on some level in a way indicative of a second-degree polynomial. So a quadratic regression model may fit this data.

WARNING

Suppose a data analyst (not you!) doesn't know about polynomial regression and just tries to fit a straight line to the quiz-score data. In Figure 8-7, you can see the data and the straight line that the analyst tried to fit to the data. The correlation as shown in the figure is –0.033, which is basically zero. This correlation means that no linear relationship lies between *x* and *y*. It doesn't mean that no relationship is present at all, just that it's not a linear relationship (see Chapter 5 for more on linear relationships). So trying to fit a straight line here was indeed a bad idea.

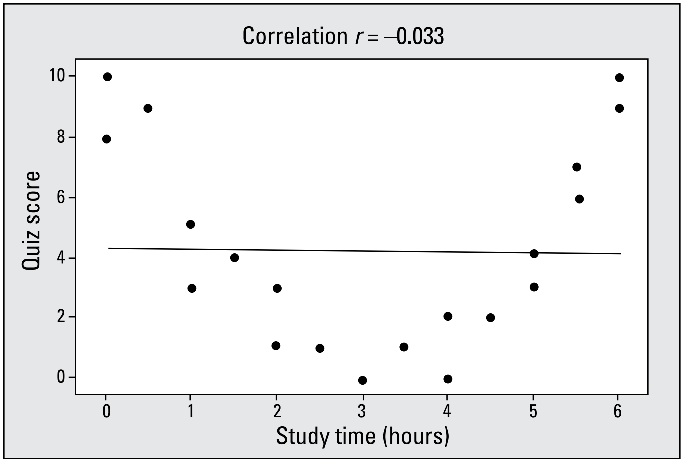
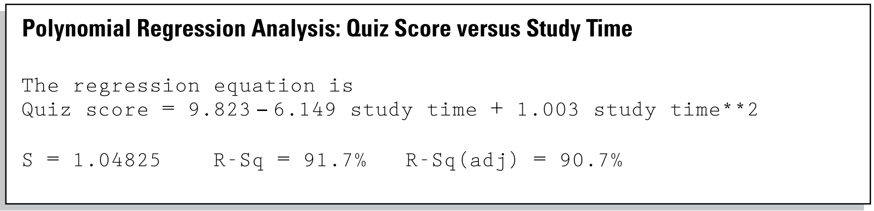


FIGURE 8-7: Trying to fit a straight line to quadratic data.

After you know that a quadratic polynomial seems to be a good fit for the data, the next challenge is finding the equation for that particular parabola that fits the data from among all the possible parabolas out there.

Remember from algebra that the general equation of a parabola is $y = ax^2 + bx + c$. Now you have to find the values of *a*, *b*, and *c* that create the best-fitting parabola to the data (just like you find the *a* and the *b* that create the best-fitting line to data in a linear regression model). That's the object of any regression analysis.

Suppose that you fit a quadratic regression model to the quiz-score data by using Minitab (see the Minitab output in Figure 8-8 and the instructions for using Minitab to fit this model in the previous section). On the top line of the output, you can see that the equation of the best-fitting parabola is $\text{quiz score} = 9.82 - 6.15 * (\text{study time}) + 1.00 * (\text{study time})^2$. (Note that y is quiz score and x is study time in this example because you're using study time to predict quiz score.)



Polynomial Regression Analysis: Quiz Score versus Study Time

```
The regression equation is
Quiz score = 9.823 - 6.149 study time + 1.003 study time**2

S = 1.04825    R-Sq = 91.7%   R-Sq(adj) = 90.7%
```

FIGURE 8-8: Minitab output for fitting a parabola to the quiz-score data.

The scatterplot of the quiz-score data and the parabola that was fit to the data via the regression model is shown in Figure 8-9. From algebra, you may remember that a positive coefficient on the quadratic term (here $a = 1.00$) means the bowl is right-side up, which you can see is the case here.

Looking at Figure 8-9, it appears that the quadratic model fits this data pretty well, because the data fall close to the curve that Minitab found. However, data analysts can't live by scatterplots alone, so the next section helps you figure out how to assess the fit of a polynomial model in more detail.

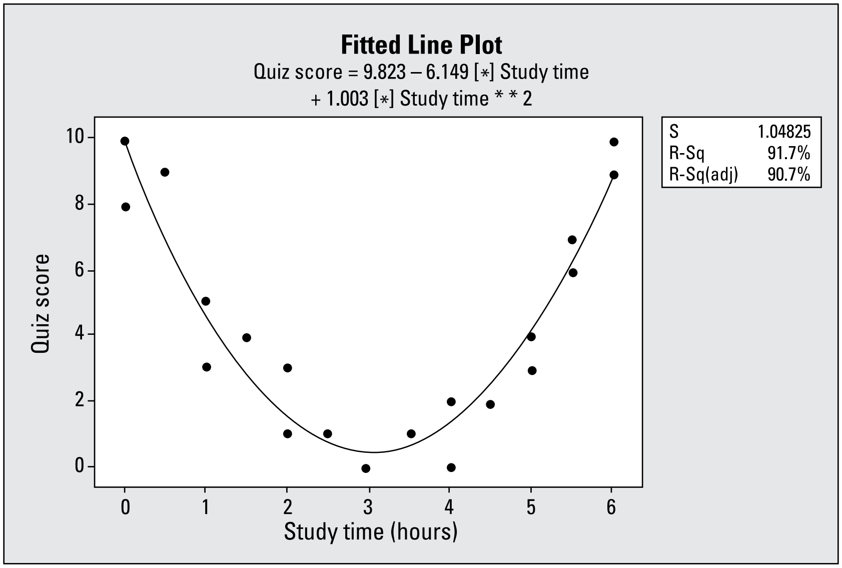


FIGURE 8-9: The parabola appears to fit the quiz-score data nicely.

Assessing the fit of a polynomial model

You make a scatterplot of your data, and you see a curved pattern. So you use polynomial regression to fit a model to the data; the model appears to fit well because the points follow closely to the curve Minitab found, but don't stop there. To make sure your results can be generalized to the population from which your data was taken, you need to do a little more checking beyond just the graph to make sure your model fits well.

To assess the fit of any model beyond the usual suspect, a scatterplot of the data, you look at two additional items, typically in this order: the value of R^2 adjusted and the residual plots.

REMEMBER

All three assessments must agree before you can conclude that the model fits. If the three assessments don't agree, you'll likely have to use a different model to fit the data besides a polynomial model, or you'll have to change the units of the data to help a polynomial model fit better. However, the latter fix is outside the scope of Stats II, and you probably won't encounter that situation.

In the following sections, you take a deeper look at the value of R^2 adjusted and the residual plots and figure out how you can use them to assess your model's fit. (You can find more information on the scatterplot in the section, "Starting Out with Scatterplots," earlier in this chapter.)

Examining R^2 and R^2 adjusted

Finding R^2, the *coefficient of determination* (see Chapter 6 for full details), is like the day of reckoning for any model. You can find R^2 on your regression output, listed as "R-Sq" in the Model Summary portion of the regression output just below where the regression equation appears. Figure 8-8 shows the Minitab output for the quiz-score data example; the value of R^2 in this case is 91.7 percent.

The value of R^2 tells you what percentage of the variation in the y-values the model can explain. To interpret this percentage, note R^2 is the square of r, the correlation coefficient (see Chapter 6). Because values of r beyond ± 0.80 are considered to be good, R^2 values above 0.64 are also considered pretty good, especially for models with only one x variable.

You can consider values of R^2 over 80 percent good, and values under 60 percent not so good. Those in between I'd consider so-so; they could be better. (This assessment is just my rule-of-thumb; opinions may vary a bit from one statistician to another.)

However, you can find such a thing in statistics as too many variables spoiling the pot. Every time you add another x variable to a regression model, the value of R^2 automatically goes up, whether the variable really helps or not (this is just a mathematical fact). Right beside R^2 on the computer output from any regression analysis is the value of R^2 adjusted, which adjusts the value of R^2 down a notch for each variable (and each power of each variable) entered into the model. You can't just throw a ton of variables into a model whose tiny increments all add up to an acceptable R^2 value without taking a hit for throwing everything in the model but the kitchen sink.

REMEMBER

To be on the safe side, you should always use R^2 adjusted to assess the fit of your model, rather than R^2, especially if you have more than one x variable in your model (or more than one power of an x variable). The values of R^2 and R^2 adjusted are close if you have only a couple of different variables (or powers) in the model, but as the number of variables (or powers) increases, so does the gap between R^2 and R^2 adjusted. In that case, R^2 adjusted is the most fair and consistent coefficient to use to examine model fit.

In the quiz-score example analysis (shown in Figure 8-8), the value of R^2 adjusted is 90.7 percent, which is still a very high value, meaning that the quadratic model fits this data very well. (See Chapter 7 for more on R^2 and R^2 adjusted.)

Checking the residuals

You've looked at the scatterplot of your data, and the value of R^2 is high. What's next? Now you examine how well the model fits each individual point in the data to make sure you can't find any spots where the model is way off or places where you missed another underlying pattern in the data.

A *residual* is the amount of error, or leftover, that occurs when you fit a model to a data set. The residuals are the distances between the predicted values in the model and the observed values of the data themselves. For each observed y-value in the data set, you also have a predicted value from the model, typically called *y-hat*, denoted $\hat{y}$. The residual is the difference between the values of y and y-hat. Each y-value in the data set has a residual; you examine all the residuals together as a group, looking for patterns or unusually high values (indicating a big difference between the observed y and the predicted y at that point; see Chapter 5 for the full information on residuals and their plots).

In order for the model to fit well, the residuals need to meet two conditions:

- **The residuals are independent.** The independence of residuals means that you don't see any pattern as you plot the residuals. The residuals don't affect each other and should be random.

- **The residuals have a normal distribution centered at zero, and the standardized residuals follow suit.** Having a normal distribution with mean zero means that most of the residuals should be centered around zero, with fewer of them occurring the farther from zero you get. You should observe about as many residuals above the zero line as below it. If the residuals are standardized, this means that as a group their standard deviation is 1; you should expect about 95 percent of them to lie between –2 and +2, following the 68-95-99.7% Rule (see your Stats I text).

You determine whether or not these two conditions are met for the residuals by using a series of four graphs called *residual plots.* Most statisticians prefer to standardize the residuals (meaning they convert them to *Z*-scores by subtracting their mean and dividing by their standard deviation) before looking at them, because then they can compare the residuals with values on a *Z*-distribution. If you also take this step, you can ask Minitab to give you a series of four standardized residual plots with which to check the conditions. (See Chapter 5 for full details on standardized residuals and residual plots.)

Figure 8-10 shows the standardized residual plots for the quadratic model, using the quiz-score data from the previous sections. Looking at the figure, you should notice the following:

- The upper-left plot shows that the standardized residuals resemble a normal distribution because your data and the normal distribution match up pretty well, point for point.
- The upper-right plot shows that most of the standardized residuals fall between –2 and +2 (see Chapter 5 for more on standardized residuals).
- The lower-left plot shows that the residuals bear some resemblance to a normal distribution.
- The lower-right plot demonstrates how the residuals have no pattern. They appear to occur at random.

When taken together, all these plots suggest that the conditions on the residual are met to apply the selected quadratic regression model.

Making predictions

After you've found the model that fits well, you can use that model to make predictions for *y* given *x*: Simply plug in the desired *x*-value, and out comes your predicted value for *y*. (Make sure any values you plug in for *x* occur within the range of where data were collected; if not, you can't guarantee the model will hold.)

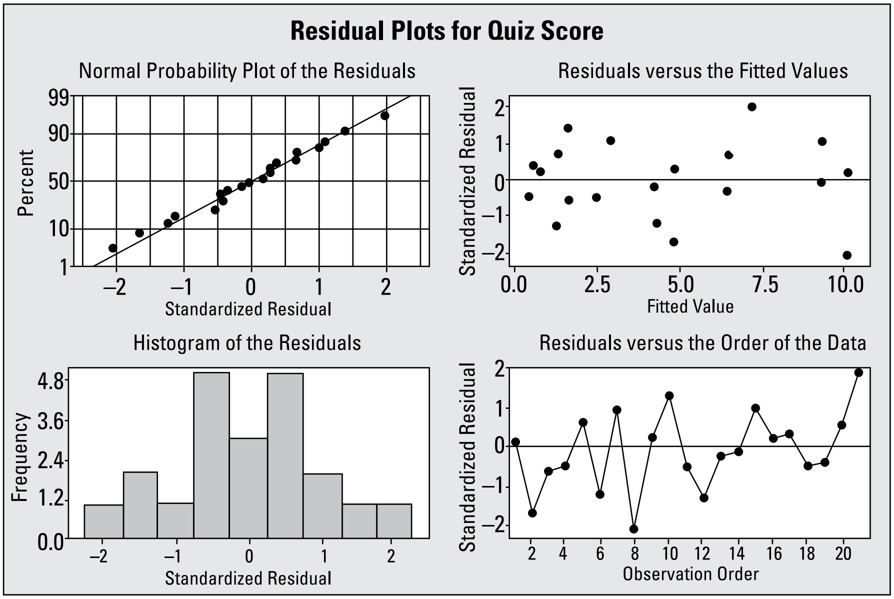


FIGURE 8-10: Standardized residual plots for the quiz-score data, using the quadratic model.

Returning to the quiz-score data from previous sections, can you use study time to predict quiz score by using a quadratic regression model? By looking at the scatterplot and the value of R^2 adjusted (review Figures 8-8 and 8-9, respectively), you can see that the quadratic regression model appears to fit the data well. (Isn't it nice when you find something that fits?) The residual plots in Figure 8-10 indicate that the conditions seem to be met to fit this model; you can find no major patterns in the residuals, they appear to center at 1, and most of them stay within the normal boundaries of standardized residuals of −2 and +2.

Considering all this evidence together, study time does appear to have a quadratic relationship with quiz score in this case. You can now use the model to make estimates of quiz score given study time. For example, because the model (shown in Figure 8-8) is $y = 9.82 - 6.15x + 1.00x^2$, if your study time is 5.5 hours, then your estimated quiz score is $9.82 - 6.15 * 5.5 + 1.00 * 5.5^2 = 9.82 - 33.83 + 30.25 = 6.25$.

This value makes sense according to what you see on the graph in Figure 8-6 if you look at the place where $x = 5.5$; the y-values are in the vicinity of 6 to 7.

REMEMBER

As with any regression model, you can't estimate the value of y for x-values outside the range of where data was collected. If you try to do this, you commit a no-no called *extrapolation.* It refers to trying to make predictions beyond where your data allows you to. You can't be sure that the model you fit to your data actually continues ad infinitum for any old value of x. In the quiz-score example (see Figure 8-6), you really can't estimate quiz scores for study times higher than six hours using this model because the data don't show anyone studying more than six hours. The model likely levels off after six hours to a score of ten, indicating that studying more than six hours is overkill. (You didn't hear that from me, though!)

Going Up? Going Down? Go Exponential!

Exponential models work well in situations where a *y* variable either increases or decreases exponentially over time. That means the *y* variable either starts out slow and then increases at a faster and faster rate or starts out high and decreases at a slower and slower rate.

Many processes in the real world behave like an exponential model: for example, the change in population size over time, average household incomes over time, the length of time a product lasts, or the level of patience one has as the number of statistics homework problems goes up.

In this section, you familiarize yourself with the exponential regression model and see how to use it to fit data that either rise or fall at an exponential rate. You also discover how to build and assess exponential regression models in order to make accurate predictions for a response variable *y*, using an explanatory variable *x*.

Recollecting exponential models

Exponential models have the form $y = \alpha\beta^x$. These models involve a constant, β, raised to higher and higher powers of *x* multiplied by a constant, α. The constant β represents the amount of curvature in the model. The constant α is a multiplier in front of the model that shows where the model crosses the *y*-axis because when $x = 0$, y becomes $\alpha\beta^0 = \alpha$.

REMEMBER

An exponential model generally looks like the upper part of a hyperbola (remember those from advanced algebra?). A *hyperbola* is a curve that crosses the *y*-axis at a point and curves downward toward zero or starts at some point and curves upward to infinity (see Figure 8-11 for examples). If β is greater than 1 in an exponential model, the graph curves upward toward infinity. If β is less than 1, the graph curves downward toward zero. All exponential models stay above the *x*-axis.

For example, the model $y = 1 * 3^x$ is an exponential model. Here, suppose you make $\alpha = 1$, indicating that the model crosses the *y*-axis at 1 (because plugging $x = 0$ into the equation gives you 1). You set the value of β equal to 3, indicating that you want a bit of curvature to this model. The *y*-values curve upward quickly from the point

(0, 1). For example, when $x = 1$, you get $1 * 3^1 = 3$; for $x = 2$, you get $1 * 3^2 = 9$; for $x = 3$, you get $1 * 3^3 = 27$; and so on. Figure 8-11a shows a graph of this model. Notice the huge scale needed on the *y*-axis when *x* is only 10.

Now suppose you let $\alpha = 1$ and $\beta = 0.5$. These values give you the model $y = 1 * 0.5^x$. This model takes 0.5 (a fraction between 0 and 1) to higher and higher powers starting at $1 * 0.5^0 = 1$, which makes the *y*-values smaller and smaller, never reaching zero but always getting closer. (For example, 0.5 to the second power is 0.25, which is less than 0.50, and 0.50 to the tenth power is 0.00098.) Figure 8-11b shows a graph of this model.

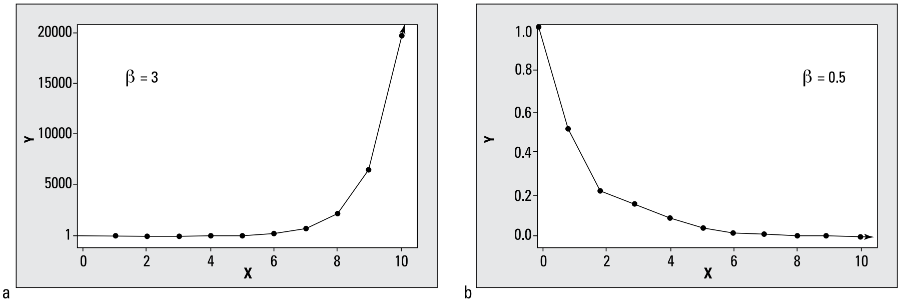


FIGURE 8-11: The exponential regression model for different values of β.

Searching for the best exponential model

Finding the best-fitting exponential model requires a bit of a twist compared to finding the best-fitting line by using simple linear regression (see Chapter 5). Because fitting a straight-line model is much easier than fitting an exponential model directly from data, you transform the data into something for which a line fits. Then you fit a straight-line model to that transformed data. Finally you undo the transformation, getting you back to an exponential model.

For the transformation, you use *logarithms* because they're the inverse of exponentials. But before you start sweating, don't worry; these math gymnastics aren't something you do by hand — the computer does most of the grunt work for you.

The exponential model looks like this (if you're using base 10): $y = 10^{b_0 + b_1 x}$; note the equation of the line is in the exponent. Follow these steps for fitting an exponential model to your data and using it to make predictions:

REMEMBER

The math magic used in these steps is courtesy of the definition of logarithm, which says $\log_b(a) = y \Leftrightarrow b^y = a$. Suppose you have the equation $\log_{10} y = 2 + 3x$.

If you take ten to the power of each side, you get $10^{\log_{10}(y)} = 10^{2+3x}$. By the definition of logarithm, the tens cancel out on the left side and you get $y = 10^{2+3x}$. This model is exponential because *x* is in the exponent. You can take step two up another notch to include the general form of the straight-line model $y = \mathrm{b}_0 + \mathrm{b}_1 x$. Using the definition of logarithm on this line, you get $\log_{10}(y) = b_0 + b_1 x \Leftrightarrow 10^{b_0 + b_1 x} = y$.

1. **Make a scatterplot of the data and see whether the data appear to have a curved pattern that resembles an exponential curve.**

   If the data follow an exponential curve, proceed to Step 2; otherwise, consider alternative models (such as multiple regression in Chapter 6).

   Chapter 5 tells you how to make a scatterplot in Minitab. For more details on what shape to look for, refer to the previous section.

2. **Use Minitab to fit a line to the log(*y*) data.**

   COMPUTER OUTPUT

   In Minitab, go to the regression model (curve fit) by clicking Stat>Regression> Fitted Line Plot and within that window clicking the Linear selection. Under the Options, select Logten of *y*. Then select Using scale of logten to give you the proper units for the graph.

   Understanding the basic idea of what Minitab does during this step is important; being able to calculate it by hand isn't. Here's what Minitab does:

   1. Minitab applies the log (base 10) to the *y*-values. For example, if *y* is equal to 100, $\log_{10}100$ equals 2 (because 10 to the second power equals 100). Note that if the *y*-values fell close to an exponential model before, the log(*y*) values will fall close to a straight-line model. This phenomenon occurs because the logarithm is the inverse of the exponential function, so they basically cancel each other out and leave you with a straight line.
   2. Minitab fits a straight line to the log(*y*) values by using simple linear regression (see Chapter 5). The equation of the best-fitting straight line for the log(*y*) data is $\log(y) = b_0 + b_1 x$. Minitab passes this model on to you in its output, and you take it from there.

3. **Transform the model back to an exponential model by starting with the straight-line model, $\boldsymbol{\log(y) = b_0 + b_1 x}$, that was fit to the $\boldsymbol{\log_{10}(y)}$ data and then applying ten to the power of the left side of the equation and ten to the power of the right side.**

   By the definition of logarithm, you get *y* on the left side of the model and ten to the power of $b_0 + b_1 x$ on the right side. The resulting exponential model for *y* is $y = 10^{b_0 + b_1 x}$.

4. **Use the exponential model in Step 3 to make predictions for *y* (your original variable) by plugging your desired value of *x* into the model.**

   Only plug in values for *x* that are in the range of where the data are located.

5. **Assess the fit of the model by looking at the scatterplot of the log(*y*) data, checking out the value of R^2 (adjusted) for the straight-line model for log(*y*), and checking the residual plots for the log(*y*) data.**

   The techniques and criteria you use to do this are the same as those I discuss in the previous section, "Assessing the fit of a polynomial model."

If these steps seem dubious to you, stick with me. The example in the next section lets you see each step firsthand, which helps a great deal. In the end, actually finding predictions by using an exponential model is a lot easier to do than it is to explain.

Spreading secrets at an exponential rate

Often, the best way to figure something out is to see it in action. Using the secret-spreading example from Figure 8-2, you can work through the series of steps from the preceding section to find the best-fitting exponential model and use it to make predictions.

Step one: Check the scatterplot

Your goal in Step 1 is to make a scatterplot of the secret-spreading data and determine whether the data resemble the curved function of an exponential model. Figure 8-2 shows the data for the spread of a number of people knowing the secret, as a function of the number of days. You can see that the number of people who know the secret starts out small, but then as more and more people tell more and more people, the number grows quickly until the secret isn't a secret anymore. This is a good situation for an exponential model, due to the amount of upward curvature in this graph.

Step two: Let Minitab do its thing to log(y)

In Step 2, you let Minitab find the best-fitting line to the log(*y*) data (see the section, "Searching for the best exponential model," to find out how to do this in Minitab). The output for the analysis of the secret-spreading data is in Figure 8-12; you can see that the best-fitting line is $\log(y) = -0.27 + 0.29 * x$, where *y* is the number of people knowing the secret and *x* is the number of days.

Step three: Go exponential

After you have your Minitab output, you're ready for Step 3. You transform the model $\log(y) = -0.27 + 0.29 * x$ into a model for *y* by taking 10 to the power of the left-hand side and 10 to the power of the right-hand side. Transforming the log(*y*) equation for the secret-spreading data, you get $y = 10^{-.27+.29x}$.

Regression Analysis: y (Number of People) versus x (Day)

The regression equation is
log10(y (Number of People)) = – 0.2732 + 0.2949 x (Day)

Model Summary

S	R-sq	R-sq(adj)
0.0577401	99.13%	98.91%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1.52146	1.52146	456.36	0.000
Error	4	0.01334	0.00333		
Total	5	1.53480			

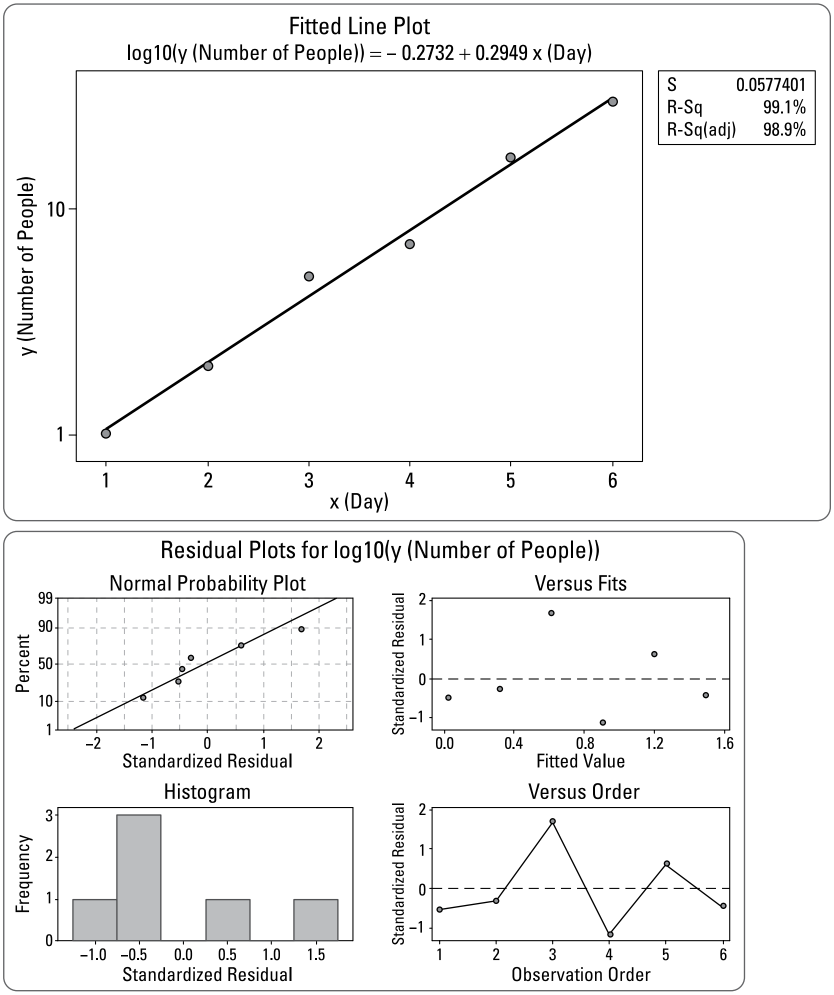


FIGURE 8-12: Minitab fits a line to the log(*y*) for the secret-spreading data.

Step four: Make predictions

By using the exponential model from Step 3, you can move on to Step 4: making predictions for appropriate values of *x* (within the range of where data was collected). Continuing to use the secret-spreading data, suppose you want to estimate the number of people knowing the secret on day five (see Figure 8-2). Just plug $x = 5$ into the exponential model to get $y = 10^{-0.27+0.29*5} = 10^{1.21} = 15.14$ or about 15 people. Looking back at Figure 8-2, you can see that this estimate falls right in line with the data on the graph.

Step five: Assess the fit of your exponential model

Now that you've found the best-fitting exponential model, you have the worst behind you. You've arrived at Step 5 and are ready to further assess the model fit (beyond the scatterplot of the original data) to make sure no major problems arise.

In general, to assess the fit of an exponential model, you're really looking at the straight-line fit of log(*y*). Just use these three items (in any order) in the same way as described in the earlier section, "Assessing the fit of a polynomial model":

- **Check the scatterplot of the log(*y*) data to see how well it resembles a straight line.** You assess the fit of the log(*y*) for the secret-spreading data first through the scatterplot shown in Figure 8-13. The scatterplot shows that the model appears to fit the data well, because the points are scattered in a tight pattern around a straight line.

  REMEMBER

  During this process, the data were also transformed. You started with *x* and *y* data, and now you have *x* and log(*y*) for your data. You see *x*, *y*, and log(*y*) for the secret-spreading data in Table 8-2.

TABLE 8-2 Log(*y*) Values for the Secret-Spreading Data

x (Day)	y (Number of People)	log(y)
1	1	0.00
2	2	0.30
3	5	0.70
4	7	0.85
5	17	1.23
6	30	1.48

- **Examine the value of R^2 adjusted for the model of the best-fitting line for log(*y*), done by Minitab.** The value of R^2 adjusted for this model is found in Figure 8-13 to be 98.91 percent. This value also indicates a good fit because it's very close to 100 percent. Therefore, 98.91 percent of the variation in the number of people knowing the secret is explained by how many days it has been since the secret-spreading started. (Makes sense.)
- **Look at the residual plots from the fit of a line to the log(*y*) data.** The residual plots from this analysis (see Figure 8-14) show no major departures from the conditions that the errors are independent and have a normal distribution. Note that the histogram in the lower-left corner doesn't look all that bell-shaped, but you don't have a lot of data in this example, and the rest of the residual plots seem okay. So, you have little cause to really worry.

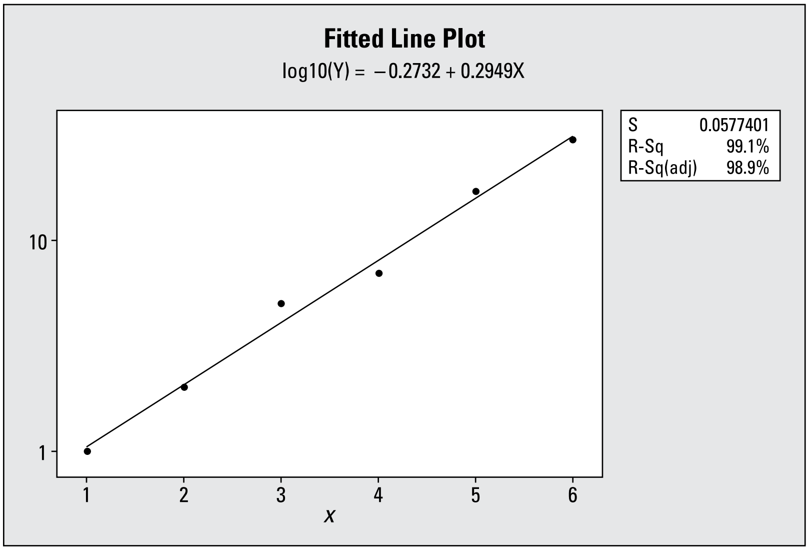


FIGURE 8-13: A scatterplot showing the fit of a straight line to log(*y*) data.

All in all, it appears that the secret's out on the secret-spreading data, now that you have an exponential model that explains how it happens.

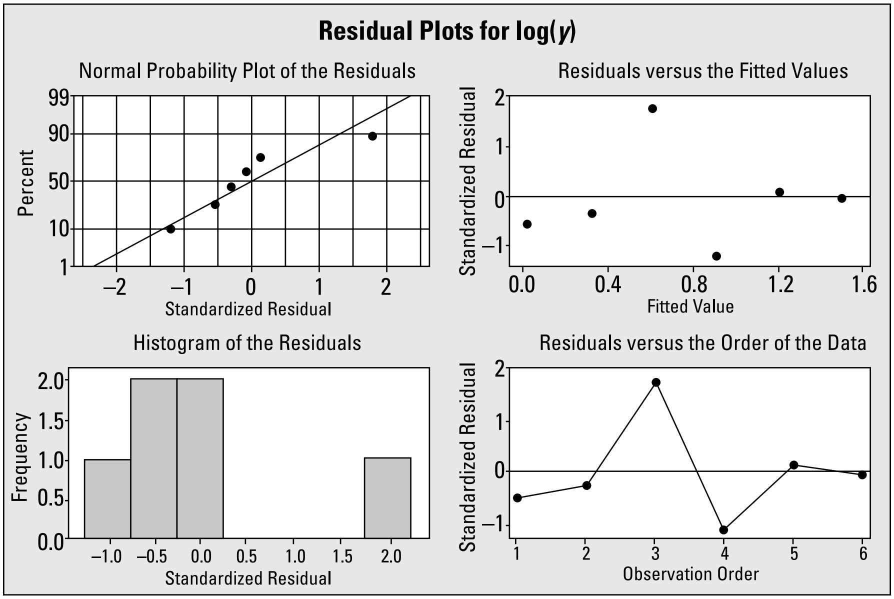


FIGURE 8-14: Residual plots showing the fit of a straight line to log(*y*) data.

IN THIS CHAPTER

- Knowing when logistic regression is appropriate
- Building logistic regression models for yes or no data
- Checking model conditions and making the right conclusions

Chapter 9
Yes, No, Maybe So: Making Predictions by Using Logistic Regression

Everyone (even yours truly) tries to make predictions about whether or not a certain event is going to happen. For example, what's the chance it's going to rain this weekend? What are your team's chances of winning the next game? What's the chance that I'll have complications during this surgery? These predictions are often based on *probability,* the long-term likelihood that an event is expected to happen.

In the end, you want to estimate *p*, the probability of an event occurring. In this chapter, you see how to build and test models for *p* based on a set of explanatory (*x*) variables. This technique is called *logistic regression,* and in this chapter, I explain how to put it to use.

Understanding a Logistic Regression Model

In a logistic regression, you're estimating the probability that an event occurs for a randomly selected individual versus the probability that the event doesn't occur. In essence, you're looking at yes or no data: yes, it occurred (probability $= p$); or no, it didn't occur (probability $= 1 - p$). Yes or no data that come from a random sample have a binomial distribution with probability of success (the event occurring) equal to *p*.

In the binomial problems you saw in Stats I (and reviewed in Chapter 3 of this book), you had a sample of size *n* trials, you had yes or no data, and you had a probability of success on each trial, denoted by *p*. In your Stats I course, for any binomial problem, the value of *p* was somehow given to be a certain value, like a fair coin has a probability $p = 0.50$ for coming up heads. But in Stats II, you operate under the much more realistic scenario that it doesn't. In fact, because *p* isn't known, your job is to estimate what it is and to use a model to do that.

REMEMBER

To estimate *p*, the chance of an event occurring, you need data that come in the form of yes or no, indicating whether or not the event occurred for each individual in the data set.

Because yes or no data don't have a normal distribution, which is a condition needed for other types of regression, you need a new type of regression model to do this job; that model is *logistic regression.*

How is logistic regression different from other regressions?

You use logistic regression when you use a quantitative variable to predict or guess the outcome of some categorical variable with only two outcomes (for example, using barometric pressure to predict whether or not it will rain). (You can also use categorical variables in logistic regression but that is outside the scope of this book.)

A logistic regression model ultimately gives you an estimate for *p*, the probability that a particular outcome will occur in a yes or no situation (for example, the chance that it will rain versus not). The estimate is based on information from one or more explanatory variables; you can call them x_1, x_2, x_3,. . .x_k. (For example, $x_1 =$ humidity, $x_2 =$ barometric pressure, $x_3 =$ cloud cover,. . . and $x_k =$ wind speed.)

Because you're trying to use one variable (*x*) to make a prediction for another variable (*y*), you may think about using regression — and you would be right. However, you have many types of regression to choose from, and you need to determine what kind is most appropriate here. You need the type of regression that uses a quantitative variable (*x*) to predict the outcome of some categorical variable (*y*) that has only two outcomes (yes or no).

So being the good Stats II student that you are, you go to your trusty list of statistical techniques, and you look under regression — and immediately see more than one type. Do you use:

- **Simple linear regression?** No, you use it when you have one quantitative variable predicting another (see Chapter 5).
- **Multiple regression?** No, that method just expands simple linear regression to add more *x* variables (see Chapter 6).
- **Nonlinear regression?** Well, no, that still works with two quantitative variables; it's just that the data form a curve, not a line.

But then you come across logistic regression, and . . . eureka! You see that logistic regression handles situations where the *x* variable is numerical and the *y* variable is categorical with two possible categories. Just what you're looking for!

REMEMBER

Logistic regression, in essence, estimates the probability of *y* being in one category or the other, based on the value of some quantitative variable, *x*. For example, suppose you want to predict a baby's length based on sex. Because a baby's sex at birth is a categorical variable, you use logistic regression to make these predictions. Suppose a 1 indicates a male. Babies who receive a probability of more than 0.5 of being male (based on their lengths) are predicted to be male, and babies who receive a probability of less than 0.5 of being male (based on their lengths) are predicted to be female.

TIP

In this chapter, I present only the case where you use one explanatory variable to predict the outcome. You can extend the ideas in exactly the same way as you can extend the simple linear regression model to a multiple regression model.

Using an S-curve to estimate probabilities

In a simple linear regression model, the general form of a straight line is $y = \beta_0 + \beta_1 x$ where *y* is a quantitative variable. In the logistic regression model, the *y* variable is categorical, not quantitative. What you're estimating, however, is not which category the individual lies in, but rather what the probability is that the individual lies in a certain category. So, the model for logistic regression is based on estimating this probability, called *p*.

If you were to estimate p using a simple linear regression model, you might think that you should try to fit a straight line, $p = \beta_0 + \beta_1 x$. However, it doesn't make sense to use a straight line to estimate the probability of an event occurring based on another variable, due to the following reasons:

- **The estimated values of *p* can never be outside of [0, 1], which goes against the idea of a straight line (a straight line continues on in both directions).**
- **It doesn't make sense to force the values of *p* to increase in a linear way based on *x*.** For example, an event may occur very frequently with a range of large values of *x* and very frequently with a range of small values of *x*, with very little chance of the event happening in an area in between. This type of model would have a "U" shape rather than a straight-line shape.

To come up with a more appropriate model for p, statisticians created a new function of p whose graph is called an S-curve. The *S-curve* is a function that involves p, but it also involves e (the natural logarithm) as well as a ratio of two functions.

The values of the S-curve always fit between 0 and 1, which allows the probability, p, to change from low to high or high to low, according to a curve that's shaped like an S. The general form of the logistic regression model based on an S-curve is $p = \dfrac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$.

Interpreting the coefficients of the logistic regression model

REMEMBER

The sign on the parameter β_1 tells you the direction of the S-curve. If β_1 is positive, the S-curve goes from low to high (see Figure 9-1a); if β_1 is negative, the S-curve goes from high to low (see Figure 9-1b).

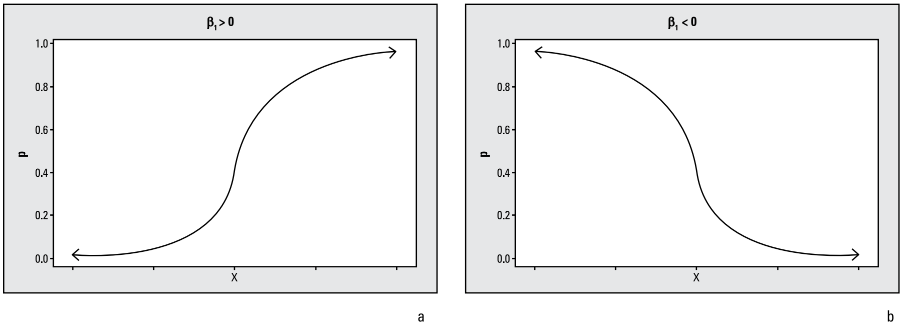


FIGURE 9-1: Two basic types of S-curves.

The magnitude of β_1 (indicated by its absolute value) tells you how much curvature is in the model. High values indicate a steep curvature, and low values indicate a gradual curvature. The parameter β_0 just shifts the S-curve to the proper location to fit your data. It shows you the cutoff point where x-values change from high to low probability and vice versa.

The logistic regression model in action

Often, the best way to figure something out is to see it in action. In this section, I give you an example of a situation where you can use a logistic regression model to estimate a probability. (I expand on this example later in this chapter; for now, I'm just setting up a scenario for logistic regression.)

Suppose movie marketers want to estimate the chance that someone will enjoy a certain family movie, and you believe age may have something to do with it. Translating this research question into x's and y's, the response variable (y) is whether or not a person will enjoy the movie, and the explanatory variable (x) is the person's age. You want to estimate p, the chance of someone enjoying the movie.

You collect data on a random sample of 40 people, shown in Table 9-1. Based on your data, it appears that younger people enjoyed the movie more than older people and that at a certain age, the trend switches from liking the movie to disliking it. Armed with this data, you can build a logistic regression model to estimate p based on the person's age.

TABLE 9-1 Movie Enjoyment (Yes or No Data) Based on Age

Age	Enjoyed the Movie	Total Number Sampled
10	3	3
15	4	4
16	3	3
18	2	3
20	2	3
25	2	4
30	2	4
35	1	5
40	1	6
45	0	3
50	0	2

Carrying Out a Logistic Regression Analysis

The basic idea of any model-fitting process is to look at all possible models you can have under the general format and find the one that fits your data best.

The general form of the best-fitting logistic regression model is $\hat{p} = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}}$, where $\hat{p}$ is the estimate of p, b_0 is the estimate of β_0, and b_1 is the estimate of β_1 (from the earlier section, "Using an S-curve to estimate probabilities"). The only values you have a choice about to form your particular model are the values of b_0 and b_1. These values are the ones you're trying to estimate through the logistic regression analysis.

To find the best-fitting logistic regression model for your data, complete the following steps:

1. **Run a logistic regression analysis on the data you collected (see the next section).**
2. **Find the coefficients of the constant and *x*, where *x* is the name of your explanatory variable.**

   These coefficients are b_0 and b_1, the estimates of β_0 and β_1 in the logistic regression model.
3. **Plug the coefficients from Step 1 into the logistic regression model:**

   $$\hat{p} = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}}$$

   This equation is your best-fitting logistic regression model for the data. Its graph is an S-curve (for more on the S-curve, see the section, "Using an S-curve to estimate probabilities," earlier in this chapter).

In the sections that follow, you see how to ask Minitab to do these steps for you. You also see how to interpret the resulting computer output, find the equation of the best-fitting logistic regression model, and use that model to make predictions (being ever mindful that all conditions are met).

Running the analysis in Minitab

COMPUTER OUTPUT

Here's how to perform a logistic regression using Minitab (other statistical software packages are similar):

1. **Input your data in the spreadsheet as a table that lists each value of the *x* variable in column one, the number of yeses for that value of *x* in column two, and the total number of trials at that *x*-value in column three.**

   These last two columns represent the outcome of the response variable *y*. (For an example of how to enter your data, see Table 9-1, which is based on the movie and age data.)

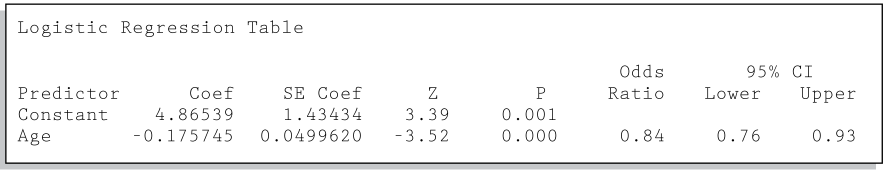
2. **Go to Stat>Regression>Binary Logistic Regression>Fit Binary Logistic Model.**
3. **Select the pull-down menu at the top and choose Response in Event/Trial Format.** In the Event Name box, type the name of what represents a yes in your data (for example, likes the movie).
4. **For the number of events box, select your variable name from column two, and beside the number of trials box, select your variable name from column three.**
5. **Under Predictors, select your variable name from column one, because that's the column containing the explanatory (*x*) variable in your model.** In this case, the *x* variable is age, so it goes in the Continuous Predictors box.
6. **Click OK.** You get your logistic regression output.

When you fit a logistic regression model to your data, the computer output is composed of two major portions:

- **The model-building portion.** In this part of the output, you can find the coefficients b_0 and b_1. (I describe coefficients in the next section.)
- **The model-fitting portion.** You can see the results of a Chi-square goodness-of-fit test (see Chapter 16); large *p*-values mean the model fits. In addition, you see the percentage of concordant and discordant pairs in this section of the output. (A *concordant pair* means the predicted outcome from the model matches the observed outcome from the data. A *discordant pair* is one that doesn't match.) ***Note:*** To see the concordant/discordant pairs, make sure the Measures of Association option is selected under the Results section of the original binary logistic regression screen.

In the case of the movie and age data, the model-building part of the Minitab output is shown in Figure 9-2. The model-fitting part of the Minitab output from the logistic regression analysis is in Figure 9-4.

In the following sections, you see how to use this output to build the best-fitting logistic regression model for your data and to check the model's fit.



```
Logistic Regression Table

                                                        Odds      95% CI
Predictor       Coef     SE Coef       Z        P     Ratio   Lower   Upper
Constant    4.86539     1.43434     3.39    0.001
Age       -0.175745   0.0499620    -3.52    0.000      0.84    0.76    0.93
```

FIGURE 9-2: The model-building part of the movie and age data's logistic regression output.

Finding the coefficients and making the model

After you have Minitab run a logistic regression analysis on your data, you can find the coefficients b_0 and b_1 and put them together to form the best-fitting logistic regression model for your data.

Figure 9-2 shows part of the Minitab output for the movie enjoyment and age data. (I discuss the remaining output in the section, "Checking the fit of the model.") The first column of numbers is labeled Coef, which stands for the coefficients in the model. The first coefficient, b_0, is labeled Constant. The second coefficient is in the row labeled by your explanatory variable, x. (In the movie and age data, the explanatory variable is age. This age coefficient represents the value of b_1 in the model.)

According to the Minitab output in Figure 9-2, the value of b_0 is 4.87 and the value of b_1 is −0.18. After you've determined the coefficients b_0 and b_1 from the Minitab output to find the best-fitting S-curve for your data, you put these two coefficients into the general logistic regression model: $\hat{p} = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$.

For the movie and age data, you get $\hat{p} = \frac{e^{4.87 - 0.18x}}{1 + e^{4.87 - 0.18x}}$, which is the best-fitting logistic regression model for this data set.

The graph of the best-fitting logistic regression model for the movie and age data is shown in Figure 9-3. Note that the graph is a downward-sloping S-curve because higher probabilities of liking the movie are affiliated with lower ages and lower probabilities are affiliated with higher ages.

The movie marketers now have the answer to their question. This movie has a higher chance of being well liked by kids (and the younger, the better) and a lower chance of being well liked by adults (and the older they are, the lower the chance of them liking the movie).

The point where the probability changes from high to low (that is, at the $\hat{p} = 0.50$ mark) is between ages 25 and 30. That means that the tide of probability of liking the movie appears to turn from higher to lower in that age range. Using calculus terms, this point is called the *saddle point* of the S-curve, which is the point where the graph changes from concave up to concave down, or vice versa.

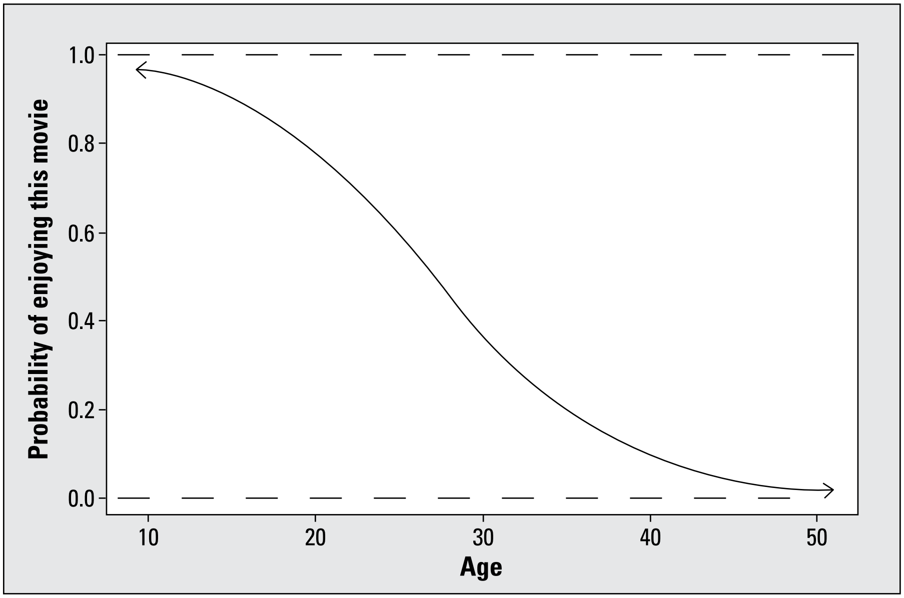


FIGURE 9-3: The best-fitting S-curve for the movie and age data.

Estimating p

You've determined the best-fitting logistic regression model for your data, obtained the values of b_0 and b_1 from the logistic regression analysis, and know the precise S-curve that fits your data best (check out the previous sections). You're now ready to estimate *p* and make predictions about the probability that the event of interest will happen, given the value of the explanatory variable *x*.

To estimate *p* for a particular value of *x*, plug that value of *x* into your equation (the best-fitting logistic regression model) and simplify it by using your algebra skills. The number you get is the estimated chance of the event occurring for that value of *x*, and it should be a number between 0 and 1, being a probability and all.

Continuing with the movie and age example from the preceding sections, suppose you want to predict whether a 15-year-old would enjoy the movie. To estimate *p*, plug 15 in for *x* in the logistic regression model $\hat{p} = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$ to get $\hat{p} = \frac{e^{4.87 - 0.18 * 15}}{1 + e^{4.87 - 0.18 * 15}} = \frac{e^{2.17}}{1 + e^{2.17}} = \frac{8.76}{9.76} = 0.90$.

That answer means you estimate there's a 90 percent chance that a 15-year-old will like the movie. You can see in Figure 9-3 that when *x* is 15, *p* is approximately 0.90. On the other hand, if the person is 50 years old, the chance they'll like this movie is $\hat{p} = \frac{e^{4.87 - 0.18 * 50}}{1 + e^{4.87 - 0.18 * 50}}$, or 0.02 (shown in Figure 9-3 for $x = 50$), which is only a 2 percent chance.

Checking the fit of the model

REMEMBER

The results you get from a logistic regression analysis, as with any other data analysis, are all subject to the model fitting appropriately.

To determine whether or not your logistic regression model fits, follow these steps (which I cover in more detail later in this section):

1. **Locate the *p*-value of the goodness-of-fit test (found in the Goodness-of-Fit portion of the computer output; see Figure 9-4 for an example).** If the *p*-value is larger than 0.05, you conclude that your model fits, and if the *p*-value is less than 0.05, you conclude that your model doesn't fit.
2. **Find the *p*-value for the b_1 coefficient (it's listed under *P* in the row for your column one (explanatory) variable in the model-building portion of the output; see Figure 9-2 for an example).** If the *p*-value is less than 0.05, the *x* variable is statistically significant in the model, so it should be included. If the *p*-value is greater than or equal to 0.05, the *x* variable isn't statistically significant and shouldn't be included in the model.
3. **Look later in the output at the percentage of concordant pairs.** This percentage reflects the proportion of time that the data and the model actually agree with each other. The higher the percentage, the better the model fits.

REMEMBER

The conclusion in Step 1 based on the *p*-value may seem backward to you, but here's what's happening: Chi-square goodness-of-fit tests measure the overall difference between what you expect to see via your model and what you actually observe in your data. (Chapter 16 gives you the lowdown on Chi-square tests.) The null hypothesis (H_0) for this test says you have a difference of zero between what you observed and what you expected from the model; that is, your model fits. The alternative hypothesis, denoted H_a, says that the model doesn't fit. If you get a small *p*-value (under 0.05), you reject H_0 and conclude the model doesn't fit. If you get a larger *p*-value (above 0.05), you can stay with your model.

WARNING

Failure to reject H_0 here (having a large *p*-value) only means that you can't say your model doesn't fit the population from which the sample came. It doesn't necessarily mean the model fits perfectly. Your data could be unrepresentative of the population just by chance.

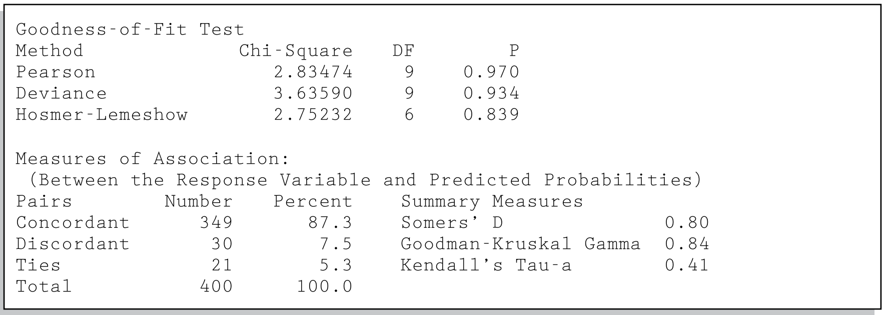
Fitting the movie model

You're ready to check out the fit of the movie data to make sure you still have a job when the box office totals come in.

Step one: p-value for Chi-squared

Using Figure 9-4 to complete the first step of checking the model's fit, you can see many different goodness-of-fit tests. The particulars of each of these tests are beyond the scope of this book; however, in this case (as with most cases), each test has only slightly different numerical results and the same conclusions.

All the *p*-values in column four of Figure 9-4 are over 0.80, which is much higher than the 0.05 you need to reject the model. After looking at the *p*-values, you can say that the model using age to predict movie likeability appears to fit this data.



```
Goodness-of-Fit Test
Method               Chi-Square   DF        P
Pearson                 2.83474    9    0.970
Deviance                3.63590    9    0.934
Hosmer-Lemeshow         2.75232    6    0.839

Measures of Association:
 (Between the Response Variable and Predicted Probabilities)
Pairs        Number    Percent    Summary Measures
Concordant      349       87.3    Somers' D                0.80
Discordant       30        7.5    Goodman-Kruskal Gamma    0.84
Ties             21        5.3    Kendall's Tau-a          0.41
Total           400      100.0
```

FIGURE 9-4: The model-fitting part of the movie and age data's logistic regression output.

Step two: p-value for the x variable

For step two, you look at the significance of the *x* variable age. Back in Figure 9-2, you can see the constant for age, −0.18, and farther along in its row, you can see that the *Z*-value is −3.52; this *Z*-value is the test statistic for testing $H_o: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$. The *p*-value is listed as 0.000, which means it's smaller than 0.001 (a highly significant number). So you know that the coefficient in front of *x*, also known as β_1, is statistically significant (not equal to zero), and you should include *x* (age) in the model.

Step three: Concordant pairs

To complete step three of the fit-checking process, look at the percentage of concordant pairs reported in Figure 9-4. This value shows the percentage of times the data actually agreed with the model (87.3). To determine concordance, the computer makes predictions as to whether the event should have occurred for each individual based on the model and compares those results to what actually happened.

TIP

The logistic regression model is for *p*, the probability of the event occurring, so if *p* is estimated to be > 0.50 for some value of *x*, the computer predicts that the event will occur (versus not occurring). If the estimated value of *p* is < 0.50 for a particular *x*-value, the computer predicts that it won't occur.

For the movie and age data, the percentage of concordant pairs (that is, the percentage of times the model made the right decision in predicting what would happen) is 87.3 percent, which is quite high.

REMEMBER

The percentage of concordant pairs was obtained by taking the number of concordant pairs and dividing by the total number of pairs. I'd start getting excited if the percentage of concordant pairs got over 75 percent; the higher, the better.

Figure 9-5 shows the logistic regression model for the movie and age data, with the actual values of the observed data added as circles. The S-curve shows the probability of liking the movie for each age level, and the computer will predict "1" = they will like the movie, if $\hat{p} > 0.50$. Circles indicate whether the people of those age levels actually liked the movie $(y = 1)$ or not $(y = 0)$.

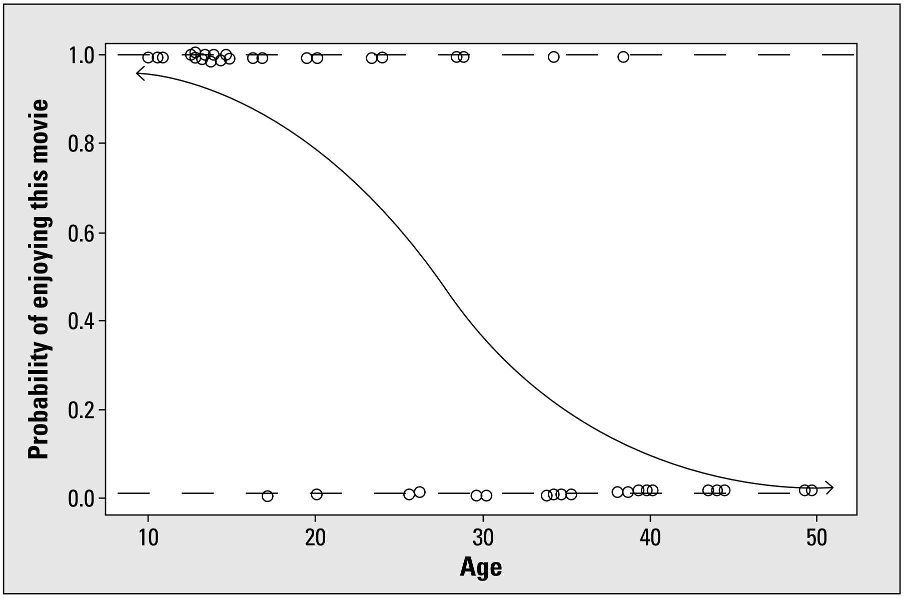


FIGURE 9-5: Actual observed values (0 and 1) compared to the model.

Much of the time, the model made the right decision; probabilities above 0.50 are associated with more circles at the value of 1, and probabilities below 0.50 are associated with more circles at the value of 0 (zero). It's the outcomes that have p near 0.50 that are hard to predict because the results can go either way.

All this evidence helps confirm that your model fits your data well. You can go ahead and make predictions based on this model for the next individual that comes up, whose outcome you don't know (see the section, "Estimating p," earlier in this chapter).

WHICH METHOD TO USE TO COMPARE? SORTING OUT SIMILAR SITUATIONS

Data come in a variety of forms, and each form has its own analysis to use to make comparisons. It can be difficult to decide which type of analysis to use when.

It may help to sort out some situations that sound similar but have subtle differences that lead to very different analyses. You can use the following list to compare these subtle, but important, differences:

- **If you want to compare three or more groups of numerical variables,** use ANOVA (see Chapter 11). For only two groups, use a *t*-test (see Chapters 4 and 10).
- **If you want to estimate one numerical variable based on another,** use simple linear regression (see Chapter 5).
- **If you want to estimate one numerical variable using many other numerical variables,** use multiple regression (see Chapter 6).
- **If you want to estimate a categorical variable with two categories by using a numerical variable,** use logistic regression, which is the focus of this chapter, of course.
- **If you want to compare two categorical variables to each other,** head straight for a Chi-square test (see Chapter 15).

3

Analyzing Variance with ANOVA

IN THIS PART . . .

Get the basics of analysis of variance (ANOVA).

Compare the means using multiple comparisons.

Work with two-way ANOVA.

Relate regression and ANOVA.

IN THIS CHAPTER

- Extending the *t*-test for comparing two means by using ANOVA
- Utilizing the ANOVA process
- Carrying out an *F*-test
- Navigating the ANOVA table

Chapter 10

Testing Lots of Means? Come On Over to ANOVA!

One of the most commonly used statistical techniques at the Stats II level is *analysis of variance* (affectionately known as ANOVA). Because the name has the word *variance* in it, you may think that this technique has something to do with variance — and you would be right. Analysis of variance is all about examining the amount of variability in a *y* (response) variable and trying to understand where that variability is coming from.

One way you can use ANOVA is to compare several populations regarding some quantitative variable, *y*. The populations you want to compare constitute different groups (denoted by an *x* variable), such as political affiliations, age groups, or different brands of a product. ANOVA is also particularly suitable for situations involving an experiment in which you apply certain treatments (*x*) to subjects and measure a response (*y*).

In this chapter, you start with the *t*-test for two population means, the precursor to ANOVA. Then you move on to the basic concepts of ANOVA to compare more than two means: sums of squares, the *F*-test, and the ANOVA table. You apply these basics to the *one-factor* or *one-way ANOVA*, where you compare the responses based only on one treatment variable. (In Chapter 12, you can see the basics applied to a two-way ANOVA, which has two treatment variables.)

Comparing Two Means with a t-Test

The *two-sample t-test* is designed to test whether two population means are different. The conditions for the two-sample *t*-test are as follows:

- The two populations are independent. In other words, their outcomes don't affect each other. (And the observations within each sample are independent as well.)
- The response variable (*y*) is a quantitative variable, meaning its values have numerical meaning and represent quantities of some kind.
- The *y*-values for each population have a normal distribution. However, their means may be different; that's what the *t*-test determines.
- The variances of the two normal distributions are equal.

TIP

For large sample sizes when you know the variances, you use a *Z*-test for the two population means. However, a *t*-test allows you to test two population means when the variances are unknown or the sample sizes are small. This occurs quite often in situations where an experiment is performed and the number of subjects is limited. (See your Stats I text or *Statistics For Dummies*, 2nd Edition [Wiley] for information on the *Z*-test.)

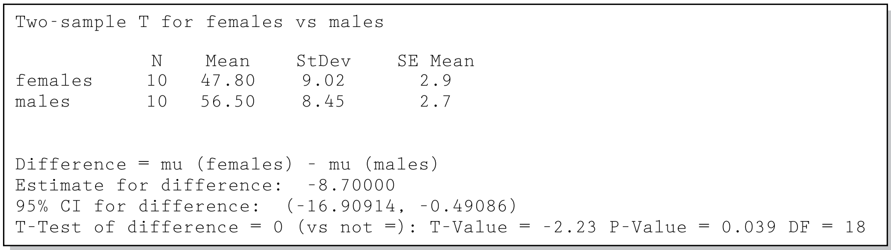
Although you've seen *t*-tests before in your Stats I class, it may be good to review the main ideas. The *t*-test tests the hypotheses H_o: $\mu_1 = \mu_2$ versus H_a: μ_1 is <, >, or $\neq \mu_2$, where the situation dictates which of these hypotheses you use. (***Note:*** With ANOVA, you extend this idea to *k* different means from *k* different populations, and the only version of H_a of interest is ≠.)

REMEMBER

To conduct the two-sample *t*-test, you collect two data sets from the two populations, using two independent samples. To form the test statistic (the *t*-statistic), you subtract the two sample means and divide by the *standard error* (a combination of the two standard deviations from the two samples and their sample sizes). You compare the *t*-statistic to the *t*-distribution with $n_1 + n_2 - 2$ degrees of freedom and find the *p*-value. If the *p*-value is less than the predetermined α level, say 0.05, you have enough evidence to say the population means are different. (For information on hypothesis tests, see Chapter 4.)

For example, suppose you're at a watermelon seed-spitting contest where contestants each put watermelon seeds in their mouths and spit them as far as they can. Results are measured in inches and are treated with the reverence of the shot-put results at the Olympics. You want to compare the watermelon seed-spitting distances of female and male adults. Your data set includes ten people from each group.

You can see the results of the *t*-test in Figure 10-1. The mean spitting distance for females was 47.8 inches; the mean for males was 56.5 inches; and the difference (females – males) is −8.71 inches, meaning the females in the sample spit seeds at shorter distances, on average, than the males. The *t*-statistic for the difference in the two means (females – males) is $t = -2.23$, which has a *p*-value of 0.039 (see the last line of the output in Figure 10-1). At a level of $\alpha = 0.05$, this difference is significant (because $0.039 < 0.05$). You conclude that males and females differ with respect to their mean watermelon seed-spitting distance. And you can say males are likely spitting farther because their sample mean was higher.



```
Two-sample T for females vs males

            N    Mean    StDev    SE Mean
females    10   47.80    9.02       2.9
males      10   56.50    8.45       2.7


Difference = mu (females) - mu (males)
Estimate for difference:  -8.70000
95% CI for difference:  (-16.90914, -0.49086)
T-Test of difference = 0 (vs not =): T-Value = -2.23 P-Value = 0.039 DF = 18
```

FIGURE 10-1: A *t*-test comparing mean watermelon seed-spitting distances for females versus males.

Evaluating More Means with ANOVA

When you can compare two independent populations inside and out, at some point two populations will not be enough. Suppose you want to compare more than two populations regarding some response variable (*y*). This idea kicks the *t*-test up a notch into the territory of ANOVA. The ANOVA procedure is built around a hypothesis test called the *F*-test, which compares how much the groups differ from each other compared to how much variability is within each group. In this section, I set up an example of when to use ANOVA and show you the steps involved in the ANOVA process. You can then apply the ANOVA steps to the following example throughout the rest of the chapter.

Spitting seeds: A situation just waiting for ANOVA

Before you can jump into using ANOVA, you must figure out what question you want answered and collect the necessary data.

Suppose you want to compare the watermelon seed-spitting distances for four different age groups: 6–8 years old, 9–11, 12–14, and 15–17. The hypotheses for this example are H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ versus H_a: At least two of these means are different, where the population means μ represent those from the age groups, respectively.

Over the years that this contest was held, you've collected data on more than 200 children from each age group, so you have some prior ideas about what the distances typically look like. This year, you have 20 entrants, 5 in each age group. You can see the data from this year, in inches, in Table 10-1.

TABLE 10-1 Watermelon Seed-Spitting Distances for Four Age Groups of Children (Measured in Inches)

6–8 Years	9–11 Years	12–14 Years	15–17 Years
38	38	44	44
39	39	43	47
42	40	40	45
40	44	44	45
41	43	45	46

Do you see a difference in distances for these age groups based on this data? If you were to just combine all the data, you would see quite a bit of difference (the range of the combined data goes from 38 inches to 47 inches). And you may suspect that older kids can spit farther.

Perhaps accounting for which age group each contestant is in does explain at least some of what's going on. But don't stop there. The next section walks you through the official steps you need to perform to answer your question.

Walking through the steps of ANOVA

You've decided on the quantitative response variable (y) you want to compare for your k various population (or treatment) means, and you've collected a random sample of data from each population (refer to the preceding section). Now you're ready to conduct ANOVA on your data to see whether the population means are different for your response variable, y.

The characteristic that distinguishes these populations is called the *treatment variable*, x. Statisticians use the word *treatment* in this context because one of the biggest uses of ANOVA is for designed experiments where subjects are randomly assigned to treatments, and the responses are compared for the various treatment groups. So statisticians often use the word *treatment* even when the study isn't an experiment and they're comparing regular populations. Hey, don't blame me! I'm just following the proper statistical terminology.

Here are the general steps in a one-way ANOVA:

1. **Check the ANOVA conditions, using the data collected from each of the *k* populations.**

   See the next section, "Checking the Conditions," for the specifics on these conditions.

2. **Set up the hypotheses H_o:** $\mu_1 = \mu_2 = \ldots = \mu_k$ versus H_a: At least two of the population means are different.

   Another way to state your alternative hypothesis is by saying H_a: At least two of $\mu_1, \mu_2, \ldots \mu_k$ are different.

3. **Collect data from *k* random samples, one from each population.**

4. **Conduct an *F*-test on the data from Step 3, using the hypotheses from Step 2, and find the *p*-value.**

   See the section, "Doing the *F*-Test," later in this chapter for these instructions.

5. **Make your conclusions.**

   If you reject H_o (when your *p*-value is less than 0.05 or your predetermined α level), you conclude that at least two of the population means are different; otherwise, you conclude that you didn't have enough evidence to reject H_o (you can't say the means are different).

If these steps seem like a foreign language to you, don't worry — I describe each in detail in the sections that follow.

Checking the Conditions

Step 1 of ANOVA involves checking to be sure all necessary conditions are met before diving into the data analysis. The conditions for using ANOVA are just an extension of the conditions for a *t*-test (see the section, "Comparing Two Means with a *t*-Test"). The following conditions all need to hold in order to conduct ANOVA:

- The *k* populations are independent. In other words, their outcomes don't affect each other. (And the observations within each sample are independent also.)
- The *k* populations each have a normal distribution.
- The variances of the *k* normal distributions are equal.

Verifying independence

To check the first condition, examine how the data were collected from each of the separate populations. In order to maintain independence, the outcomes from one population can't affect the outcomes of the other populations. If the data have been collected by using a separate random sample from each population (*random* here meaning that each individual in the population had an equal chance of being selected), this factor ensures independence at the strongest level.

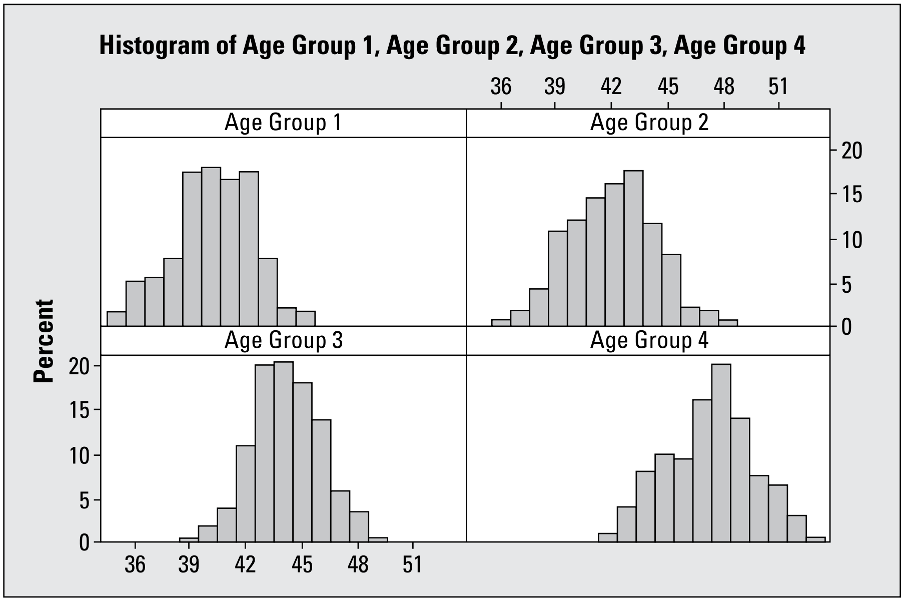
In the watermelon seed-spitting data (see Table 10-1), the data aren't randomly sampled from each age group because the data represent everyone who participated in the contest. But, you can argue that in most cases, the seed-spitting distances from one age group don't affect the seed-spitting distances from the other age groups, so the independence assumption is relatively okay.

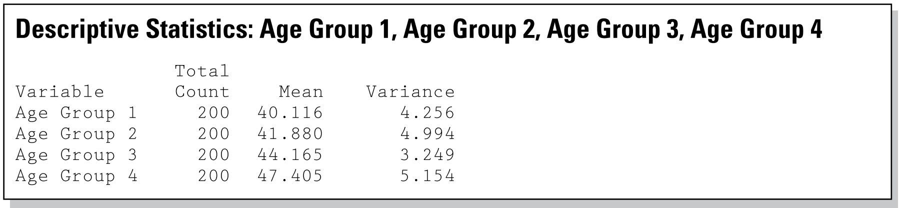
Looking for what's normal

The second ANOVA condition is that each of the *k* populations has a normal distribution. To check this condition, make a separate histogram of the data from each group and see whether it resembles a normal distribution. Data from a normal distribution should look symmetric (in other words, if you split the histogram down the middle, it looks the same on each side) and have a bell shape. Don't expect the data in each histogram to follow a normal distribution exactly (remember, it's only a sample), but it shouldn't be extremely different from a normal, bell-shaped distribution.

Because the seed-spitting data contain only five children per age group, checking conditions can be iffy. But in this case, you have past years' data for 200 children in each age group, so you can use that to check the conditions. The histograms and descriptive statistics of the seed-spitting data for the four age groups are shown in Figure 10-2, all in one panel, so you can easily compare them to each other on the same scale.

Looking at the four histograms in Figure 10-2, you can see that each graph resembles a bell shape; the normality condition isn't being severely violated here. (Red flags should come up if you see two peaks in the data, a skewed shape where the peak is off to one side, or a flat histogram, for example.)





```
Descriptive Statistics: Age Group 1, Age Group 2, Age Group 3, Age Group 4

                Total
Variable        Count     Mean    Variance
Age Group 1       200   40.116       4.256
Age Group 2       200   41.880       4.994
Age Group 3       200   44.165       3.249
Age Group 4       200   47.405       5.154
```

FIGURE 10-2: Checking ANOVA conditions by using histograms and descriptive statistics.

COMPUTER OUTPUT

You can use Minitab to make histograms for each of your samples and have all of them appear on one large panel, all using the same scale. To do this, go to Graph>Histogram and click OK. Choose the variables that represent data from each sample by highlighting them in the left-hand box and clicking Select.

Then click on Multiple Graphs, and a new window opens. Under the Show Graph Variables option, check the box that says, "In separate panels of the same graph." On the Same Scales for Graphs option, check the box for *x* and the box for *y*. This option gives you the same scale on both the *x* and *y* axes for all the histograms. Then click OK. The descriptive statistics come from a separate analysis (Stat>Basic Statistics>Display Descriptive Statistics).

Taking note of spread

The third condition for ANOVA is that the variance in each of the *k* populations is the same; statisticians call this the *equal variance condition.* You have two ways to check this condition on your data:

- Calculate each of the variances from each sample and see how they compare.
- Create one graph showing all the boxplots of each sample sitting side by side. This type of graph is called a *side-by-side boxplot.* (See your Stats I text or my book, *Statistics For Dummies,* 2nd Edition [Wiley], for more information on boxplots.)

If one or more of the calculated variances is significantly different from the others, the equal variance condition is not likely to be met. What does *significantly different* mean? A hypothesis test for equal variances is the statistical tool used to handle this question; however, it falls outside the scope of most Stats II courses, so for now you can make a judgment call. I always say that if the differences in the calculated variances are enough for you to write home about (say, they differ by 10 percent or more), then the equal variance condition is likely not to be met.

Similarly, if the lengths of one or more of the side-by-side boxplots look different enough for you to write home about, the equal variance condition is not likely to be met. (But listen, if you really do write home about any of your statistical issues, you may want to spice up your life a bit.)

TECHNICAL STUFF

The length of the box portion of a boxplot is called the *interquartile range.* You calculate it by taking the third quartile (the 75th percentile) minus the first quartile (the 25th percentile.) See your Stats I text or *Statistics For Dummies,* 2nd Edition for more information.

Table 10-2 shows an example of four small data sets, with each of their calculated variances shown in the last row. Note that the variance of Data Set 4 is significantly smaller than for the others. In this case, it's safe to say that the equal variance condition is not met.

TABLE 10-2 **Comparing Variances of Four Data Sets to Check the Equal Variance Condition**

Data Set 1	Data Set 2	Data Set 3	Data Set 4
1	32	4	3
2	24	3	4
3	27	5	5
4	32	10	5
5	31	7	6
6	28	4	6
7	30	8	7
8	26	12	7
9	31	9	8
10	24	10	9
Variance = 9.167	Variance = 9.833	Variance = 9.511	Variance = 3.333

Figure 10-3 shows the side-by-side boxplots for these same four data sets. You see that the boxplot for Data Set 4 has an interquartile range (length of the box) that's significantly smaller than the others. I calculated the actual interquartile ranges for these four data sets; they're 5.50, 5.75, 6.00, and 2.50, respectively. These findings confirm the conclusion that the equal variance condition is not met, due to group 4's much smaller variability.

COMPUTER OUTPUT

To find descriptive statistics (including the variance and interquartile range) for each sample, go to Stat>Basic Statistics>Display Descriptive Statistics. Click on each variable in the left-hand box for which you want the descriptive statistics, and click Select. Click on the Statistics option; a new window appears with tons of different types of statistics. Click on the ones you want and click off the ones you don't want. Click OK. Then click OK again. Your descriptive statistics are calculated.

To find side-by-side boxplots in Minitab, go to Graph>Boxplot. A window appears. Click on the picture for Multiple Y's, Simple, and then click OK. Highlight the variables from the left-hand side that you want to compare, and click Select. Then click OK.

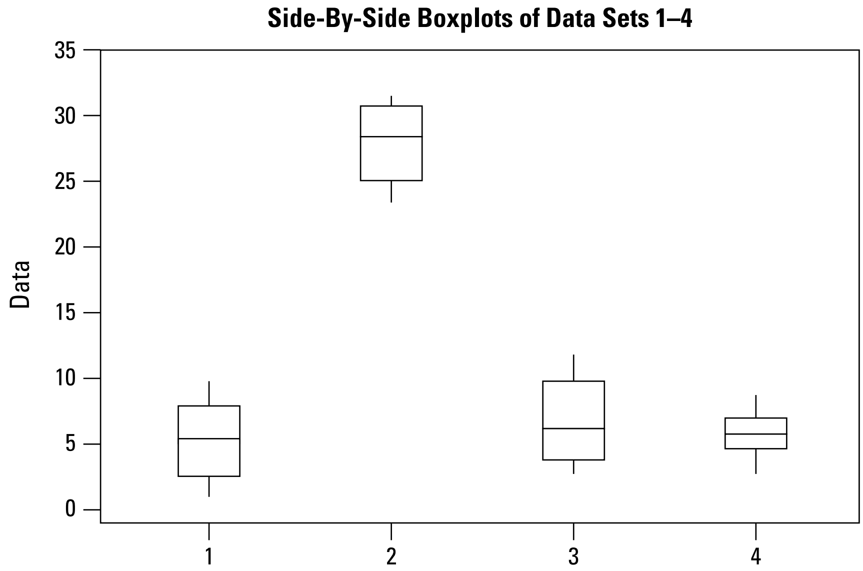


FIGURE 10-3: Examining side-by-side boxplots to check the equal variance condition.

REMEMBER

Note that you don't need the sample sizes in each group to be equal to carry out ANOVA; however, in Stats II, you'll typically see what statisticians call a *balanced design,* where each sample from each population has the same sample size. (As I explain in Chapter 4, for more precision in your data, the larger the sample sizes, the better.)

For the seed-spitting data, the variances for each age group are listed in Figure 10-2. These variances are close enough to say the equal variance condition is met.

Setting Up the Hypotheses

Step 2 of ANOVA involves setting up the hypotheses to be tested. You're testing to see if all the population means can be deemed equal to each other. The null hypothesis for ANOVA is that all the population means are equal. That is, H_o: $\mu_1 = \mu_2 = \ldots = \mu_k$, where μ_1 is the mean of the first population, μ_2 is the mean of the second population, and so on until you reach μ_k (the mean of the k^{th} population).

TIP

What appears in the alternative hypothesis (H_a) must be the opposite of what's in the null hypothesis (H_o). What's the opposite of having all *k* of the population's means equal to each other? You may think the opposite is that they're all different. But that's not the case. In order to blow H_o wide open, all you need is for at least two of those means to not be equal. So, the alternative hypothesis, H_a, is that at least two of the population means are different from each other. That is, H_a: At least two of μ_1, μ_2, . . . μ_k are different.

Note that H_o and H_a for ANOVA are an extension of the hypotheses for a two-sample *t*-test (which only compares two independent populations). And even though the alternative hypothesis in a *t*-test may be that one mean is greater than, less than, or not equal to the other, you don't consider any alternative other than ≠ in ANOVA. (Statisticians use more in-depth models for the others. Aren't you glad someone else is doing it?)

REMEMBER

You only want to know whether or not the means are equal — at this stage of the game, anyway. After you reach the conclusion that H_o is rejected in ANOVA, you can proceed to figure out how the means are different, which ones are bigger than others, and so on, using multiple comparisons. Those details appear in Chapter 11.

Doing the F-Test

Step 3 of ANOVA involves collecting the data, and it includes taking *k* random samples, one from each population. Step 4 of ANOVA involves doing the *F*-test on this data, which is the heart of the ANOVA procedure. This test is the actual hypothesis test of H_o: $\mu_1 = \mu_2 = \ldots = \mu_k$ versus H_a: At least two of μ_1, μ_2, . . . μ_k are different.

You have to carry out three major steps in order to complete the *F*-test (***Note:*** Don't get these steps confused with the main ANOVA steps; consider the *F*-test a few steps within a step):

1. **Break down the variance of *y* into different sums of squares.**
2. **Find the mean sums of squares.**
3. **Put the mean sums of squares together to form the *F*-statistic.**

I describe each step of the *F*-test in detail and apply it to the example of comparing watermelon seed-spitting distances (see Table 10-1) in the following sections.

TIP

Data analysts rely heavily on computer software to conduct each step of the *F*-test, and you can do the same. All computer software packages organize and summarize the important information from the *F*-test into a table format for you.

This table of results for ANOVA is called (what else?) the *ANOVA table.* Because the ANOVA table is a critical part of the entire ANOVA process, I start the following sections out by describing how to run ANOVA in Minitab to get the ANOVA table, and I continue to reference this section as I describe each step of the ANOVA process.

Running ANOVA in Minitab

COMPUTER OUTPUT

In using Minitab to run ANOVA, you first have to enter the data from the *k* samples. You can enter the data in one of two ways:

- **As stacked data.** You enter all the data into two columns. Column one includes the number indicating what sample the data value is from (1 to *k*), and column two includes the responses (*y*). To analyze this data, go to Stat>ANOVA>One-Way. In the pull-down menu, choose the option, "Response data are in one column for all factor levels." Highlight the response (*y*) variable, and click Select. Highlight the factor (population) variable, and click Select. Then click OK.
- **As unstacked data.** You enter (only) the *y*-values from the response (*y*) data into separate columns. That is, you include the *y*-values from sample 1 in column one, the *y*-values from sample 2 in column two, and so on. You will have *k* columns of data. To analyze the data entered this way, go to Stat>ANOVA>One-Way. In the pull-down menu, choose the option, "Response data are in a separate column for each factor level." Highlight the names of the columns where your response data are located. Then click OK.

I typically use the unstacked version of data entry just because I think it helps visualize the data. However, the choice is up to you, and the results come out the same no matter which method you choose, as long as you're consistent.

Breaking down the variance into sums of squares

Step 1 of the *F*-test involves splitting up the variability in the *y* variable into portions that define where the variability is coming from. Each portion of variability is called a *sum of squares.* The term *analysis of variance* is a great description for exactly how you conduct a test of *k* population means. With the overall goal of testing whether *k* population (or treatment) means are equal, you take a random sample from each of the *k* populations.

You first put all the data together into one big group and measure how much total variability there is; this variability is called the *sums of squares total*, or SSTO. If the data are really diverse, SSTO is large. If the data are very similar, SSTO is small.

You can split the total variability in the combined data set (SSTO) into two parts:

- **SST.** The variability between the groups, known as the *sums of squares for treatment.*
- **SSE.** The variability within the groups, known as the *sums of squares for error.*

Splitting up the variability in your data results in one of the most important equalities in ANOVA:

$$\text{SSTO} = \text{SST} + \text{SSE}$$

REMEMBER

The formula for SSTO is the numerator of the formula for s^2, the variance of a single data set, so $\text{SSTO} = \sum_i \sum_j \left(x_{ij} - \bar{x}\right)^2$, where *i* and *j* represent the *j*th value in the sample from the *i*th population and $\bar{x}$ is the *overall sample mean* (the mean of the entire data set). So, in terms of ANOVA, SSTO is the total squared distance between the data values and their overall mean.

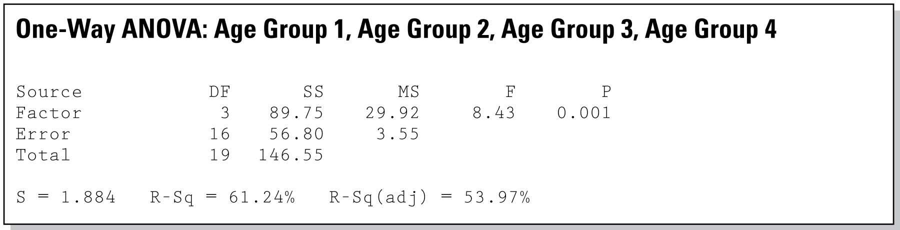
The formula for SST is $\sum_i n_i \left(\bar{x}_i - \bar{x}\right)^2$, where n_i is the size of the sample coming from the *i*th population and $\bar{x}$ is the overall sample mean. SST represents the total squared distance between the means from each sample and the overall mean.

The formula for SSE is $\sum_i \sum_j \left(x_{ij} - \bar{x}_i\right)^2$, where x_{ij} is the *j*th value in the sample from the *i*th population and $\bar{x}_i$ is the mean of the sample coming from the *i*th population. This formula represents the total squared distance between the values in each sample and their corresponding sample means. Using algebra, you can confirm (with some serious elbow grease) that $\text{SSTO} = \text{SST} + \text{SSE}$.

COMPUTER OUTPUT

The Minitab output for the watermelon seed-spitting contest for the four age groups is shown in Figure 10-4. Under the Source column of the ANOVA table, you see Factor listed in row one. The factor variable (as described by Minitab) represents the treatment or population variable. In column three of the Factor row, you see the SST, which is equal to 89.75. In the Error row (row two), you locate the SSE in column three, which equals 56.80. In column three of the Total row (row three), you see the SSTO, which is 146.55. Using the values of SST, SSE, and SSTO from the Minitab output, you can verify that $\text{SST} + \text{SSE} = \text{SSTO}$.

Now you're ready to use these sums of squares to complete the next step of the *F*-test.



One-Way ANOVA: Age Group 1, Age Group 2, Age Group 3, Age Group 4

```
Source          DF      SS      MS      F       P
Factor           3   89.75   29.92   8.43   0.001
Error           16   56.80    3.55
Total           19  146.55

S = 1.884   R-Sq = 61.24%   R-Sq(adj) = 53.97%
```

FIGURE 10-4: ANOVA Minitab output for the watermelon seed-spitting example.

Locating those mean sums of squares

After you have the sums of squares for treatment, SST, and the sums of squares for error, SSE (see the preceding section for more on these), you want to compare them to see whether the variability in the y-values due to the model (SST) is large compared to the amount of error left over in the data after the groups have been accounted for (SSE). So you ultimately want a ratio that somehow compares SST to SSE.

To make this ratio form a statistic that they know how to work with (in this case, an F-statistic), statisticians decided to find the means of SST and SSE and work with that. Finding the mean sums of squares is the second step of the F-test, and the mean sums are as follows:

- **MST** is the *mean sums of squares for treatments,* which measures the mean variability that occurs between the different treatments (the different samples in the data). What you're looking for is the amount of variability in the data as you move from one sample to another. A great deal of variability between samples (treatments) may indicate that the populations are different as well.

  You can find MST by taking SST and dividing by $k-1$ (where k is the number of treatments).

- **MSE** is the *mean sums of squares for error,* which measures the mean within-treatment variability. The *within-treatment variability* is the amount of variability that you see within each sample itself, due to chance and/or other factors not included in the model.

  You can find MSE by taking SSE and dividing by $n-k$ (where n is the total sample size and k is the number of treatments). The values of $k-1$ and $n-k$ are called the *degrees of freedom* (or DF) for SST and SSE, respectively.

COMPUTER OUTPUT

Minitab calculates and posts the degrees of freedom for SST, SSE, MST, and MSE in the ANOVA table in columns two and four, respectively.

From the ANOVA table for the seed-spitting data in Figure 10-4, you can see that column two has the heading DF, which stands for degrees of freedom. You can find the degrees of freedom for SST in the Factor row (row two); this value is equal to $k-1=4-1=3$. The degrees of freedom for SSE is found to be $n-k=20-4=16$. (Remember, you have four age groups and five children in each group for a total of $n=20$ data values.) The degrees of freedom for SSTO is $n-1=20-1=19$ (found in the Total row under DF). You can verify that the degrees of freedom for $\text{SSTO} = \text{degrees of freedom for SST} + \text{degrees of freedom for SSE}$.

The values of MST and MSE are shown in column four of Figure 10-4, with the heading MS. You can see the MST in the Factor row, which is 29.92. This value was calculated by taking $\text{SST}=89.75$ and dividing it by degrees of freedom, 3. You can see MSE in the Error row, equal to 3.55. MSE is found by taking $\text{SSE}=56.80$ and dividing it by its degrees of freedom, 16.

By finding the mean sums of squares, you've completed step two of the F-test, but don't stop here! You need to continue to the next section in order to complete the process.

Figuring the F-statistic

The test statistic for the test of the equality of the k population means is $F=\frac{\text{MST}}{\text{MSE}}$. The result of this formula is called the *F-statistic.* The F-statistic has an F-distribution, which is equivalent to the square of a t-test (when the numerator degrees of freedom is 1; see more on this interesting connection between the t- and F-distributions in Chapter 12). All F-distributions start at zero and are skewed to the right. The degree of curvature and the height of the curvature of each F-distribution is reflected in two degrees of freedom, represented by $k-1$ and $n-k$. (These come from the denominators of MST and MSE, respectively, where n is the total sample size and k is the total number of treatments or populations.) A shorthand way of denoting the F-distribution for this test is $F_{(k-1,\, n-k)}$.

In the watermelon seed-spitting example, you're comparing four means and have a sample size of five from each population. Figure 10-5 shows the corresponding F-distribution, which has degrees of freedom $4-1=3$ and $20-4=16$; in other words $F_{(3,\, 16)}$.

COMPUTER OUTPUT

You can see the F-statistic on the Minitab ANOVA output (see Figure 10-4) in the Factor row, under the column indicated by F. For the seed-spitting example, the value of the F-statistic is 8.43. This number was found by taking $\text{MST}=29.92$ divided by $\text{MSE}=3.55$. Now locate 8.43 on the F-distribution in Figure 10-5 to see where it stands in terms of its p-value. (Turns out it's waaay out there; more on that in the next section.)

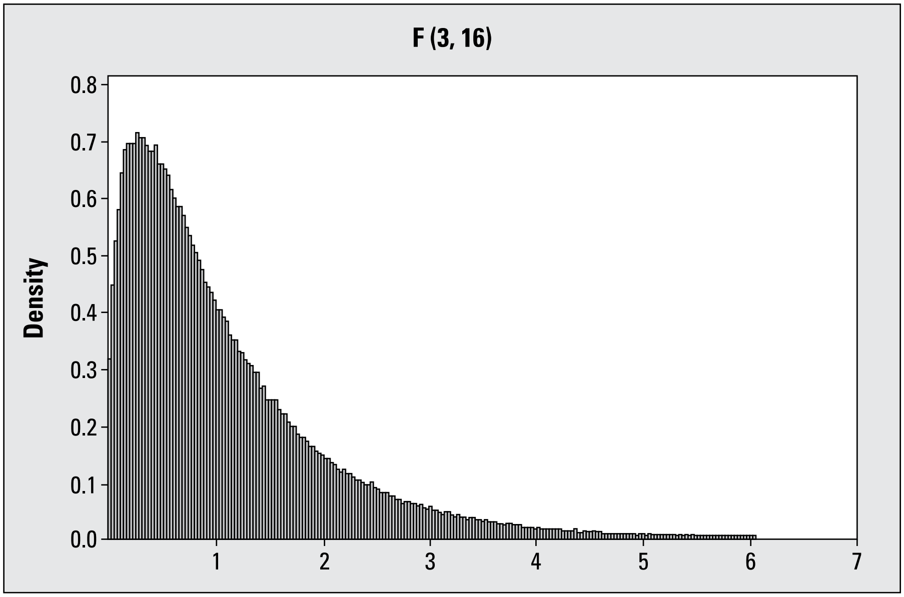


FIGURE 10-5: *F*-distribution with (3, 16) degrees of freedom.

WARNING

Be sure not to exchange the order of the degrees of freedom for the *F*-distribution. The difference between $F_{(3,\,16)}$ and $F_{(16,\,3)}$ is a big one.

Making conclusions from ANOVA

If you've completed the *F*-test and found your *F*-statistic (Step 4 in the ANOVA process), you're ready for Step 5 of ANOVA: making conclusions for your hypothesis test of the *k* population means. If you haven't already done so, you can compare the *F*-statistic to the corresponding *F*-distribution with $(k-1,\ n-k)$ degrees of freedom to see where it stands, and make a conclusion. You can make the conclusion in one of two ways: the *p*-value approach or the critical-value approach. The approach you use depends primarily on whether you have access to a computer, especially during exams. I describe these two approaches in the following sections.

Using the *p*-value approach

COMPUTER OUTPUT

On Minitab ANOVA output (see Figure 10-4), the value of the *F*-statistic is located in the Factor row, under the column noted by F. The associated *p*-value for the *F*-test is located in the Factor row under the column headed by P. The *p*-value tells you whether or not you can reject H_0.

- **If the *p*-value is less than your predetermined α (typically 0.05), reject H_0.** You conclude that the *k* population means aren't all equal and that at least two of them are different.

- **If the *p*-value is greater than α, then you can't reject H_o.** You don't have enough evidence in your data to say the *k* population means have any differences.

The *F*-statistic for comparing the mean watermelon seed-spitting distances for the four age groups is 8.43. The *p*-value as indicated in Figure 10-4 is 0.001. That means the results are highly statistically significant. You reject H_o and conclude that at least one pair of age groups differs in its mean watermelon seed-spitting distances. (You would hope that a 17-year-old could do a lot better than a 6-year-old, but maybe those 6-year-olds have a lot more spitting practice than 17-year-olds do.)

Using Figure 10-5, you see how the *F*-statistic of 8.43 stands on the *F*-distribution with $(4-1,\ 20-4)=(3,\ 16)$ degrees of freedom. You can see that it's way off to the right, out of sight. It makes sense that the *p*-value, which measures the probability of being beyond that *F*-statistic, is 0.001.

Using critical values

If you're in a situation where you don't have access to a computer (as is still the case in many statistics courses today when it comes to taking exams), finding the exact *p*-value for the *F*-statistic isn't possible using a table. You just choose the *p*-value of the *F*-statistic that's closest to yours. However, if you do have access to a computer while doing homework or an exam, statistical software packages automatically calculate all *p*-values exactly, so you can see them on any computer output.

REMEMBER

To approximate the *p*-value from your *F*-statistic in the event you don't have a computer or computer output available, you find a cutoff value on the *F*-distribution with $(k-1,\ n-k)$ degrees of freedom that draws a line in the sand between rejecting H_o and not rejecting H_o. This cutoff, also known as the *critical value,* is determined by your predetermined α (typically 0.05). You choose the critical value so that the area to its right on the *F*-distribution is equal to α.

F-distribution tables for other values of α are available in various statistics textbooks and on the web; however, $\alpha = 0.05$ is by far the most common α level used for the *F*-distribution and is sufficient for your purposes.

This table of values for the *F*-distribution is called the *F-table,* and students typically receive an *F*-table with their exams. For the seed-spitting example, the *F*-statistic has an *F*-distribution with degrees of freedom (3, 16), where $3 = k-1$, and $16 = n-k$. To find the critical value, consult an *F*-table (Table A-5 in the Appendix). Look up the degrees of freedom (3, 16), and you'll find that the critical value is 3.2389 (or 3.24). Your *F*-statistic for the seed-spitting example is 8.43, which is well beyond this critical value (you can see how 8.43 compares to 3.24 by looking at Figure 10-5). Your conclusion is to reject H_o at level α. At least two of the age groups differ on mean seed-spitting distances.

REMEMBER

With the critical value approach, any *F*-statistic that lies beyond the critical value results in rejecting H_o, no matter how far from or close to the line it is. If your *F*-statistic is beyond the value found in the *F*-table you consult, then you reject H_o and say at least two of the treatments (or populations) have different means.

What's next?

After you've rejected H_o in the *F*-test and concluded that not all the population means are the same, your next question may be, "Which ones are different?" You can answer that question by using a statistical technique called *multiple comparisons.* Statisticians use many different multiple comparison procedures to further explore the means themselves after the *F*-test has been rejected. I discuss and apply some of the more common multiple comparison techniques in Chapter 11.

Checking the Fit of the ANOVA Model

As with any other model, you must determine how well the ANOVA model fits before you can use its results with confidence. In the case of ANOVA, the model basically boils down to a treatment variable (also known as the population you're in) plus an error term. To assess how well that model fits the data, see the values of R^2 and R^2 adjusted on the last line of the ANOVA output below the ANOVA table. For the seed-spitting data, you see those values at the bottom of Figure 10-4.

- **The value of R^2 measures the percentage of the variability in the response variable (*y*) explained by the explanatory variable (*x*).** In the case of ANOVA, the *x* variable is the factor due to treatment (where the treatment can represent a population being compared). A high value of R^2 (say, above 80 percent) means this model fits well.
- **The value of R^2 adjusted, the preferred measure, takes R^2 and adjusts it for the number of variables in the model.** In the case of one-way ANOVA, you have only one variable, the factor due to treatment, so R^2 and R^2 adjusted won't be very far apart. For more on R^2 and R^2 adjusted, see Chapter 6.

For the watermelon seed-spitting data, the value of R^2 adjusted (as found in the last row of Figure 10-4) is only 53.97 percent. That means age group (shown to be statistically significant by the *F*-test; see the section, "Making conclusions from

ANOVA") explains just over half of the variability in the watermelon seed-spitting distances. Because of that connection, you may find other variables you can examine in addition to age group, making an even better model for predicting how far those seeds will go.

As you see in Figure 10-1, the results of the t-test done to compare the spitting distances of males and females in the section, "Comparing Two Means with a t-Test," show that males and females were significantly different on mean seed-spitting distances (p-value $= 0.039 < 0.05$). So I would venture a guess that if you include gender as well as age group, thereby creating what statisticians call a *two-factor ANOVA* (or *two-way ANOVA*), the resulting model would fit the data even better, resulting in higher values of R^2 and R^2 adjusted. (Chapter 12 walks you through the two-way ANOVA.)

UPFRONT REJECTION IS THE BEST POLICY FOR MOST REFUSAL LETTERS

Many medical and psychological studies use designed experiments to compare the responses of several different treatments, looking for differences. A *designed experiment* is a study in which subjects are randomly assigned to treatments (experimental conditions) and their responses are recorded. The results are used to compare treatments to see which one(s) work best, which ones work equally well, and so on.

Ohio State University researchers conducted one such experiment using ANOVA to determine the most effective way to write a rejection letter. (Is there really a best way to say "no" to someone? Turns out the answer is "yes.") The experiment tested three traditional principles of writing refusal letters:

- Using a buffer, which is a neutral or positive sentence that delays the negative information.
- Placing the reason before the refusal.
- Ending the letter on a positive note as a way of reselling the business.

Subjects were randomly assigned to treatments, and their responses to the rejection letters were compared (likely on some sort of scale such as 1 = very negative to 7 = very positive, with 4 being a neutral response).

You can analyze this scenario by using ANOVA because it compares three treatments (forms of the rejection letters) on some quantitative variable (response to the letter).

(continued)

(continued)

You can argue that response to the letter isn't a continuous variable; however, it has enough possible values that ANOVA isn't unreasonable. The data were also shown to have a bell shape.

The null hypothesis would be H_o: Mean responses to the three types of rejection letters are equal versus H_a: At least two forms of the rejection letter resulted in different mean responses.

In the end, the researchers did find some significant results: the different ways the rejection letter was written affected the participants differently (so the F-test was rejected). Using multiple comparison procedures (see Chapter 11), you could go in and determine which forms of the rejection letters gave different responses and how the responses differed.

In case you have to write a rejection letter at some point, the researchers recommend the following guidelines:

- Don't use buffers to begin negative messages.
- Give a reason for the refusal when it makes the sender's boss look good.
- Present the negative positively but clearly; offer an alternative or compromise if possible.
- A positive ending isn't necessary.

IN THIS CHAPTER

- » Knowing when and how to follow up ANOVA with multiple comparisons
- » Comparing two well-known multiple comparison procedures
- » Taking additional procedures into consideration

Chapter 11
Sorting Out the Means with Multiple Comparisons

Imagine this: You're comparing the means of not two, but *k* independent populations, and you find out (using ANOVA; see Chapter 10) that you reject H_o: All the population means are equal, and you conclude H_a: At least two of the population means are different. Now you gotta know — which of those populations are different? Answering this question requires a follow-up procedure to ANOVA called *multiple comparisons,* which makes sense because you want to compare the multiple means you have to see which ones are different.

In this chapter, you figure out when you need to use a multiple comparison procedure. Two of the most well-known multiple comparison procedures are Fisher's LSD (least significant difference) and Tukey's test. They can help you answer that burning question: So some of the means are different, but *which* ones? In this chapter, I also tell you about other comparison procedures that you may encounter or want to try.

Note: For those individuals who come up with new multiple comparison procedures, the procedures are generally named after them. (It's like having a star named after you, but less romantic and a whole lot more work.)

Following Up after ANOVA

The main reason folks use ANOVA to analyze data is to find out whether there are any differences in a group of population means. Your null hypothesis is that there are no differences, and the alternative hypothesis is that there's at least one difference somewhere between two of the means. (***Note:*** It doesn't say that all the means have to be different.)

If it's established that at least two of the population means are different, then the next natural question is, "Okay, which ones are different?" Although this is a very simple-sounding question, it doesn't have a simple answer. The concept of means being different can be interpreted in hundreds of ways. Is one larger than all the others? Are three pairs of them different from each other and the rest all the same? Statisticians have worked long and hard to come up with a wide range of choices of procedures to explore and find differences of all types in two or more population means. This family of procedures is called *multiple comparisons.*

This section starts off with an example in which the ANOVA procedure was used and H_0 was rejected, leading you to the next step: multiple comparisons. You then get an overview of how and why multiple comparison procedures work.

Comparing cellphone minutes: An example

Suppose you want to compare the average number of cellphone minutes used per month for various age groups, where the age groups are defined as follows.

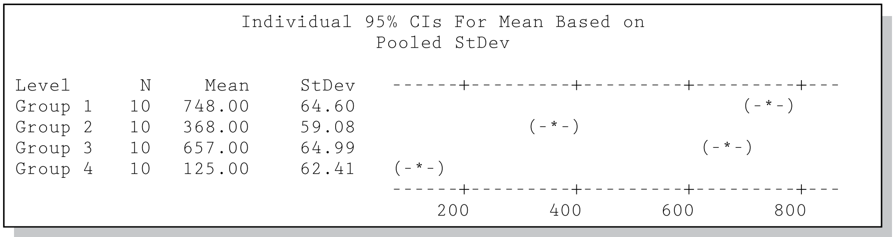
- Group 1: 19 years old and under
- Group 2: 20–39 years old
- Group 3: 40–59 years old
- Group 4: 60 years old and over

You collect data on a random sample of ten people from each group (where no one knows anyone else, to ensure independence), and you record the number of minutes each person used their cellphone in one month. The first ten lines of a hypothetical data set are shown in Table 11-1.

The means and standard deviations of the sample data are shown in Figure 11-1, as well as confidence intervals for each of the population means separately (see Chapter 4 for information on confidence intervals). Looking at Figure 11-1, it appears that all four means are different, with the 19-and-under group heading the pack, 40- to 59-year-olds not far behind, and 20- to 39-year-olds and those over 60 bringing up the rear (in that order).

TABLE 11-1 **Cellphone Minutes Used in One Month**

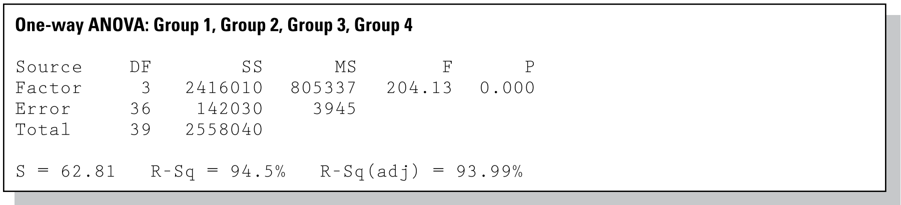
19 and Under (Group 1)	20–39 (Group 2)	40–59 (Group 3)	60 and Over (Group 4)
800	250	700	200
850	350	700	120
800	375	750	150
650	320	650	90
750	430	550	20
680	380	580	150
800	325	700	200
750	410	700	130
690	450	590	160
710	390	650	30



```
                          Individual 95% CIs For Mean Based on
                                         Pooled StDev

Level      N    Mean    StDev   ------+---------+---------+---------+---
Group 1   10  748.00    64.60                                   (-*-)
Group 2   10  368.00    59.08                     (-*-)
Group 3   10  657.00    64.99                               (-*-)
Group 4   10  125.00    62.41   (-*-)
                                ------+---------+---------+---------+---
                                    200       400       600       800
```

FIGURE 11-1: Basic statistics and confidence intervals for the cellphone data.

Knowing that you can't live by sample results alone, you decide that ANOVA is needed to see whether any differences that appear in the samples can be extended to the population (see Chapter 10). By using the ANOVA procedure, you test whether the average cell minutes used is the same across all groups. The results of the ANOVA, using the data from Table 11-1, are shown in Figure 11-2.



One-way ANOVA: Group 1, Group 2, Group 3, Group 4

```
Source    DF        SS      MS       F      P
Factor     3   2416010  805337  204.13  0.000
Error     36    142030    3945
Total     39   2558040

S = 62.81   R-Sq = 94.5%   R-Sq(adj) = 93.99%
```

FIGURE 11-2: ANOVA results for comparing cellphone use for four age groups.

Looking at Figure 11-2, the *F*-test for equality of all four population means has a *p*-value of 0.000, meaning it's less than 0.001. That says at least two of these age groups have a significant difference in their cellphone use (see Chapter 10 for information on the *F*-test and its results).

Okay, so what's your next question? You just found out that the average number of cellphone minutes used per month isn't the same across these four groups. This doesn't mean all four groups are different (see Chapter 10), but it does mean that at least two groups are significantly different in their cellphone use. So your questions are

- Which groups are different?
- How are they different?

Setting the stage for multiple comparison procedures

Determining which populations have differing means after the ANOVA *F*-test has been rejected involves a new data-analysis technique called *multiple comparisons.* The basic idea of multiple comparison procedures is to compare various means and report where and what the differences are. For example, you may conclude from a multiple comparison procedure that the first population had a mean that was statistically lower than the second population, but it was statistically higher than the mean of the third population.

There are myriad different multiple comparison procedures out there; how do you know which one you should use, and when? Two basic elements distinguish multiple comparison procedures from each other; I call them purpose and price.

- **Purpose.** When you know that a group of means aren't all equal, you zoom in to explore the relationships between them, depending on the purpose of your research. Maybe you just want to figure out which means are equivalent and which are not. Maybe you want to sort them into statistically equivalent groups from smallest to largest. Or it may be important to compare the average of one group of means to the average of another group of means. Different multiple comparison procedures were built for different purposes; for the most part, if you use them for their designed purposes, you have a better chance of finding specific differences you're looking for, if those differences are actually there.

- **Price.** Any statistical procedure you use comes with a price: the probability of making a Type I error in your conclusions somewhere during the procedure, due to chance. (A *Type I error* is committed when H_0 is rejected when it shouldn't be; in other words, you think two means are different but they really aren't. See Chapter 4 for more information.) This probability of making at least one Type I error during a multiple comparisons procedure is called the *overall error rate* (also known as the *experiment-wise error rate* (EER), or the *family-wise error rate*). Small overall error rates are, of course, desirable. Each multiple comparison procedure has its own overall error rate; generally, the more specific the relationships are that you're trying to find, the smaller your overall error rate is, assuming you're using a procedure that was designed for your purpose.

  In the next section, I describe two all-purpose multiple comparison procedures: Fisher's LSD and Tukey's test.

WARNING

Don't attempt to explore the data with a multiple comparison procedure if the test for equality of the populations isn't rejected. In this case, you must conclude that you don't have enough evidence to say the population means aren't all equal, so you must stop there. Always look at the *p*-value of the *F*-test on the ANOVA output before moving on to conduct any multiple comparisons.

Pinpointing Differing Means with Fisher and Tukey

You've conducted ANOVA to see whether a group of *k* populations has the same mean, and you've rejected H_0. You conclude that at least two of those populations have different means. But you don't have to stop there; you can go on to find out how many and which means are different by conducting multiple comparison tests.

In this section, you see two of the most well-known multiple comparison procedures: *Fisher's LSD* (also known as *Fisher's protected LSD* or *Fisher's test*) and *Tukey's test* (also known as *Tukey's simultaneous confidence intervals*).

REMEMBER

Although I only discuss two procedures in detail in this chapter, tons of other multiple comparison procedures exist (see the section, "So Many Other Procedures, So Little Time!" at the end of this chapter). Although the other procedures' methods differ a great deal, their overall goal is the same: to figure out which population means differ by comparing their sample means.

Fishing for differences with Fisher's LSD

In this section, I outline the original least significant difference (LSD) procedure and R. A. Fisher's improvement on it (aptly called *Fisher's least significant difference procedure,* or *Fisher's LSD*). The LSD and Fisher's LSD procedures both compare pairs of means using some form of *t*-tests, but they do so in different ways (see Chapter 4 or your Stats I textbook for more on the *t*-test).

The original LSD procedure

To use the original (pre-Fisher) LSD, simply choose certain pairs of means in advance and conduct a *t*-test on each pair at level $\alpha = 0.05$ to look for differences. LSD doesn't require an ANOVA test first (which is a problem that Fisher later noticed). If *k* population means are all to be compared to each other in pairs using LSD, then the number of *t*-tests performed will be represented by $\frac{k(k-1)}{2}$.

TECHNICAL STUFF

Here's how to count the number of *t*-tests when all means are compared. To start, you compare the first mean and the second mean, the first mean and the third mean, and so on until you compare the first mean and the k^{th} mean. Then you compare the second and third, second and fourth, and so on all the way down to the $(k-1)^{\text{th}}$ mean and the k^{th} mean. The total number of pairs of means to compare equals $k*(k-1)$. Because comparing the two means in either order (mean one and mean two versus mean two and mean one) gives you the same result regarding which one is largest, you divide the total by 2 to avoid double counting. For example, if you have four populations labeled A, B, C, and D, you have $\frac{4(4-1)}{2} = 6$ *t*-tests to perform: A versus B; A versus C; A versus D; B versus C; B versus D; and C versus D.

WARNING

The original LSD procedure is very straightforward, easy to conduct, and easy to understand. However, the procedure has some issues. Because each *t*-test is conducted at α level 0.05, each test done has a 5 percent chance of making a Type I error (rejecting H_0 when you shouldn't have, as I explain in Chapter 4).

Although a 5 percent error rate for each test doesn't seem too bad, the errors have a multiplicative effect as the number of tests increases. For example, the chance of making at least one Type I error with six *t*-tests, each at level $\alpha = 0.05$, is 26.50 percent, which is your overall error rate for the procedure.

TECHNICAL STUFF

If you want or need to know how I arrived at the number 26.50 percent as the overall error rate in that last example, here goes: The probability of making a Type I error for each test is 0.05. The chance of making at least one error in six tests equals 1 minus the probability of making no errors in six tests. The chance of not making an error in one test is $1-\alpha=0.95$. The chance of no error in six tests is this quantity times itself six times, or $(0.95)^6$, which equals 0.735. Now take 1 minus this quantity to get $1-0.735=0.2650$ or 26.50 percent.

Using Fisher's new and improved LSD

Fisher suggested an improvement over the regular LSD procedure, and his procedure is called *Fisher's LSD*, or *Fisher's protected LSD.* It adds the requirement that an ANOVA *F*-test must be performed first and must be rejected before any pairs of means can be compared individually or collectively. By requiring the *F*-test to be rejected, you're concluding that at least one difference exists in the means. Adding this requirement, the overall error rate of Fisher's LSD is somewhere in the area of α, which is much lower than what you get from the regular LSD procedure.

WARNING

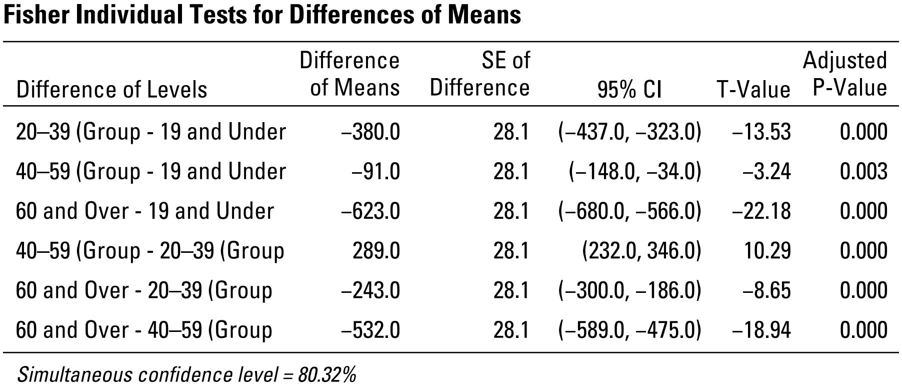
The downside of Fisher's LSD is that because each *t*-test is made at level α and the overall error rate is also near α, it's good at finding differences that really do exist, but it also makes some false alarms in the process (mainly saying there's a difference when there really isn't).

COMPUTER OUTPUT

To conduct Fisher's LSD in Minitab, go to Stat>ANOVA>One-way and indicate in the pull down menu how your data are entered in Minitab. Highlight the data for the groups you're comparing, and click Select. Then click on Comparisons, and then Fisher's. (Make sure to check the tests option if it's not already checked as well.) The individual error rate is listed at 5 (percent), which is typical. If you want to change it, type in the desired error rate (between 0.5 and 0.001), and click OK. You may type in your error rate as a decimal, 0.05, or as a number greater than 1, such as 5. Numbers greater than 1 are interpreted as a percentage.

An ANOVA procedure was done on the cellphone data presented in Table 11-1 to compare the mean number of minutes used for four age groups. Looking at the output in Figure 11-2, you see that H_0 (all the population means are equal) was rejected. The next step is to conduct multiple comparisons by using Fisher's LSD to see which population means differ. Figure 11-3 shows the selected Minitab output for those tests.

The first 3 results show Group 1's mean subtracted from the others, where Group 1 = age 19 and under. Each line after that represents the other age groups (Group 2 = 20- to 39-year-olds, Group 3 = 40- to 59-year-olds, and Group 4 = age 60 and over). Each line shows the results of comparing the mean for some other group minus the mean for Group 1.



Fisher Individual Tests for Differences of Means

Difference of Levels	Difference of Means	SE of Difference	95% CI	T-Value	Adjusted P-Value
20–39 (Group - 19 and Under	−380.0	28.1	(−437.0, −323.0)	−13.53	0.000
40–59 (Group - 19 and Under	−91.0	28.1	(−148.0, −34.0)	−3.24	0.003
60 and Over - 19 and Under	−623.0	28.1	(−680.0, −566.0)	−22.18	0.000
40–59 (Group - 20–39 (Group	289.0	28.1	(232.0, 346.0)	10.29	0.000
60 and Over - 20–39 (Group	−243.0	28.1	(−300.0, −186.0)	−8.65	0.000
60 and Over - 40–59 (Group	−532.0	28.1	(−589.0, −475.0)	−18.94	0.000

Simultaneous confidence level = 80.32%

FIGURE 11-3: Output showing Fisher's LSD applied to the cellphone data.

For example, the first row shows Group 2 being compared with Group 1. Moving to the right in that same row, you see the confidence interval for the difference in these two means, which turns out to be −437 to −323. Because zero isn't contained in this interval, you conclude that these two means are also different in the populations. Because this difference $(\mu_2 - \mu_1)$ is negative, you can also say that μ_2 is less than μ_1. Or, a better way to think of it may be that μ_1 is greater than μ_2. That is, Group 1's mean is greater than Group 2's mean.

If two means are equal, their difference equals zero, and a confidence interval for the difference should contain zero. If zero isn't included, you say the means are different.

In this case, each subsequent row in the "Group 1 subtracted from" section of Figure 11-3 shows similar results. None of the confidence intervals contain zero, so you conclude that the mean cellphone use for Group 1 is different from the mean cellphone use for any other group.

Moreover, because all confidence intervals are in negative territory, you can conclude that the mean cellphone use for those users age 19 and under is greater than for all the others. (Remember, the mean for this group is subtracted from those of the others, so a negative difference indicates that its mean is greater.)

This process continues as you move down through the output until all six pairs of means are compared to each other. Then you put them all together into one conclusion. For example, in the second portion of the output, Group 2 is subtracted from Groups 3 and 4. You see the confidence interval for the "Group 3" line is (232, 346); this gives possible values for Group 3's mean minus Group 2's mean. The interval is entirely positive, so you conclude that Group 3's mean is greater than Group 2's mean (according to this data).

On the next line, the interval for Group 4 minus Group 2 is −300 to −186. All these numbers are negative, so you conclude that Group 4's mean is less than Group 2's. You combine conclusions to say that Group 3's mean is greater than Group 2's, which is greater than Group 4's.

REMEMBER

In this example, none of the means are equal to each other, and based on the signs of confidence intervals and the results of all the individual pairwise comparisons, the following order of cellphone mean usage prevails: $\mu_1 > \mu_3 > \mu_2 > \mu_4$. (Hypothetical data aside, it may be the case that 40- to 59-year-olds use a lot of cellphone time because of their jobs.) Comparing these results to the sample means in Figure 11-1, this ordering makes sense and the means are separated enough to be declared statistically significant.

Notice near the bottom of Figure 11-3 that you see "$\text{Simultaneous confidence level} = 80.32\%$." That means the overall error rate for this procedure is $1 - 0.8032 = 0.1968$, which is close to 20 percent, a bit on the high side.

Separating the turkeys with Tukey's test

The basic idea behind Tukey's test is to provide a series of simultaneous tests for differences in the means. It still examines all possible pairs of means and keeps the overall error rate at α and also keeps the individual Type I error rate for each pair of means at α. Its distinguishing feature is that it performs the tests all at the same time.

Although the details of the formulas used for Tukey's test are beyond the scope of this book, they're not based on the *t*-test but rather something called a *studentized range statistic,* which is based on the highest and lowest means in the group and their difference. The individual error rates are held at 0.05 because Tukey developed a cutoff value for his test statistic that's based on all pairwise comparisons (no matter how many means are in each group).

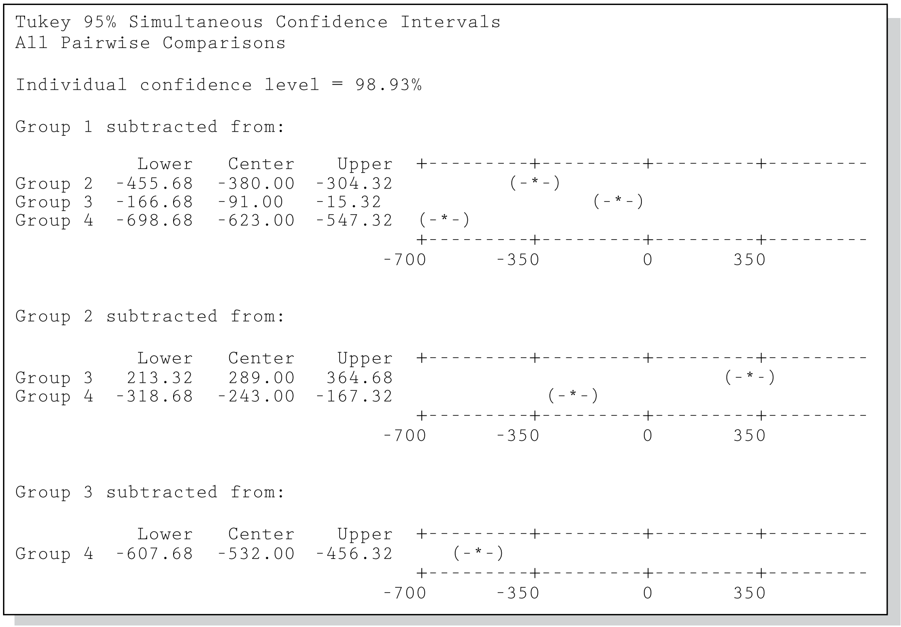
COMPUTER OUTPUT

To conduct Tukey's test, go to Stat>ANOVA>One-way. In the pull-down menu at the top of the window, choose whether your data are all in two columns (stacked), with one column for the sample number and one column for the response value, or in multiple columns, with each column containing the responses of that particular sample (unstacked). Highlight the data for the groups you're comparing, and click Select. Then click on Comparisons, and then Tukey's. The family-wise (overall) error rate is listed at 5 (percent), which is typical. If you want to change it, type in the desired error rate (between 0.5 and 0.001) and click OK. You may type in your error rate as a decimal, such as 0.05, or as a number greater than 1, such as 5. Numbers greater than 1 are interpreted as a percentage.

The Minitab output for comparing the groups regarding cellphone use by using Tukey's test appears in Figure 11-4. You can interpret its results in the same ways as those in Figure 11-3. Some of the numbers in the confidence intervals are different, but in this case, the main conclusions are the same: Those age 19 and under use their cellphones the most, followed by 40- to 59-year-olds, then 20- to 39-year-olds, and finally those age 60 and over.

REMEMBER

The results of Fisher and Tukey don't always agree, usually because the overall error rate of Fisher's procedure is larger than Tukey's (except when only two means are involved). Most statisticians I know prefer Tukey's procedure over Fisher's. That doesn't mean they don't have other procedures they like even better than Tukey's, but Tukey's is a commonly used procedure, and many people like to use it.



```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons

Individual confidence level = 98.93%

Group 1 subtracted from:

           Lower   Center   Upper  +---------+---------+---------+---------
Group 2  -455.68  -380.00  -304.32          (-*-)
Group 3  -166.68  -91.00   -15.32                    (-*-)
Group 4  -698.68  -623.00  -547.32  (-*-)
                                   +---------+---------+---------+---------
                                -700      -350         0       350


Group 2 subtracted from:

           Lower   Center   Upper  +---------+---------+---------+---------
Group 3   213.32   289.00   364.68                                (-*-)
Group 4  -318.68  -243.00  -167.32              (-*-)
                                   +---------+---------+---------+---------
                                -700      -350         0       350


Group 3 subtracted from:

           Lower   Center   Upper  +---------+---------+---------+---------
Group 4  -607.68  -532.00  -456.32     (-*-)
                                   +---------+---------+---------+---------
                                -700      -350         0       350
```

FIGURE 11-4: Output for Tukey's test used to compare cellphone usage.

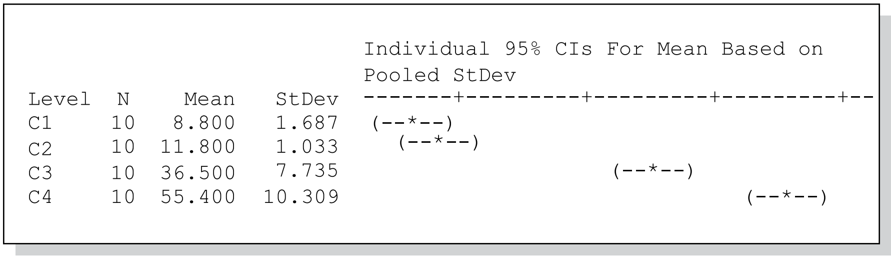
Examining the Output to Determine the Analysis

Sometimes, the process of answering questions is flipped around in your stats courses. Instead of asking you a question that you use computer output to answer, your professor may give you computer output and ask you to determine the question that the analysis answers. (Kind of like *Jeopardy*.) To work your way backward to the question, you look for clues that tell you what type of analysis was done, and then fill in the details using what you already know about that particular type of analysis.

For example, your professor may give you computer output comparing the ages of ten consumers of each of four cereal brands, labeled C1 to C4 (see Figure 11-5). On the analysis, you can see the mean consumer ages for the four cereals being compared to each other, and the analysis also shows and compares the confidence intervals for the averages. The comparison of confidence intervals tells you that you're dealing with a multiple comparison procedure.

Remember, you're looking to see whether the confidence intervals for each cereal group overlap; if they don't, those cereals have different average ages of consumers. If they do overlap, those cereals have mean ages that can't be declared different.

Based on the data in Figure 11-5, you can see that cereals one (C1) and two (C2) aren't significantly different, but for cereal three (C3), consumers have a higher average age than for C1 and C2. Cereal four (C4) has a significantly higher age than the three others. After the multiple comparison procedure, you know which cereals are different and how they compare to the others.



```
                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level  N    Mean    StDev -------+---------+---------+---------+--
C1    10   8.800    1.687  (--*--)
C2    10  11.800    1.033    (--*--)
C3    10  36.500    7.735                     (--*--)
C4    10  55.400   10.309                               (--*--)
```

FIGURE 11-5: Multiple comparison results for the cereal example.

Sometimes multiple comparison procedures give you groups of means that are equivalent to each other, different from each other, or overlapping. In this case, the final result is $\mu_{C1} = \mu_{C2} < \mu_{C3} < \mu_{C4}$.

So Many Other Procedures, So Little Time!

Many more multiple comparison procedures exist beyond Fisher's and Tukey's imaginations. Those that I discuss in this section are a little more specialized in what they were designed to look for, compared to Tukey's and Fisher's. For example, you may want to know whether a certain combination of means is larger than another combination of means; or you may want to only compare specific means to each other, not all the pairs of means.

REMEMBER

One thing to note, however, is that in many cases you don't know exactly what you're looking for when comparing means — you're just looking for differences, period. If that's the case, one of the more general procedures, like Fisher's or Tukey's, is the way to go. They're built for general exploration and do a better job of it than more-specialized procedures.

This section provides an overview of other multiple comparison procedures that exist and briefly describes each one, including the people who developed them. Given the years when these procedures were developed, I think you'll agree with me that the 1950s was the golden age of the multiple comparison procedure.

Controlling for baloney with the Bonferroni adjustment

The *Bonferroni adjustment* (or *Bonferroni correction*) is a technique used in a host of situations, not just for multiple comparisons. It was basically created to stop people from over-analyzing data. There's a limit to what you should do when analyzing data; there's a line that, when crossed, results in something statisticians call data snooping. And the Bonferroni adjustment curbs that.

Data snooping occurs when someone analyzes their data over and over again until they get a result that they can say is *statistically significant* (meaning the result is said to have been unlikely to have happened by chance; see Chapter 4). Because the number of tests completed by the data snooper is so high, they are likely to find something significant just by chance. And that result is highly likely to be bogus.

For example, suppose researchers want to find out what variable is related to sales of bedroom slippers. They collect data on everything they can think of, including the size of people's feet, the frequency with which they go out to get the paper in their slippers, even their favorite colors. Not finding anything significant, the researchers go on to examine marital status, age, and income.

Still coming up short, they go out on a limb and look at hair color, whether or not the subjects have seen a circus, and where they like to sit on an airplane (aisle or window, sir?). Then, wouldn't you know, they strike gold. Turns out that, according to their data, people who sit in the aisle seats on planes are more likely to buy bedroom slippers than those who sit by the window or in the middle of a row.

What's wrong with this picture? Too many tests. Each time a researcher examines one variable and conducts a test on it, they choose an α level at which to conduct the test. (Recall that the α level is the amount of chance you're willing to take of rejecting the null hypothesis and making a false alarm.) As the number of tests increases, the α's pile up.

Suppose α is chosen to be 0.05. The researcher then has a 5 percent chance of being wrong in finding a significant conclusion, just by chance. So if they do 100 tests, each with a 5 percent chance of an error, then on average, 5 of those 100 tests will result in a statistically significant result, just by chance. However, researchers who don't know that (or who know and go ahead regardless) find results that they claim are significant even though they're really bogus.

REMEMBER

An Italian mathematician named Carlo Emilio Bonferroni (1892–1960) said "enough already" and created something statisticians call the Bonferroni adjustment in 1950 to control the madness. The *Bonferroni adjustment* simply says that if you're doing *k* tests of your data, you can't do each one at level $\alpha = 0.05$; you need to have an α level for each test equal to $0.05 \div k$.

For example, someone who conducts 20 tests on one data set needs to do each one at level $\alpha = 0.05 \div 20 = 0.0025$. This adjustment makes it harder to find a conclusion that's significant because the *p*-value for any test must be less than 0.0025. The Bonferroni adjustment curbs the chance of data snooping until you find something bogus.

WARNING

The downside of Bonferroni's adjustment is that it's very conservative. Although it reduces the chance of concluding two means differ when they really don't, it fails to catch some differences that really *are* there. In statistical terms, Bonferroni has power issues. (See your Stats I text or *Statistics For Dummies*, 2nd Edition [Wiley] for a discussion on power.)

Comparing combinations by using Scheffé's method

Scheffé's method was developed in 1953 by Henry Scheffé (1907–1977). This method doesn't just compare two means at a time, like Tukey's and Fisher's tests do; it compares all different combinations (called *contrasts*) of the means. For example, if you have the means from four populations, you may want to test to see if their sum equals a certain value, or if the average of two of them equals the average of the other two.

Finding out whodunit with Dunnett's test

Dunnett's test was developed in 1955 by Charles Dunnett (1921–1977). *Dunnett's test* is a special multiple comparison procedure used in a designed experiment that contains a control group. The test compares each treatment group to the control group and determines which treatments do better than others.

Compared to other multiple comparison procedures, Dunnett's test is better able to find real differences in this situation because it focuses only on the differences between each treatment and the control — not on the differences between every single pair of treatments in the entire study.

Staying cool with Student Newman-Keuls

Student Newman-Keuls test is a different approach from Tukey and Fisher in comparing pairs of means in a multiple comparison procedure. This test comes from the work of three people: "Student," Newman, and Keuls.

The *Student Newman-Keuls procedure* is based on a stepwise or layer approach. You order sample means from the smallest to the largest and then examine the differences between the ordered means.

You first test the largest minus smallest difference, and if that turns out to be statistically significant, you conclude that their two respective populations are different in terms of their means. Of the remaining means, the ones that are farthest apart in the order are tested for a significant difference, and so on. You stop when you don't find any more differences.

Duncan's multiple range test

David B. Duncan (1916–2006) designed the *Duncan's multiple range test* (MRT) in 1955. The test is based on the Student Newman-Keuls test but has increased power in its ability to detect when the null hypothesis is not true (see Chapter 4), because it increases the value of α at each step of the Student Newman-Keuls test. Duncan's test is used especially in agronomy (crop and farm land management) and other types of agricultural research. One of the neatest things about being a statistician is that you never know what kinds of problems you'll be working on or who will use your methods and results.

LACING UP WITH HSU'S MCB

Hsu's MCB (Multiple Comparisons with the Best) method (1996) is much more recent compared to the others. It is a multiple comparison method that is designed to identify factor levels that are the best, insignificantly different from the best, and significantly different from the best. You can define "best" as having either the highest or lowest mean. Hsu's MCB method only compares a subset of all possible pairwise comparisons, unlike Tukey's method, which does all comparisons. This makes Hsu's method more powerful in the case where you only want to make certain comparisons.

Although Duncan won the favor of many researchers who used his test (and still do), he wasn't without his critics. Both John Tukey (who developed Tukey's test) and Henry Scheffé (who developed Scheffé's test) accused Duncan's test of being too liberal by not controlling the rate of an overall error (called a *family-wise error rate* in the big leagues). But Duncan stood his ground. He said that means are usually never equal anyway, so he wanted to err on the side of making a false alarm (Type I error) rather than missing an opportunity (Type II error) to find out when means are different.

REMEMBER

Every procedure in statistics has some chance of making the wrong conclusion, not because of an error in the process but because results vary from data set to data set. You just have to know your situation and choose the procedure that works best for that situation. When in doubt, consult a statistician for help in sorting it all out.

THE SECRET LIVES OF STATISTICIANS

Sometimes it's hard to imagine famous people having real lives, and it may be especially hard to picture statisticians doing anything but sitting in the back room calculating numbers. But the truth is, famous statisticians are interesting folks with interesting lives, just like you and me. Consider these stellar statisticians.

- **Henry Scheffé:** Scheffé was a very distinguished statistician at University of California, Berkeley. One of his five books *The Analysis of Variance,* written in 1959, is the classic book on the subject and is still used today. (I used it in grad school and still have a copy in my office.) Scheffé enjoyed backpacking, swimming, cycling, reading, and music, having learned to play the recorder during his adult life. Sadly, he died from a bicycle accident on his way to the university in 1977.

(continued)

(continued)

- **Charles Dunnett:** Nicknamed "Charlie" (did you ever think of famous statisticians as having nicknames?), Dunnett was a distinguished, award-winning professor in the Departments of Mathematics, Statistics, Clinical Epidemiology, and Biostatistics at McMaster University in Ontario, Canada. He wrote many papers, two of which were so important that they made it onto the list of the top 25 most-cited statistical papers of all time.

- **William Sealy Gosset, or "Student":** The first name included on the Student Newman-Keuls test is a story in itself. "Student" is a pseudonym of the English statistician William Sealy Gosset (1876–1937). Gosset was a statistician working for the Guinness brewery in Dublin, Ireland, when he became famous for developing the *t*-test, also known as the Student *t*-distribution (see Chapter 4), one of the most commonly used hypothesis tests in the statistical world. Gosset devised the *t*-test as a way to cheaply monitor the quality of beer. He published his work in the best of statistical journals, but his employer regarded his use of statistics in quality control to be a trade secret and wouldn't let him use his real name on his publications (although all his cronies knew exactly who "Student" was). So if not for Guinness beer, the Student's *t*-test would have been called the Gosset *t*-test (or you'd be drinking "Gosset beer").

IN THIS CHAPTER

- Building and carrying out ANOVA with two factors
- Getting familiar with (and looking for) interaction effects and main effects
- Putting the terms to the test
- Demystifying the two-way ANOVA table

Chapter 12 Finding Your Way through Two-Way ANOVA

Analysis of variance (ANOVA) is often used in experiments to see whether different levels of an explanatory variable (*x*) get different results on some quantitative variable *y.* The *x* variable in this case is called a *factor,* and it has certain levels to it, depending on how the experiment is set up.

For example, suppose you want to compare the average change in blood pressure on certain dosages of a drug. The factor is drug dosage. Suppose it has three levels: 10mg per day, 20mg per day, or 30mg per day. Suppose someone else studies the response to that same drug and examines whether the times taken per day (one time or two times) has any effect on blood pressure. In this case, the factor is number of times per day, and it has two levels: once and twice.

Suppose you want to study the effects of dosage *and* number of times taken together because you believe both factors may have an effect on the response. So what you have is called a *two-way ANOVA,* using two factors together to compare the average response. It's an extension of one-way ANOVA (refer to Chapter 10) with a twist, because the two factors you use may operate on the response differently together than they would separately.

In this chapter, first I give you an example of when you'd need to use a two-way ANOVA. Then I show you how to set up the model, make your way through the ANOVA table, take the *F*-tests, and draw the appropriate conclusions.

Setting Up the Two-Way ANOVA Model

The two-way ANOVA model extends the ideas of the one-way ANOVA model and adds an interaction term to examine how various combinations of the two factors affect the response. In this section, you see the building blocks of a two-way ANOVA: the treatments, the main effects, the interaction term, and the sums of squares equation that puts everything together.

Determining the treatments

The two-way ANOVA model contains two factors, A and B, and each factor has a certain number of levels — say, *i* levels of Factor A and *j* levels of Factor B.

In the drug study example from the chapter introduction, you have A = drug dosage with $i = 1, 2,$ or 3, and B = number of times taken per day with $j = 1$ or 2. Each person involved in the study is subject to one of the three different drug dosages and will take the drug in one of the two methods given. That means you have $3 * 2 = 6$ different combinations of Factors A and B that you can apply to the subjects, and you can study these combinations and their effects on blood pressure changes in the two-way ANOVA model.

REMEMBER

Each different combination of levels of Factors A and B is called a *treatment* in the model. Table 12-1 shows the six treatments in the drug study. For example, Treatment 4 is the combination of 20mg of the drug taken in two doses of 10mg each per day.

TIP

If Factor A has *i* levels and Factor B has *j* levels, you have $i * j$ different combinations of treatments in your two-way ANOVA model.

TABLE 12-1 Six Treatment Combinations for the Drug Study Example

Dosage Amount	One Dose Per Day	Two Doses Per Day
10mg	Treatment 1	Treatment 2
20mg	Treatment 3	Treatment 4
30mg	Treatment 5	Treatment 6

Stepping through the sums of squares

The two-way ANOVA model contains the following three terms.

- **The main effect A:** Term for the effect of Factor A on the response.
- **The main effect B:** Term for the effect of Factor B on the response.
- **The interaction of A and B:** The effect of the combination of Factors A and B (denoted AB).

The sums of squares equation for the one-way ANOVA (which I cover in Chapter 10) is $SSTO = SST + SSE$, where SSTO is the total variability in the response variable, *y*; SST is the variability explained by the treatment variable (call it Factor A); and SSE is the variability left over as error.

The purpose of a one-way ANOVA model is to test to see whether the different levels of Factor A produce different responses in the *y* variable. The way you do it is by using H_o: $\mu_1 = \mu_2 = \ldots = \mu_i$, where *i* is the number of levels of Factor A (the treatment variable). If you reject H_o, Factor A (which separates the data into the groups being compared) is significant. If you can't reject H_o, you can't conclude that Factor A is significant.

REMEMBER

In the two-way ANOVA, you add another factor to the mix (B) plus an interaction term (AB). The sums of squares equation for the two-way ANOVA model is $SSTO = SSA + SSB + SSAB + SSE$. Here, SSTO is the total variability in the *y*-values; SSA is the sums of squares due to Factor A (representing the variability in the *y*-values explained by Factor A); and similarly for SSB and Factor B. SSAB is the sums of squares due to the interaction of Factors A and B, and SSE is the amount of variability left unexplained, and deemed error.

Although the mathematical details of all the formulas for these terms are unwieldy and beyond the focus of this book, they just extend the formulas for one-way ANOVA found in Chapter 10. The computer software handles the calculations for you, so you don't have to worry about that part.

COMPUTER OUTPUT

To carry out a two-way ANOVA in Minitab, enter your data in three columns:

- Column 1 contains the responses (the actual data).
- Column 2 represents the level of Factor A (Minitab calls it the *row factor*).
- Column 3 represents the level of Factor B (Minitab calls it the *column factor*).

Go to Stat>ANOVA>General Linear Model>Fit General Linear Model. In Responses, enter **C1**. In Factors, highlight and select C2 and C3. Click Model. To the right of Interactions through order, highlight C2 and C3, then choose the number 2 and click Add. (This allows you to include the "second order interaction term" AB in the model.) Then click OK in each dialog box.

For example, suppose you have six data values in Column 1: 11, 21, 38, 14, 15, and 62. Suppose Column 2 contains 1, 1, 1, 2, 2, 2, and Column 3 contains 1, 2, 3, 1, 2, 3. This means that Factor A has two levels (1, 2), and Factor B has three levels (1, 2, 3). Table 12-2 shows a breakdown of the data values and which combinations of levels and factors are affiliated with them.

TABLE 12-2

Data and Its Respective Levels from Two Factors

Data Value	Level of Factor A	Level of Factor B
11	1	1
21	1	2
38	1	3
14	2	1
15	2	2
62	2	3

Suppose Factor A has i levels and Factor B has j levels, with a sample of size m collected on each combination of A and B. The degrees of freedom for Factor A, Factor B, and the interaction term AB are $(i-1)$, $(j-1)$, and $(i-1)*(j-1)$, respectively. This formula is just an extension of the degrees of freedom for the one-way model for Factors A and B. The degrees of freedom for SSTO is $(i*j*m)-1$, and the degrees of freedom for SSE is $i*j*(m-1)$. (See Chapter 10 for details on degrees of freedom.)

Understanding Interaction Effects

The interaction effect is the heart of the two-way ANOVA model. Knowing that the two factors may act together in a different way than they would separately is important and must be taken into account. In this section, you see the many ways in which the interaction term AB and the main effects of Factors A and B affect the response variable in a two-way ANOVA model.

What is interaction, anyway?

Interaction is when two factors meet, or interact with each other, on the response in a way that's different from how each factor affects the response separately.

For example, before you can test to see whether dosage of medicine (Factor A) or number of times taken (Factor B) are important in explaining changes in blood pressure, you have to look at how they operate together to affect blood pressure. That is, you have to examine the interaction term.

Suppose you're taking one type of medicine for cholesterol and another medicine for a heart problem. Suppose researchers only looked at the effects of each drug alone, saying each one was good for managing the problem for which it was designed with little or no side effects. Now you come along and mix the two drugs in your system. As far as the individual study results are concerned, all bets are off. With only those separate studies to go on, no one knows how the drugs will interact with each other, and you can find yourself in a great deal of trouble very quickly if you take them together.

Fortunately, drug companies and medical researchers do a great deal of work studying drug interactions, and your pharmacist knows which drugs interact as well. You can bet a statistician was involved in this work from day one!

Baking is another good example of how interaction works. Slurp down one raw egg, drink a cup of milk, and eat a cup of sugar, a cup of flour, and a stick of margarine. Then eat a cup of chocolate chips. Each one of these items has a certain taste, texture, and effect on your taste buds that, in most cases, isn't all that great. But mix them all together in a bowl and voilà! You have a batch of chocolate chip cookie dough, thanks to the magical effects of interaction. The taste is totally different (but be careful of the raw eggs!).

REMEMBER

In any two-way ANOVA, you must check out the interaction term first. If A and B interact with each other and the interaction is statistically significant, you can't examine the effects of either factor separately. Their effects are intertwined and can't be separated.

Interacting with interaction plots

In the two-way ANOVA model, you're dealing with two factors and their interaction. A number of results could come out of this model in terms of significance of the individual terms, as you can see in the following list:

- Factors A and B are both significant.
- Factor A is significant but not Factor B.
- Factor B is significant but not Factor A.
- Neither Factor A nor B is significant.
- The interaction term AB is significant, so you don't examine A or B separately.

Figure 12-1 depicts data that I made up to illustrate each of these five situations in terms of a diagram using the drug study example. Plots that show how Factors A and B react separately and together on the response variable *y* are called *interaction plots.* In the following sections, I describe each of these five situations in detail in terms of what the plots tell you and what the results mean in the context of the drug study example.

Factors A and B are significant, but not AB

Figure 12-1a shows the situation when both A and B are significant in the model and no interaction (AB) is present. The lines represent the levels of the times-per-day factor (B); the *x*-axis represents the levels of the dosage factor (A); and the *y*-axis represents the average value of the response variable *y*, which is change in blood pressure, at each combination of treatments.

In order to interpret these interaction plots, you first look at the general trends each line is making. The top line in Figure 12-1a is moving uphill from left to right, meaning that when the drug is taken two times per day, the changes in blood pressure increase as dosage level increases. The bottom line shows a similar result when the drug is taken once per day; blood pressure changes increase as dosage level increases. Assuming these differences are large enough, you conclude that dosage level (Factor A) is significant.

Now you look at how the lines compare to each other. Note that the lines, although parallel, are quite far apart. In particular, the amounts of blood pressure changes are higher overall when taking the drug twice per day (top line) than they are when taking the drug once per day (bottom line). Again, assuming these differences are large enough, you conclude that times per day (Factor B) is significant.

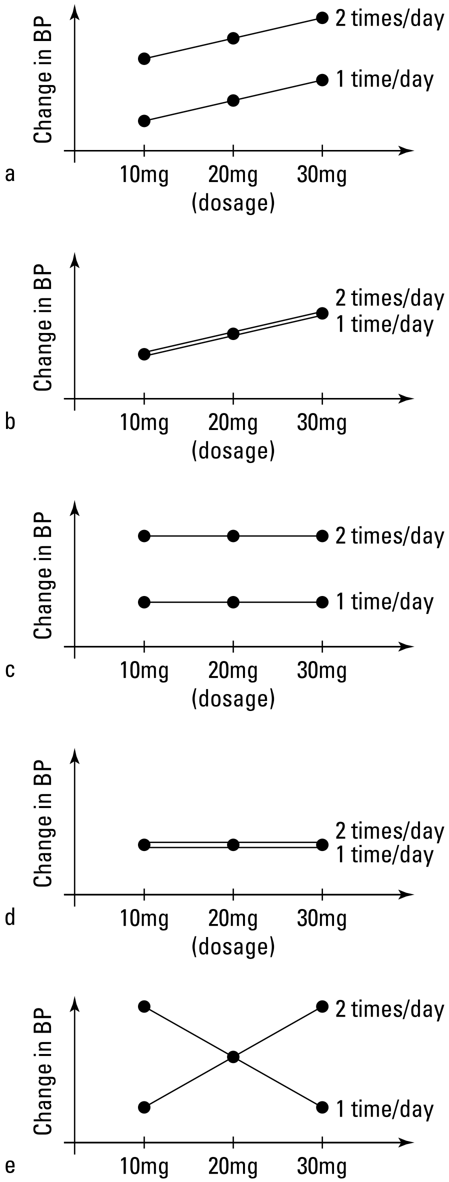


FIGURE 12-1: Five examples of the results from a two-way ANOVA with interaction.

In this case, the different combinations of Factors A and B don't affect the overall trends in blood pressure changes in opposite ways (that is, the lines don't cross each other), so there's no interaction effect between dosage level and times per day.

REMEMBER

Two parallel lines in an interaction plot indicate a lack of an interaction effect. In other words, the effect of Factor A on the response doesn't change as you move across different levels of Factor B. In the drug study example, the levels of A don't change blood pressure differently for different levels of B.

Factor A is significant but not Factor B

Figure 12-1b shows that blood pressure changes increase across dosage levels for people taking the drug once or twice a day. However, the two lines are so close together that it makes no difference whether you take the drug once or twice a day. So Factor A (dosage) is significant, and Factor B (times per day) isn't. Parallel lines indicate no interaction effect.

Factor B is significant but not Factor A

Figure 12-1c shows where Factor B (times per day) is significant but Factor A (dosage level) isn't. The lines are flat across dosage levels, indicating that dosage has no effect on blood pressure. However, the two lines for times per day are spread apart, so their effect on blood pressure is significant. Parallel lines indicate no interaction effect.

Neither factor is significant

Figure 12-1d shows two flat lines that are very close to each other. From the previous discussions about Figures 12-1b and 12-1c, you can guess that this figure represents the case where neither Factor A nor Factor B is significant, and you don't have an interaction effect because the lines are parallel.

Interaction term AB is significant

Finally you get to Figure 12-1e, the most interesting interaction plot of all. The big picture is that because the two lines cross, Factors A and B interact with each other in the way that they operate on the response. If they didn't interact, the lines would be parallel.

Start with the line in Figure 12-1e that increases from left to right (the one for 2 times per day). This line shows that when you take the drug two times per day at the low dose, you get a low change in blood pressure; as you increase dosage, blood pressure change also increases. But when you take the drug once per day, the opposite result happens, as shown by the other line that decreases from left to right in Figure 12-1e.

COMPUTER OUTPUT

To make this graph in Minitab, go to Stat>ANOVA>Interaction Plots. In the Response box, click the first column variable; in the Factors box, click the two other column variables. Then click OK.

I made an interaction plot of the data in Table 12-2, and it falls into the 12-1e category, where the interaction is present.

WARNING

If you didn't look for a possible interaction effect before you examined the main effects, you may have thought that no matter how many times you took this drug per day, the effects would be the same. Not so! Always check out the interaction term first in any two-way ANOVA. If the interaction term is significant, you have no way to pull out the effects due to just Factor A or just Factor B; they're moot.

Checking the main effects of Factor A or B without checking out the interaction AB term is considered a no-no in the two-way ANOVA world. Another taboo is examining the factors individually (also known as analyzing *main effects*) if the interaction term is significant.

Testing the Terms in Two-Way ANOVA

In a one-way ANOVA, you have only one overall hypothesis test; you use an *F*-test to determine whether the means of the *y* values are the same or different as you go across the levels of the one factor. In two-way ANOVA, you have more items to test besides the overall model. You have the interaction term AB to examine first, and possibly the main effects of A and B. Each test in a two-way ANOVA is an *F*-test based on the ideas of one-way ANOVA (see Chapter 10 for more on this).

To conduct the *F*-tests for these terms, you basically want to see whether more of the total variability in the *y*'s can be explained by the term you're testing compared to what's left in the error term. A large value of *F* means that the term you're testing is significant.

First, you test whether the interaction term AB is significant. To do this, you use the test statistic $F = \frac{MS_{AB}}{MSE}$, which has an *F*-distribution with $(i-1)*(j-1)$ degrees of freedom from MS_{AB} (mean sum of squares for the interaction term of A and B) and $i*j*(m-1)$ degrees of freedom from MSE (mean sum of squares for error), respectively. (Recall that *i* and *j* are the number of levels of A and B, and *m* is the sample size at each combination of A and B.)

If the interaction term isn't significant, you take the AB term out of the model, and you can explore the effects of Factors A and B separately regarding the response variable *y*.

The test for Factor A uses the test statistic $F = \frac{MS_A}{MSE}$, which has an F-distribution with $i - 1$ degrees of freedom from MS_A (mean sum of squares for Factor A) and $i * j * (m - 1)$ degrees of freedom from MSE (mean sum of squares for error), respectively.

Testing for Factor B uses the test statistic $F = \frac{MS_B}{MSE}$, which has an F-distribution with $j - 1$ and $i * j * (m - 1)$ degrees of freedom. (See Chapter 10 for all the details on F-tests, MSE, and degrees of freedom.)

REMEMBER

The results you can get from testing the terms of the ANOVA model are the same as those represented in Figure 12-1. They're all provided in Minitab output outlined in the next section, including their sum of squares, degrees of freedom, mean sum of squares, and p-values for their appropriate F-tests.

Running the Two-Way ANOVA Table

The ANOVA table for two-way ANOVA includes the same elements as the ANOVA table for one-way ANOVA (see Chapter 10). But whereas in the one-way ANOVA you have one line for Factor A's contributions, now you add lines for the effects of Factor B and the interaction term AB. Minitab calculates the ANOVA table for you as part of the output from running a two-way ANOVA.

In this section, you figure out how to interpret the results of a two-way ANOVA, assess the model's fit, and use a multiple comparisons procedure, all using the drug data study.

Interpreting the results: Numbers and graphs

The drug study example involves four people in each treatment combination of three possible dosage levels (10mg, 20mg, and 30mg per day) and two possible times for taking the drug (one time per day and two times per day). The total sample size is $4 * 3 * 2 = 24$. I made up five different data sets in which the analyses represent each of the five scenarios shown in Figure 12-1. Their ANOVA tables, as created by Minitab, are shown in Figure 12-2.

Notice that each ANOVA table in Figure 12-2 shows that the degrees of freedom for dosage is $3 - 1 = 2$; the degrees of freedom for times per day is $2 - 1 = 1$; the degrees of freedom for the interaction term is $(3 - 1) * (2 - 1) = 2$; the degrees of freedom for error is $3 * 2 * (4 - 1) = 18$; and the degrees of freedom for total is $3 * 2 * 4 - 1 = 23$.

The order of the graphs in Figure 12-1 and the ANOVA tables in Figure 12-2 isn't the same. Can you match them up? (I promise to give you the answers, so keep reading.)

Here's how the graphs from Figure 12-1 match up with the output in Figure 12-2:

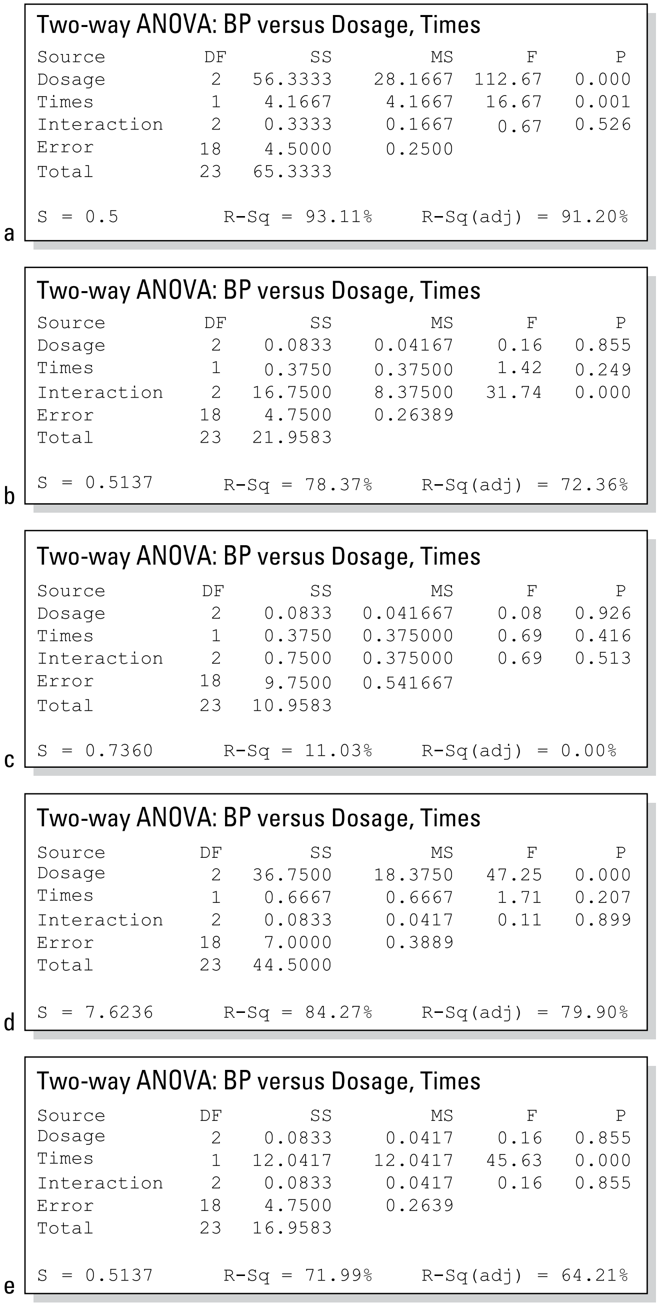
- In the ANOVA table for Figure 12-2a, you see that the interaction term isn't significant ($p\text{-value} = 0.526$), so the main effects can be studied. The *p*-values for dosage (Factor A) and times taken (Factor B) are 0.000 and 0.001, indicating both Factors A and B are significant; this matches the plot in Figure 12-1a.
- In Figure 12-2b, you see that the *p*-value for interaction is significant ($p\text{-value} = 0.000$), so you can't examine the main effects of Factors A and B (in other words, don't look at their *p*-values). This represents the situation in Figure 12-1e.
- Figure 12-2c shows nothing is significant. The *p*-value for the interaction term is 0.513; *p*-values for main effects of Factors A (dosage) and B (times taken) are 0.926 and 0.416, respectively. These results coincide with Figure 12-1d.
- Figure 12-2d matches Figure 12-1b. It has no interaction effect ($p\text{-value} = 0.899$); dosage (Factor A) is significant ($p\text{-value} = 0.000$), and times per day (Factor B) isn't ($p\text{-value} = 0.207$).
- Figure 12-2e matches Figure 12-1c. The interaction term, dosage * times per day, isn't significant ($p\text{-value} = 0.855$); times per day is significant with *p*-value 0.000, but dosage level isn't significant ($p\text{-value} = 0.855$).

Assessing the fit

To assess the fit of the two-way ANOVA models, you can use the R^2 adjusted (see Chapter 7). The higher this number is, the better (the maximum is 100 percent or 1.00). Notice that all the ANOVA tables in Figure 12-2 show a fairly high R^2 adjusted except for Figure 12-2c. In this table, none of the terms were significant.

Multiple comparisons

In the case where you find that an interaction effect is statistically significant, you can conduct multiple comparisons to see which combinations of Factors A and B create different results in the response. The same ideas hold here as they do for multiple comparisons (covered in Chapter 11), except the tests can be performed on all $i * j$ interactions.



```
Two-way ANOVA: BP versus Dosage, Times
Source         DF       SS        MS       F       P
Dosage          2  56.3333   28.1667  112.67   0.000
Times           1   4.1667    4.1667   16.67   0.001
Interaction     2   0.3333    0.1667    0.67   0.526
Error          18   4.5000    0.2500
Total          23  65.3333

S = 0.5          R-Sq = 93.11%    R-Sq(adj) = 91.20%
```

a

```
Two-way ANOVA: BP versus Dosage, Times
Source         DF       SS        MS       F       P
Dosage          2   0.0833   0.04167    0.16   0.855
Times           1   0.3750   0.37500    1.42   0.249
Interaction     2  16.7500   8.37500   31.74   0.000
Error          18   4.7500   0.26389
Total          23  21.9583

S = 0.5137       R-Sq = 78.37%    R-Sq(adj) = 72.36%
```

b

```
Two-way ANOVA: BP versus Dosage, Times
Source         DF       SS        MS       F       P
Dosage          2   0.0833  0.041667    0.08   0.926
Times           1   0.3750  0.375000    0.69   0.416
Interaction     2   0.7500  0.375000    0.69   0.513
Error          18   9.7500  0.541667
Total          23  10.9583

S = 0.7360       R-Sq = 11.03%    R-Sq(adj) = 0.00%
```

c

```
Two-way ANOVA: BP versus Dosage, Times
Source         DF       SS        MS       F       P
Dosage          2  36.7500   18.3750   47.25   0.000
Times           1   0.6667    0.6667    1.71   0.207
Interaction     2   0.0833    0.0417    0.11   0.899
Error          18   7.0000    0.3889
Total          23  44.5000

S = 7.6236       R-Sq = 84.27%    R-Sq(adj) = 79.90%
```

d

```
Two-way ANOVA: BP versus Dosage, Times
Source         DF       SS        MS       F       P
Dosage          2   0.0833    0.0417    0.16   0.855
Times           1  12.0417   12.0417   45.63   0.000
Interaction     2   0.0833    0.0417    0.16   0.855
Error          18   4.7500    0.2639
Total          23  16.9583

S = 0.5137       R-Sq = 71.99%    R-Sq(adj) = 64.21%
```

e

FIGURE 12-2: ANOVA tables for the interaction plots from Figure 12-1.

COMPUTER OUTPUT

To perform multiple comparisons for a two-way ANOVA using Minitab, you have to do it indirectly. You can number the interactions from 1 to $i * j$, and for each data pair, enter the interaction number in Column 1 (C1) and enter the response data with that interaction in Column 2 (C2). Continue down Columns 1 and 2. Then do a one-way ANOVA with multiple comparisons to see which ones are different and which are the same. See Chapter 11 for instructions.

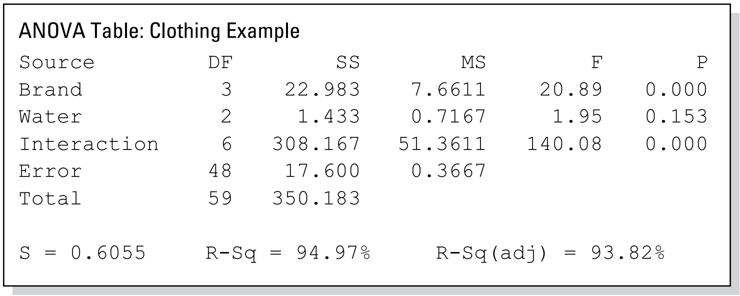
Are Whites Whiter in Hot Water? Two-Way ANOVA Investigates

You use two-way ANOVA when you want to compare the means of *n* populations that are classified according to two different categorical variables (factors). For example, suppose you want to see how four brands of detergent (Brands A, B, C, D) and water temperature ($1 = \text{cold}$, $2 = \text{warm}$, $3 = \text{hot}$) work together to affect the whiteness of dirty T-shirts being washed. (Both product-testing groups and detergent companies can use this information to investigate or advertise, respectively, how a detergent measures up to its competitors.)

Because this question involves two different factors and their effects on some numerical (quantitative) variable, you know that you need to do a two-way ANOVA. You can't assume that water temperature affects whiteness of clothes in the same way for each brand, so you need to include an interaction effect of brand and temperature in the two-way ANOVA model. Because brand of detergent has four possible types (or levels) and water temperature has three possible values (or levels), you have $4 * 3 = 12$ different combinations to examine in terms of how brand and temperature interact. Those combinations are as follows: Brand A in cold water, Brand A in warm water, Brand A in hot water, Brand B in cold water, Brand B in warm water, Brand B in hot water, and so on.

The resulting two-way ANOVA model looks like this: $y = b_i + w_j + bw_{ij} + e$, where *b* represents the brand of detergent, *w* represents the water temperature, *y* represents the whiteness of the clothes after washing, and bw_{ij} represents the interaction of brand *i* of detergent ($i = \text{A, B, C, D}$) and temperature *j* of the water ($j = 1, 2, 3$). (Note that *e* represents the amount of variation in the *y* values [whiteness] that isn't explained by either brand or temperature.)

Suppose you decide to run the experiment five times on each of the 12 combinations, which means 60 observations. (That's 60 T-shirts to wash — hey, it's a dirty job but someone's got to do it!) The results of the two-way ANOVA are shown in Figure 12-3.



```
ANOVA Table: Clothing Example
Source          DF       SS        MS         F        P
Brand            3   22.983    7.6611     20.89    0.000
Water            2    1.433    0.7167      1.95    0.153
Interaction      6  308.167   51.3611    140.08    0.000
Error           48   17.600    0.3667
Total           59  350.183

S = 0.6055     R-Sq = 94.97%     R-Sq(adj) = 93.82%
```

FIGURE 12-3: ANOVA table for the clothing example.

Note that the degrees of freedom (DF) for Brand, Water, Interaction, Error, and Total were arrived at from the following:

- DF for brand: $4-1=3$
- DF for water temperature: $3-1=2$
- DF for interaction term: $(4-1)*(3-1)=6$
- DF for error: $60-(4*3)=48$
- DF for total: $n-1=60-1=59$

Looking at the ANOVA table in Figure 12-3, you can see that the model fits the data very well, with R^2 *adjusted* equal to 93.82 percent. The interaction term (brand of detergent interacting with water temperature) is significant, with a *p*-value of 0.000. This means you can't look separately at the effect of brand of detergent or water temperature separately. One brand of detergent isn't always best, and one water temperature isn't always best; it's the combination of the two that has different effects.

Your next question may be, "Okay, which combination of detergent brand and water temperature is best?" To answer this question, I did multiple comparisons on the means from all 12 combinations. (To do this, I followed the Minitab directions from the previous section.) Luckily, Tukey gives me an overall error rate of only 5 percent, so doing this many tests doesn't lead to making a lot of incorrect conclusions.

Because of the high number of combinations to compare, making sense of all the results on Tukey's output was a little difficult. Instead, I opted to first make boxplots of the data for each combination of brand and water temperature to help me see what was going on. The results of my boxplots are shown in Figure 12-4.

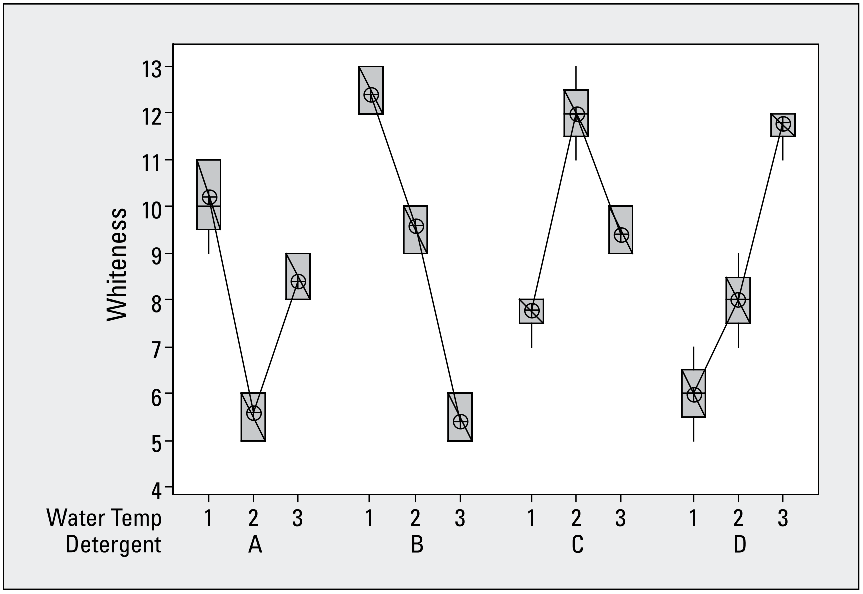


FIGURE 12-4: Boxplots showing how brand of detergent and water temperature interact to affect clothing whiteness.

COMPUTER OUTPUT

To create one set of boxplots for the data from each of the combinations in a two-way ANOVA, go to Graph>Boxplots, and choose One Y With Groups. Click OK. Then, in Graph Variables, choose Column 1 (C1) and in Categorical Variables, choose Columns 2 and 3 (C2 and C3). Click OK.

Figure 12-4 shows four groups of three connected boxes; each group of three represents data from one brand of detergent, tested under each of the three water temperatures $(1 = \text{cold}, 2 = \text{warm}, \text{and } 3 = \text{hot})$. For example, the first group of three shows the data from Brand A under each of the three water temperatures 1, 2, and 3, respectively. Each boxplot shows the results of the whiteness levels for the five shirts washed under that combination of detergent and water temperature.

Looking at these plots, you can see that each detergent reacts differently with different water temperatures. For example, Brand A does best in cold water (water temp level 1) and worst in warm water (water temp level 2), while Brand C is just the opposite, having the highest scores in warm water and the lowest in cold water. Each detergent does best or worst under a different combination of water temperatures. You can really see why the interaction term in this model is significant!

Now which combination of detergent and water temperature does the best? If you look at the plots, Brand B in cold water looks really good, and so does Brand C in warm water, closely followed perhaps by Brand D in hot water. This is where Tukey's multiple comparisons come in.

Running multiple comparisons on all 12 combinations of detergent and water temperature, you confirm that the three top combinations identified are all significantly higher than all the others (because their sample means were higher and their differences from all the other means had *p*-values less than 0.05). But the top three can't be distinguished from each other (because the *p*-values for the differences between them all exceed 0.05). Tukey also tells you that the three worst combinations are Brand A in warm water, Brand B in hot water, and Brand D in cold water. And they're all at the bottom of the barrel together (their means are significantly lower than all the rest but can't be distinguished from each other). So no single combination can claim all the bragging rights or shoulder all the blame.

You can imagine the many other comparisons that you could make from here to put the other combinations in some sort of order, but I think the best and worst are the most interesting for this case. It's like the fashion police commenting on what the stars wear on awards night. (Whatever they do wear, let's hope their statistician told them which brand of detergent to use and what water temperature to wash it in!)

IN THIS CHAPTER

- Rewriting a regression line as an ANOVA model
- Connecting regression equations to the ANOVA table

Chapter 13
Regression and ANOVA: Surprise Relatives!

So you're motoring on in your Stats II course, working your way through regression (where you estimate y using one or more x-variables; see Chapter 5). Then you hit a new topic, ANOVA, which stands for *analysis of variance* and refers to comparing the means of several populations (see Chapter 10). That seems to be no problem. But wait a minute; now your professor starts talking about how ANOVA is related to regression, and suddenly everything starts to spin out of control. How do you reconcile two techniques that appear to be as different as apples and oranges? That's what this chapter is all about.

Think of this chapter as your bridge across the gap between simple linear regression and ANOVA, allowing you to walk smoothly across, answering any questions that a professor may throw your way. Keep in mind that you don't actually apply these two techniques in this chapter (you can find that information in Chapters 5 and 10); the goal of this chapter is to determine and describe the relationship between regression and ANOVA so they don't look quite so much like an apple and an orange.

Seeing Regression through the Eyes of Variation

Every basic statistical model tries to explain why the different outcomes (*y*) are what they are. It tries to figure out what factors or explanatory variables (*x*) can help explain that variability in those *y*'s. In this section, you start with the *y*-values by themselves and see how their variability plays a central role in the regression model. This is the first step toward applying ANOVA to the regression model.

REMEMBER

No matter what *y*-variable you're interested in predicting, you'll always have variability in those *y*-values. If you want to predict the length of a fish, for example, you know that fish have many different lengths (indicating a great deal of variability). Even if you put all the fish of the same age and species together, you still have some variability in their lengths (it's less than before but still there nonetheless). The first step in understanding the basic ideas of regression and ANOVA is to understand that variability in the *y*'s is to be expected, and your job is to try to figure out what can explain most of it.

Spotting variability and finding an "x-planation"

Both regression and ANOVA work to get a handle on explaining the variability in the *y*-variable using an *x*-variable. After you collect your data, you can find the standard deviation in the *y*-variable to get a sense of how much the data varies within the sample. From there, you collect data on an *x*-variable and see how much it contributes to explaining that variability.

Suppose you notice that people spend different amounts of time on the Internet, and you want to explore why that may be. You start by taking a small sample of 20 people and record how many hours per month they spend on the Internet. The results (in hours) are 20, 20, 22, 39, 40, 19, 20, 32, 33, 29, 24, 26, 30, 46, 37, 26, 45, 15, 24, and 31. The first thing you notice about this data is the large amount of variability in it. The *standard deviation* (average distance from the data values to their mean) of this data set is 8.93 hours, which is quite large given the size of the numbers in the data set.

So you figured out that the *y*-values — the amount of time someone uses the Internet — have a great deal of variability in them. What can help explain this? Part of the variability is due to chance. But you suspect some variable is out there (call it *x*) that has some connection to the *y*-variable, and that *x*-variable can help you make more sense out of this seemingly wide range of *y*-values.

Suppose you have a brainstorm that number of years of education could possibly be related to Internet use. In this case, the explanatory variable (input variable, x) is years of education, and you want to use it to try to estimate y, the number of hours spent on the Internet in a month. You ask a larger random sample of 250 Internet users how many years of education they have (so $n = 250$). You can check out the first ten observations from your data set containing the (x, y) pairs in Table 13-1. If a significant connection of some sort exists between the x-values and the y-values, then you can say that x is helping to explain some of the variability in the y's. If it explains enough variability, you can place x into a simple regression model and use it to estimate y.

TABLE 13-1 First Ten Observations from the Education and Internet Use Example

Years of Education	Hours Spent on the Internet (In One Month)
15	41
15	32
11	33
10	42
10	28
10	21
10	17
10	14
9	18
9	14

Getting results with regression

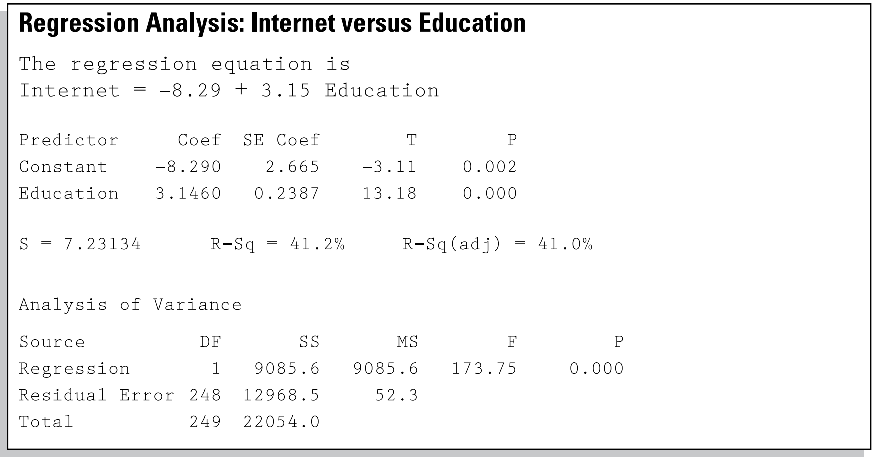
After you have a possible x-variable picked, you collect pairs of data (x, y) on a random sample of individuals from the population, and you look for a possible linear relationship between them. Looking at the small snippet of 10 out of the 250-person data set in Table 13-1, you can begin to see that you may have a pattern between education and Internet use. It looks like as education increases, so does Internet use.

To delve deeper, you make a scatterplot of the data and calculate the correlation (*r*). If the data appear to follow a straight line (as shown on the scatterplot), go ahead and perform a simple linear regression of the response variable *y* based on the *x*-variable. The *p*-value of the *x*-variable in the simple linear regression analysis tells you whether or not the *x*-variable does a significant job in predicting *y*. (For the details on simple linear regression, see Chapter 5.)

COMPUTER OUTPUT

To do a simple linear regression using Minitab, enter your data in two columns: the first column for your *x*-variable and the second column for your *y*-variable (as in Table 13-1). Go to Stat>Regression>Regression>Fit Regression Model. Click on your *y*-variable in the left-hand box; the *y*-variable then appears in the Response box on the right-hand side. Click on your *x*-variable in the left-hand box; the *x*-variable then appears in the (Continuous) Predictor box on the right-hand side. Click OK, and your regression analysis is done. As part of every regression analysis, Minitab also provides you with the corresponding ANOVA results, found at the bottom of the output.

The simple linear regression output that Minitab gives you for the education and Internet example is in Figure 13-1. (Notice the ANOVA output at the bottom; you can see the connection in the upcoming section, "Regression and ANOVA: A Meeting of the Models.")



Regression Analysis: Internet versus Education

```
The regression equation is
Internet = -8.29 + 3.15 Education

Predictor      Coef  SE Coef       T       P
Constant     -8.290    2.665   -3.11   0.002
Education    3.1460   0.2387   13.18   0.000

S = 7.23134      R-Sq = 41.2%     R-Sq(adj) = 41.0%


Analysis of Variance

Source           DF       SS       MS       F       P
Regression        1   9085.6   9085.6  173.75   0.000
Residual Error  248  12968.5     52.3
Total           249  22054.0
```

FIGURE 13-1: Output for simple linear regression applied to education and Internet use data.

Looking at Figure 13-1, you see that the *p*-value on the row marked *Education* is 0.000, which means the *p*-value is less than 0.0005. Therefore, the relationship between years of education and Internet use is statistically significant. A scatterplot of the data (not shown here) also indicates that the data appear to have a positive linear relationship, so as you increase number of years of education, Internet use also tends to increase (on average).

Assessing the fit of the regression model

Before you go ahead and use a regression model to make predictions for *y* based on an *x*-variable, you must first assess the fit of your model. You can do this with a scatterplot and correlation or with R^2.

Using a scatterplot and correlation

One way to get a rough idea of how well your regression model fits is by using a *scatterplot,* which is a graph showing all the pairs of data plotted in the *x*-*y* plane. Use the scatterplot to see whether the data appear to fall in the pattern of a line. If the data appear to follow a straight-line pattern (or even something close to that — anything but a curve or a scattering of points that has no pattern at all), you calculate the *correlation, r*, to see how strong the linear relationship between *x* and *y* is. The closer *r* is to $+1$ or -1, the stronger the relationship; the closer *r* is to zero, the weaker the relationship. Minitab can do scatterplots and correlations for you; see Chapter 5 for more on simple linear regression, including making a scatterplot and finding the value of *r*.

TIP

If the data don't have a significant correlation and/or the scatterplot doesn't look linear, stop the analysis; you can't go further to find a line that fits a relationship that doesn't exist.

Using R^2

The more general way of assessing not only the fit of a simple linear regression model but many other models too is to use R^2, also known as the *coefficient of determination.* (For example, you can use this method in multiple, nonlinear, and logistic regression models in Chapters 6, 8, and 9, to name a few.) In simple linear regression, the value of R^2 (indicated by Minitab and statisticians as a capital *R* squared) is equal to the square of the Pearson correlation coefficient, *r* (indicated by Minitab and statisticians by a small *r*). In all other situations, R^2 provides a more general measure of model fit. (Note that *r* only measures the fit of a straight-line relationship between one *x*-variable and one *y*-variable; see Chapter 5.) An even better statistic, *R^2 adjusted,* modifies R^2 to account for the number of variables in the model. (For more information on R^2 and its use and interpretation, see Chapter 7.)

The value of *R^2 adjusted* for the model using education to estimate Internet use (see Figure 13-1) is equal to 41 percent. This value reflects the percentage of variability in Internet use that can be explained by a person's years of education. This number isn't close to 1, but note that *r*, the square root of 41 percent, is 0.64, which in the case of linear regression indicates a moderate relationship.

This evidence gives you the green light to use the results of the regression analysis to estimate number of hours of Internet use in a month by using years of education. The regression equation, as it appears in the top part of the Figure 13-1 output, is $\text{Internet use} = -8.29 + 3.15 * \text{years of education}$. So if you have 16 years of education, for example, your estimated Internet use is $-8.29 + 3.15 * 16 = 42.11$, or about 42 hours per month (about 10.5 hours per week).

But wait! Look again at Figure 13-1 and zoom in on the bottom part. I didn't ask for anything special to get this information on the Minitab output, but you can see an ANOVA table there. That seems like a fish out of water, doesn't it? The next section connects the two, showing you how an ANOVA table can describe regression results (albeit it in a different way).

Regression and ANOVA: A Meeting of the Models

After you've broken down the regression output into all its pieces and parts, the next step toward understanding the connection between regression and ANOVA is to apply the sums of squares from ANOVA to regression (something that's typically not done in a regression analysis). Before you start, think of this process as going to a 3-D movie, where you have to wear special glasses in order to see all the special effects!

In this section, you see the sums of squares in ANOVA applied to regression and how the degrees of freedom work out. You build an ANOVA table for regression and discover how the *t*-test for a regression coefficient is related to the *F*-test in ANOVA.

Comparing sums of squares

Sums of squares is a term you may remember from ANOVA (see Chapter 10), but it certainly isn't a term you normally use when talking about regression (see Chapter 5). Yet, you can break down both types of models into sums of squares, and that similarity gets at the true connection between ANOVA and regression.

REMEMBER

In step-by-step terms, you first partition out the variability in the *y*-variable by using formulas for sums of squares from ANOVA (sums of squares for total, treatment, and error). Then you find those same sums of squares for regression — this is the twist on the process. You compare the two procedures through their sums of squares. This section explains how this comparison is done.

Partitioning variability by using SSTO, SSE, and SST for ANOVA

ANOVA is all about partitioning the total variability in the y-values into sums of squares (find all the information you'll ever need on one-way ANOVA in Chapter 10). The key idea is that $\text{SSTO} = \text{SST} + \text{SSE}$, where SSTO is the total variability in the y-values; SST measures the variability explained by the model (also known as the treatment, or x-variable in this case); and SSE measures the variability due to error (what's left over after the model is fit).

Following are the corresponding formulas for SSTO, SSE, and SST, where $\bar{y}$ is the mean of the y's, y_i is each observed value of y, and $\hat{y}_i$ is each predicted value of y from the ANOVA model:

$$\begin{aligned}\text{SSTO} &= \sum\left(y_i - \bar{y}\right)^2 \\ \text{SSE} &= \sum\left(y_i - \hat{y}_i\right)^2 \\ \text{SST} &= \sum\left(\hat{y}_i - \bar{y}\right)^2\end{aligned}$$

Use these formulas to calculate the sums of squares for ANOVA. (Minitab does this for you when it performs ANOVA.) Keep these values of SSTO, SST, and SSE. You'll use them to compare to the results from regression.

Finding sums of squares for regression

In regression, you measure the deviations in the y-values by taking each y_i minus its mean, $\bar{y}$. Square each result and add them all up, and you have SSTO. Next, take the residuals, which represent the difference between each y_i and its estimated value from the model, $\hat{y}_i$. Square the residuals and add them up, and you get the formula for SSE.

After you calculate SSTO and SSE, you need the bridge between them — that is, you need a formula that connects the variability in the y_i's (SSTO) and the variability in the residuals after fitting the regression line (SSE). That bridge is called the *sum of squares for regression*, or SSR (equivalent to SST in ANOVA). In regression, $\hat{y}_i$ represents the predicted value of y_i based on the regression model. These are the values on the regression line. To assess how much this regression line helps to predict the y-values, you compare it to the model you'd get without any x-variable in it.

Without any other information, the only thing you can do to predict y is to look at the average, $\bar{y}$. So, SSR compares the predicted value from the regression line to the predicted value from the flat line (the mean of the y's) by subtracting them. The result is $\left(\hat{y}_i - \bar{y}\right)$. Square each result and sum them all up, and you get the formula for SSR, which is the same as the formula for SST in ANOVA. *Voilà!*

REMEMBER

Instead of calling the sum of squares for the regression model SST as is done in ANOVA, statisticians call it SSR for *sum of squares regression.* Consider SSR to be equivalent to the SST from ANOVA. You need to know the difference because computer output (Minitab and otherwise) always lists the sums of squares for the regression model as SSR, not SST.

To summarize the sums of squares as they apply to regression, you have $\text{SSTO} = \text{SSR} + \text{SSE}$ where

- **SSTO** measures the variability in the observed *y*-values around their mean. This value represents the variance of the *y*-values.
- **SSE** represents the variability between the predicted values for *y* (the values on the line) and the observed *y*-values. SSE represents the variability left over after the line has been fit to the data.
- **SSR** measures the variability in the predicted values for *y* (the values on the line) from the mean of *y*. SSR is the sum of squares due to the regression model (the line) itself.

Minitab calculates all the sums of squares for you as part of the regression analysis. You can see this calculation in the section, "Bringing regression to the ANOVA table."

Dividing up the degrees of freedom

In ANOVA, you test a model for the treatment (population) means by using an F-test, which is $F = \frac{MST}{MSE}$. To get MST (the mean sum of squares for treatment), you take SST (the sum of squares for treatment) and divide by its degrees of freedom. You do the same with MSE (that is, take SSE, the sum of squares for error, and divide by its degrees of freedom). The questions now are, what do those degrees of freedom represent, and how do they relate to regression?

Degrees of freedom in ANOVA

In ANOVA, the degrees of freedom for SSTO is $n-1$, which represents the sample size minus one. In the formula for SSTO, $\sum\left(y_i - \bar{y}\right)^2$, you see there are *n* observed y-values minus one mean. In a very general way, that's where the $n-1$ comes from.

TIP

Note that if you divide SSTO by $n-1$, you get $\frac{\sum\left(y_i - \bar{y}\right)^2}{n-1}$, the variance in the y-values. This calculation makes good sense because the variance measures the *total* variability in the y-values.

Degrees of freedom in regression

The degrees of freedom for SST in ANOVA equal the number of treatments minus 1. How does the degrees of freedom idea relate to regression? The number of treatments in regression is equivalent to the number of parameters in a model (a *parameter* being an unknown constant in the model that you're trying to estimate).

When you test a model, you're always comparing it to a different (simpler) model to see whether it fits the data better. In linear regression, you compare your regression line $y = a + bx$, to the horizontal line $y = \bar{y}$. This second, simpler model just uses the mean of y to predict y all the time, no matter what x is. In the regression line, you have two coefficients: one to estimate the parameter for the y-intercept (a) and one to estimate the parameter for slope (b) in the model. In the second, simpler model, you have only one parameter: the value of the mean. The degrees of freedom for SSR in simple linear regression is the difference in the number of parameters from the two models: $2 - 1 = 1$.

The degrees of freedom for SSE in ANOVA is $n - k$. In the formula for SSE, $\sum(\hat{y}_i - \bar{y})^2$, you see there are n predicted y-values, and k is the number of treatments in the model. In regression, the number of parameters in the model is $k = 2$ (the slope and the y-intercept). So you have degrees of freedom $n - 2$ associated with SSE when you're doing regression.

REMEMBER

Putting all this together, the degrees of freedom for regression must add up for the equation $\text{SSTO} = \text{SSR} + \text{SSE}$. The degrees of freedom corresponding to this equation are $(n-1) = (2-1) + (n-2)$, which is true if you do the math. So the degrees of freedom for regression, using the ANOVA approach, all check out. Whew!

In Figure 13-1, you can see the degrees of freedom for each sum of squares listed under the DF column of the ANOVA part of the output. You see SSR has $2 - 1 = 1$ degree of freedom, SSE has $250 - 2 = 248$ degrees of freedom (because $n = 250$ observations were in the data set and $k = 2$, and you find $n - k$ to get degrees of freedom for SSE). The degrees of freedom for SSTO is $250 - 1 = 249$.

Bringing regression to the ANOVA table

In ANOVA, you test your model H_o (all k population means are equal) versus H_a (at least two population means are different) by using an F-test. You build your F-test statistic by relating the sums of squares for treatment to the sum of squares for error. To do this, you divide SSE and SST by their degrees of freedom ($n - k$ and $k - 1$, respectively, where n is the sample size and k is the number of treatments) to get the mean sums of squares for error (MSE) and mean sums of squares for treatment (MST). In general, you want MST to be large compared to MSE, indicating that the model fits well. The results of all these statistical gymnastics are summarized by Minitab in a table called (cleverly) the ANOVA table.

The ANOVA table shown in the bottom part of Figure 13-1 for the Internet use data example represents the ANOVA table you get from using the regression line as your model. Under the Source column, you may be used to seeing treatment, error, and total. For regression, the treatment is the regression line, so you see *regression* instead of treatment. The error term in ANOVA is labeled *residual error,* because in regression, you measure error in terms of residuals. Finally you see *total*, which is the same the world around.

The SS column represents the sums of squares for the regression model. The three sums of squares listed in the SS column are SSR (for regression), SSE (for residuals), and SST (total). These sums of squares are calculated using the formulas from the previous section; the degrees of freedom, DF in the table, are also found by using the formulas from the previous section.

The MS column takes the value of SS [you fill in the blank] and divides it by the respective degrees of freedom, just like ANOVA. For example, in Figure 13-1, SSE is 12968.5, and the degrees of freedom is 248. Take the first value divided by the second one to get 52.29 or 52.3, which is listed in the ANOVA table for MSE.

The value of the *F*-statistic, using the ANOVA method, is $F = \frac{MST}{MSE} = \frac{9085.6}{52.3} = 173.7$ in the Internet use example, which you can see in column five of the ANOVA part of Figure 13-1 (subject to rounding). The *F*-statistic's *p*-value is calculated based on an *F*-distribution with $k-1=2-1=1$ and $n-k=250-2=248$ degrees of freedom, respectively. (In the Internet use example, the *p*-value listed in the last column of the ANOVA table is 0.000, meaning the regression model fits.) But remember, in regression you don't use an *F*-statistic and an *F*-test. You use a *t*-statistic and a *t*-test. (Whoa. . .)

Relating the F- and t-statistics: The final frontier

In regression, one way of testing whether the best-fitting line is statistically significant is to test H_0: Slope $= 0$ versus H_a: Slope $\neq 0$. To do this, you use a *t*-test (see Chapter 4). The slope is the heart and soul of the regression line, because it describes the main part of the relationship between *x* and *y*. If the slope of the line equals zero (you can't reject H_0), you're just left with $y = a$, a horizontal line, and your model $y = a + bx$ isn't doing anything for you.

In ANOVA, you test to see whether the model fits by testing H_o: The means of the populations are all equal versus H_a: At least two of the population means aren't equal. To do this, you use an *F*-test (taking MST and dividing it by MSE; see Chapter 10).

REMEMBER

The sets of hypotheses in regression and ANOVA seem totally different, but in essence, they're both doing the same general thing: testing whether a certain model fits. In the regression case, the model you want to see fit is the straight line, and in the ANOVA case, the model of interest is a set of (normally distributed) populations with at least two different means (and the same variance). Here each population is labeled as a treatment by ANOVA.

You can also think of it this way: Suppose you took all the populations from the ANOVA and lined them up side by side on an *x*-*y* plane (see Figure 13-2). If the means of those distributions are all connected by a flat line (representing the mean of the *y*'s), then you have no evidence against H_o in the *F*-test, so you can't reject it — your model isn't doing anything for you (simply put, it doesn't fit). This idea is similar to the idea of fitting a flat horizontal line through the *y*-values in regression; a straight-line model with a non-zero slope. This also indicates no relationship between *x* and *y*.

The big thing is that statisticians can prove (so you don't have to) that an *F*-statistic is equivalent to the square of a *t*-statistic and that the *F*-distribution is equivalent to the square of a *t*-distribution when the SSR has $DF = 2 - 1 = 1$. And when you have a simple linear regression model, the degrees of freedom is exactly 1! (Note that *F* is always greater than or equal to zero, which is needed if you're making it the square of something.) So there you have it! The *t*-statistic for testing the regression model is equivalent to an *F*-statistic for ANOVA when the ANOVA table is formed for the simple regression model.

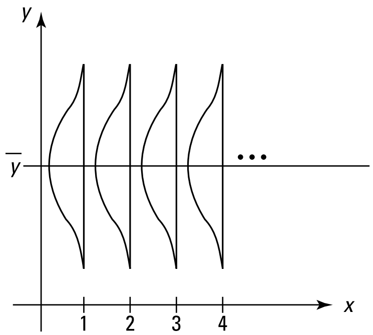


FIGURE 13-2: Connecting means of populations to the slope of a line.

Indeed (the stats professor's way of saying, "and this is the *really* cool part. . ."), if you look at the value of the t-statistic for testing the slope of the education variable in Figure 13-1, you see that it's 13.18 (look at the row marked Education and the column marked T). Square that value, and you get 173.71, the F-statistic in the ANOVA table of Figure 13-1. The F-statistic from ANOVA and the square of the t-statistic from regression are equal to each other in Figure 13-2, subject to a little round-off error done by Minitab on the output. (Just like magic! I still get chills just thinking about it.)

4

Building Strong Connections with Chi-Square Tests and Nonparametrics

IN THIS PART . . .

Relate two categorical variables with two-way tables.

Look for independence using Chi-square tests.

Do goodness-of-fit testing.

Understand when to use nonparametrics and the sign test.

Find out about the signed rank test.

IN THIS CHAPTER

- Reading and interpreting two-way tables
- Figuring probabilities and checking for independence
- Watching out for Simpson's Paradox

Chapter 14

Forming Associations with Two-Way Tables

Looking for relationships between two categorical variables is a very common goal for researchers. For example, many medical studies center on how some characteristic about a person either raises or lowers their chance of getting some disease. Marketers ask questions like, "Who's more likely to buy our product: young people or older folks?" Sports stats freaks ask questions like, "Does winning the coin toss at the beginning of a football game increase a team's chance of winning the game?" (I believe it does!)

To answer each of the preceding questions, you must first collect data (from a random sample) on the two categorical variables being compared — call them x and y. Then you organize that data into a table that contains columns and rows, showing how many individuals from the sample appear in each combination of x and y. Finally, you use the information in the table to conduct a hypothesis test (called the *Chi-square test*). Using the Chi-square test, you can determine whether you can see a relationship between x and y in the population from which the data were drawn. You need the machinery from Chapter 15 to accomplish this last step.

The goals of this chapter are to help you to understand what it means for two categorical variables (x and y) to be associated and to discover how to use percentages to determine whether a sample data set appears to show a relationship between x and y.

Suppose you're collecting data on cellphone users, and you want to find out whether more adults use cellphones for personal use than teenagers. A study of 508 randomly selected teen cellphone users and 508 randomly selected adult cellphone users conducted by a cell phone provider found that adults tend to use their phones for personal calls more than teens (big shocker). The survey showed that 427 of the adults said they used their cellphones primarily to talk with friends and family, while only 325 of the teens admitted to doing so (they prefer to text and to be on the socials, right, kids?).

But you can't stop there. You need to break down this information, calculate some percentages, and compare those percentages to see how close they really are. Results vary from sample to sample, and differences can appear by chance.

In this chapter, you find out how to organize data from categorical variables (that's data based on categories rather than measurements) into a table format. This skill is especially useful when you're looking for relationships between two categorical variables, such as using a cellphone for personal calls (a yes or no category) and *age group* (teens versus adults). You also summarize the data to answer your questions. And, finally, you also get to figure out, once and for all, what's going on with that Simpson's Paradox thing.

Breaking Down a Two-Way Table

A *two-way table* contains rows and columns and helps you organize data from categorical variables in the following ways:

- **Rows** represent the possible categories for one categorical variable, such as teens and adults.
- **Columns** represent the possible categories for a second categorical variable, such as using your cellphone for personal calls, or not.

Organizing data into a two-way table

To organize your data into a two-way table, first set up the rows and columns. Table 14-1 shows the setup for the cellphone data example from the chapter introduction.

Notice that Table 14-1 has four empty cells inside of it (not counting the empty space in the upper-left corner). Because age group here has two choices (adult or teen) and personal cellphone use has two choices (yes or no), the resulting two-way table has $2 * 2 = 4$ cells.

TABLE 14-1 Two-Way Table for the Cellphone Data

	Personal Calls: Yes	Personal Calls: No
Teens		
Adults		

REMEMBER

To figure out the number of cells in any two-way table, multiply the number of possible categories for the row variable times the number of possible categories for the column variable.

Filling in the cell counts

After you set up the table with the appropriate number of rows and columns, you need to fill in the appropriate numbers in each of the cells of the two-way table. The number in each cell of a two-way table is called the *cell count* for that cell. Of the four cells in the two-way table shown in Table 14-1, the upper-left cell represents the number of teens who use their cellphones for personal calls. With the information you have in the cellphone example, the cell count for this cell is 325. You also know that 427 adults use their cellphones for personal calls, and this number goes into the lower-left cell.

To figure out the numbers in the remaining two cells, you do a bit of subtraction. You know from the information given that the total number of teen cellphone users in the survey is 508. Each teen either uses their cellphone for personal calls (falling into the *yes* group) or doesn't use it for personal calls (falling into the *no* group). Because 325 teens fall into the *yes* group, and you have 508 teens total, 183 teens $(508-325=183)$ don't use their cellphones for personal calls. This number is the cell count for the upper-right cell of the two-way table. Finally, because 508 adults took the survey, and 427 of them use their cellphones for personal calls, you know that the rest of them $(508-427=81)$ don't. Therefore, 81 is the cell count for the lower-right cell of the table. Table 14-2 shows the completed table for the cellphone user example, with the four cell counts filled in.

TABLE 14-2 Completed Two-Way Table for the Cellphone Data

	Personal Calls: Yes	Personal Calls: No
Teens	325	$183=(508-325)$
Adults	427	$81=(508-427)$

REMEMBER

Just to save you a little time, if you have the total number in a group and the number of individuals who fall into one of the categories of the two-way table, you can determine the number falling into the remaining category by subtracting the total number in the group minus the number in the given category. You can complete this process for each remaining group in the table.

Making marginal totals

One of the most important characteristics of a two-way table is that it gives you easy access to all the pertinent totals. Because every two-way table is made up of rows and columns, you can imagine that the totals for each row and the totals for each column are important. Also, the grand total is important to know.

If you take a single row and add up all the cell counts in the cells of that row, you get a *marginal row total* for that row. Where does this marginal row total go on the table? You guessed it — out in the margin at the end of that row. You can find the marginal row totals for every row in the table and put them into the margins at the end of the rows. This group of marginal row totals for each row represents what statisticians call the *marginal distribution* for the row variable.

REMEMBER

The marginal row totals should add up to the *grand total*, which is the total number of individuals in the study. (The individuals may be people, cities, dogs, companies, and so on, depending on the scenario of the problem at hand.)

Similarly, if you take a single column and add up all the cell counts in the cells of that column, you get the *marginal column total* for that column. This number goes in the margin at the bottom of the column. Follow this pattern for each column in the table, and you have the marginal distribution for the column variable. Again, the sum of all the marginal column totals equals the grand total. The grand total is always located in the lower-right corner of the two-way table.

The marginal row total, marginal column totals, and the grand total for the cellphone example are shown in Table 14-3.

TABLE 14-3 Marginal and Grand Totals for the Cellphone Data

	Personal Calls: Yes	Personal Calls: No	**Marginal Row Totals**
Teens	325	183 = (508 – 325)	508
Adults	427	81 = (508 – 427)	508
Marginal Column Totals	752	264	1,016 **(Grand Total)**

WARNING

The marginal row totals add the cell counts in each row; yet the marginal row totals show up as a column in the two-way table. This phenomenon occurs because when summing the cell counts in a row, you put the result in the margin at the end of the row, and when you do this for each row, you're stacking the row totals into a column. Similarly, the marginal column totals add the cell counts in each column; yet they show up as a row in the two-way table. Don't let this result be a source of confusion when you're trying to navigate or set up a two-way table. I recommend that you label your totals as marginal row, marginal column, or grand total to help keep it all clear.

Breaking Down the Probabilities

In the context of a two-way table, a percentage can be interpreted in one of two ways: in terms of a group or an individual. Regarding a group, a percentage represents the portion of the group that falls into a certain category. However, a percentage also represents the probability that an individual selected at random from the group falls into a certain category.

A two-way table gives you the opportunity to find many different kinds of probabilities, which help you to find the answers to different questions about your data or to look at the data another way. In this section, I cover the three most important types of probabilities found in a two-way table: marginal probabilities, joint probabilities, and conditional probabilities. (For more complete coverage of these types of probabilities, check out *Probability For Dummies,* by yours truly and published by Wiley.)

WARNING

When you find probabilities based on a sample, as you do in this chapter, you have to realize that those probabilities pertain to that sample only. They don't transfer automatically to the population being studied. For example, if you take a random sample of 1,000 adults and find that 55 percent of them watch reality TV, this study doesn't mean that 55 percent of all adults in the entire population watch reality TV. (The media makes this mistake every day.) You need to take into account the fact that sample results vary; in Chapters 15 and 16, you do just that. But this chapter zeros in on summarizing the information in your sample, which is the first step toward that end (but not the last step in terms of making conclusions about your corresponding population).

Marginal probabilities

A *marginal probability* makes a probability out of the marginal total, for either the rows or the columns. A marginal probability represents the proportion of the entire group that belongs in that single row or column category. Each marginal probability represents only one category for only one variable — it doesn't consider the other variable at all. In the cellphone example, you have four possible marginal probabilities (refer to Table 14-3):

- Marginal probability of adult $\left(\frac{508}{1{,}016}=0.50\right)$, meaning that 50 percent of all the cellphone users in this sample were adults.
- Marginal probability of teen $\left(\frac{508}{1{,}016}=0.50\right)$, meaning that 50 percent of all the cellphone users in this sample were teens.
- Marginal probability of using a cellphone for personal calls $\left(\frac{752}{1{,}016}=0.74\right)$, meaning that 74 percent of all cellphone users in this sample make personal calls with their cellphones.
- Marginal probability of not using a cellphone for personal calls $\left(\frac{264}{1{,}016}=0.26\right)$, meaning that 26 percent of all the cellphone users in this sample don't make personal calls with their cellphones.

Statisticians use shorthand notation for all probabilities. If you let $\text{A}=\text{adult}$, $\text{T}=\text{teen}$, $\text{Yes}=\text{personal cellphone use}$, and $\text{No}=\text{no personal cellphone use}$, then the preceding marginal probabilities are written as follows:

- $P(A)=0.50$
- $P(T)=0.50$
- $P(\text{Yes})=0.74$
- $P(\text{No})=0.26$

Notice that P(T) and P(A) add up to 1.00. This result is no coincidence because these two categories make up the entire age group variable. Similarly, P(Yes) and P(No) sum up to 1.00 because those choices are the only two for the personal cellphone use variable. Everyone has to be classified somewhere.

WARNING

Be advised that some probabilities aren't useful in terms of discovering information about the population in general. For example, $P(A)=0.50$ because the researchers determined ahead of time that they wanted exactly 508 adults and exactly 508 teens. The fact that 50 percent of the sample is adult and 50 percent of the sample is teen doesn't mean that in the entire population of cellphone users 50 percent are adults and 50 percent are teens. If you want to study what proportion of cellphone users are adults and teens, you need to take a combined sample instead of two separate ones and see how many adults and teens appear in the combined sample.

Joint probabilities

A *joint probability* gives the probability of the intersection of two categories, one from the row variable and one from the column variable. It's the probability that someone selected from the entire group has two particular characteristics at the same time. In other words, both characteristics happen jointly, or together. You find a joint probability by taking the cell count for those having both characteristics and dividing by the grand total.

Here are the four joint probabilities in the cellphone example:

- The probability that someone from the entire group is a teen and uses their cellphone for personal calls is $\frac{325}{1{,}016} = 0.32$, meaning that 32 percent of all the cellphone users in this sample are teens using their cellphones for personal calls.
- The probability that someone from the entire group is a teen and doesn't use their cellphone for personal calls is $\frac{183}{1{,}016} = 0.18$.
- The probability that someone from the entire group is an adult and makes personal calls with their cellphone is $\frac{427}{1{,}016} = 0.42$.
- The probability that someone from the entire group is an adult and doesn't make personal calls with their cellphone is $\frac{81}{1{,}016} = 0.08$.

The notation for the joint probabilities listed is as follows, where $\cap$ represents the intersection of the two categories listed:

- $P(T \cap Yes) = 0.32$
- $P(T \cap No) = 0.18$
- $P(A \cap Yes) = 0.42$
- $P(A \cap No) = 0.08$

TIP

The sum of all the joint probabilities for any two-way table should be 1.00, unless you have a little round-off error, which makes it very close to 1.00 but not exactly. The sum is 1.00 because everyone in the group is classified somewhere with respect to both variables. It's like dividing the entire group into four parts and showing which proportion falls into each part.

Conditional probabilities

A *conditional probability* is what you use to compare subgroups in the sample. In other words, if you want to break down the table further, you turn to a conditional probability. Each row has a conditional probability for each cell within the row, and each column has a conditional probability for each cell within that column.

Note: Because conditional probability is one of the sticking points for a lot of students, I spend extra time on it. My goal with this section is for you to have a good understanding of what a conditional probability really means and how you can use it in the real world (something many statistics textbooks neglect to mention, I have to say).

Figuring conditional probabilities

To find a conditional probability, you first look at a single row or column of the table that represents the known characteristic about the individuals. The marginal total for that row (column) now represents your new grand total, because this group becomes your entire universe when you examine it. Then take the cell counts from that row (column) and divide the sum by the marginal total for that row (column).

Consider the cellphone example in Table 14-3. Suppose you want to look at just the teens who took the survey. The total number of teens is 508. You can break down this group into two subgroups by using conditional probability: You can find the probability of using cellphones for personal calls (teens only) and the probability of not using cellphones for personal calls (teens only). Similarly, you can break down the adults into those adults who use cellphones for personal calls and those adults who don't.

In the cellphone example, you have the following conditional probabilities when you break down the table by age group:

- The conditional probability that a teen uses a cellphone for personal calls is $\frac{325}{508}=0.64$.
- The conditional probability that a teen doesn't use a cellphone for personal calls is $\frac{183}{508}=0.36$.
- The conditional probability that an adult uses a cellphone for personal calls is $\frac{427}{508}=0.84$.
- The conditional probability that an adult doesn't use a cellphone for personal calls is $\frac{81}{508}=0.16$.

To interpret these results, you say that within this sample, if you're a teen, you're more likely than not to use your cellphone for personal calls (64 percent compared to 36 percent). However, the percentage of personal-call makers is higher for adults (84 percent versus 16 percent).

TIP

Notice that for the teens in the previous example, the two conditional probabilities (0.64 and 0.36) add up to 1.00. This is no coincidence. The teens have been broken down by cellphone use for personal calls, and because everyone in the study is a cellphone user, each teen has to be classified into one group or the other. Similarly, the two conditional probabilities for the adults sum to 1.00.

Notation for conditional probabilities

You denote conditional probabilities with a straight vertical line that lists and separates the event that's known to have happened (what's given) and the event for which you want to find the probability. You can write the notation like this: P(XX|XX). You place the given event to the right of the line and the event for which you want to find the probability to the left of the line. For example, suppose you know someone is adult (A) and you want to find out the chance they are a Democrat (D). In this case, you're looking for P(D|A). On the other hand, say you know a person is a Democrat and you want the probability that person is an adult — you're looking for P(A|D).

REMEMBER

The vertical line in the conditional probability notation isn't a division sign; it's just a line separating events A and B. Also, be careful of the order in which you place A and B into the conditional probability notation. In general, $P(A \mid B) \neq P(B \mid A)$.

Following is the notation used for the conditional probabilities in the cellphone example:

- **P(Yes | T) = 0.64.** You can say it this way: "The probability of Yes given teen is 0.64."
- **P(No | T) = 0.36.** In human terms, say, "The probability of No given teen is 0.36."
- **P(Yes | A) = 0.84.** Say this one with gusto: "The probability of Yes given adult is 0.84."
- **P(No | A) = 0.16.** You translate this notation by saying, "The probability of No given adult is 0.16."

WARNING

You can see that $P(\text{Yes} \mid T) + P(\text{No} \mid T) = 1.00$ because you're breaking all teens into two groups: those using cellphones for personal calls (Y) and those not (N). Notice, however, that $P(\text{Yes} \mid T) + P(\text{Yes} \mid A)$ doesn't sum to 1.00. In the first term, you're looking only at the teens, and in the second term, only at the adults.

Comparing two groups with conditional probabilities

One of the most common questions regarding two categorical variables is this: Are they related? To answer this question, you compare their conditional probabilities.

REMEMBER

To compare the conditional probabilities, follow these steps:

1. **Take one variable and find the conditional probabilities based on the other variable.**
2. **Repeat Step 1 for each category of the first variable.**
3. **Compare those conditional probabilities (you can even graph them for the two groups) and see whether they're the same or different.**

   If the conditional probabilities are the same for each group, the variables aren't related in the sample. If they're different, the variables are related in the sample.
4. **Generalize the results to the entire population by using the sample results to draw a conclusion from the overall population involved by doing a Chi-square test (see Chapter 15).**

Revisiting the cellphone example from the previous section, you can ask specifically, "Is personal use related to age group?" You know that you want to compare cellphone use for teens and adults to find out whether use is related to age group. However, it's very difficult to compare cell counts; for example, 325 teens use their phones for personal calls, compared to 427 adults. In fact, it's impossible to compare these numbers without using some total for perspective (325 out of what?).

REMEMBER

You have no way of comparing the cell counts in two groups without creating percentages (achieved by dividing each cell count by the appropriate total). Percentages give you a means of comparing two numbers on equal terms. For example, suppose you gave a one-question opinion survey (yes, no, and no opinion) to a random sample of 1,099 people; 465 respondents said yes, 357 said no, and 277 had no opinion. To truly interpret this information, you're probably trying to compare these numbers to each other in your head. That's what percentages do for you. Showing the percentage in each group in a side-by-side fashion gives you a relative comparison of the groups with each other.

But first, you need to bring conditional probabilities into the mix. In the cellphone example, if you want the percentage of adults who use their cellphones for personal calls, you take 427 divided by the total number of adults (508) to get 84 percent. Similarly, to get the percentage of teens who use their cellphones for personal calls, take the cell count (325) and divide it by that row total for teens (508), which gives you 64 percent. This percentage is the conditional probability of using a cellphone for personal calls, given the person is a teen.

Now you're ready to compare the teens and adults by using conditional probabilities. Take the percentage of adults who use their cellphones for personal calls and compare it to the percentage of teens who use their cellphones for personal calls. By finding these conditional probabilities, you can easily compare the two groups and say that in this sample at least, more adults use their cellphones (84 percent) for personal calls than teens (64 percent).

Using graphs to display conditional probabilities

One way to highlight conditional probabilities as a tool for comparing two groups is to use graphs, such as a pie chart comparing the results of the other variable for each group or a bar chart comparing the results of the other variable for each group. (For more information on pie charts and bar charts, see my book, *Statistics For Dummies,* 2nd Edition [Wiley] or your Stats I textbook.)

TIP

You may be wondering how close the two pie charts need to look (in terms of how close the slice amounts are for one pie compared to the other) in order to say the variables are independent. This question isn't one you can answer completely until you conduct a hypothesis test for the proportions themselves (see the Chi-square test in Chapter 15). For now, with respect to your sample data, if the difference in the appearance of the slices for the two graphs is enough that you would write a newspaper article about it, then go for dependence. Otherwise, conclude independence.

Figures 14-1a and 14-1b use two pie charts to compare cellphone use of teens and adults. Figure 14-1a shows the conditional distribution of cellphone use for (given) teens. Figure 14-1b shows the conditional distribution of cellphone use for (given) adults. A comparison of Figures 14-1a and 14-1b reveals that the slices for cellphone use aren't equal (or even close) for teens compared to adults, meaning that age group and cellphone use for personal calls are dependent in this sample. This confirms the previous conclusions.

Another way you can make comparisons is to break down the two-way table by the column variable. (You don't always have to use the row variable for comparisons.) In the cellphone example (Table 14-3), you can compare the group of personal-call makers to the group of nonpersonal-call makers and see what percentage in each group is teen or adult. This type of comparison puts a different spin on the information because you're comparing the behaviors to each other in terms of age group.

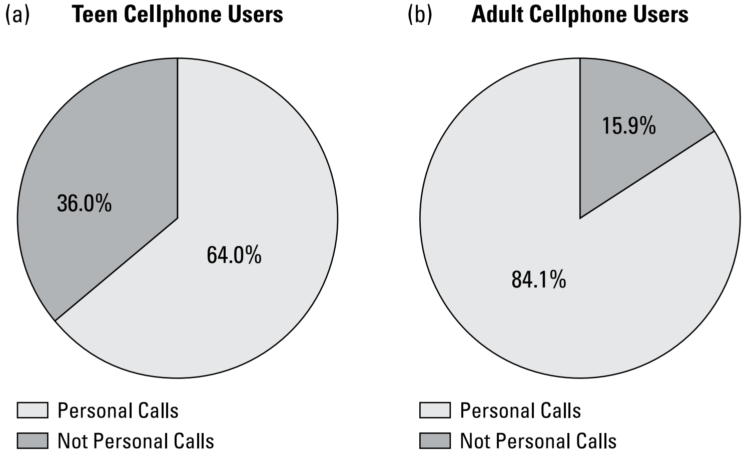


FIGURE 14-1: Pie charts comparing teen versus adult personal cellphone use.

With this new breakdown of the two-way table, looking first at the personal callers, you get the following:

- The conditional probability of being a teen, given you use your cellphone for personal calls, is $\mathrm{P(T \mid Yes)} = \frac{325}{752} = 0.43$.

  Note: The denominator is 752, the total number of people who make personal calls with their cellphones.

- The conditional probability of being an adult, given you use your cellphone for personal calls, is $\mathrm{P(A \mid Yes)} = \frac{427}{752} = 0.57$.

Again, these two probabilities add up to 1.00 because you're breaking down the personal-call makers according to age group (adult or teen).

Then, finding the conditional probabilities for the nonpersonal cellphone users, you get the following: $\mathrm{P(T \mid No)} = \frac{183}{264} = 0.69$ and $\mathrm{P(A \mid No)} = \frac{81}{264} = 0.31$.

These two probabilities also sum to 1.00 because you're breaking down the nonpersonal-call makers by age group (adult or teen).

The overall conclusions are similar to those found in the previous section, but the specific percentages and the interpretation are different. Interpreting the data this way, if you primarily use your cellphone for personal calls, you're more likely to be adult than teen (57 percent compared to 43 percent). And if you don't really use your cellphone to make personal calls, you're more likely to be teen (69 percent compared to 31 percent).

WHAT SHOULD YOU DIVIDE BY? THAT IS THE QUESTION!

To get the correct answer for any probability in a two-way table, here's the trick: Always identify the group being examined. What's the probability "out of"? In the cellphone example (refer to Table 14-3),

- If you want the percentage of *all* users who are teens and using their phones for personal calls, you take the cell count 325 divided by 1,016, the grand total.
- If you want the percentage of teens who are using their cellphones for personal calls, you take 325 divided by 508, the total number of teens.
- If you want the percentage of *personal-call makers* who are teens, you take 325 divided by 752, the total number of people who make personal calls with their cellphones.

In each of these three cases, the numerator is the same but the denominators are different, leading you to very different answers. Deciding which number to divide by is a very common source of confusion for people, and this trick can really give you an edge on keeping it straight.

Trying To Be Independent

Independence is a big deal in statistics. The term generally means that two items have outcomes whose probabilities don't affect each other. The items could be events A and B, variables *x* and *y*, survey results from two people selected at random from a population, and so on. If the outcomes of the two items do affect each other, statisticians call those two items *dependent* (or not independent). In this section, you check for and interpret independence of individual categories, one from each categorical variable in a sample, and you check for and interpret independence of two categorical variables in a sample.

Checking for independence between two categories

Statistics instructors often have students check to see whether two categories (one from a categorical variable *x* and the other from a categorical variable *y*) are independent. I prefer to just compare the two groups and talk about how similar or different the percentages are, broken down by another variable. However, to cover all the bases and make sure you can answer this very popular question, here's the official definition of independence, straight from the statistician's mouth: Two categories are *independent* if their joint probability equals the product of their marginal probabilities. The only caveat here is that neither of the categories can be completely empty.

For example, if being adult is independent of being a Democrat, then $P(A \cap D) = P(A) * P(D)$, where $D = \text{Democrat}$ and $A = \text{Adult}$. You don't have any conditional probabilities involved. So, to show that two categories are independent, find the joint probability and compare it to the product of the two marginal probabilities. If you get the same answer both times, the categories are independent. If not, then the categories are dependent.

You may be wondering, "Don't all probabilities work this way, where the joint probability equals the product of the marginals?" No, they don't. For example, if you draw a card from a standard 52-card deck, you get a red card with probability $\frac{1}{2}$. You draw a heart with probability $\frac{1}{4}$. The chance of drawing both a heart and a red card with one draw is still $\frac{1}{4}$ (because all hearts are red). However, the product of the individual probabilities for red and heart comes out to $\frac{1}{2} * \frac{1}{4} = \frac{1}{8}$ which is not equal to $\frac{1}{4}$. This tells you that the categories "red" and "heart" aren't independent (that is, they're dependent). Now the joint probability of a red two is $\frac{2}{52}$, or $\frac{1}{26}$. This equals the probability of a red card, $\frac{1}{2}$, times the probability of a two (because $\frac{1}{2} * \frac{4}{52} = \frac{1}{26}$). This tells you that the categories "red" and "two" are independent.

Another way to check for independence is to compare the conditional probability to the marginal probability. Specifically, if you want to check whether being adult is independent of being Democrat, check either of the following two situations (they'll both work if the variables are independent):

- **Is $P(A \mid D) = P(A)$?** If yes, then knowing someone is Democrat doesn't affect the chance of being an adult and A and D are independent. If not, then knowing someone is Democrat does change the chance of being an adult, and A and D are dependent.
- **Is $P(D \mid A) = P(D)$?** This question is asking whether knowing if someone is an adult changes their chances of being a Democrat. If the equality is true, then D and A are independent. If not, then D and A are dependent.

REMEMBER

Is knowing that you're in one category going to change the probability of being in another category? If so, the two categories aren't independent. If knowing doesn't affect the probability, then the two categories are independent.

Checking for independence between two variables

REMEMBER

The previous section focuses on checking whether two specific categories are independent in a sample. If you want to extend this idea to showing that two entire categorical variables are independent, you must check the independence conditions for every combination of categories in those variables. All of them must work, or independence is lost. The first case where dependence is found between two categories means that the two variables are dependent. If you find that the first case shows independence, you must continue checking all the combinations before declaring independence.

Suppose a doctor's office wants to know whether calling patients to confirm their appointments is related to whether they actually show up. The variables are x = called the patient (called or didn't call) and y = patient showed up for their appointment (showed or didn't show). Here are the four conditions that need to hold before you declare independence:

- P(showed) = P(showed | called)
- P(showed) = P(showed | didn't call)
- P(didn't show) = P(didn't show | called)
- P(didn't show) = P(didn't show | didn't call)

If any one of these conditions isn't met, you stop there and declare the two variables to be dependent in the sample. Only if all the conditions are met do you declare the two variables independent in the sample.

You can see the results of a sample of 100 randomly selected patients for this example scenario in Table 14-4.

TABLE 14-4 Confirmation Calls Related to Showing Up for the Appointment

	Called	Didn't Call	**Row Totals**
Showed	57	33	90
Didn't Show	3	7	10
Column Totals	60	40	100

Checking the conditions for independence, you can start at the first condition and check to see whether P(showed) = P(showed|called). From the last column of Table 14-4, you can see that P(showed) is equal to $\frac{90}{100} = 0.90$, or 90 percent. Next, look at the first column to find P(showed|called); this probability is $\frac{57}{60} = 95$ percent. Because these two probabilities aren't equal (although they're close), you say that showing up and calling first are dependent in this sample. You can also say that people come a little more often when you call them first. (To determine whether these sample results carry through to the population, which also takes care of the question of how close the probabilities need to be in order to conclude independence, see Chapter 15.)

Demystifying Simpson's Paradox

Simpson's Paradox is a phenomenon in which results appear to be in direct contradiction to one another, which can make even the best student's heart race. This situation can go unnoticed unless three variables (or more) are examined, in which case you organize the results into a *three-way table*, with columns within columns or rows within rows.

Simpson's Paradox is a favorite among statistics instructors (because it's so mystical and magical — and the numbers get so gooey and complex), but it's a nonfavorite among many students, mainly because of the following two reasons (in my opinion):

- Due to the way Simpson's Paradox is presented in most statistics courses, you can easily get buried in the details and have no hope of seeing the big picture. Simpson's Paradox draws attention to a big problem in terms of interpreting data, and you need to understand the paradox fully in order to avoid it.
- Most textbooks do a good job of showing you examples of Simpson's Paradox, but they fall short in explaining why it occurs, so it just looks like smoke and mirrors. Some even neglect to explain the why part at all!

This section helps you to get a handle on what Simpson's Paradox is, to better understand why and how it happens, and to know how to watch for it.

Experiencing Simpson's Paradox

Simpson's Paradox was discovered in 1951 by an American statistician named E. H. Simpson. He realized that if you analyze some data sets one way by breaking them down by two variables only, you can get one result, but when you break down the data further by a third variable, the results switch direction. That's why his result is called *Simpson's Paradox* — a paradox being an apparent contradiction in results.

Simpson's Paradox in action: Video games and the gender gap

The best way to sort through Simpson's Paradox is to watch it play out in an example and explain all the whys along the way. Suppose I'm interested in finding out who's better at playing video games, men or women. I watch males and females choose and play a variety of video games, and I record whether the player wins or loses. Suppose I record the results of 200 video games, as shown in Table 14-5. (Note that the females played 120 games, and the males played 80 games; the total number of games does not need to be the same in order to make fair comparisons.)

TABLE 14-5 Video Games Won and Lost for Males Versus Females

	Won	Lost	**Marginal Row Totals**
Males	44	36	80
Females	84	36	120
Marginal Column Totals	128	72	200 **(Grand Total)**

Looking at Table 14-5, you see the proportion of males who won their video games, P(Won|Male), is $\frac{44}{80}=0.55$. The proportion of females who won their video games, P(Won|Female), is $\frac{84}{120}=0.70$. So overall, the females won more of their video games than the males did. Does this finding mean that women are better than men at video games in general in the sample?

Not so fast, my friend. Notice that the people in the study were allowed to choose the video games they played. This factor blows the study wide open. Suppose females and males choose different types of video games: Can this affect the results? The answer may be yes. Considering other variables that could be related to the results but weren't included in the original study (or at least not in the original data analysis) is important. These additional variables that cloud the results are called *lurking variables.*

Factoring in difficulty level

Many people may expect the video game results from the previous section to be turned around to indicate that men are better at playing video games than women. According to the research, men spend more time playing video games, on average, and are by far the primary purchasers of video games, compared to women. So what explains the eyebrow-raising results in this study? Is there another possible explanation? Is important information missing that's relevant to this case?

One of the variables that wasn't considered when I made Table 14-5 was the difficulty level of the video game being played. Suppose I go back and include the difficulty level of the chosen game each time, along with each result (won or lost). Level one indicates easy video games, comparable to the level of Ms. Pac-Man (games that are my speed), and level two means more challenging video games (like war games or sophisticated strategy games).

Table 14-6 represents the results with the addition of this new information on difficulty level of games played. You have three variables now: level of difficulty (one or two), gender (male or female), and outcome (won or lost). That makes Table 14-6 a three-way table.

TABLE 14-6 Three-Way Table for Gender, Game Level, and Game Outcome

	Level-One Games		Level-Two Games	
	Won	***Lost***	***Won***	***Lost***
Males	9	1	35	35
Females	72	18	12	18

Note in Table 14-6 that the number of level-one video games chosen was $9+1+72+18=100$, and the number of level-two video games chosen was $35+35+12+18=100$. In order to reevaluate the data based on the game level information, you need to look at who chose which level of game. The next section probes this very issue.

Comparing success rates with conditional probabilities

To compare the success rates for males versus females using Table 14-6, you can figure out the appropriate conditional probabilities, first for level-one games and then for level-two games.

For level-one games (only), the conditional probability of winning given male is $P(\text{Won} \mid \text{Male}) = \frac{9}{10} = 0.90$. So for the level-one games, males won 90 percent of the games they played. For level-one games, the percentage of games won by the females is $P(\text{Won} \mid \text{Female}) = \frac{72}{90} = 0.80$, or 80 percent. These results mean that at level one, the males did 10 percent better than the females at winning their games. But this percentage appears to contradict the results found in Table 14-5. (Just wait — the contradictions don't end here!)

Now figure the conditional probabilities for the level-two video games won. For the men, the percentage of males winning level-two games was $\frac{35}{70} = 0.50$, or 50 percent. For the ladies, the percentage of women winning level-two games was $\frac{12}{30} = 0.40$, or 40 percent. Once again, the males outdid the females!

Step back and think about this scenario for a minute. Table 14-5 shows that females won a higher percentage of the video games they played overall. But Table 14-6 shows that males won more of the level-one games and more of the level-two games. What's going on? No need to check your math. No mistakes were made — no tricks were pulled. This inconsistency in results happens in real life from time to time in situations where an important third variable is left out of a study, a situation aptly named *Simpson's Paradox.* (See why it's called a paradox?)

Figuring out why Simpson's Paradox occurs

REMEMBER

Lurking variables are the underlying cause of Simpson's Paradox. A *lurking variable* is a third variable that's related to each of the other two variables and can affect the results if not accounted for.

In the video game example, when you look at the video game outcomes (won or lost) broken down by gender only (Table 14-5), females won a higher percentage of their overall games than males (70 percent overall winning percentage for females compared to 55 percent overall winning for males). Yet, when you split up the results by the level of the video game (level one or level two; see Table 14-6), the results reverse themselves, and you see that males did better than females on the level-one games (90 percent compared to 80 percent), and males also did better on the level-two games (50 percent compared to 40 percent).

To see why this seemingly impossible result happens, take a look at the marginal row *probabilities* versus the marginal row *totals* for the level-one games in Table 14-6. The percentage of times a male won when he played an easy video game was 90 percent. However, males chose level-one video games only 10 times out of 80 total level-one games played by men. That's only 12.5 percent.

To break this idea down further, the males' nonstellar performance on the challenging video games (50 percent — but still better than the females) coupled with the fact that the males chose challenging video games 87.5 percent of the time (that's 70 out of 80 times) really brought down their overall winning percentage (55 percent). And even though the men did really well on the level-one video games, they didn't play many of them (compared to the females), so their high winning percentage on level-one video games (90 percent) didn't count much toward their overall winning percentage.

Meanwhile, in Table 14-6, you see that females chose level-one video games 90 times (out of 120). Even though the females only won 72 out of the 90 games (80 percent, a lower percentage than the males, who won 9 out of 10 of their games), they chose to play many more level-one games, therefore boosting their overall winning percentage.

Now the opposite situation happens when you look at the level-two video games in Table 14-6. The males chose the harder video games 70 times (out of 80), while the females only chose the harder ones 30 times out of 120. The males did better than the females on level-two video games (winning 50 percent of them versus 40 percent for the females). However, level-two video games are harder to win than level-one video games. This factor means that the males' winning percentage on level-two video games, being only 50 percent, doesn't contribute much to their overall winning percentage. However, the low winning percentage for females on level-two video games doesn't hurt them much, because they didn't play many level-two video games.

REMEMBER

The bottom line is that the occurrence or nonoccurrence of Simpson's Paradox is a matter of weights. In the overall totals from Table 14-5, the males don't look as good as the females. But when you add in the difficulty of the games, you see that most of the males' wins came from harder games (which have a lower winning percentage). The females played many more of the easier games on average, and easy games carry a higher chance of winning no matter who plays them. So it all boils down to this: Which games did the males choose to play, and which games did the females choose to play? The males chose harder games, which contributed in a negative way to their overall winning percentage and made the females look better than they actually were.

Keeping one eye open for Simpson's Paradox

Simpson's Paradox shows you the importance of including data about possible lurking variables when attempting to look at relationships between categorical variables.

Level of game wasn't included in the original summary, Table 14-5, but it should have been included because it's a variable that affected the results. Level of game, in this case, was the lurking variable. More men chose to play the more difficult games, which are harder to win, thereby lowering their overall success rate.

TIP

You can avoid Simpson's Paradox by making sure that obvious lurking variables are included in a study; that way, when you look at the data, you get the relationships right the first time and there's a lower chance of reversing the results. And, as with all other statistical results, if it looks too good to be true or too simple to be correct, it probably is! Beware of someone who tried to oversimplify any result. While three-way tables are a little more difficult to examine, they're often worth using.

IN THIS CHAPTER

- Testing for independence in the population (not just the sample)
- Using the Chi-square distribution
- Discovering the connection between the *Z*-test and the Chi-square test

Chapter 15

Being Independent Enough for the Chi-Square Test

You've seen these hasty judgments before — people who collect one sample of data and try to use it to make conclusions about the whole population. When it comes to two categorical variables (where data fall into categories and don't represent measurements), the problem seems to be even more widespread.

For example, a TV news show finds that out of 1,000 presidential voters, 20 unemployed voters are voting Republican, 30 unemployed voters are voting Democrat, 570 employed voters are voting Republican, and 380 employed voters are voting Democrat. The news anchor shows the data and then states that 3 percent $(30 \div 1{,}000)$ of *all* presidential voters are unemployed voters voting Democrat (and so on for the other counts).

This conclusion is misleading. It's true that in this sample of 1,000 voters, 3 percent of them are unemployed voters voting Democrat. However, this result doesn't automatically mean that 30 percent of the entire population of voters is unemployed and voting Democrat. Results change from sample to sample.

In this chapter, you see how to move beyond just summarizing the sample results from a two-way table (discussed in Chapter 14) to using those results in a hypothesis test to make conclusions about an entire population. This process requires a new probability distribution called the *Chi-square distribution.* You also find out how to answer a very popular question among researchers: Are these two categorical variables independent (not related to each other) in the entire population?

The Chi-Square Test for Independence

Looking for relationships between variables is one of the most common reasons for collecting data. Looking at one variable at a time usually doesn't cut it. The methods used to analyze data for relationships are different depending on the type of data collected. If the two variables are quantitative (for example, study time and exam score), you use correlation and regression (see Chapter 5). If the two variables are categorical (for example, gender and political affiliation), you use a Chi-square test to examine relationships. In this section, you see how to use a Chi-square test to look for relationships between two categorical variables.

REMEMBER

If two categorical variables don't have a relationship, they're deemed to be *independent.* If they do have a relationship, they're called *dependent variables.* Many folks get confused by these terms, so it's important to be clear about the distinction right up front.

To test whether two categorical variables are independent, you need a Chi-square test. The steps for the Chi-square test follow. (Minitab can conduct this test for you, from Steps 3 to 7.)

1. **Collect your data, and summarize it in a two-way table.**

   These numbers represent the observed cell counts. (For more on two-way tables, see Chapter 14.)

2. **Set up your null hypothesis, H_o: Variables are independent; and the alternative hypothesis, H_a: Variables are dependent.**

3. **Calculate the expected cell counts under the assumption of independence.** (This refers to the number you expect to fall into that particular cell if the variables were independent.)

   The expected cell count for a cell is the row total times the column total divided by the grand total.

4. **Check the conditions of the Chi-square test before proceeding; each expected cell count must be greater than or equal to five.**

5. **Figure the Chi-square test statistic.**

   This statistic finds the observed cell count minus the expected cell count, squares the difference, and divides it by the expected cell count. Do these steps for each cell, and then add them all up.

6. **Look up your test statistic on the Chi-square table (see the Appendix) and find the *p*-value (or one that's close).**

7. **If your result is less than your predetermined cutoff (the α level), usually 0.05, reject H_o and conclude you have evidence of the dependence of the two variables.**

   If your result is greater than the α level, fail to reject H_o; we do not have enough evidence that the the variables are dependent.

COMPUTER OUTPUT

To conduct a Chi-square test in Minitab, you can enter data that is already in a table, or as raw data (not summarized yet). To enter data already in a table, enter your data in the spreadsheet exactly as it appears in your two-way table (see Chapter 14 for setting up a two-way table for categorical data). Go to Stat>Tables>Chi-Square Test for Association. In the pull-down menu, select Summarized Data in a Table. Click on the two variable names in the left-hand box corresponding to your column variables in the spreadsheet. They appear in the box labeled Columns Contained in the Table. In the Rows box, click on the column one variable where the row titles are. Then click OK. To enter the data as raw data (the first column is the variable 1 name and the second column is the variable 2 name), go to Stat>Tables>Chi-square Test for Association, choose raw data from the pull-down menu, select your row variable, select your column variable, and click OK.

Collecting and organizing the data

The first step in any data analysis is collecting your data. In the case of two categorical variables, you collect data on the two variables at the same time for each individual in the study.

A survey conducted by American Demographics asked 500 men and 500 women about the color of their next house. The results showed that 36 percent of the men wanted to paint their houses white, and 25 percent of the women wanted to paint their houses white. Keeping the data together in pairs (for example: male, white paint; female, nonwhite paint), you organize them into a two-way table where the rows represent the categories of one categorical variable (males and females for gender) and the columns represent the categories of the other categorical variable (white paint and nonwhite paint). Table 15-1 contains the results from a sample of 1,000 people (500 men and 500 women).

TABLE 15-1 Gender and House Paint Color Preference: Observed Cell Counts

	White Paint	Nonwhite Paint	**Marginal Row Totals**
Men	180	320	500
Women	125	375	500
Marginal Column Totals	305	695	1,000 **(Grand Total)**

The *marginal row totals* represent the total number in each row; the *marginal column totals* represent the total number in each column. (See Chapter 14 for more information on row and column marginal totals.)

Notice that of the males, the percentage that want to paint the house white is $180 \div 500 = 0.36$, or 36 percent, as stated previously. And the percentage of females that want to paint the house white is $125 \div 500 = 0.25$, or 25 percent. (Both of these percentages represent conditional probabilities as explained in Chapter 14.)

The American Demographics report concluded from this data that ". . .men and women generally agree on exterior house paint colors, the main exception being the top male choice, white (36 percent would paint their next house white versus 25 percent of women)." This type of conclusion is commonly formed, but it's an overgeneralization of the results at this point.

You know that in this sample, more men wanted to paint their houses white than women, but is 180 really that different from 125 when you're dealing with a sample size of 1,000 people whose results will vary the next time you do the survey? How do you know these results carry over to the population of all men and women? That question can't be answered without a formal statistical procedure called a *hypothesis test* (see Chapter 4 for the basics of hypothesis tests).

To show that men and women in the population differ according to favorite house color, first note that you have two categorical variables:

- Gender (male or female)
- Paint color (white or nonwhite)

WARNING

Making conclusions about the population based on the sample (observed) data in a two-way table is taking too big of a leap. You need to conduct a Chi-square test in order to broaden your conclusions to the entire population. The media, and even some researchers, can get into trouble by ignoring the fact that sample results vary. Stopping with the sample results only and going merrily on your way can lead to conclusions that others can't confirm when they take new samples.

TIP

You keep the connection between the two pieces of information by organizing the data into one two-way table versus two individual tables — one for gender and one for house-paint preference. With one two-way table, you can look at the relationship between the two variables. (For full details on organizing and interpreting the results from a two-way table, see Chapter 14.)

Determining the hypotheses

Every hypothesis test (whether it be a Chi-square test or some other test) has two hypotheses.

- **Null hypothesis:** You have to believe this unless someone shows you otherwise. The notation for this hypothesis is H_o.
- **Alternative hypothesis:** You want to conclude this in the event that you can't support the null hypothesis anymore. The notation for this hypothesis is H_a.

In the case where you're testing for the independence of two categorical variables, the null hypothesis is when no relationship exists between them. In other words, they're independent. The alternative hypothesis is when the two variables are related, or dependent.

For the paint color preference example from the previous section, you write H_o: Gender and paint color preference are independent versus H_a: Gender and paint color preference are dependent. And there you have it — Step 2 of the Chi-square test.

For a quick review of hypothesis testing, turn to Chapter 4. For a full discussion of the topic, see my other book, *Statistics For Dummies,* 2nd Edition (Wiley) or your Stats I textbook.

Figuring expected cell counts

When you've collected your data and set up your two-way table (for example, see Table 15-1), you already know what the observed values are for each cell in the table. Now you need something to compare them to. You're ready for Step 3 of the Chi-square test — finding expected cell counts.

The null hypothesis says that the two variables *x* and *y* are independent. That's the same as saying *x* and *y* have no relationship. Assuming independence, you can determine which numbers should be in each cell of the table by using a formula for what's called the *expected cell counts.* (Each individual square in a two-way table is called a *cell,* and the number that falls into each cell is called the *cell count;* see Chapter 14.)

Table 15-1 shows the observed cell counts from the gender and paint color preference example. To find the expected cell counts, you take the row total times the column total divided by the grand total, and do this for each cell in the table. Table 15-2 shows the calculations for the expected cell counts for the gender and paint color preference data.

TABLE 15-2 **Gender and House Paint Color Preference: Expected Cell Counts**

	White Paint	Nonwhite Paint	**Marginal Row Totals**
Men	$(500 * 305) \div 1{,}000 = 152.5$	$(500 * 695) \div 1{,}000 = 347.5$	500
Women	$(500 * 305) \div 1{,}000 = 152.5$	$(500 * 695) \div 1{,}000 = 347.5$	500
Marginal Column Totals	305	695	1,000 **(Grand Total)**

Next you compare the observed cell counts in Table 15-1 to the expected cell counts in Table 15-2 by looking at their differences. The differences between the observed and expected cell counts shown in these tables are as follows:

$$\begin{aligned} 180 - 152.5 &= 27.5 \\ 320 - 347.5 &= -27.5 \\ 125 - 152.5 &= -27.5 \\ 375 - 347.5 &= 27.5 \end{aligned}$$

Next you do a Chi-square test for independence (see Chapter 16) to determine whether the differences found in the sample between the observed and expected cell counts are simply due to chance, or whether they carry through to the population.

REMEMBER

Under independence, you conclude there is not a significant difference between what you observed and what you expected.

Checking the conditions for the test

REMEMBER

Step 4 of the Chi-square test is checking conditions. The Chi-square test has one main condition that must be met in order to test for independence on a two-way table: The expected count for each cell must be at least five — that is, greater than or equal to five. Expected cell counts that fall below five aren't reliable in terms of the variability that can take place.

We also need the observations to be independent. Married couples being counted as one couple per pair are OK because the couples themselves were chosen randomly in this example.

In the gender and paint color preference example, Table 15-2 shows that all the expected cell counts are at least five, so the conditions of the Chi-square test are met.

WARNING

If you're analyzing data and you find that your data set doesn't meet the expected cell count of at least five for one or more cells, you may be able to combine some of your rows and/or columns if it makes sense to do so. This combination makes your table smaller, but it increases the cell counts for the cells that you do have, which helps you meet the condition.

Calculating the Chi-square test statistic

Every hypothesis test uses data to make the decision about whether or not to reject H_o in favor of H_a. In the case of testing for independence in a two-way table, you use a hypothesis test based on the Chi-square test statistic. In the following sections, you can see the steps for calculating and interpreting the Chi-square test statistic, which is Step 5 of the Chi-square test.

Working out the formula

A major component of the Chi-square test statistic is the expected cell count for each cell in the table. The formula for finding the expected cell count, e_{ij}, for the cell in row i, column j is $e_{ij} = \frac{\text{row } i \text{ total} * \text{column } j \text{ total}}{\text{grand total}}$.

Note that the values of i and j vary for each cell in the table. In a two-way table, the upper-left cell of the table is in row one, column one. The cell in the upper-right corner is in row one, column two. The cell in the lower-left corner is in row two, column one, and the lower-right cell is in row two, column two.

The formula for the Chi-square test statistic is $\chi^2 = \sum_i \sum_j \frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}}$, where o_{ij} is the observed cell count for the cell in row i, column j, and e_{ij} is the expected cell count for the cell in row i, column j.

WARNING

When you calculate the expected cell count for some cells, you typically get a number that has some digits after the decimal point (in other words, the number isn't a whole number). Don't round this number off, despite the temptation to do so. This expected cell count is actually an overall-average expected value, so keep the count as it is, with decimal included.

Calculating the test statistic

Here are the major steps of how to calculate the Chi-square test statistic for independence (Minitab does these steps for you as well):

1. **Subtract the observed cell count from the expected cell count for the upper-left cell in the table.**
2. **Square the result from Step 1 to make the number positive.**
3. **Divide the result from Step 2 by the expected cell count.**
4. **Repeat this process for all the cells in the table, and add up all the results to get the Chi-square test statistic.**

TECHNICAL STUFF

The reason you divide by the expected cell count in the Chi-square test statistic is to account for cell-count sizes. If you expect a big cell count, say 100, and are off by only 5 for the observed count of that cell, that difference shouldn't count as much as if you expected a small cell count (like 10) and the observed cell count was off by 5. Dividing by the expected cell count puts a more fair weight on the differences that go into the Chi-square test statistic.

COMPUTER OUTPUT

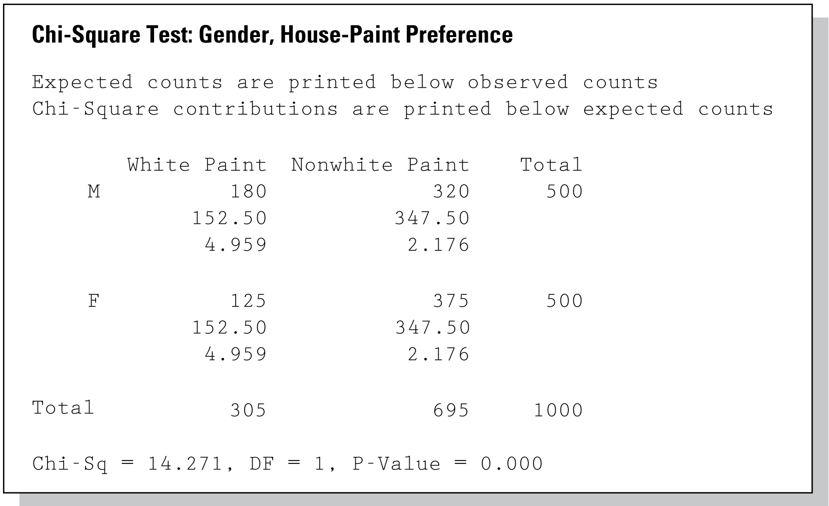
To perform a Chi-square test in Minitab, you can enter your data as a summarized table (see the previous instructions for a Chi-square test for independence) or you can first enter the raw data (the data on each person) in two columns, which is most likely to be the case. In the latter case, the first column contains the values of the first variable in your data set. (For example, if your first variable is gender, go down the first Minitab column, entering the gender of each person.) Then enter the data from your second variable in the second column, where each row represents a single person in the data set. (If your second variable is house paint color preference, for example, enter each person's paint color preference in column two, keeping the data from each person together in each row.) Go to Stat> Tables>Cross-tabulation and Chi-square.

Now Minitab needs to know which is your row variable and which is your column variable in your table. On the left-hand side, click on the variable that you want to represent the rows of your two-way table (you may click on the first variable). Click Select, and the variable name appears in the row variable portion of the table on the right. Now find the column variable blank on the right-hand side and click on it. Go to the left-hand side and click on the name of your second variable. Click Select. Then click on the Chi-square button and choose Chi-square analysis by checking the box. (You can also choose which items you want to include in your output by checking those items. For example, if you want the expected cell counts included, then also check that box.) Then click OK. Finally, click OK again to clear all the windows.

Picking through the output

The Minitab output for the Chi-square analysis for the gender and house paint color preference example (from Table 15-1) is shown in Figure 15-1. You can pick out quite a few numbers from the output in Figure 15-1 that are especially important. The following three numbers are listed in each cell:

- The first (top) number is the observed cell count for that cell; this matches the observed cell count for each cell shown in Table 15-1. (Notice that the marginal row and column totals of Figure 15-1 also match those from Table 15-1.)
- The second number in each cell of Figure 15-1 is the expected cell count for that cell; you find it by taking the row total times the column total divided by the grand total (see the section, "Figuring expected cell counts"). For example, the expected cell count for the upper-left cell (males who prefer white house paint) is $(500 * 305) \div 1{,}000 = 152.50$.
- The third number in each cell of Figure 15-1 is that part of the Chi-square test statistic that comes from that cell. (See Steps 1 through 3 of the previous section, "Working out the formula.") The sum of the third numbers in each cell equals the value of the Chi-square statistic listed in the last line of the output. (For the house paint color preference example, the Chi-square test statistic is 14.27.) (Note to get this third number in the Minitab output, click on Statistics and select Each Cell's Contribution to Chi-Square.)



Chi-Square Test: Gender, House-Paint Preference

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

       White Paint  Nonwhite Paint    Total
    M          180             320      500
            152.50          347.50
             4.959           2.176

    F          125             375      500
            152.50          347.50
             4.959           2.176

Total          305             695     1000

Chi-Sq = 14.271, DF = 1, P-Value = 0.000
```

FIGURE 15-1: Minitab output for the house paint color preference versus gender data.

Finding your results on the Chi-square table

The only way to make an assessment about your Chi-square test statistic is to compare it to all the possible Chi-square test statistics you would get if you had a two-way table with the same row and column totals, yet you distributed the numbers in the cells in every way possible. (You can do that in your sleep, right?) Some resulting tables give large Chi-square test statistics, and some give small Chi-square test statistics.

Putting all these Chi-square test statistics together gives you what's called a *Chi-square distribution.* You find your particular test statistic on that distribution (Step 6 of the Chi-square test), and see where it stands compared to the rest.

If your test statistic is large enough that it appears way out on the right tail of the Chi-square distribution (boldly going where no test statistic has gone before), you reject H_0 and conclude the two variables are not independent. If the test statistic isn't that far out, you can't reject H_0.

In the following sections, you find out more about the Chi-square distribution and how it behaves, so you can make a decision about the independence of your two variables based on your Chi-square statistic.

Determining degrees of freedom

Each type of two-way table has its own Chi-square distribution, depending on the number of rows and columns it has, and each Chi-square distribution is identified by its *degrees of freedom.*

In general, a two-way table with *r* rows and *c* columns uses a Chi-square distribution with $(r-1)*(c-1)$ degrees of freedom. A two-way table with two rows and two columns uses a Chi-square distribution with one degree of freedom. Notice that $1=(2-1)*(2-1)$. A two-way table with three rows and two columns uses a Chi-square distribution with $(3-1)*(2-1)=2$ degrees of freedom.

TECHNICAL STUFF

Understanding *why* degrees of freedom are calculated this way is likely to be beyond the scope of your statistics class. But if you really want to know, the degrees of freedom represents the number of cells in the table that are flexible, or free, given all the marginal row and column totals.

For example, suppose that a two-way table has all row and column totals equal to 100 and the upper-left cell is 70. Then the upper-right cell must be $100\text{ (row total)}-30=70$. Because the column one total is 100, and the upper-left cell count is 70, the lower-left cell count must be $100-70=30$. Similarly, the lower-right cell count must be 70.

So you have only one free cell in a two-way table after you have the marginal totals set up. That's why the degree of freedom for a two-way table is 1. In general, you always lose one row and one column because of knowing the marginal totals. That's because the last row and column values can be calculated through subtraction. That's where the formula $(r-1)*(c-1)$ comes from. (That's more than you wanted to know, isn't it? But you gotta admit, it's pretty cool!)

Discovering how Chi-square distributions behave

Figure 15-2 shows pictures of Chi-square distributions with 1, 2, 4, 6, 8, and 10 degrees of freedom, respectively. Here are some important points to keep in mind about Chi-square distributions:

- For 1 degree of freedom, the distribution looks like a hyperbola (see Figure 15-2, top left); for more than 1 degree of freedom, it looks like a mound that has a long right tail (see Figure 15-2, lower right).
- All the values are greater than or equal to zero.
- The shape is always skewed to the right (tail going off to the right).
- As the number of degrees of freedom increases, the mean (the overall average) increases (moves to the right) and the variances increase (resulting in more spread).
- No matter what the degree of freedom is, the values on the Chi-square distribution (known as the *density*) approach zero for increasingly larger Chi-square values. That means that larger and larger Chi-square values are less and less likely to happen.

Using the Chi-square table

After you find your Chi-square test statistic and its degrees of freedom, you want to determine how large your statistic is, relative to its corresponding distribution. (You're now venturing into Step 7 of the Chi-square test.)

If you think about it graphically, you want to find the probability of being beyond (getting a larger number than) your test statistic. If that probability is small, your Chi-square test statistic is something unusual — it's out there — and you can reject H_o. You then conclude that your two variables are not independent (they're related somehow).

If the probability of being to the right of your Chi-square test statistic (on a graph) isn't small enough, you don't have enough evidence to reject H_o. You then stick with H_o; you can't reject it. You conclude that your two variables are independent (unrelated).

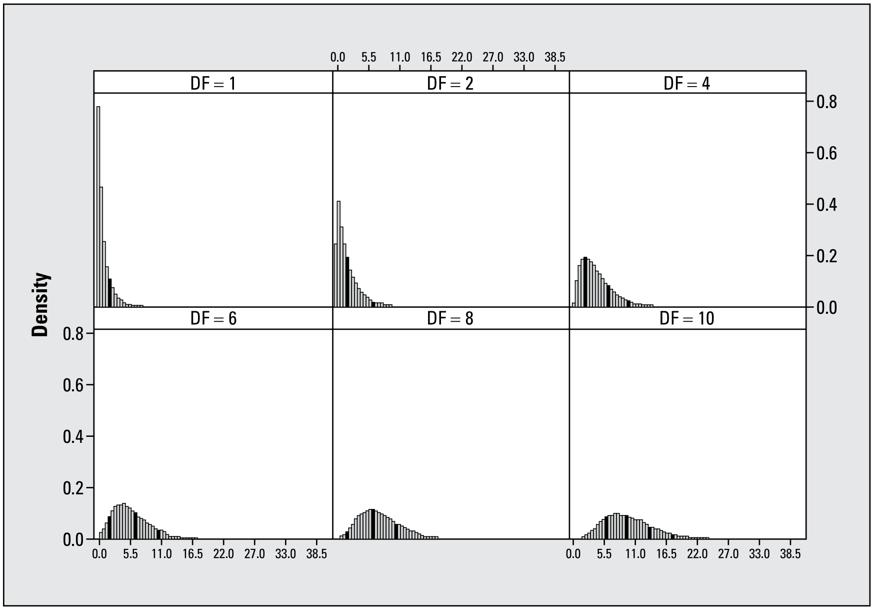


FIGURE 15-2: Chi-square distributions with 1, 2, 4, 6, 8, and 10 degrees of freedom (moving from upper left to lower right).

TIP

How small of a probability do you need to reject H_0? For most hypothesis tests, statisticians generally use 0.05 as the cutoff. (For more information on cutoff values, also known as α levels, flip to Chapter 4, or check out my other book, *Statistics For Dummies,* 2nd Edition.)

Your job now is to find the probability of being beyond your Chi-square test statistic on the corresponding Chi-square distribution with $(r-1)*(c-1)$ degrees of freedom. Each Chi-square distribution is different, and because the number of possible degrees of freedom is infinite, showing every single value of every Chi-square distribution isn't possible.

In the Chi-square table (Table A-3 in the Appendix), you see some of the most important values on each Chi-square distribution with degrees of freedom from 1 to 50.

To use the Chi-square table, you find the row that represents your degrees of freedom (abbreviated DF). Move across that row until you reach the value closest to your Chi-square test statistic, without going over. (It's like a game show where you're trying to win the showcase by guessing the price.)

Then go to the top of the column you're in. That number represents the area to the right of (above) the Chi-square test statistic you saw in the table. The area above your particular Chi-square test statistic is less than or equal to this number. This result is the approximate p-value of your Chi-square test.

In the house paint color preference example (see Figure 15-1), the Chi-square test statistic is 14.27. You have $(2-1)*(2-1)=1$ degree of freedom. In the Chi-square table, go to the row for $\mathrm{DF}=1$, and go across to the number closest to 14.27 (without going over), which is 7.88.

Drawing your conclusions

You have two alternative ways to draw conclusions from the Chi-square test statistic. You can look up your test statistic on the Chi-square table and see the probability of being greater than that. This method is known as *approximating the p-value.* (The *p*-value of a test statistic is the probability of being at or beyond your test statistic on the distribution to which the test statistic is being compared — if the null hypothesis were true (in this case, the Chi-square distribution.) Or you can have the computer calculate the exact *p*-value for your test. (For a quick review of *p*-values and α levels, turn to Chapter 4. For a full review of these topics, see my other book, *Statistics For Dummies*, 2nd Edition.)

Before you do anything though, set your α, the cutoff probability for your *p*-value, in advance. If your *p*-value is less than your α level, reject H_o. If it's more, you can't reject H_o.

Approximating p-value from the table

For the house paint color preference example (see Figure 15-1), the Chi-square test statistic is 14.27 with $(2-1)*(2-1)=1$ DF (degree of freedom). The closest number in row one of the Chi-square table (see the Appendix), without going over, is 7.88 (in the last column).

The number at the top of that column is 0.005. This number is less than your typical α level of 0.05, so you reject H_0. You know that your *p*-value is less than 0.005 because your test statistic was more than 7.88. In other words, if 7.88 is the minimum evidence you need to reject H_0, you have more evidence than that with a value of 14.28. More evidence against H_0 means a smaller *p*-value.

However, because Chi-square tables in general only give a few values for each Chi-square distribution, the best you can say using this table is that your *p*-value for this test is less than 0.005.

Here's the big news: Because your *p*-value is less than 0.05, you can conclude based on this data that gender and house paint color preference are likely to be related in the population (dependent), like the American Demographics Survey said (quoted at the beginning of this chapter). Only now, you have a formal statistical analysis that says this result found in the sample is also likely to occur in the entire population. This statement is much stronger!

REMEMBER

If your data shows you can reject H_o, you only know at that point that the two variables have some relationship. The Chi-square test statistic doesn't tell you what that relationship is. In order to explore the relationship between the two variables, you find the conditional probabilities in your two-way table (see Chapter 14). You can use those results to give you some ideas as to what may be happening in the population.

For the gender and house paint color preference example, because paint color preference is related to gender, you can examine the relationship further by comparing the male versus female paint color preferences and describing how they're different. Start by finding the percentage of men that prefer white houses, which comes out to $180 \div 500 = 0.36$, or 36 percent, calculated from Table 15-1. Now compare this result to the percentage of women who prefer white houses: $125 \div 500 = 0.25$, or 25 percent. You can now conclude that in this population (not just the sample), men prefer white houses more than women do. Hence, gender and house paint color preference are dependent.

REMEMBER

Dependent variables affect each other's outcomes, or cell counts. If the cell counts you actually observe from the sample data won't match the expected cell counts under H_o: The variables are independent, you conclude that the dependence relationship you found in the sample data carries over to the population. In other words, big differences between observed and expected cell counts mean that the variables are dependent.

Extracting the p-value from computer output

After Minitab calculates the test statistic for you, it reports the exact *p*-value for your hypothesis test. The *p*-value measures the likelihood that your results were found just by chance while H_o is still true. It tells you how much strength you have against H_o. If the *p*-value is 0.001, for example, you have much more strength against H_o than if the *p*-value is, say, 0.10.

Looking at the Minitab output for the gender-paint color preference data in Figure 15-1, the *p*-value is reported to be 0.000. This means that the *p*-value is smaller than 0.0005; for example, it may be 0.00009. That's a very small *p*-value! (Minitab only reports results to three decimal points, which is typical of many statistical software packages.)

WARNING

I've seen situations where people get a result that isn't quite what they want (like a *p*-value of 0.068), and so they do some tweaking to get what they want. They change their α level from 0.05 to 0.10 after the fact. This change makes the *p*-value less than the α level, and they feel they can reject H_o and say that a relationship exists.

But what's wrong with this picture? They changed the α after they looked at the data, which isn't allowed. That's like changing your bet in blackjack after you find out what the dealer's cards look like. (Tempting, but a serious no-no.) Always be wary of large α levels, and make sure that you always choose your α before collecting any data — and stick to it.

The good news is that when *p*-values are reported, anyone reading them can make their own conclusion; no cut-and-dried rejection and acceptance region is set in stone. But setting an α level once and then changing it after the fact to get a better conclusion is never good!

Putting the Chi-square to the test

If two variables turn out to be dependent, you can describe the relationship between them. But if two variables are independent, the results are the same for each group being compared. The following example illustrates this idea.

There has been much speculation and debate as to whether cellphone use should be banned while driving. You're interested in Americans' opinions on this issue, but you also suspect that the results may differ by age group. You decide to do a Chi-square test for independence to see if your theory plays out. Table 15-3 is a two-way table of observed data from 60 adults and 60 teenagers regarding whether they agree with the policy (banning cellphone use while driving) or not. From Table 15-3 you see that $12 \div 60 = 20$ percent of adults agree with the policy of banning cellphones while driving, compared to only $9 \div 60 = 15$ percent of teenagers. You see these percentages are different, but is this enough to say that age group and opinion on this issue are dependent? Only a Chi-square test for independence can help you decide.

TABLE 15-3 Age Group and Opinion on Cellphone Ban: Observed Cell Counts

	Agree with Cellphone Ban	Disagree with Cellphone Ban	**Marginal Row Totals**
Adults	12	48	60
Teenagers	9	51	60
Marginal Column Totals	21	99	120 **(Grand Total)**

Table 15-4 shows the expected cell counts under H_o, along with their calculations.

TABLE 15-4 Age Group and Opinion on Cellphone Ban: Expected Cell Counts

	Agree with Cellphone Ban	Disagree with Cellphone Ban	**Marginal Row Totals**
Adults	$(60*21)\div 120 = 10.5$	$(60*99)\div 120 = 49.5$	60
Teenagers	$(60*21)\div 120 = 10.5$	$(60*99)\div 120 = 49.5$	60
Marginal Column Totals	21	99	120 **(Grand Total)**

Running a Chi-square test in Minitab for this data, the degrees of freedom equals $(2-1)*(2-1)=1$; the Chi-square test statistic can be shown to be equal to 0.519, and the *p*-value is 0.471. Because the *p*-value is greater than 0.05 (the typical cut-off), you can't reject H_o; therefore, you conclude that age group and opinion on the banning of cellphones while driving are independent and therefore not related. Your theory that age group had something to do with it just doesn't pan out; there's not sufficient evidence for it.

REMEMBER

In general, *independence* means that you can find no major difference in the way the rows look as you move down a column. Put another way, the proportion of the data falling into each column across the row is about the same for each row. Because Table 15-4 has the same number of adults as teenagers, the row totals are the same, and you get the same expected cell counts for adults and teens in both the Agree column (10.5) and the Disagree column (49.5).

Comparing Two Tests for Comparing Two Proportions

You can also use the Chi-square test to check whether two population proportions are equal. For example, is the proportion of teenage cellphone users who have the latest cell phone the same as the proportion of adult cellphone users who have the latest cell phone? Or are the kids ahead of the game?

You may be thinking, "But wait a minute, don't statisticians already have a test for two proportions? I seem to remember it from my Stats I course . . . I'm thinking . . . yeah, it's the *Z*-test for two proportions. What's that test got to do with a Chi-square test?" In this section, you get an answer to that question and practice using both methods to investigate a possible age gap in cellphone use.

Getting reacquainted with the Z-test for two population proportions

The way that most people figure out how to test the equality of two population proportions is to use a *Z-test for two population proportions.* With this test, you collect a random sample from each of the two populations, find and subtract their two sample proportions, and divide by their pooled standard error (see your Stats I textbook for details on this particular test).

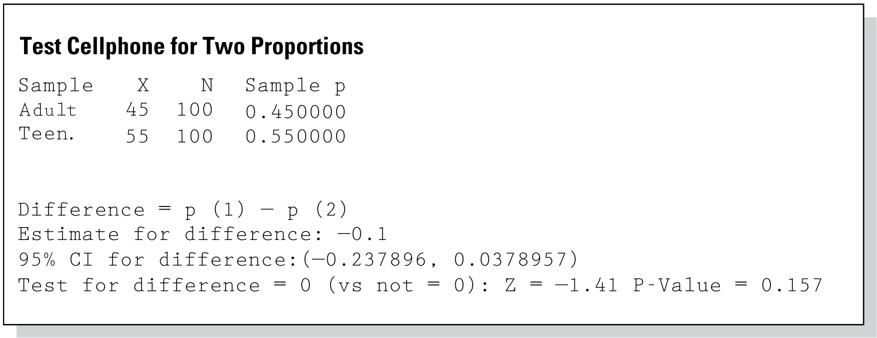
This test is possible to do as long as the sample sizes from the two populations are large — at least five successes and five failures in each sample.

The null hypothesis for the Z-test for two population proportions is H_o: $p_1 = p_2$, where p_1 is the proportion of the first population that falls into the category of interest, and p_2 is the proportion of the second population that falls into the category of interest. And as always, the alternative hypothesis is one of the following choices, H_a: Not equal to, greater than, or less than.

Suppose you want to compare the proportion of adult versus teenage cellphone users to see who is more likely to own the latest cellphone, where p_1 is the proportion of adults who own the latest cellphone, and p_2 is the proportion of all teenagers who own the latest cellphone. You collect data, find the sample proportions from each group, take their difference, and make a Z-statistic out of it using the formula $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$.

Here, x_1 and x_2 are the number of individuals from samples one and two, respectively, with the desired characteristic; n_1 and n_2 are the two sample sizes.

Suppose that you collect data on 100 adults and 100 teenagers and find 45 adult cellphone owners have the latest cellphone and 55 teen cellphone owners have the latest cellphone. This means that $\hat{p}_1$ equals $45 \div 100 = 0.45$, and $\hat{p}_2$ equals $55 \div 100 = 0.55$. Your samples have at least five successes (having the desired characteristic, in this case, latest cellphone ownership) and five failures (not having the desired characteristic, which is latest cellphone ownership). So you compute the Z-statistic for comparing the two population proportions (adults versus teens) based on this data; it's −1.41, as shown on the last line of the Minitab output in Figure 15-3.



```
Test Cellphone for Two Proportions
Sample    X    N   Sample p
Adult    45  100   0.450000
Teen.    55  100   0.550000


Difference = p (1) — p (2)
Estimate for difference: —0.1
95% CI for difference:(—0.237896, 0.0378957)
Test for difference = 0 (vs not = 0): Z = —1.41 P-Value = 0.157
```

FIGURE 15-3: Minitab output comparing the proportion of adult and teenage owners of the latest cellphone.

The p-value for the test statistic of $Z = -1.41$ is 0.157 (calculated by Minitab, or by looking at the area below the Z-value of -1.41 on the Z-table, which is found in the Appendix). This p-value (0.157) is greater than the typical α level (predetermined cutoff) of 0.05, so you can't reject H_o. You can't say that the two population proportions aren't equal, so you must conclude that the proportion of adult cellphone owners having the latest cellphone is not statistically different than teenagers. (It's nice when the adults can keep up with the kids for once!)

Even though the sample seemed to have evidence for a difference (after all, 45 percent isn't equal to 55 percent), you don't have enough evidence in the data to say that this same difference carries over to the population. So you can't lay claim to an age gap in owning the latest cellphone, at least not with this sample.

COMPUTER OUTPUT

To run this hypothesis test for two proportions in Minitab, go to Stat>Basic Statistics>Two Proportions, and in the pull-down menu, select summarized data. For each sample, type in how many "events" (individuals with the characteristic of interest) occurred, as well as the number of trials (total individuals in the group). Click OK.

Equating Chi-square tests and Z-tests for a two-by-two table

Here's the key to relating the Z-test to a Chi-square test for independence. The Z-test for two proportions and the Chi-square test for independence in a two-by-two table (one with two rows and two columns) are equivalent if the sample sizes from the two populations are large enough — that is, when the number of successes and the number of failures in each cell of the two samples is at least five.

If you use the Z-test to see whether the proportion of adult cellphone owners with the latest cell phones is equal to the proportion of teen latest cellphone owners, you're really looking at whether you can expect the same proportion of latest cellphone owners despite age group (after you take the sample sizes into account). And that means you're testing whether age group (adult or teen) is independent of latest cellphone ownership (yes or no).

If the proportion of teen latest cellphone owners equals the proportion of adult latest cellphone owners, the proportion of latest cellphone owners is the same regardless of age group, so age group and latest cellphone ownership are independent. On the other hand, if you find the proportion of adult latest cellphone owners to be unequal to the proportion of teenage latest cellphone owners, you can say that latest cellphone ownership differs by age group, so age group and latest cellphone ownership are dependent.

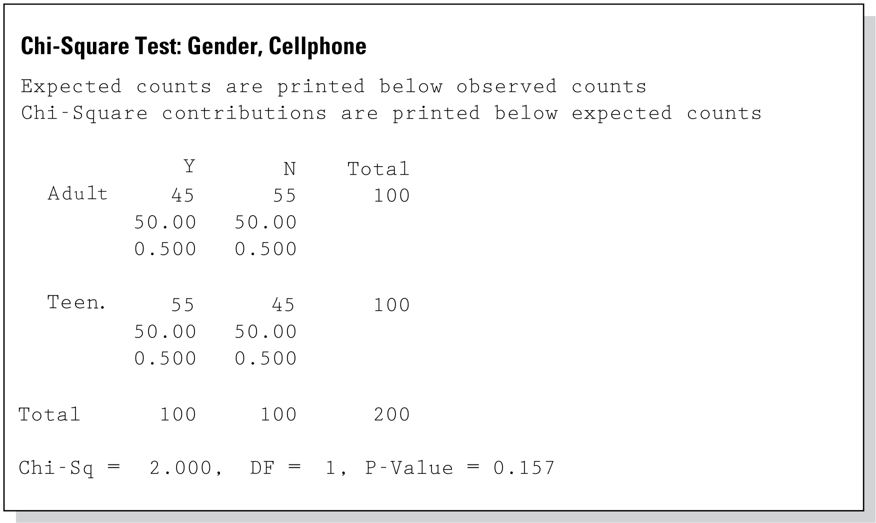
With the cellphone data, you have 45 adults using the latest cellphones (out of 100 adult cellphone owners) and 55 teenagers using the latest cellphones (out of 100 teenage cellphone owners). The Minitab output for the Chi-square test for independence (complete with observed and expected cell counts, degrees of freedom, test statistic, and p-value) is shown in Figure 15-4. The p-value for this test is 0.157, which is greater than the typical α level 0.05, so you can't reject H_o.

Because the Chi-square test for independence and the Z-test are equivalent when you have a two-by-two table, the p-value from the Chi-square test for independence is identical to the p-value from the Z-test for two proportions. If you compare the p-values from Figures 15-3 and 15-4, you can see that for yourself.

Also, note that if you take the Z-test statistic for this example (from Figure 15-3), which is -1.41, and square it, you get 2.00, which is equal to the Chi-square test statistic for the same data (last line of Figure 15-4). It's also the case that the square of the Z-test statistic (when testing for the equality of two proportions) is equal to the corresponding Chi-square test statistic for independence.

REMEMBER

The Chi-square test and Z-test are equivalent only if the table is a two-by-two table (two rows and two columns) and if the Z-test is two-tailed (the alternative hypothesis is that the two proportions aren't equal, instead of using H_a: One proportion is greater than or less than the other). If the Z-test isn't two-tailed, a Chi-square test isn't appropriate. If the two-way table has more than two rows or columns, use the Chi-square test for independence (because many categories mean you no longer have only two proportions, so the Z-test isn't applicable).



Chi-Square Test: Gender, Cellphone

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

               Y        N    Total
  Adult       45       55      100
           50.00    50.00
           0.500    0.500

  Teen.       55       45      100
           50.00    50.00
           0.500    0.500

Total        100      100      200

Chi-Sq =  2.000,  DF =  1, P-Value = 0.157
```

FIGURE 15-4: Minitab output testing independence of age group and latest cellphone ownership.

THE CAR ACCIDENT-CELLPHONE CONNECTION

Researchers are doing a great deal of study of the effects of cellphone use while driving. One study published in the *New England Journal of Medicine* observed and recorded data in 1997 on 699 drivers who had cellphones and were involved in motor vehicle collisions resulting in substantial property damage but no personal injury. Each person's cellphone calls on the day of the collision and during the previous week were analyzed through the use of detailed billing records. A total of 26,798 cellphone calls were made during the 14-month study period.

One conclusion the researchers made was that ". . . the risk of a collision when using a cellphone is four times higher than the risk of a collision when a cellphone was not being used." They basically conducted a Chi-square test to see whether cellphone use and having a collision are independent, and when they found out the events were not, the researchers were able to examine the relationship further using appropriate ratios. In particular, they found that the risk of a collision is four times higher for those drivers using cellphones than for those who aren't.

Researchers also found out that the relative risk was similar for drivers who differed in personal characteristics, such as age and driving experience. (This finding means that they conducted similar tests to see whether the results were the same for drivers of different age groups and drivers of different levels of experience, and the results always came out about the same. Therefore, age and the experience of the driver weren't related to the collision outcome.)

The research also shows that ". . . calls made close to the time of the collision were found to be particularly hazardous $(p < 0.001)$. Hands-free cellphones offered no safety advantage over hand-held units (*p*-value not significant)." ***Note:*** The items in parentheses show the typical way that researchers report their results: using *p*-values. The *p* in both cases of parentheses represents the *p*-value of each test.

In the first case, the *p*-value is very tiny, less than 0.001, indicating strong evidence for a relationship between collisions and cellphone use at the time. The second *p*-value in parentheses was stated to be insignificant, meaning that it was substantially more than 0.05, the usual α level. This second result indicates that using hands-free equipment didn't affect the chances of a collision happening; the proportion of collisions using hands-free cellphones versus using regular cellphones were found to be statistically the same (they could have easily occurred by chance under independence). Whether you use a regular or hands-free cellphone, this study may provide you with good information!

IN THIS CHAPTER

- Understanding what goodness-of-fit really means
- Using the Chi-square model to test for goodness-of-fit
- Looking at the conditions for goodness-of-fit tests

Chapter 16

Using Chi-Square Tests for Goodness-of-Fit (Your Data, Not Your Jeans)

Many phenomena in life may appear to be haphazard in the short term, but they actually occur according to some preconceived, preselected, or predestined model over the long term. For example, even though you don't know whether it will rain tomorrow, your local meteorologist can give you their model for the percentage of days that it rains, snows, is sunny, or cloudy, based on the last five years. Whether or not this model is still relevant this year is anyone's guess, but it's a model nonetheless. As another example, a biologist can produce a model for predicting the number of goslings raised by a pair of geese per year, even though you have no idea what the pair in your backyard will do. Is his model correct? Here's your chance to find out.

In this chapter, you build models for the proportion of outcomes that fall into each category for a categorical variable. You then test these models by collecting data and comparing what you observe in your data to what you expect from the model. You do this evaluation through a goodness-of-fit test that's based on the Chi-square distribution. In a way, a goodness-of-fit test is likened to a reality check of a model for categorical data.

Finding the Goodness-of-Fit Statistic

The general idea of a *goodness-of-fit* procedure involves determining what you expect to find and comparing it to what you actually observe in your own sample through the use of a test statistic. This test statistic is called the *goodness-of-fit test statistic* because it measures how well your model (what you expected) fits your actual data (what you observed).

In this section, you see how to figure out the numbers that you should expect in each category given your proposed model, and you also see how to put those expected values together with your observed values to form the goodness-of-fit test statistic.

What's observed versus what's expected

For an example of something that can be observed versus what's expected, look no further than a bag of tasty M&M'S Milk Chocolate Candies. A ton of different kinds of M&M'S are out there, and each kind has its own variation of colors and tastes. For this study, any reference I give to M&M'S is to the original milk chocolate candy — my favorite.

The percentage of each color of M&M'S that appear in a bag is something Mars (the company that makes M&M'S) spends a lot of time thinking about. Mars wants specific percentages of each color in its bags of M&M'S, which it determines through comprehensive marketing research based on what people like and want to see. Mars then posts its current percentages for each color of M&M'S on its website. Table 16-1 shows the percentage of M&M'S of each color as of this writing.

TABLE 16-1 **Expected Percentage of Each Color of M&M'S Milk Chocolate Candies**

Color	Percentage
Brown	13%
Yellow	14%
Red	13%
Cyan Blue	24%
Orange	20%
Green	16%

Now that you know what to expect from a bag of M&M'S, the next question is, how does Mars deliver? If you were to open a bag of M&M'S right now, would you get the percentages of each color that you're supposed to get? You know from your previous studies in statistics that sample results vary (for a quick review of this idea, see Chapter 4). So you can't expect each bag of M&M'S to have exactly the correct number of each color of M&M'S as listed in Table 16-1. However, in order to keep customers happy, Mars should get close to the expectations. How can you determine how close the company does get?

Table 16-1 tells you what percentages are expected to fall into each category in the entire population of all M&M'S (that means every single M&M'S Milk Chocolate Candy that's currently being made). This set of percentages is called the *expected model* for the data. You want to see whether the percentages in the expected model are actually occurring in the packages you buy. To start this process, you can take a sample of M&M'S (after all, you can't check every single one in the population) and make a table showing what percentage of each color you observe. Then you can compare this table of observed percentages to the expected model.

TIP

Some expected percentages are known, as they are for the M&M'S, or you can figure them out by using math techniques. For example, if you're examining a single die to determine whether or not it's a fair die, you know that if the die is fair, you should expect $\frac{1}{6}$ of the outcomes to fall into each category of 1, 2, 3, 4, 5, and 6.

As an example, I examined one 1.69-ounce bag of plain, milk-chocolate M&M'S (tough job, but someone had to do it), and you can see my results in Table 16-2, column two. (Think of this bag as a random sample of 56 M&M'S, even though it's not technically the same as reaching into a silo filled with M&M'S and pulling out a true random sample of 1.69 ounces. For the sake of argument, one bag is okay.)

Compare what I observed in my sample (column two of Table 16-2) to what I expected to get (column three of Table 16-2). Notice that I observed a lower percentage of brown and red M&M'S than expected and a lower percentage of blues than expected. I also observed a higher percentage of yellow, orange, and green M&M'S than expected. Sample results vary by random chance, from sample to sample, and the difference I observed may just be due to this chance variation. But could the differences indicate that the expected percentages reported by Mars aren't being followed?

It stands to reason that if the differences between what you observed and what you expected are small, you should attribute those differences to chance and let the expected model stand. On the other hand, if the differences between what you observed and what you expected are large enough, you may have enough evidence to indicate that the expected model has some problems. How do you know which conclusion to make? The operative phrase is, "if the differences are large enough."

You need to quantify this term *large enough.* Doing so takes a bit more machinery, which I cover in the next section.

TABLE 16-2 **Percentage of M&M'S Observed in One Bag (1.69 oz.) versus Percentage Expected**

Color	Percentage Observed	Percentage Expected
Brown	$4/56 = 7.14$	13.00
Yellow	$10/56 = 17.86$	14.00
Red	$4/56 = 7.14$	13.00
Blue	$10/56 = 17.86$	24.00
Orange	$15/56 = 26.79$	20.00
Green	$13/56 = 23.21$	16.00
TOTAL	100.00	100.00

Calculating the goodness-of-fit statistic

The goodness-of-fit statistic is one number that puts together the total amount of difference between what you expect in each cell compared to the number you observe. The term *cell* is used to express each individual category within a table format. With the M&M'S example, the first columns of Tables 16-1 and 16-2 contain six cells, one for each color of M&M'S. For any cell, the number of items you observe in that cell is called the *observed cell count.* The number of items you expect in that cell (under the given model) is called the *expected cell count.* You get the expected cell count by multiplying the expected cell percentage by the sample size.

The expected cell count is just a proportion of the total, so it doesn't have to be a whole number. For example, if you roll a fair die 200 times, you should expect to roll ones $\frac{1}{6}$, or 16.67 percent, of the time. In terms of the number of ones you expect, it should be $0.1667 * 200 = 33.33$. Use the 33.33 in your calculations for goodness-of-fit; don't round to a whole number. Your final answer is more accurate that way.

REMEMBER

The reason the goodness-of-fit statistic is based on the *number* in each cell rather than the *percentage* in each cell is because percents are a bit deceiving. If you know that 8 out of 10 people support a certain view, that's 80 percent. But 80 out of 100 is also 80 percent. Which one would you feel is a more-precise statistic? The 80 out of 100 percent because it uses more information. Using percents alone disregards the sample size. Using the counts (the number in each group) keeps track of the amount of precision you have.

For example, if you roll a fair die, you expect the percentage of ones to be $\frac{1}{6}$. If you roll that fair die 600 times, the expected number of ones will be $\frac{1}{6} * 600 = 100$. That number (100) is the expected cell count for the cell that represents the outcome of one. If you roll this die 600 times and get 95 ones, then 95 is the observed cell count for that cell. The formula for the goodness-of-fit statistic is given by the following: $\sum_{all\ cells} \frac{(O-E)^2}{E}$, where E is the expected number in a cell and O is the observed number in a cell. The steps for this calculation are as follows:

1. **For the first cell, find the expected number for that cell (*E*) by multiplying the percentage expected in that cell by the sample size.**
2. **Take the observed value in the first cell (*O*) minus the number of items that are expected in that cell (*E*).**
3. **Square that difference.**
4. **Divide the answer by the number that's expected in that cell (*E*).**
5. **Repeat Steps 1 through 4 for each cell.**
6. **Add up the results to get the goodness-of-fit statistic.**

REMEMBER

The reason you divide by the expected cell count in the goodness-of-fit statistic (Step 4) is to take into account the magnitude of any differences you find. For example, if you expect 100 items to fall in a certain cell and you get 95, the difference is 5. But in terms of a percentage, this difference is only $\frac{5}{100} = 5$ percent. However, if you expect 10 items to fall into that cell and you observe 5 items, the difference is still 5, but in terms of a percentage, it's $\frac{5}{10} = 50$ percent. This difference is much larger in terms of its impact. The goodness-of-fit statistic operates much like a percentage difference. The only added element is to square the difference to make it positive. (That's done because whether you expect 10 and get 15 or expect 10 and get 5 makes no difference to others; you're still off by 50 percent.)

Table 16-3 shows the step-by-step calculation of the goodness-of-fit statistic for the M&M'S example, where O indicates observed cell counts and E indicates expected cell counts. To get the expected cell counts, you take the expected percentages shown in Table 16-1 and multiply by 56 because 56 is the number of M&M'S I had in my sample. The observed cell counts are the ones found in my sample, shown in Table 16-2.

TABLE 16-3 Goodness-of-Fit Statistic for M&M'S Example

Color	O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Brown	4	$0.13 * 56 = 7.28$	$4 - 7.28 = -3.28$	10.76	1.48
Yellow	10	$0.14 * 56 = 7.84$	$10 - 7.84 = 2.16$	4.67	0.60
Red	4	$0.13 * 56 = 7.28$	$4 - 7.28 = -3.28$	10.76	1.48
Blue	10	$0.24 * 56 = 13.44$	$10 - 13.44 = -3.44$	11.83	0.88
Orange	15	$0.20 * 56 = 11.20$	$15 - 11.20 = 3.80$	14.44	1.29
Green	13	$0.16 * 56 = 8.96$	$13 - 8.96 = 4.04$	16.32	1.82
TOTAL	56	56			**7.55**

The goodness-of-fit statistic for the M&M'S example turns out to be 7.55, the bolded number in the lower-right corner of Table 16-3. This number represents the total squared difference between what I expected and what I observed, adjusted for the magnitude of each expected cell count. The next question is how to interpret this value of 7.55. Is it large enough to indicate that colors of M&M'S in the bag aren't following the percentages posted by Mars? The next section addresses how to make sense of these results.

Interpreting the Goodness-of-Fit Statistic Using a Chi-Square

After you get your goodness-of-fit statistic, your next job is to interpret it. To do this, you need to figure out the possible values you could have gotten and where your statistic fits in among them. You can accomplish this task with a Chi-square goodness-of-fit test.

The values of a goodness-of-fit statistic actually follow a Chi-square distribution with $k-1$ degrees of freedom, where k is the number of categories in your particular population (see Chapter 15 for full details on the Chi-square). You use the

Chi-square table (Table A-3 in the Appendix) to find the *p*-value of your Chi-square test statistic.

If your Chi-square goodness-of-fit statistic is large enough, you conclude that the original model doesn't fit and you have to chuck it; there's too much of a difference between what you observed and what you expected under the model. However, if your goodness-of-fit statistic is relatively small, you don't reject the model. (What constitutes a large or small value of a Chi-square test statistic depends on the degrees of freedom.)

TIP

The goodness-of-fit statistic follows the main characteristics of the Chi-square distribution. The smallest-possible value of the goodness-of-fit statistic is zero. Continuing the example from the previous section, if the M&M'S in my sample followed the exact percentages found in Table 16-1, the goodness-of-fit statistic would be zero. That's because the observed counts and the expected counts would be the same, so the values of the observed cell count minus the expected cell count would all be zero.

The largest-possible value of Chi-square isn't specified, although some values are more likely to occur than others. Each Chi-square distribution has its own set of likely values, as you can see in Figure 16-1. This figure shows a simulated Chi-square distribution with $6-1=5$ degrees of freedom (relevant to the M&M'S example). It basically gives a breakdown of all the possible values you could have for the goodness-of-fit statistic in this situation and how often they occur. You can see in Figure 16-1 that a Chi-square test statistic of 7.55 isn't unusually high, indicating that the model for M&M'S colors probably can't be rejected. However, more particulars are needed before you can formally make that conclusion.

Checking the conditions before you start

Every statistical technique seems to have a catch, and this case is no exception. In order to use the Chi-square distribution to interpret your goodness-of-fit statistic, you have to be sure you have enough information to work with in each cell. The stats gurus usually recommend that the expected count for each cell turn out to be greater than or equal to five. If it doesn't, one option is to combine categories to increase the numbers.

In the M&M'S example, the expected cell counts are all above seven (see Table 16-3), so the conditions are met. If this weren't the case, you should have taken a larger sample size, because you calculate the expected cell counts by multiplying the expected percentage in that cell by the sample size. If you increase the sample size, you increase the expected cell count. A higher sample size also increases your chances of detecting a real deviation from the model. This idea is related to the power of the test (see Chapter 4 for information on power).

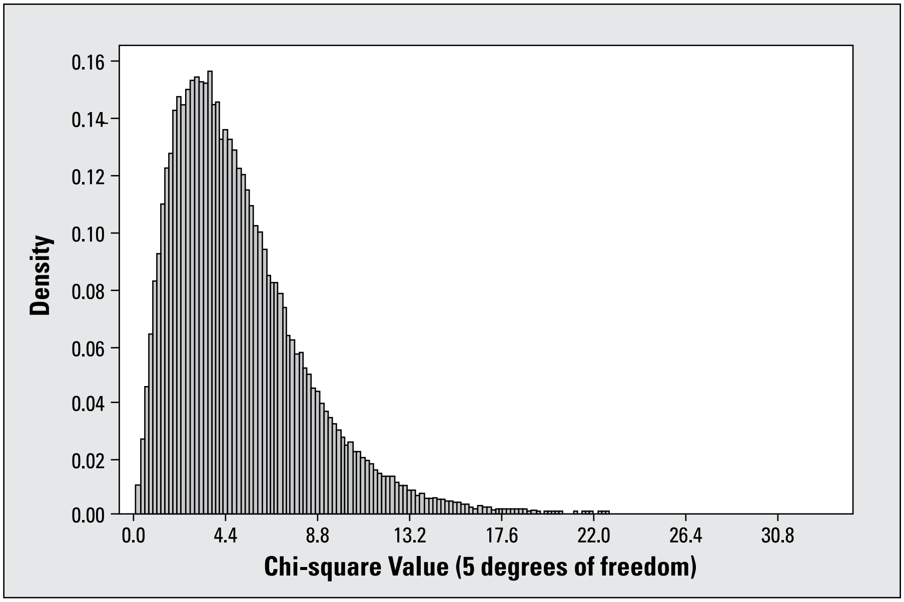


FIGURE 16-1: Chi-square distribution with 5 degrees of freedom.

WARNING

After you collect your data, it's not right to go back and take a new and larger sample. It's best to set up the appropriate sample size ahead of time, and you can do this by determining what sample size you need to get the expected cell counts to be at least five. For example, if you roll a fair die, you expect $\frac{1}{6}$ of the outcomes to be ones. If you only take a sample of six rolls, you have an expected cell count of $\frac{1}{6} * 6 = 1$, which isn't enough. However, if you roll the die 30 times, your expected cell count is $\frac{1}{6} * 30 = 5$, which is just enough to meet the condition.

The steps of the Chi-square goodness-of-fit test

Assuming the necessary condition is met (see the previous section), you can get down to actually conducting a formal goodness-of-fit test.

The general version of the null hypothesis for the goodness-of-fit test is H_o: The model holds for all categories; versus the alternative hypothesis H_a: The model doesn't hold for at least one category. Each situation will dictate what proportions should be listed in H_o for each category. For example, if you're rolling a fair die, you have H_o: Proportion of ones $= 1s = \frac{1}{6}$; proportion of twos $= 2s = \frac{1}{6}$; . . .; proportion of sixes $= 6s = \frac{1}{6}$.

Following are the general steps for the Chi-square goodness-of-fit test, with the M&M'S example illustrating how you can carry out each step:

1. **Write down H_o using the percentages that you expect in your model for each category.**

   Using a subscript to indicate the proportion (*p*) of M&M'S you expect to fall into each category (see Table 16-1), your null hypothesis is H_o: $p_{\text{brown}} = 0.13$, $p_{\text{yellow}} = 0.14$, $p_{\text{red}} = 0.13$, $p_{\text{blue}} = 0.24$, $p_{\text{orange}} = 0.20$, and $p_{\text{green}} = 0.16$. All these proportions must hold in order for the model to be upheld.

2. **Write your H_a: This model doesn't hold for at least one of the percentages.**

   Your alternative hypothesis, H_a, in this case, would be: One (or more) of the probabilities given in H_o isn't correct. In other words, you conclude that at least one of the colors of M&M'S has a different proportion than what's stated in the model.

3. **Calculate the goodness-of-fit statistic using the steps in the earlier section, "**Calculating the goodness-of-fit statistic."

   The goodness-of-fit statistic for M&M'S, from the earlier section, is 7.55. As a reminder, you take the observed number in each cell minus the expected number in that cell, square it, and divide by the expected number in that cell. Do that for every cell in the table and add up the results. For the M&M'S example, that total is equal to 7.55, the goodness-of-fit statistic.

4. **Look up the Chi-square distribution with $k-1$** degrees of freedom, where k is the number of categories you have.

   You compare this statistic (7.55) to the Chi-square distribution with $6-1=5$ degrees of freedom (because you have $k=6$ possible colors of M&M'S). (See Table A-3 in the Appendix.)

   Looking at Figure 16-1, you can see that the value of 7.55 is nowhere near the high end of this distribution, so you likely don't have enough evidence to reject the model provided by Mars for M&M'S colors.

5. **Find the *p*-value of your goodness-of-fit statistic.**

   You use a Chi-square table to find the *p*-value of your test statistic (see Table A-3 in the Appendix). (For more information on the Chi-square distribution, refer to Chapter 15.)

   REMEMBER

   Because the Chi-square table can only list a certain number of results for each of the degrees of freedom, the exact *p*-value for your test statistic may fall between two *p*-values listed on the table.

To find the *p*-value for the test statistic in the M&M'S example (7.55), find the row for 5 degrees of freedom on the Chi-square table (Table A-3 in the Appendix) and look at the numbers (the degrees of freedom is $k-1=6-1=5$, where *k* is the number of categories). You see that the number 7.55 is less than the first value in the row (9.24), which has a *p*-value of 0.10. (Find the *p*-value by looking at the column heading above the number.) So the *p*-value for 7.55, which is the area to the right of 7.55 in Figure 16-1, must be greater than 0.10, because 7.55 is to the left of 9.24 on that Chi-square distribution.

TIP

Many computer programs exist (online or via a graphing calculator) that will find exact *p*-values for a Chi-square test, saving time and headaches when you have access to them (the technology, not the headaches). Using one such online *p*-value calculator, I found that the exact *p*-value for the goodness-of-fit test for the M&M'S example (test statistic 7.55 with 5 degrees of freedom for Chi-square) is 0.1828. To find online *p*-value calculators, simply type the name of the distribution and the word *p-value* in an Internet search engine. For this example search, you would type **Chi-square p-value**.

6. **If your *p*-value is less than your predetermined cutoff (α), reject H_o; the model doesn't hold. If your *p*-value is greater than α, you can't reject the model.**

   A typical value of α is 0.05. Some data analysts may use a higher value (up to 0.10), and others may go lower (for example, 0.010). See Chapter 4 for more information on choosing α and comparing your *p*-value to it.

   Going again to the M&M'S example, the *p*-value, 0.18, is greater than 0.05, so you fail to reject H_o. You can't say the model is wrong. So, Mars does appear to deliver on the percentages of M&M'S of each color as advertised. At least, you can't say it doesn't. (I'm sure Mars already knew that.)

COMPUTER OUTPUT

To run a Chi-square goodness-of-fit test in Minitab, enter the category names (M&M'S colors) in column one, the observed counts (number of M&M'S of each color) in column two, and the expected percents of each color (from Table 16-1) in column three. Then go to Stat>Tables>Chi-Square goodness-of-fit test. Click on Observed Counts and select column two (C2); in the Category Names box, select column one (C1); and under Test, click on Specific Proportions, and select column three (C3). (If you are in a situation where you want all the expected proportions to be the same, like a fair 6-sided die, just click that option; you don't need a column three in that case.) Then click OK.

I ran a Minitab Chi-Square Goodness-of-Fit analysis on the M&M'S data and got a nice graph of the observed versus expected values for each color in Figure 16-2, and the output in Figure 16-3. Subject to my round-off error, the result of 7.55 is the same as what is found in Table 16-3.

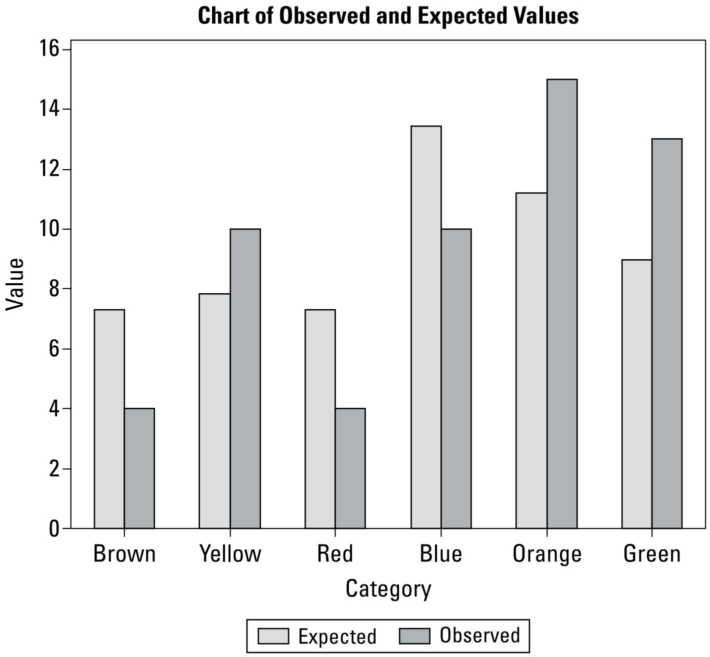
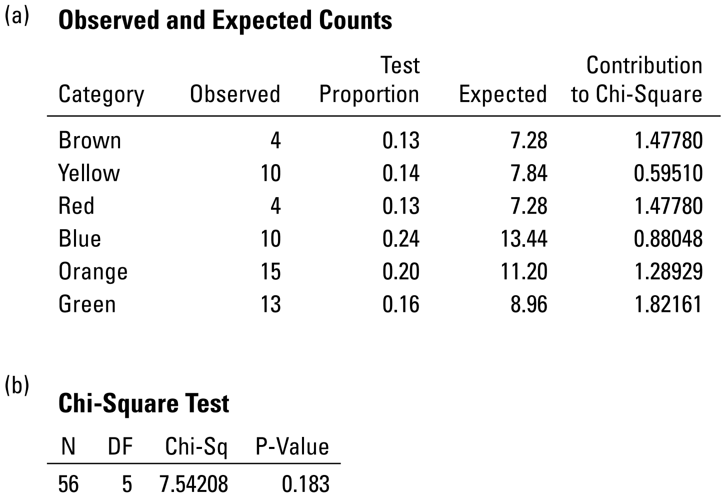


FIGURE 16-2: Graph of observed versus expected counts for the M&M'S colors data.



(a) **Observed and Expected Counts**

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
Brown	4	0.13	7.28	1.47780
Yellow	10	0.14	7.84	0.59510
Red	4	0.13	7.28	1.47780
Blue	10	0.24	13.44	0.88048
Orange	15	0.20	11.20	1.28929
Green	13	0.16	8.96	1.82161

(b) **Chi-Square Test**

N	DF	Chi-Sq	P-Value
56	5	7.54208	0.183

FIGURE 16-3: Minitab output of Chi-square goodness-of-fit test on M&M'S colors data.

REMEMBER

Although some hypothesis tests are two-sided tests, the goodness-of-fit test is always a *right-tailed test.* You're only looking at the right tail of the Chi-square distribution when you're doing a goodness-of-fit test. That's because a small value for the goodness-of-fit statistic means that the observed data and the expected model don't differ much, so you stick with the model. If the value of the goodness-of-fit statistic is way out on the right tail of the Chi-square distribution, however, that's a different story. That situation means the difference between what you observed and what you expected is larger than what you should get by chance, and therefore, you have enough evidence to say the expected model is wrong.

TIP

You use the Chi-square goodness-of-fit test to check to see whether a specified model fits. A *specified model* is a model in which each possible value of the variable x is listed, along with its associated probability according to the model. For example, if you want to test whether three local hospitals take in the same percentage of emergency room patients, you test H_o: $p_1 = p_2 = p_3$, where each p represents the percentage of ER patients going to each hospital, respectively. In this case, each p must equal 0.30 if the hospitals share the ER load equally.

IN THIS CHAPTER

- » Understanding the need for nonparametric techniques
- » Distinguishing regular methods from nonparametric methods
- » Laying the groundwork: The basics of nonparametric statistics

Chapter 17

Rebels Without a Distribution — Nonparametric Procedures

Many researchers do analyses involving hypothesis tests, confidence intervals, Chi-square tests, regression, and ANOVA. But nonparametric statistics doesn't seem to gain the same popularity as the other methods. It's more in the background — an unsung hero, if you will. However, nonparametric statistics is, in fact, a very important and very useful area of statistics because it gives you accurate results when other, more common methods fail.

In this chapter, you see the importance of nonparametric techniques and why they should have a prominent place in your data-analysis toolbox. You also discover some of the basic terms and techniques involved with nonparametric statistics.

Arguing for Nonparametric Statistics

Nonparametric statistics plays an important role in the world of data analysis in that it can save the day when you can't use other methods. The problem is that researchers often disregard, or don't even know about, nonparametric techniques and don't use them when they should. In that case, you never know what kind of results you get; what you do know is they could very well be wrong.

In the following sections, you see the advantages and the flexibility of using a nonparametric procedure. You also find out just how minimal the downside is, which makes it a win-win situation most of the time.

No need to fret if conditions aren't met

Many of the techniques that you typically use to analyze data, including many shown in this book, have one very strong condition on the data that must be met in order to use them: The populations from which your data are collected typically require a normal distribution (see Chapter 3). Methods requiring a certain type of distribution (such as a normal distribution) in order to use them are called *parametric* methods.

The following are ways to help you decide whether a population has a normal distribution, based on your sample:

- You can graph the data using a histogram, and see whether it appears to have a bell shape (a mound of data in the middle, trailing down on each side).

  COMPUTER OUTPUT

  To make a histogram in Minitab, enter your data into a column. Go to Graph>Histogram, click OK, then choose With Fit and Groups, and click OK. Click on your variable in the left-hand box, click Select, and it appears in the Graph Variables box. Click OK, and check out your histogram.

- You can make a normal probability plot, which compares your data to that of a normal distribution, using an x, y graph (similar to the ones used when you graph a straight line). If the data do follow a normal distribution, your normal probability plot will show a straight line. If the data don't follow a normal distribution, the normal probability plot won't show a straight line; it may show a curve off to one side or the other, for example.

  COMPUTER OUTPUT

  To make a normal probability plot in Minitab, enter your data in a column. Go to Graph>Probability Plot, and click OK. Click Distributions and make sure Normal is selected. Click on your variable in the left-hand column, click Select, and it appears in the Graph Variables column. Click OK, and you see your normal probability plot.

When you find that the normal distribution condition is clearly not met, that's where nonparametric methods come in. *Nonparametric methods* are those data-analysis techniques that don't require the data to have a specific distribution. Nonparametric procedures may require one of the following two conditions (and these are only in certain situations):

- The data come from a symmetric distribution (which looks the same on each side when you cut it down the middle).
- The data from two populations come from the same type of distribution (they have the same general shape).

Note also that the normal distribution centers solely on the mean as its main statistic (for example, the Z-value for the hypothesis test for one population mean is calculated by taking the data value, subtracting the mean, and dividing by the standard deviation). So the condition that the population has a normal distribution automatically says you're working with the mean. However, many nonparametric procedures work with the *median,* which is a much more flexible statistic because it isn't affected by *outliers* (extreme values either above or below the mean) or *skewness* (a peak on one side and a long tail on the other side), as the mean is.

The median's in the spotlight for a change

Often, a particular statistics question will revolve around the center of a population — that is, the number that represents a typical value, or a central value, in the population. One of those measures of center is the *mean.* The *population mean* is the average value over the entire population, which is typically not known (that's why you take a sample). Many data analysts focus heavily on the population mean; they want to estimate it, test it, compare the means of two or more populations, or predict the mean value of a y variable given an x variable. However, the mean isn't the only measure of the center of a population; you also have the good ol' median.

You may recall that the *median* of a data set is the value that represents the exact middle when you order the data from smallest to largest. For example, in the data set 1, 5, 4, 2, 3, you order the data to get 1, 2, 3, 4, 5 and find that the number in the middle is 3, the median. If the data set has an even number of values — for example, suppose your data set is 2, 4, 6, 8 — then you average the two middle numbers to get the median, which in this case is $(4+6)\div 2=5$.

As you may recall from Stats I, you can find the mean and the median of a data set and compare them to each other. You first organize your data into a histogram, and look at its shape.

- **If the data set is symmetric,** meaning it looks the same on either side when you draw a line down the middle, the mean and median are the same (or close). Figure 17-1a shows an example of this situation. In this case, the mean and median are both 5.
- **If the histogram is skewed to the right,** meaning that you have a lot of smaller values and a few larger values, the mean increases due to those few larger values, but the median isn't affected. In this case, the mean is larger than the median. Figure 17-1b shows an example of this situation, in which the mean is 4.5 and the median is 4.0.
- **If the histogram is skewed to the left,** you have many larger values that pile up, but only a few smaller values. The mean goes down because of the few small values, but the median still isn't affected. In this case, the mean is lower than the median. Figure 17-1c illustrates this situation with a 6.5 mean and a 7.0 median.

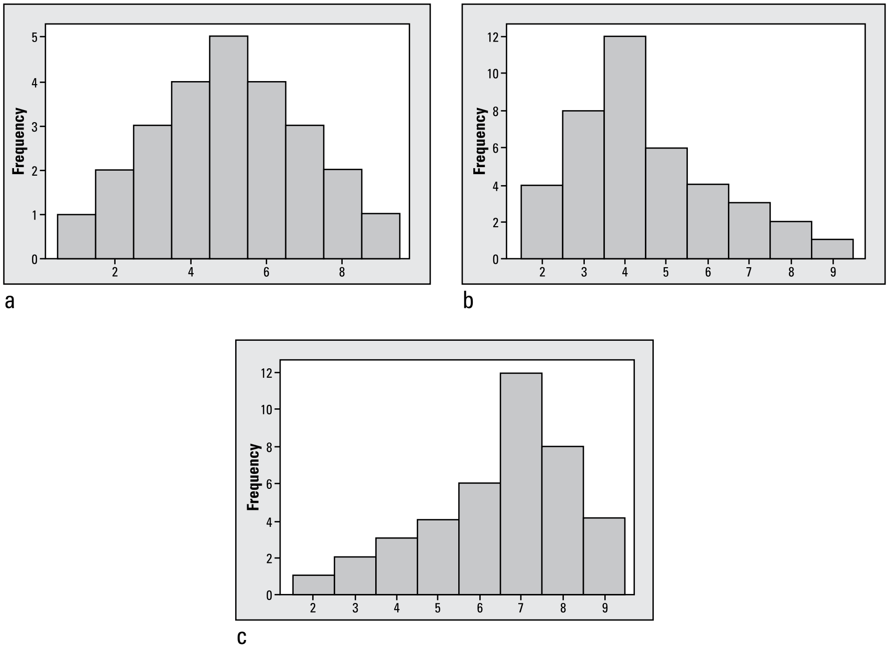


FIGURE 17-1: Symmetric and skewed histograms.

REMEMBER

My point is that the median is important! It's a measure of the center of a population or a sample data set. The median competes with the mean and often wins. Researchers use nonparametric procedures when they want to estimate, test, or compare the median(s) of one or more populations. They also use the median in cases where their data are symmetric but don't necessarily follow a normal distribution, or when they want to focus on a measure of center that's not influenced by outliers or skewness.

For example, if you look at house prices in your neighborhood, you may find a large number of houses within a certain relatively small price range as well as a few homes that cost a great deal more. If a real estate agent wants to sell a house in your neighborhood and intends to justify a high price for it, they may report the mean price of homes in your neighborhood because the mean is affected by outliers. The mean is higher than the median in this case. But if the agent wants to help someone buy a house in your 'hood, they look at the median of the house prices in the neighborhood because the median isn't affected by those few higher-priced homes and is lower than the mean.

Now suppose you want to come up with a number that describes the typical house price in your entire county. Should you use the mean or the median? You gathered techniques in Stats I for estimating the mean of a population (see Chapter 4 for a quick review), but you probably didn't hear about how to come up with a confidence interval for the median of a population. Oh sure, you can take a random sample and calculate the median of that sample. But you need a margin of error to go with it. And I'll tell you something — the formula for the margin of error for the mean doesn't work for the margin of error associated with the median. (Hang on, this book has you covered on that score.)

So, what's the catch?

You may be wondering, what's the catch if I use a nonparametric technique? A downside must be lurking around here somewhere. Well, many researchers believe that nonparametric techniques water down statistical results; for example, suppose you find an actual difference between two population means, and the populations really do have a normal distribution. A parametric technique, the hypothesis test for two means, would likely detect this difference (if the sample size was large enough).

The question is, if you use a nonparametric technique (which doesn't need the populations to be normal), do you risk the chance of not finding the difference? The answer is maybe, but the risk isn't as big as you think. More often than not, nonparametric procedures are only slightly less efficient than parametric procedures (meaning they don't work quite as well at detecting a significant result or at estimating a value as parametric procedures) when the normality condition is met, but this difference in efficiency is small.

But the big payoff occurs when the normal distribution conditions aren't met. Parametric techniques can make the wrong conclusion, and corresponding nonparametric techniques can lead to a correct answer. Many researchers don't know this, so spread the word!

REMEMBER

The bottom line: Always check for normality of the residuals first. If you're very confident that the normality condition is met, go ahead and use parametric procedures because they're more precise. Or even if it's pretty close, you'll be ok, because the procedures allow for some bending of that rule. But if you have doubts about the normality condition, or if you know the normality condition just does not hold, consider using nonparametric procedures. Even if the normality condition is met, nonparametric procedures are only a little less precise than parametric procedures. If the normality condition isn't met, nonparametrics provide appropriate and justifiable results where parametric procedures may not.

Mastering the Basics of Nonparametric Statistics

Because you may not have run into nonparametric statistics during your Stats I class, your first step toward using these techniques is figuring out some of the basics. In this section, you get to know some of the terminology and beginning concepts involved in nonparametric statistics. (Stats II textbooks may contain more on this topic, depending on which route the instructor decides to take.)

Sign

The *sign* is a value of 0 or 1 that's assigned to each number in the data set. The sign for a value in the data set represents whether that data value is larger or smaller than some specified number. The value of +1 is given if the data value is greater than the specified number, and the value of 0 is given if the data value is less than or equal to the specified number. For example, suppose your data set is 10, 12, 13, 15, 20, and your specified number for comparison is 16. Because 10, 12, 13, and 15 are all less than 16, they each receive a sign of 0. Because 20 is greater than 16, it receives a sign of +1.

Several uses of the sign statistic appear in nonparametric statistics. You can use signs to test whether the median of a population equals some specified value. Or you can use signs to analyze data from a matched-pairs experiment (where subjects are matched up according to some variable and a treatment is applied and compared).

In the following sections, you see exactly how to use the sign statistic to test the median of a population and analyze data in a matched-pairs experiment.

Testing the median

You can use signs to test whether the median of a population is equal to some value *m*. You do this by conducting a hypothesis test based on signs. You have $H_o: \text{ Median} = m$ versus $H_a: \text{ Median} \neq m$ (or, you can use a > or < sign in H_a). Your test statistic is the sum of the signs for all the data. If this sum is significantly greater or significantly smaller than what's expected if H_o is true, then you reject H_o. Exactly how large or how small the sum of the signs must be to reject H_o is given by the sign test (refer to Chapter 18).

Suppose you're testing whether the median of a population is equal to 5. That is, you're testing $H_o: \text{ Median} = 5$ versus $H_a: \text{ Median} \neq 5$. You collect the following data: 4, 4, 3, 3, 2, 6, 4, 3, 3, 5, 7, 5. Ordering the data, you get 2, 3, 3, 3, 3, 4, 4, 4, 5, 5, 6, 7. Now you find the sign for each value in the data set, determined by whether the value is greater than 5. The sign of the first data value, 2, is 0, because it's below 5. Each of the 3s receives a sign of 0, as do the 4s and 5s, for the same reason. Only the numbers 6 and 7 receive a sign of +1, being the only values in the data set that are greater than 5 (the number of interest for the median).

By summing the signs, you're in essence counting the number of values in the data set that are greater than the given quantity in H_o. For example, the total of all the signs of the ordered data values is

$$0+0+0+0+0+0+0+0+0+0+1+1=2$$

You can see that the total number of data values above 5 (the number of interest for the median) is 2. The fact that the total of the signs (2) is much less than half the sample size gives you some evidence that the median is probably not 5 here because the median represents the middle of the population. If the median were truly 5 in the population, your sample should yield about six values below it and six values above.

Doing a matched-pairs experiment

You can use signs in a *matched-pairs experiment*, where you use the same subject twice or pair up subjects on some important variables. For example, you can use signs to test whether or not a certain treatment resulted in an improvement in patients, compared to a control. In the cases where the sign statistic is used, improvement is measured not by the mean of the differences in the responses for treatment versus control (as in a paired t-test), but rather by the median of the differences in the responses.

Suppose you're testing a new antihistamine for allergy patients. You take a sample of 100 patients and have each patient assess the severity of their allergy symptoms before and after taking the medication on a scale from 1 (best) to 10 (worst). (Of course, you do a controlled experiment in which some of the patients get a placebo to adjust for the fact that some people may perceive their symptoms to be going away just because they took something.)

In this study, you're not interested in what level the patients' symptoms are at, but rather in how many patients had a lower level of symptoms after taking the medicine. So you take the symptom level before the experiment minus the symptom level after the experiment for each subject.

- If that difference is positive, the medicine appears to have helped, and you give that person a sign of +1 (in other words, count them as a success).
- If the difference is zero, the medicine had no effect, and you give that person a sign of 0.

Remember, though, that the difference could be negative, indicating that the symptoms before were lower than the symptoms after; in other words, the medicine made their symptoms worse. This scenario results in a sign of 0 as well.

After you've found the sign for each value or pair in the data set, you're ready to analyze it by using the sign test (see Chapter 18).

IN THIS CHAPTER

- **Testing and estimating the median: The sign test**
- **Examining matched pairs with the sign test**

Chapter 18

All Signs Point to the Sign Test

The hypothesis tests you see in Stats I use well-known distributions like the normal distribution or the *t*-distribution (see Chapter 4). Using these tests requires that certain conditions be met, such as the type of data you're using, the distribution of the population the data came from, and the size of your data set. Such procedures that involve such conditions are called *parametric procedures.* In general, parametric procedures are very powerful and precise, and statisticians use them as often as they can.

But, situations do arise in which your data don't meet the conditions for a parametric procedure. Perhaps you just don't have enough data (the biggest hurdle is whether the data come from a population with a normal distribution), or your data are just of a different type than quantitative data, such as ranks (where you don't collect numerical data, but instead just order the data from low to high or vice versa).

In these situations, your best bet is a *nonparametric procedure* (see Chapter 17 for background information). In general, nonparametric procedures aren't as powerful as parametric procedures, but they have very few assumptions tied to them. Moreover, nonparametric procedures are easy to carry out, and their formulas make sense. Most importantly, nonparametric procedures give accurate results compared to the use of parametric procedures when the conditions of parametric procedures aren't met or aren't appropriate.

In this chapter, you use the sign test to test or estimate the median of one population. This nonparametric procedure is the counterpart to the one-sample and matched-pairs *t*-tests, which require data from a normal population (see Chapter 3 for more information on the normal distribution). Your Stats I textbook may describe more nonparametric procedures, depending on your instructor. But at least this chapter will get you started.

Reading the Signs: The Sign Test

You use the one-sample *t*-test from Stats I to test whether or not the population mean is equal to a certain value. It requires the data to have a normal distribution. When this condition isn't met, the *sign test* is a nonparametric alternative for the one-sample *t*-test. It tests whether or not the population median is equal to a certain value.

What makes the sign test so nice is that it's based on a very basic distribution, the binomial. You use the binomial distribution when you have a sequence of *n* trials of an experiment, with only two possible outcomes each time (success or failure). The probability of success is denoted by *p*, and is the same for each trial. The variable is *x*, the number of successes in the *n* trials. (For more information on the binomial distribution, see Chapter 3.)

The only condition of the sign test is that the data are ordinal or quantitative — not categorical. However, this is no big deal because if you're interested in the median, you don't collect categorical data anyway.

Here are the steps for conducting the sign test. Note that Minitab can do Steps 4 through 8 for you; however, understanding what Minitab does behind the scenes is important, as always.

1. **Set up your null hypothesis: $\mathrm{H_o} : m = m_o$.**

   The true value of the median is *m*, and m_o is the claimed value of the median (the value you're testing).

2. **Set up your alternative hypothesis. Your choices are $\mathrm{H_a} : m \neq m_o$; or $\mathrm{H_a} : m > m_o$; or $\mathrm{H_a} : m < m_o$.**

   Which $\mathrm{H_a}$ you choose depends on what conclusion you want to make in the case that $\mathrm{H_o}$ is rejected. For example, if you only want to know when the median is greater than some number *m*, use $\mathrm{H_o} : m > m_o$. Chapter 4 tells you more about setting up alternative hypotheses.

3. **Collect a random sample of (ordinal or quantitative) data from the population.**

4. **Assign a plus or minus sign to each value in the data set.**

   If an observation is less than m_o, assign it a minus (–) sign. If the observation is greater than m_o, give it a plus (+) sign. If the observation equals m_o, disregard it and let the sample size decrease by one.

   In terms of the binomial distribution, you have *n* values in the data set, and each value has one of two outcomes: It falls either below m_o or above it. (This is akin to success and failure.)

5. **Count up all the plus signs. This sum is your test statistic, noted by *k*.**

   In terms of the binomial, this sum represents the total number of successes, where a plus (+) sign is the designated success.

6. **Locate the test statistic *k* (from Step 5) on the binomial distribution (using Table A-2 in the Appendix).**

   You determine where your test statistic falls on the binomial distribution by looking it up in a binomial distribution table (see Table A-2 in the Appendix). To do this, you need to know *n*, *k*, and *p*.

   Your sample size is *n*, your test statistic is *k* from Step 5, but what's your value of *p,* the probability of success? If the null hypothesis H_o is true, 50 percent of the data should lie below m_o and 50 percent should lie above it. This corresponds to a success (+) having a probability of $p = 0.50$ on the binomial distribution.

7. **Find the *p*-value of your test statistic:**

   - If H_a has a < sign, add up all the probabilities on the binomial table for $x \le k$.
   - If H_a has a > sign, add up all the probabilities on the binomial table for $x \ge k$.
   - If H_a has a ≠ sign, add up the probabilities on the binomial table of *x* being greater than or equal to *k* and double this value. This gives you the *p*-value of the test.

8. **Make your conclusion.**

   If the *p*-value from Step 7 is less than the predetermined value of α (typically 0.05), reject $H_o$and say the median is greater than, less than, or $\neq m_o$, depending on H_a. Otherwise, you can't reject H_o.

COMPUTER OUTPUT

To run a sign test in Minitab, enter your data in a single column. Go to Stat>Nonparametric>One-sample Sign. Click on your variable in the left-hand box, and click Select. The variable will appear in the Variables box. Click Test Median and type in the box whichever value you have in H_o. For example, if you want to test that the median is 7, type **7** in the box. A pull-down menu indicates which H_a you want to choose (less-than, greater-than, or not-equal-to). Choose your H_a. Then click OK, and you get the results of the sign test.

In the sections that follow, I show you two different ways in which you can use the sign test:

- To test or estimate the median of one population
- To test or estimate the median difference of data where the observations come in pairs, either from the same individual (pretest versus post-test) or individuals paired up according to relevant characteristics

Testing the median in real estate

Situations arise in which you aren't interested in the mean, but rather the median of a population. (Chapter 17 has more on the median.) For example, perhaps the data don't have a normal, or even a symmetric, distribution. When you want to estimate or test the median of a population (call it *m*), the sign test is a great option.

Suppose you're a real estate agent selling homes in a particular neighborhood, and you hear from other agents that the median house price in that neighborhood is $110,000. You think the median is actually higher. Because you're interested in the median price of a home rather than the mean price, you decide to test the claim by using a sign test. Follow the steps of the sign test:

1. **Set up your null hypothesis.** Because the original claim is that the median price of a home is $110,000, you have $H_o : m = \$110{,}000$.
2. **Set up the alternative hypothesis.** Because you believe the median is higher than $110,000, your alternative hypothesis is $H_a : m > \$110{,}000$.
3. **Take a random sample of ten homes in the neighborhood.** You can see the data in Table 18-1; its histogram is shown in Figure 18-1.

   Now the question is, is the median selling price of all homes in the neighborhood equal to $110,000, or is it more than that (as you suspect)?

TABLE 18-1 Sample of House Prices in a Neighborhood

House	Price	Sign (Compared to $110,000)
1	$132,000	+
2	$107,000	–
3	$111,000	+
4	$105,000	–
5	$100,000	–
6	$113,000	+
7	$135,000	+
8	$120,000	+
9	$125,000	+
10	$126,000	+

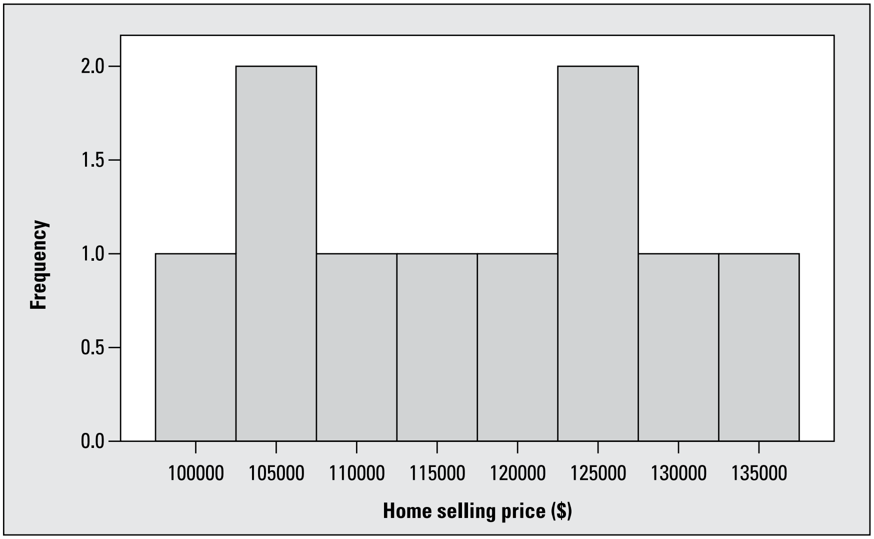


FIGURE 18-1: Histogram of the selling prices of ten houses.

4. **Assign a plus sign to any house price more than $110,000 and a minus sign to any house less than $110,000. (See column three of Table 18-1.)**
5. **Find your test statistic. Your test statistic is 7, the number of "+" signs in your data set (see Table 18-1), representing the number of houses in your sample whose prices are above $110,000.**

6. **Compare your test statistic to the binomial distribution (refer to a binomial distribution table) to find the *p*-value.**

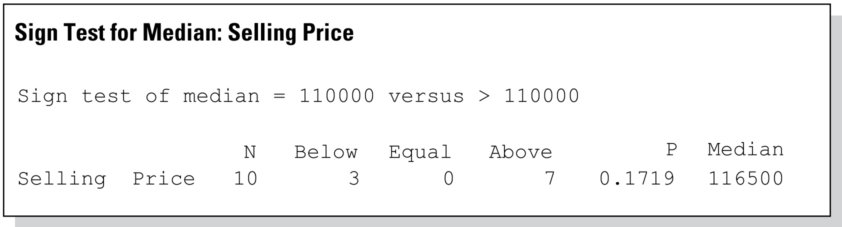
   For this case, look at the row in the binomial table where $n = 10$ (the sample size) and $k = 7$ (the test statistic) and the column where $p = 0.50$ (because if the population median equals m_o, 50 percent of the values in the population should be above it and 50 percent below it). According to the table, you find the probability that *x* equals 7 is 0.117.

   Because you have a right-tailed test (meaning H_a has a > sign in it), you add up the probabilities of being at or beyond 7 to get the *p*-value. The *p*-value in this case is $0.117 + 0.044 + 0.010 + 0.001 = 0.172$.

7. **To conclude, compare the *p*-value (0.172) to the predetermined α (I always use 0.05).**

   Because the *p*-value is greater than 0.05, you can't reject H_o. You don't have enough evidence to say the median house selling price is more than $110,000.

Figure 18-2 shows these results as calculated by Minitab.



Sign Test for Median: Selling Price

```
Sign test of median = 110000 versus > 110000

                    N   Below  Equal   Above         P   Median
Selling  Price     10       3      0       7    0.1719   116500
```

FIGURE 18-2: Sign test for house prices conducted by Minitab.

WARNING

If your data are close to normal and the mean is the more appropriate measure of center for your situation, don't use the sign test. Instead, use the one-sample *t*-test (or *Z*-test). The sign test isn't quite as powerful (able to reject H_o when it should) as the *t*-test in situations where the conditions for the *t*-test are met. More importantly, though, don't run to the *t*-test to reanalyze your data if the sign test doesn't reject H_o. That would be improper and unethical. In general, statisticians consider the idea of following a nonparametric procedure with a parametric procedure in hopes of getting more significant results to be *data fishing*, which is analyzing data in different ways until a statistically significant result appears.

Estimating the median

You can also use the sign test to find a confidence interval for one population median. This comes in handy when you're interested in estimating what the median value of a population is, such as the median income of a household in the United States or the median salary of people fresh out of an MBA program.

Following are the steps for conducting a confidence interval for the median by using the test statistic for the sign test, assuming your random sample of data has already been collected. Note that Minitab can calculate the confidence interval for you (Steps 2 to 5), but knowing how Minitab does the steps is important:

1. **Determine your level of confidence, $1-\alpha$ (that is, how confident you want to be that this process will correctly estimate *m* over the long term).**

   The typical confidence level that data analysts use is 95 percent (see Chapter 3 for more information).

2. **On the binomial table (Table A-2 in the Appendix), find the section for *n* equal to your sample size, and the column where $p = 0.50$ (because the median is the point where 50 percent of the data lies below and 50 percent lies above).**

   You'll find probabilities for values of *x* from 0 to *n* in that section.

3. **Starting at each end ($x = 0$ and $x = n$) and moving one step at a time toward the middle of the x values, add up the probabilities for those values of x until you pass the total of α (which is one minus your confidence level).**
4. **Record the number of steps that you had to make just before you passed the value of $1-\alpha$. Call this number c.**
5. **Order your data set from smallest to largest. Starting at each end, work your way to the middle until you reach the *c*th number from the bottom and the *c*th number from the top.**
6. **Use these numbers as the low end and the high end of an interval. This result is your confidence interval for the median.**

You can use these steps to find a confidence interval for the median in the house-price example from the preceding section. Here's how this example breaks down:

1. **Let your confidence level be set at $1-\alpha = 0.95$.**
2. **On the binomial table (see the Appendix), look at the section where $n = 10$ (the sample size) and $p = 0.50$.** These values are listed in Table 18-2.
3. **Start with the outermost values of $x\left(x = 0 \text{ and } x = 10\right)$ and sum those probabilities to get $0.001 + 0.001 = 0.002$.** Because you haven't yet passed 0.05 (the value of α), you go to the second-innermost values of $x\left(x = 1 \text{ and } x = 9\right)$. Add their probabilities to what you have so far to get $0.002\left(\text{old total}\right) + 0.010 + 0.010 = 0.022$. You're still not past 0.05 (α), so go one more step. Add the third-innermost probabilities for $x = 2$ and $x = 8$ to the grand total to get $0.022\left(\text{old total}\right) + 0.044 + 0.044 = 0.110$. You've now passed the value of $\alpha = 0.05$. The value of *c* equals 2 because you passed 0.05 at the third-innermost values of *x*, and you back off one step from there to get your value of *c*.

TABLE 18-2 **Binomial Probabilities to Help Calculate a Confidence Interval for the Median ($n = 10$, $p = 0.50$)**

x	p(x)
0	0.001
1	0.010
2	0.044
3	0.117
4	0.205
5	0.246
6	0.205
7	0.117
8	0.044
9	0.010
10	0.001

4. **Order your data (Table 18-1) from smallest to largest, giving you (in dollars): 100,000, 105,000, 107,000, 111,000, 113,000, 120,000, 125,000, 126,000, 132,000, and 135,000.**
5. **Work your way in from each end of the data set to take the second-innermost values (because $c = 2$): the numbers \$105,000 and \$132,000.** Put these two numbers together to form an interval, and you conclude that a 95 percent confidence interval for the median selling price for a home in this neighborhood is between \$105,000 and \$132,000.

COMPUTER OUTPUT

To find a $1 - \alpha$ percent confidence interval for the median using Minitab based on the sign test, enter your data into a single column. Go to Stat>Nonparametrics>One-sample Sign. Click on the variable in the left-hand column for which you want the confidence interval, and it appears in the Variables column. Click the circle that says Confidence Interval, and type in the value of $1 - \alpha$ you want for your confidence level. (The default is 95 percent, written as 95, to get a significance level of 5 percent, or $\alpha = .05$.) Click OK to get the confidence interval.

Testing matched pairs

The most useful application of the sign test is in testing matched pairs of data — that is, data that come in pairs and represent two observations from the same person (pretests versus post-tests, for instance), or one set of data from each pair of people who are matched according to relevant characteristics. In this section, you see how you can compare data from a matched-pairs study to look for a treatment effect, using a sign test for the median.

TIP

The idea of using a sign test for the median difference with matched-pairs data is similar to using a *t*-test for the mean differences with matched-pairs data. (For details on matched-pairs data and the *t*-test, see your Stats I text or *Statistics for Dummies*, 2nd Edition [Wiley].) You use a test of the median (rather than the mean) when the data don't necessarily have a normal distribution, or if you're only interested in the median difference rather than the mean difference.

First, you set up your hypothesis, H_o: The median is zero (indicating no difference between the pairs). Your alternative hypothesis is H_a: The median is $\neq 0, > 0,$ or < 0, depending on whether you want to know if the treatment made any difference, made a positive difference, or made a negative difference compared to the control. Then you collect your data (two observations per person or a pair of observations from each pair of people you've matched up). After that, you use Minitab to conduct Steps 4 to 7 of the sign test.

COMPUTER OUTPUT

To use Minitab to test for matched pairs, find the differences between the values in each pair, and enter them in a column in Minitab. Then go to Stat>Nonparametrics>One-sample Sign. Click on the differences variable in the left-hand column, and it appears in the Variables column. Click Test the Median, and type **0** in the box (because your H_o is that the difference is zero). From the pull-down menu, select your alternative hypothesis, and click OK.

For example, suppose you wonder whether taking a test while chewing gum decreases test anxiety. You pair 20 students according to relevant factors such as GPA, score on previous exams, and so on. One member of each pair is randomly selected to chew gum during the exam, and the other member of the pair doesn't. You measure test anxiety of each person via a very short survey right after they turn in their exams. You measure the results on a scale of 1 (lowest anxiety level) to 10 (highest anxiety level). Table 18-3 shows the data based on a sample of ten pairs.

REMEMBER

The actual levels of test anxiety aren't important here; what matters is the difference between anxiety levels within each pair. So, instead of looking at all the individual anxiety levels, you can look at the difference in anxiety levels for each pair. This method gives you one data set, not two. (In this case, to calculate the differences in each pair, you can use the formula test anxiety without gum minus text anxiety with gum, and look for an overall difference that's positive.) Typically, in the case of matched-pairs data, you're testing whether the median difference equals zero. In other words, $H_o : m = 0$; the same holds in the test anxiety example.

TABLE 18-3 Testing the Effectiveness of Chewing Gum in Lowering Test Anxiety

Pair	Anxiety Level — Gum	Anxiety Level — No Gum	Difference (Gum/No Gum)	Sign
1	9	10	–1	–
2	6	8	–2	–
3	3	1	+2	+
4	3	5	–2	–
5	4	4	0	none
6	2	7	–5	–
7	2	6	–4	–
8	8	10	–2	–
9	6	8	–2	–
10	1	3	–2	–

The differences in anxiety levels for each pair in your data set now become a single data set (see column four of Table 18-3). You can now use the regular sign test methods to analyze this data, using $H_o : m = 0$ (no median difference in test anxiety of gum versus no gum) versus $H_o : m < 0$ (chewing gum reduces test anxiety).

Assign each difference a plus or minus sign, depending on whether it's greater than zero (plus sign) or less than zero (minus sign). Your test statistic is the total number of plus signs, 1, and the relevant sample size is $10 - 1 = 9$. (You don't count the data that hit the median of zero right on the head.)

Now compare this test statistic to the binomial distribution with $p = 0.50$ and $n = 9$, using the binomial table (see the Appendix). You have a test statistic of $k = 1$, and you want to find the probability that $x \leq 1$ (because you have a left-tailed test, see Step 6 of the sign test from the earlier section, "Reading the Signs: The Sign Test"). Under the column for $p = 0.50$ in the section for $n = 9$, you get the probability of 0.018 for $x = 1$ and 0.002 for $x = 0$. Add these values to get 0.020, your p-value. This result means that you reject H_o at the predetermined α level of 0.05. This tells you the anxiety levels for gum versus no gum are different. Now, how are they different? Based on this data, you conclude that chewing gum on an exam appears to decrease test anxiety because there are more negative differences than positive differences. Although, it may increase the anxiety of those around you, depending on how loudly you chew!

5

Putting it All Together: Multi-Stage Analysis of a Large Data Set

IN THIS PART . . .

See the stages of analyzing a large data set.

Analyze a large data set of movies.

Compare refrigerator brands with a large data set.

IN THIS CHAPTER

- » Getting an introduction to data science
- » Collecting and wrangling the data
- » Conducting exploratory data analysis
- » Making inferences and modeling
- » Telling the story of what you found

Chapter 19

Conducting a Multi-Stage Analysis of a Large Data Set

You've no doubt heard about "Big Data" and "Data Science." With today's data boom, where millions of pieces of data are collected per second around the world, a new science has emerged to tackle such crazily large and unwieldy data sets. Even one of the terms they use in data science is called *data wrangling*, to reflect the rodeo that is today's data. In this chapter you go through the steps involved in working with a big data set, and in the following two chapters you see examples of this process from the real world.

Steps Involved in Working with a Large Data Set

In this section, you see the six steps you need to follow when working with a large data set. Each step has its own challenges and requires certain skills. And while some steps may seem more glamorous than others (everyone loves the part where you run the analyses), they are all equally important and useful in contributing to your data's story.

1. **Wrangle the data.**

   Data wrangling is a term used in data science to describe the transformation of raw data into something usable in terms of what's in it, how it's formatted, and what might be missing. Some data sets take much more work to wrangle than others. Real-world data usually takes more wrangling than textbook data. And it often requires the power of sophisticated and flexible statistical software such as SAS or R.

2. **Visualize the data.**

   Data visualization means to show pictures of the data; you organize and display the data so you can begin to see what's going on. Graphs are the main tool used here, as you know from Stats I. Histograms and boxplots are great for single quantitative variables, while pie charts and bar graphs are helpful for visualizing categorical variables.

3. **Explore the data.**

   In exploring a data set, you summarize it using descriptive statistics. You start by looking at each variable separately, and finding measures of center (mean, median) and spread or variability (standard deviation, interquartile range, variance, and so on). You also look at things like the basic size of the data set. You compare and contrast these statistics, maybe breaking them down further by other variables, such as average test score broken down by which class the students were in.

4. **Look for relationships.**

   In this step you start to look at pairs of data. For example, for two quantitative variables that might be related, you look for correlations; for two categorical variables that are related, you make two-way tables. You look to see which variables are independent (unrelated), and which are dependent (related).

5. **Build models and make inferences.**

   Once you get a rough idea of what kinds of relationships you have in the data, and where the data are located in terms of their descriptive statistics, it's time to build some models and make inferences so you can make predictions, draw conclusions, and make decisions based on your data.

6. **Share what you learned with others.**

   After finding out as much as you can about your data, and being able to draw conclusions and make decisions, you'll most likely want to (or at least have to) share what you learned with others: your boss, your colleagues, people who will apply your information, and people who will consume your information and build on it.

Wrangling Data

It is estimated that statisticians and data analysts spend about 80 percent of their time getting the data ready to analyze, and only 20 percent of their time actually analyzing it. So if your data is in good shape before you start, be thankful! And if it's not, you're in good company.

What takes so much time? Statisticians use six steps to wrangle a large data set that's been collected:

1. Discovery
2. Structuring
3. Cleaning
4. Enriching
5. Validating
6. Publishing

Discovery

The discovery step in wrangling data is just taking a look at all the data you have and seeing what you are actually dealing with. For example, if I wanted to make a data set showing the top money-making movies this past year, I might start by looking at the data provided by IMDb (Internet Movie Database). When I go to their website looking for information on their movie data sets, I find the following information, just for starters, along with a lot of other datasets that are available (to find this information, I had to Google "IMDb datasets"). Much more information and data are available for each movie and TV show on the crew, writers and producers, actors, and so on. And this is for thousands and thousands of movies.

From the title.basics.tsv.gz data set described here, I chose to take the original title of the movie, the start year (release year), the runtime in minutes, and the genres, for the list of movies that made more than $100,000,000 in U.S. revenue in 2018. That list of movies was found on a website called The Numbers (`www.the-numbers.com`).

IMDb Data Set Details

The data set **title.basics.tsv.gz** contains the following information for each title:

- tconst — the alphanumeric unique identifier of the title.
- titleType — the type/format of the title (for example, movie, short, TV series, TV episode, video, and so on).
- primaryTitle — the more popular title / the title used by the filmmakers on promotional materials at the point of release.
- originalTitle — the original title, in the original language.
- isAdult — 0: non-adult title; 1: adult title.
- startYear (YYYY) — the release year of a title. In the case of a TV series, it is the series' start year.
- endYear (YYYY) — TV series end year. '\N' is used for all other title types.
- runtimeMinutes — the primary runtime of the title, in minutes.
- genres — genres associated with the title (includes up to three genres).

Structuring

In this step you figure out how you want your data to be set up. For example, if your data was taken from one large website by peeling off information one piece at a time, you want to figure out how to pull it all together. Do you want it in one big spreadsheet or program? Do you want to split it up? Which data variables do you want to list first and how do you want to list them? Which ones do you want to leave out? This job can be either very small, in the case where the data is already done for you (for example, in a spreadsheet), or it can be very time-consuming, as in the case where the data has been pulled off an existing website or database with many variables at a time.

In my example, I pulled the data from the IMDb dataset that already had it set up in a spreadsheet format, and I pulled it into Minitab to work with it. So it wasn't too bad.

Cleaning

Cleaning your data set is just what it sounds like: making sure it's usable. The biggest culprit in causing problems with a data set is a letter or number that's supposed to represent a null value, which means missing data. Many statistical software packages freak out when they come across letters like NA or n/a for missing data; an empty space is typically read as a missing value so that's a better plan. You don't necessarily want to replace a null value with zero, because, for example, a missing budget on a movie doesn't make sense when you just put $0 for it. It will also certainly affect the average of the budgets for all the movies in your data set. So what you need to do is go back and see if you can find the budget for that movie and plug it in, or if you can't, then look at the statistics package you are using and see what notation it recognizes for a missing value. For example, it may be an asterisk (*); put that in for the missing value.

You may also come across data that does not make sense; after all, people are still usually involved in the process of entering some of today's data (although even that notion is probably going to be up for grabs soon). What if you are looking over your survey data and find that someone entered 600 for their age in years? You know that can't be right, so you must clean up this data. Maybe you have the year they were born and can figure out their age, but maybe you don't, in which case an asterisk is better than the value 600 for their age.

Another situation is consistency in *coding* (typing in) the data. For example, if you ask people in a survey which state they are from, they may have different ways to abbreviate it, such as WI or Wis or Wisc for Wisconsin. This has to be cleaned up, so the computer doesn't think WI and Wis stand for two different things.

REMEMBER

The point is that you must look carefully at your data for values that are missing, don't make sense, or are just plain wrong. For example, it is very easy in the data entry process for someone to miss a column and suddenly have everything be one column off, and so instead of seeing a state in the column, you see a number. This is a very time-consuming and tedious step, but you have to do it to get good results.

Enriching

I look through my existing data, and I'm realizing that I've got some good information here, a lot of information on the movie title, that's for sure, but I still may not have everything I need. For example, what about how much money the movie made? Is this information available? If so, where? This is the enriching step, where you add to your developing data set if you so desire.

Back to the web. On the IMDb website you can type a movie name in the search box, and all kinds of data comes up on that movie, not in a data set, but on a single page. For example, take the movie *Godzilla vs. Kong*. When you type its title in, you immediately see it was made in 2021, it's rated PG-13, its runtime is 1 hour and 53 minutes, its genre is science-fiction action thriller, and it was released on March 31, 2021.

The budget for the movie *Godzilla vs. Kong* was estimated to be $200 million, and in the opening weekend it made $31,625,971. Its gross revenue in the U.S. was $100,102,879, and its cumulative worldwide gross revenue was $442,502,879. Its Metacritic score is 59/100 based on 57 critic reviews, and the audience rating is 6.4/10 based on 143,000 reviews. I could go on, but that's enough to get me started.

Adding this information to my data set will really enrich it, which is what this step is about. I found the list of the top money-making movies in 2018 from the original data set I was given, and I added these new variables to it: movie rating (for example, PG-13); release data, budget, opening weekend revenue, domestic revenue, worldwide revenue, critic rating, and audience rating. I added this information line by line to my data set in my Minitab spreadsheet, but you could use a statistical software package/language like R to pull this data off the website as well.

Validating

Validating is related to data cleaning, but on a deeper level. It is the activity that you do to ensure that your data quality is optimal, and that your data is consistent. This is an especially important step once you have enriched or added to your data set. Some inconsistencies may arise, and you need to check for them.

Survey data is a good example where validating, or cross-checking, is very important. When people fill out surveys, it's easy to make mistakes, round things off, guess at the information, or just put in something that is incorrect for various reasons. In the case of age, you can find out the age of the oldest person on earth (as of this writing, it's Kane Tanaka of Japan, who is 118 years old). You could use this as your upper limit on the age column of your data and look for any numbers that are over 118 years or under, say, 1 year, and flag them for further investigation or removal (or a new world record, perhaps!).

Or maybe you are collecting data on age and year of birth. It would be good to check that they match. And if they don't, you need to decide what to do: Do you trust the age in years or the birthdate? Me, I'd trust the birthdate — I can remember my birthdate, but I often forget how old I am exactly (conveniently!).

In my movie data set that I was building, I checked to make sure that the data made sense — for example, that the worldwide revenue was at least as large as the U.S. revenue, and the critics' and audience ratings were not over the limit.

TIP

Note that you can write programs to look for problems in your data using R, or some other language, so your validation doesn't have to be all by hand. But either way, validating your data is important because it ensures your data is of good quality, and is consistent and ready to go.

Publishing

Publishing your data means getting it into a finished form where it can be used by others, transformed easily into another language or format, and/or uploaded into a program by someone to analyze directly. My movie data is in a Minitab spreadsheet, ready to be analyzed in Chapter 20.

TIP

It is very important to document your data set in a separate page within the spreadsheet. In the documentation section, you list all the column names, and for each column name you give the full name of the column if it's different than what appears in the spreadsheet, and you give the type of data that's in the column and what it stands for (such as profit = whether or not the movie made a profit, 0 = no, 1 = yes). You also want to list the sources of your data, so someone else can validate your data, go back and retrace your steps, or just look at it and extract their own information from it.

As you can see, data wrangling is a tough job, and it can easily take 80 percent of your time or more, all before you start doing something with the data. It can feel like a rodeo at times, and I'm sure that's where the term came from, but it's well worth the time and effort, especially if other people are going to use it downstream, as they say.

Visualizing Data

Once your data has been wrangled, the next step is to visualize the data. This is usually simple if you have one or two variables, like budget and type of movie. But as the number of variables gets higher, so does the data's sophistication and complexity, and you must develop a system for looking at the data.

The first step is to look at the variables separately, dividing them into two groups: categorical data and quantitative data.

REMEMBER

Categorical data is data that falls into groups, such as gender identification, region, or whether a movie made a profit. *Quantitative data* is measurements, where the numerical values actually make sense. Some examples include time until you reach a real person on a helpline (endless!), length of time between accidents at an intersection, and movie budget.

From your Stats I book (or *Statistics for Dummies*, 2nd Edition [Wiley]), you know how to graph both categorical data and quantitative data. Categorical data basically uses pie charts and bar graphs, and quantitative data uses histograms and boxplots for the most part. The hard part isn't making the graphs, it's determining what type of graph to use, so be sure to check the type of data that you have for each variable first before making a visual representation of it.

Once you get a look at single variables, you can start to pair up data and look at it that way. If you want to look at two categorical variables together — for example, genre of movie and whether it made a profit — you can use what is called a *stacked bar graph*, which is covered in Stats I. In a stacked bar graph, each bar represents a genre, and the bar is broken into two colors, one color representing the percentage of those movies making a profit, and the other color representing the percentage of those movies not making a profit. For example, maybe action movies are more likely to make a profit than horror movies. See how the graphs can make you ask questions and think of more items to analyze?

If you want to compare two quantitative variables, such as movie budget and revenue, you use a *scatterplot*, a two-dimensional graph where each (x,y) point represents a movie, and the value of x represents one variable (say, budget) and the value of y represents the other variable (say, revenue). See Chapter 5 for how to make a scatterplot. You would think that budget might be related to revenue, in that the movies with the bigger budgets tend to make more money. Is it true? I look at this question in Chapter 20.

What about the situation where you have one quantitative variable, such as movie budget, and one categorical variable, such as whether the movie made a profit? Here you can use side-by-side boxplots from Stats I. In the movie case, you would have two boxplots, one for movies that made a profit and one for movies that didn't make a profit. Each boxplot would show the budgets for the movies in that particular group. Then you could compare the boxplots and look for relationships; perhaps the movies with the bigger budgets were more likely to make a profit (at least, that's what the moviemakers were hoping!).

Exploring the Data

The next step of working with a data set is exploration. Here's where you begin to uncover the numerical story behind the data. Start with single variables and move on from there.

From Stats I you know how to do the basics of exploring data. If you have a categorical variable, such as movie genre, you can find the percentage of movies in each genre. Note that some movies fall into more than one genre, so your total percent won't necessarily sum to one. You can fix this by choosing the most predominant genre for each movie, and then explore that genre with a pie chart or *relative frequency bar graph* (one that uses percents and sums to one, from Stats I). For example, more action and drama films are made than horror movies and biographies.

You also know from Stats I how to explore quantitative data; many more options exist here. You can find the measures of center (the mean and median), and the measures of variability (the standard deviation and interquartile range). You can also explore the idea of skewness using histograms. If the histogram trails off to the right when you explore movie budgets, you know the data is skewed right, and a few movies were made that cost a lot more money than the others (a *Star Wars* movie, perhaps?). And if the histogram trails off to the left when you explore movie revenues, you know a few movies made much less money than the rest, making the data skewed to the left. If the data looks about the same on each side, the data is deemed to be symmetric, and the mean and median will be close to each other.

WARNING

Make sure you are using legit graphs for the type of data that you have. You can't use a histogram on categorical data such as region. Those values can be interchanged in the bars of the graph, and looking for characteristics like skewness would not make sense.

Looking for Relationships

Everyone is interested in relationships within data sets. You want to do the same with your data set. If you only analyze single variables, you will miss out on some good juicy information about who likes to be with whom, who you thought would make a good pair but it wasn't meant to be, and who looks like they may be heading for romance.

If you are examining two quantitative variables, such as movie budget and revenue, your visual tool, as you saw previously, is a scatterplot. As far as statistics go, the area to work in is correlation and regression. As you see in Chapter 5, correlation is the statistic that measures the strength and direction of the linear relationship between two quantitative variables, so you need to find the correlation between budget and revenue. If it's strong, you take it to the next step (see the next section) and build a model for predicting one from the other.

In the case of two categorical variables such as genre and whether a movie made a profit, you can use the Chi-square test for independence, described in Chapter 15; here you are looking to see if knowing a movie's genre can help you decide whether or not the movie made a profit. You can also predict the chance of making a profit using logistic regression (see Chapter 9).

If you have one categorical and one quantitative variable, you can compare the percentages from two categories using the hypothesis test for two population percentages (see Chapter 4), or more categories with the Chi-square goodness-of-fit test (see Chapter 16). Or, you can compare the averages from two or more categories using the *t*-test for two group means (Chapter 5), or ANOVA for multiple groups' means (see Chapter 10). For example, you might want to compare the percentage of brown M&M'S to the percentage of M&M'S of all other colors, to see if they even out, or you might want to compare all the colors of the M&M'S separately using Chi-square goodness-of-fit. You might want to compare the average revenue for two types of movies using the *t*-test, or many types of movies using ANOVA.

REMEMBER

Practice thinking about how to analyze data in different situations right from the get-go. Always be thinking, "How do I attack this problem? How do I know it's supposed to be done this way?" Then, when you encounter a brand-new data set with many variables, you can be calm and cool, knowing that you know where to start and how to break it down. It's all in what type of variables you have, and how many. Also remember that in very large data sets, you may be exploring many different relationships at once; it's important to report on which relationships you looked at, whether or not they were "statistically significant," and which ones were found to be significant, but not meaningful (for example the *p*-value was small enough to reject H_o, but the actual values found weren't really practical in the context of the problem).

Building Models and Making Inferences

Along with looking for relationships, you might want to build models and make inferences. Building models involves coming up with a way to express a relationship using an equation. For example, $y = 2x + 13$ expresses a linear relationship between y and x; it says that to predict y, you take the value of x, multiply it by 2, and add 13. How you come up with that equation is called *model building*. In this case, the model is a simple linear regression model, found in Chapter 5.

If you have one or more quantitative variables, you can use simple linear regression (Chapter 5), multiple linear regression (Chapters 6 and 7), or nonlinear regression (Chapter 8) to build your model. If you want to predict a categorical variable with two possibilities, you can use logistic regression (Chapter 9). If you have one or two categorical variables, you can fit models using Chi-square (Chapters 15 and 16).

To make inferences, you ask questions and use your data to find answers and make decisions. For example, you might hypothesize that the average budget for a top money-making movie in 2018 was $200,000,000. You might think it was less than that. So, you would run a hypothesis test for one population mean to see if you are right. You could also use the movie data to find a likely range of where the average movie budget was for the top money-making movies; this would involve a confidence interval. See Chapter 4 for confidence intervals and hypothesis tests for one or two means, and one or two proportions.

If you want to answer a question about a group of means, such as “Are the average budgets different for different movie genres?”, that would involve analysis of variance (ANOVA). If you found an overall difference, you might want to find out which genres make more money, and which ones make less money by using multiple comparison procedures (see Chapters 10 and 11). And if you want to see if the means differ on a combination of two variables, such as genre and whether they made a profit, you can use two-way (or two-factor) ANOVA (see Chapter 12).

TIP

To help you decide which analysis to use in each situation, you might make an if-then sheet for yourself. The sheet would have one column for “if you have THIS situation” along with a second column that says “then do THIS.” You would also include an example. For instance, you might say “If I want to estimate the average” in column one, and in column two, write “then I use a confidence interval for a population mean.” Or, in column one, “If I want to compare the means of several populations,” and in column two, “then I use ANOVA.”

Sharing the Story

Once you have wrangled your data, visualized it, explored the individual variables, looked for relationships, built models, and made inferences and decisions, you are able to tell quite a story about the data set. It’s important to do so in a way that makes the most of what you found in the data set while also recognizing and acknowledging limitations of the data at the same time. If a story about a data set is too good to be true well, you know the rest.

You may very well be asked to write a report about your data, fulfilling a certain purpose. Maybe your job is to answer a series of scientific questions. Maybe you found out something new and want to spread the word. Or maybe you are asked to do a big project for a class where you analyze a large data set. Where do you start?

Who is the audience?

First, ask who your audience is and think about how much they know about statistics. Are they fellow students who are on your level? Are they high-level CEOs who might have taken statistics a long time ago and are so into other topics by now that they may not remember the details? Are they colleagues who are going to peer-review your paper? Or are they an audience that is more interested in the actual subject you are studying (like movies) and less interested in the gory details of the stats? It's important to figure out who your audience is, and try to reach them where they're at; don't make it boring with things they already know, and don't snow them with too much information. It's a fine line, but with time and practice, you'll learn to find it.

Make an outline

Once you know who your audience is, set up an outline. I set up very detailed outlines every time I write a paper, a report, or a book (especially a book!). It's so much easier to fill in the information once you have a detailed outline available. Like data wrangling, writing an outline might take a lot more of your time, but the rest goes so much faster once you've gotten that good foundation. You want to set up an introduction that covers enough background to put your data into context, a body that describes the following elements:

1. The questions you want to answer or the purpose of your study.
2. How you went about answering those questions using your data.
3. What you found in your data in terms of visuals, exploration, relationships, model building, and inferences.
4. What it all means within the context of the problem (in other words, your conclusions and decisions based on your data).
5. Whether there are any limitations in your data (for example, if the data was only collected over a two-year period, then inferences can't be made beyond that).
6. What your next steps or ideas for future work are, based on what you learned in this study.

Include an executive summary

Reports generally contain an executive summary, which quickly puts out the main points: “Here’s what I did, here’s why I did it, and here is what I found.” The purpose of an executive summary is to give the reader something to grab and go. They might only have a few minutes and need the big picture as quickly as possible. They may just be perusing information on their desk, and only have a few minutes to devote to your report, so make those minutes count!

Check your writing

No matter who your audience is, or what your purpose is, make sure that you check your writing very carefully for grammatical errors, sections that are accidentally repeated, lines out of context, punctuation issues, tense issues, and the like. It’s very off-putting when someone who is by now quite well-versed regarding this data set makes mistakes that simple proofreading would have caught.

REMEMBER

Make sure that what you say is correct statistically. If you ever have a question, my advice is always, “Ask a Statistician!” Hopefully your textbooks will be as helpful to you as your reference books, like this one. Make sure all conditions are checked (how do you know those test scores have a normal distribution?), and make sure your limitations are acknowledged. In this way, you will gain trust and respect from your audience, no matter who you are.

IN THIS CHAPTER

- Looking at top money-making movies
- Exploring the data and looking for relationships
- Checking out some models and inferences
- Writing it up

Chapter 20
A Statistician Watches the Movies

Everyone goes to the movies, and like everything else, people collect data on them. In this chapter I needed a data set to look at and analyze from beginning to end so I could work through it, and I decided to build my own movie-themed data set from scratch. You can find this data set by going to `www.dummies.com/go/statisticsIIfd2e`. Look at it yourself and play with it if you want. I go through all the details of my analyses in this chapter, so no worries if you don't want to look up the actual data set — you certainly don't need to.

The data set contains information on all movies that made at least $100,000,000 in U.S. revenue at the box office in 2018. The source is IMDb (Internet Movie Database; `www.imdb.com`), a wonderful encyclopedia of information on thousands of movies — it's very precise and trustworthy, and I highly recommend it. I also used a website called The Numbers, found at `www.the-numbers.com`, another good site for information on movies.

I started the process by going through the columns of the data set (if you want some information on the process of putting a data set together, head to Chapter 19). Then I visualized and explored the data, and I looked for relationships between variables of interest. I worked up a model that best predicted how well a movie was going to do in the U.S. revenue-wise, and I made some inferences comparing numbers as of this writing to 2018, when the movies in the data set were made. Finally, at the end of the chapter, I included a short write-up in report format to give you some ideas of how you might write your reports when you analyze data sets.

REMEMBER

Having a step-by-step plan like the one outlined in Chapter 19 is critical for dealing with large data sets, or you could be swimming in a sea of numbers pretty quickly.

Examining the Movie Variables and Asking Questions

I chose to include movies that made $100,000,000 or more in total U.S. box office revenue in 2018; 34 movies made the cut. After locating data from the movie websites and wrangling it (that's an actual statistical term; see Chapter 19), I settled on14 variables for my final movie data set. The variables I chose are listed here.

Name: Official name of the movie.

Rank (revenue): Where 1 = highest U.S. box office revenue for 2018. Note that all movies in this data set made over $100,000,000 in U.S. box office revenue.

Release date: Day and month the movie was released (the year was 2018). I used two columns for this data, one for day and one for month. I counted them as one variable.

Rating: The Motion Picture Association's rating of the movie's content (for example, PG means parental guidance suggested; see motionpictures.org for more details).

Genre: Predominant type of movie category (for example, action, drama, horror, comedy).

Runtime: Number of minutes the movie lasts.

Days: Number of days the movie was in U.S. theaters.

Theaters: Number of theaters the movie was shown in throughout the U.S. (A few movies did not have this information, hence an * was placed in the data.)

Budget: Amount of money the movie cost to make.

Opening weekend: Amount of money made in the first weekend that the movie was out in U.S. theaters.

U.S. revenue: Amount of money made by the movie during its time in U.S. theaters (note that some movies were in theaters from 2018 to 2019).

Critics: Movie critics' average rating of the movie, based on a scale from 0 to 100.

Audience: Audience average rating of the movie, based on a scale from 0 to 100.

Profit: 1 = movie made a profit in the U.S. because U.S. revenue > budget; or 0 = movie did not make a profit in the U.S. because revenue ≤ budget. I constructed this variable myself, based on the data in the budget and U.S. box office revenue columns already given.

The reason I chose these 14 variables is that my own questions about movies centered on the topic of predicting U.S. box office revenue, and I thought some of these variables might be helpful. I was also interested in finding out how much money the hottest movies really made, and what it takes to be in that category (it turns out it takes a lot!).

Looking quickly down the list and thinking about the variables, you might ask questions like, "Do the audience and the critics think alike when it comes to liking a movie?" or "Do movies with bigger budgets tend to make more money?" or "Is there such a thing as a movie that's too long? Maybe length of movie is negatively related to audience ratings." One of my main questions was, "What variable or set of variables best predicts how well a movie is going to do in the U.S.?" Can you take a guess?

TIP

So many questions can come to mind with a large data set, which can lead you to examine a lot more than you'd originally planned to; that's why large data sets are so important to study! Don't view statistics as being all about data sets with one or two variables; the real world is full of data, and looking at it one or two variables at a time is only a start.

Visualizing the Movie Data

Before I get too far down the road, you might want to get a feel for what the actual data set looks like. The top eight U.S. box office revenue movies of 2018 are listed in an abbreviated Table 20-1. (Note that RT stands for Runtime. The names of the movies are also shortened, and only a few variables are listed, to give you ageneral idea.) You can see the whole data set at `www.dummies.com/go/statisticsIIfd2e` if you want (not required).

The first step in the data analysis process (as explained in Chapter 19) is visualizing the data; this is where the fun begins. Some of the preconceived ideas you had about the data will be right on target, and some will go out the window. Some patterns will emerge, and some will never show up. If you start with an open mind, anything can happen. Believe me, statisticians still get surprised every day when looking at new data sets — it's one of the fun things about this field.

TABLE 20-1 **The Top Eight Movies of 2018 in Terms of U.S. Box Office Revenue**

Name of Movie	Rated	RT	Days	Budget	Opening Weekend	U.S. Revenue
Black Panther	PG-13	134	175	$200,000,000	$202,003,951	$700,059,566
Avengers	PG-13	149	140	$321,000,000	$257,698,183	$678,815,482
Incredibles 2	PG	118	182	$200,000,000	$182,687,905	$608,581,744
Jurassic World.	PG-13	128	106	$170,000,000	$148,024,610	$417,719,760
Aquaman	PG-13	143	105	$160,000,000	$67,873,522	$335,061,807
Deadpool 2	R	119	154	$110,000,000	$125,507,153	$318,491,426
The Grinch	PG	85	98	$75,000,000	$67,572,855	$270,620,950
Mission Impossible	PG-13	147	84	$178,000,000	$61,236,534	$220,159,104

Categorical movie variables

I start with the categorical variables in the data set that make sense to make graphs of:release date (using month only), rating (by the Motion Picture Association), genre, and profit. Pie charts and bar graphs (from Stats I) are used to display each of these variables; the results (which were made in Minitab) are shown in Figure 20-1.

Looking at these graphs, you can see a few results right off the bat. In Figure 20-1a, which focuses on the release date variable, none of the top movies were released in January. (January does not show up on the graph, but you could add it with a zero for the count if you want.) You also note that July, November, and December were the hottest months, which kind of makes sense, with their being in the middle of summer and around the end-of-year holidays, respectively.

Figure 20-1b, which looks at the rating variable, shows you that over half of the top movies were rated PG-13, and the remainder were split between R-rated movies and PG-rated movies. The graph in Figure 20-1c looking at the genre variable is especially interesting — action movies made a huge splash here, with 15 of the 34 top movies (44%) being action movies, distantly followed by family movies with 7 and drama with 6.

The profit variable in Figure 20-1d is a little deceptive unless you think more deeply about it. The pie chart shows that only 85.3% of the top movies made a profit — but remember, that's only in the U.S. It makes me want to ask about the international/global market; perhaps that's another column I can add to my data set in the future.

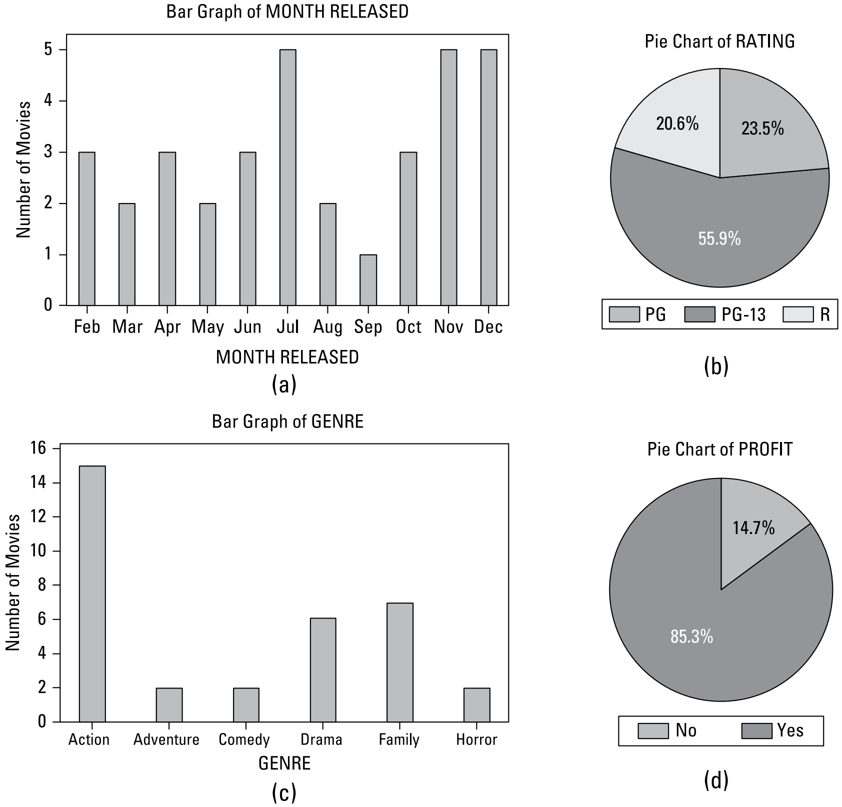


FIGURE 20-1: Pie charts and bar graphs of a) release date (month), b) rating, c) genre, and d) profit.

TIP

One of the things you'll find is that while you are examining your heaps and piles of data — what's the item you end up wanting? More data! And that's just fine; just build in time and funds to do that research in case it becomes necessary.

Quantitative movie variables

Next, I examine the quantitative variables from my list that make sense to visualize: runtime, days (in theaters), (number of) theaters, budget, opening weekend, U.S. revenue, critics' (ratings), and audience (ratings). These variables are graphed in Figure 20-2 using histograms and boxplots (from Stats I) in Minitab.

These graphs show you a great deal of general information about the center, variability, and patterns in the quantitative variables in this data set. Starting with runtime (Figure 20-2a), you see that the middle 50 percent of the top movies are within about 30 minutes of each other (between 105 and 135 minutes), with the sweet spot (the median) being just a shade under 120 minutes. Going back to the data, I found that *Avengers: Infinity War* and *Mission Impossible — Fallout* were the longest-running top movies, both being almost 150 minutes long, while

Mamma Mia! Here We Go Again was the shortest top movie, running only 84 minutes. The overall range of runtimes is large, while the middle 50% of movies are pretty close together, which I found to be interesting.

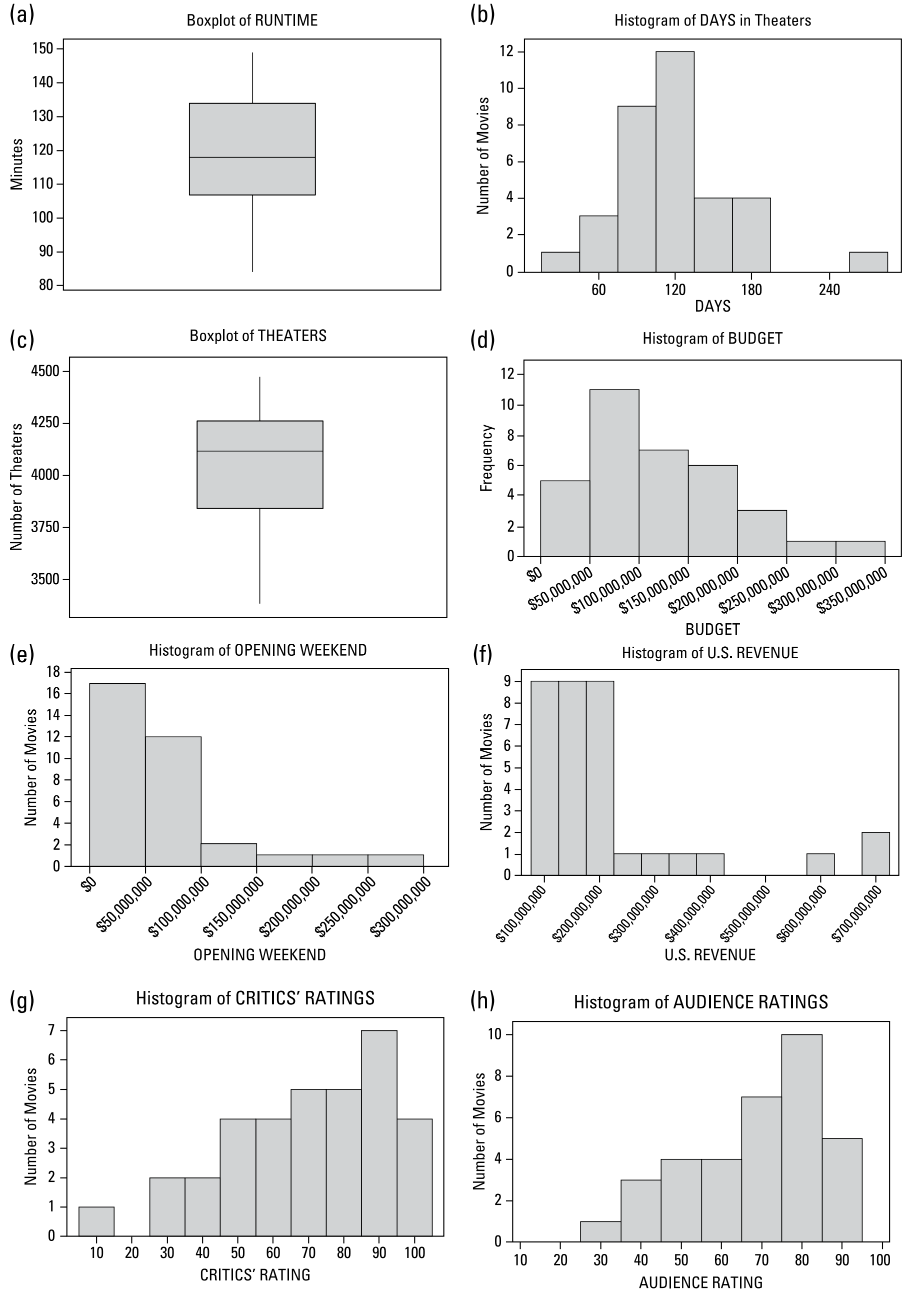


FIGURE 20-2: Histograms and boxplots of a) runtime through h) audience (ratings).

Figure 20-2b shows the number of days the movies were in U.S. theaters. I was surprised at the range of values found here as well. Movies went from around 30 days in theaters to around 185 days, except for one movie that outlasted all the others by a long shot. This movie lasted about 260 days (that's over 37 weeks or 70% of a year!). Which movie was it? (See how fun statistics can get?) Digging into the data set, I found that it was *The Mule*, with Clint Eastwood (my favorite actor and director!), at 262 days. The shortest-running top movie was *Rampage*, at 37 days. What a difference between 37 days and 37 weeks. Note, however, that while *Rampage* (starring Dwayne "The Rock" Johnson) didn't really get off the ground in the U.S., it made big money internationally, and he did just fine with it. By the way, it's interesting to note that exactly half of the top movies (17 out of 34) lasted between 100 and 120 days, so between 3 and 4 months, before everyone moved on to the next big thing.

Number of theaters is shown in Figure 20-2c. This variable has less variation than I anticipated; the top movies ranged from being shown in 3,400 to 4,500 theaters overall, with the middle 50% of them being shown in 3,800 to 4,250 theaters. Remember, these are the top movies, so they will tend to be shown in more places. If this data set had all movies made in 2018 (what a list!), the range would probably be huge!

Two movies almost tied for highest number of theaters: *Jurassic World: Fallen Kingdom* with 4,475 and *Avengers: Infinity War* with 4,474. I thought *Solo: A Star Wars Story* would be the top here; it was close, coming in with 4,381 theaters. Questions you might ask: "Is number of theaters related to number of days, and how accurate are these numbers at predicting U.S. box office revenue?"

Budgets of the top movies are displayed in Figure 20-2d. This graph is also skewed to the right, with most of the top movies having a budget of between $100 million and $200 million, and with a couple of outliers on the high end ($321 million for *Avengers: Infinity War*, and $275 million for *Solo: A Star Wars Story*). Does budget help predict box office revenue? You will have to build a model later in this chapter to find out!

Opening weekend U.S. box office revenue and total U.S. revenue are shown in Figures 20-2e and f. They are both skewed to the right (from your Stats I knowledge). They also look somewhat similar in pattern, with both having an outlier on the right (high) side.

WARNING

The question, though, is this: "Are the movies that do well on opening weekend the same movies that do well overall?" Just because two variables look similar, it doesn't mean the position of a movie on one graph matches the position of that movie on the other. That's where looking for relationships has to come in with a scatterplot (which I get to in the next section).

In case you were wondering, though, one movie does stand out with an opening weekend revenue of $270,000,000, and one movie stands out with an overall U.S. box office revenue of around $700,000,000. Are they the same movie? (Admit it, statistics is fun.) Looking into the data set, I find the answer is no, but they are close. *Black Panther* wins the overall U.S. box office revenue spot with just over $700,000,000, while *Avengers: Infinity War* is just $20,000,000, behind. And as for opening weekend, *Avengers: Infinity War* wins the top spot with over $250,000,000, with *Black Panther* following with around $200,000,000. Coincidence or pattern? You find out later in this chapter.

Critics' and audience ratings are skewed left, as shown in Figures 20-2g and h. Remember, I am talking about top movies here, all making over $100,000,000 at the U.S. box office, so they are going to have a tendency to be well-liked. You can see that the variability in critics' ratings is larger than the variability in audience ratings. The question is, do audiences and critics generally agree on movies? I've always wondered that. This is another question for later in the chapter, where I look for relationships between variables. Just by looking through the data, I note that the most highly rated movie by audiences was *Spider-Man: Into the Spider-Verse*, receiving an average rating of 93/100, and the most highly rated movie by the critics had a three-way tie between *Black Panther*, *Mission Impossible — Fallout*, and *Spider-Man: Into the Spider-Verse*, with all three receiving an average rating of 97/100.

Doing Descriptive Dirty Work

The next phase focuses on finding more details about center and variability in the quantitative variables, using descriptive statistics such as mean, median, standard deviation, and inter-quartile range (IQR), all from Stats I. To address this in an efficient way, I used Minitab to build a table showing the basic descriptive statistics for runtime through audience ratings. The results are shown in Table 20-2.

TABLE 20-2 Some Descriptive Statistics for Runtime through Audience Ratings

Statistics

Variable	Mean	StDev	Minimum	Median	Maximum	IQR
RUNTIME	118.06	17.23	84.00	118.00	149.00	27.25
DAYS	115.82	43.79	37.00	110.50	262.00	47.50
THEATERS	4042.2	310.2	3388.0	4118.0	4475.0	414.0
BUDGET	113676471	74325541	10000000	100000000	321000000	119750000
OPENING WKD	67120349	54785996	17509431	47802879	257698183	40392664
U.S. REVENUE	222494739	156950095	100407760	173245680	700059566	97687600
CRITICS	69.29	22.82	12.00	71.00	97.00	37.50
AUDIENCE	67.09	17.18	27.00	71.00	93.00	26.00

From Table 20-2, you can see the mean, standard deviation, minimum, median, maximum, and IQR of the quantitative variables that I made graphs of in the previous section. You can check the statistics against the graphs to help fill out the story; for example, if the median is close to the mean (as it is with runtime and days in theaters), the data should be fairly symmetric (and it is in Figures 20-2a and b, except for one outlier). If $\text{mean} < \text{median}$, the data are skewed left, as you can see in number of theaters, critics' ratings, and audience ratings in Figures 20-2c, g, and h. If mean > median, the data are skewed right, as you can see in the monetary variables found in Figures 20-2d, e, and f. You can use many of the tools from Stats I to look at and think about the results in Table 20-2.

WARNING

You may also notice that while many of the statistics for critics'ratings and audience ratings are similar, the data themselves do not necessarily match movie for movie. It's very tempting to think so, but that's not the case.

Looking for Relationships

After getting the *univariate* (single, separate variable) results, you look for relationships between pairs of variables. As mentioned in Chapter 19, you look for relationships between two quantitative variables using scatterplots and correlation (Chapter 5), and you look for regression in the section, "Building a Model for Predicting U.S. Revenue," later in this chapter. You look for relationships between two categorical variables using two-way tables and stacked bar graphs first, and then you apply Chi-square tests for independence (Chapter 15), also shown in the section, "Building a Model for Predicting U.S. Revenue." Finally, you look for relationships between categorical and quantitative variables using side-by-side boxplots first, then hypothesis tests (for two means) or ANOVA (for more than two means) later in this chapter, when I do hypothesis testing and model building.

Relationships between quantitative movie variables

Following the directions in Chapter 19, you look for relationships between the variables *runtime* through *audience ratings* using scatterplots and correlations (Chapter 5). Figure 20-3 is a matrix showing all pairs of relationships between these eight quantitative variables. (The bottom part of the matrix shows the same pairs of variables in reverse order from the top part of the matrix.) This matrix was made using the Minitab/Graph/Matrix Plot feature.

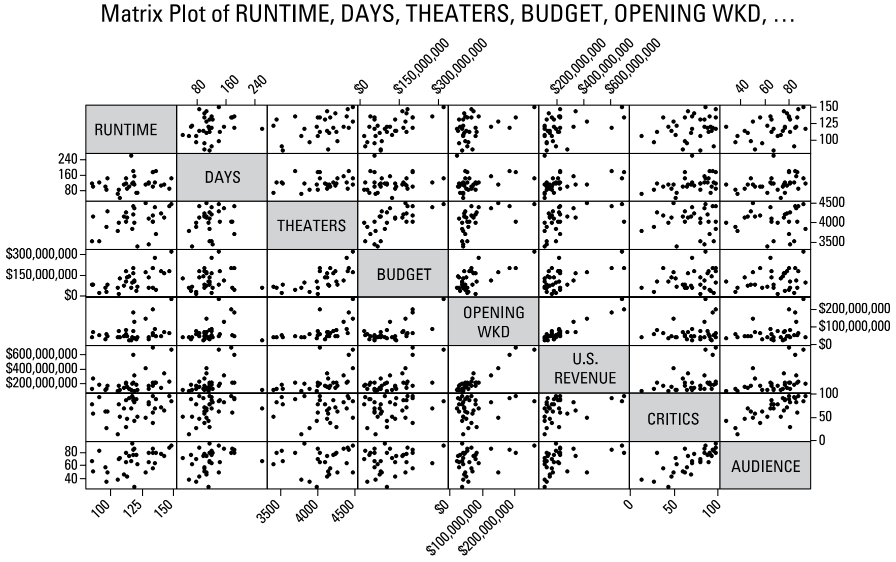


FIGURE 20-3: Matrix of scatterplots of runtime through audience ratings.

I'm interested in what variable or variables might help me to predict U.S. box office revenue, so I'm looking at column six of the matrix, which shows the relationships between U.S. revenue and runtime, U.S. revenue and days in theaters, U.S. revenue and budget, all the way through to U.S. revenue and audience ratings. I look for any strong relationships that might appear in that column of scatterplots, linear or nonlinear, and I find that opening weekend revenue seems to have a strong positive linear relationship with total U.S. revenue. The values seem to increase for both opening weekend and total U.S. revenue as you move from left to right, in a linear fashion (see Chapter 5).

Interestingly, what I don't see is a strong relationship between total U.S. revenue and budget — that's a little surprising. There may be some type of curved (nonlinear) relationship between total U.S. revenue and days in theaters and/or total U.S. revenue and number of theaters, but those relationships are not as strong as the one between total U.S. revenue and opening weekend revenue. So opening weekend revenue might be a good predictor of total U.S. box office revenue; more work needs to be done, but at least you've got something to look into.

Other interesting relationships (or non-relationships as the case may be) include audience and critics' ratings. If you look at column eight and row seven of the matrix in Figure 20-3, you see there may be a moderate positive linear relationship between the ratings. Perhaps audiences and critics aren't at odds as much as some people think!

If you look at row six (U.S. revenue) and the scatterplots in columns seven (critics' ratings) and eight (audience ratings), you see that both of those scatterplots have a flat appearance. That means no matter what the U.S. total revenue was for a movie, it wasn't related to either of the ratings. This might surprise you, but that's what the data are telling you, at least for these top movies of 2018.

Other thoughts as I look at these scatterplots: Budget may have at least a moderate positive linear relationship with runtime and also with number of theaters. These relationships both make sense, because moviemakers have to spend money to get their movies out, and the longer the movie, the more it may cost to make.

WARNING

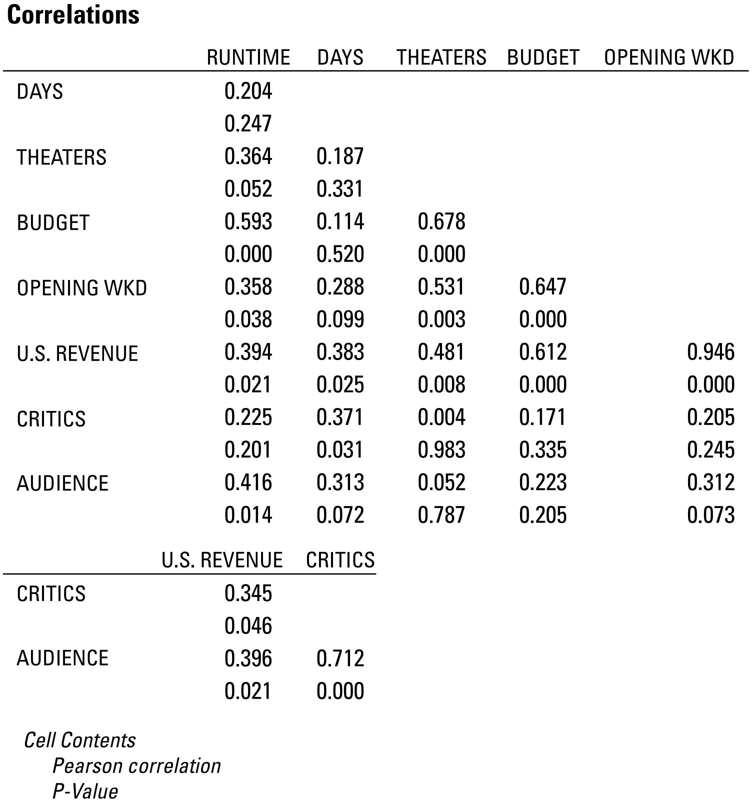
You have a large number of scatterplots here, and you don't want to start seeing relationships that aren't really present in the data, so it's great to do some looking and to pick out moderate to strong-looking relationships for further analysis. Just don't overdo it.

So, to summarize, here is what I think is worth looking into on a deeper level based on the scatterplots (that is, finding correlations in the case of linear relationships):

- Opening weekend and U.S. box office revenue
- Critics' ratings and audience ratings
- Budget and number of theaters
- Budget and runtime

As outlined in Chapter 19, the next step is to find correlations for relationships that are potentially moderate to strong and linear in nature. Figure 20-4 shows the results of finding the correlations for pairs of variables, from runtime through audience ratings.

Figure 20-4 has two entries for each cell: one is the sample correlation, r, between the two variables in the data, and appears in that particular row and column combination; the number below it is the *p*-value for testing to see whether the correlation is zero in the population. If the *p*-value is small, you conclude the correlation is not zero, and is statistically significant, according to the hypothesis test. (Note that a *p*-value listed as 0.000 means it is something less than 0.0005.) If the *p*-value is not small, you cannot say the linear relationship is statistically significant.



```
Correlations

                RUNTIME    DAYS   THEATERS   BUDGET   OPENING WKD
DAYS              0.204
                  0.247
THEATERS          0.364   0.187
                  0.052   0.331
BUDGET            0.593   0.114      0.678
                  0.000   0.520      0.000
OPENING WKD       0.358   0.288      0.531    0.647
                  0.038   0.099      0.003    0.000
U.S. REVENUE      0.394   0.383      0.481    0.612         0.946
                  0.021   0.025      0.008    0.000         0.000
CRITICS           0.225   0.371      0.004    0.171         0.205
                  0.201   0.031      0.983    0.335         0.245
AUDIENCE          0.416   0.313      0.052    0.223         0.312
                  0.014   0.072      0.787    0.205         0.073

             U.S. REVENUE   CRITICS
CRITICS           0.345
                  0.046
AUDIENCE          0.396     0.712
                  0.021     0.000

Cell Contents
    Pearson correlation
    P-Value
```

FIGURE 20-4: Matrix of correlations for runtime through audience ratings.

Looking at Figure 20-4 and picking out the correlations for the pairs of variables on my list, you find the following:

- **Opening weekend and U.S. box office revenue:** $r = 0.946$; p-value <0.0005
- **Critics' ratings and audience ratings:** $r = 0.712$; p-value < 0.0005
- **Budget and number of theaters:** $r = 0.678$; p-value < 0.0005
- **Budget and runtime:** $r = 0.593$; p-value < 0.0005

WARNING

Many hypothesis tests are being done here(28 to be exact). Each test is being conducted at level 0.05 (see Chapter 4); this means $5\% = 1$ or 2 of the hypothesis tests could come out significant just by random chance. You want to make that chance smaller, so you use the Bonferroni adjustment (Chapters 7 and 11) to change your p-values from 0.05 (a traditional p-value) to $0.05/28 = 0.0018$. So p-values that are deemed to be less than 0.0005 would be acceptable here.

So far, by looking at the scatterplots and correlations, you have found strong positive linear relationships between U.S. opening weekend and total U.S. box office revenue; critics' ratings and audience ratings; budget and number of theaters; and budget and runtime. And they all make good sense.

If you scan the matrix of correlations further in Figure 20-4, you see a couple of additional correlations of interest. First, a strong correlation exists between budget and opening weekend. However, when you look at the scatterplot in Figure 20-3 (column four, row five), you see that there seem to be two different groups of lines forming, both going upward, but with different slopes. I hesitate to say this relationship is a positive linear relationship based on the scatterplot.

The same thing happens when I see the correlation for U.S. revenue and budget (0.612) in Figure 20-4. The correlation itself is strong, and the p-value is <0.0005, but the scatterplot is not convincing in Figure 20-3. A blob of points appears on the left-hand side of the graph, while a few dots trail upwards. To me, that's not a strong-looking linear relationship.

REMEMBER

Always judge a positive or negative linear relationship based on both the scatterplot and the correlation, not just one or the other.

Relationships between two categorical variables

The categorical variables you have available are release date (day and month), rating (content), genre, and profit (yes or no). A couple of thoughts come to mind. First, you could examine content rating and genre to see if there is a relationship. You could also look at profit and rating, and profit and genre; however, because the profit variable only represents profit made in the U.S. (not international sales), I chose not to examine it here.

Figure 20-5 shows stacked bar graphs made in Minitab for content rating and genre. A stacked bar graph, as you know from Stats I, is a bar graph where each bar is broken down into pieces by a second variable.

As you can see from Figure 20-5, each bar in the graph represents a genre, and is broken down by content rating. Because none of the bars look alike, you can say there is a relationship between content rating and genre in your data set. You can also use the graph to describe that relationship. You can say that all the comedy and most of the action movies are rated PG-13 for a mixed age group, while all the family movies are rated PG for younger audiences. Adventure movies are split between PG and PG-13 for younger children and teens, dramatic movies are split between PG-13 and R, and all the horror movies are rated R.

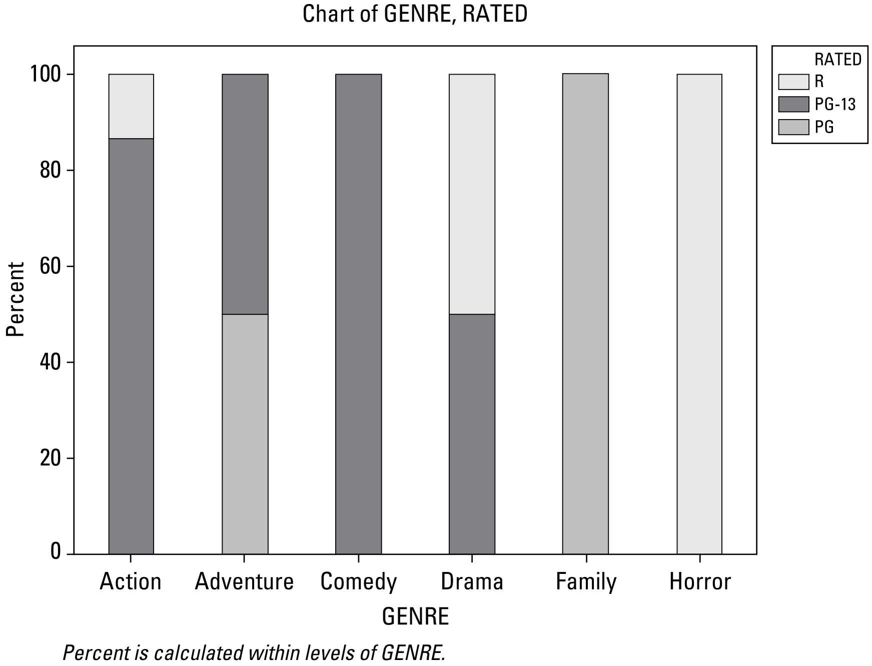


FIGURE 20-5: Stacked bar graph to explore content rating and genre.

Similarly, I created a stacked bar graph showing month of release and genre. I don't show the results here, but they were as you'd expect: more family movies were released in the summer, horror movies had big release months in September and October (of course), and more dramas came out in the winter. So, yes, there is a relationship between month of release and genre.

Relationships between quantitative and categorical variables

For this situation, I want to examine relationships where one variable is quantitative, and one is categorical. Side-by-side boxplots do a good job of looking for relationships of this type. In Minitab, I made side-by-side boxplots to help answer some interesting questions such as the following:

- Which genre costs the most to make?
- Which genre gets the highest audience ratings?

Figure 20-6 shows the side-by-side boxplots (from Stats I) that I made in Minitab to help answer these questions.

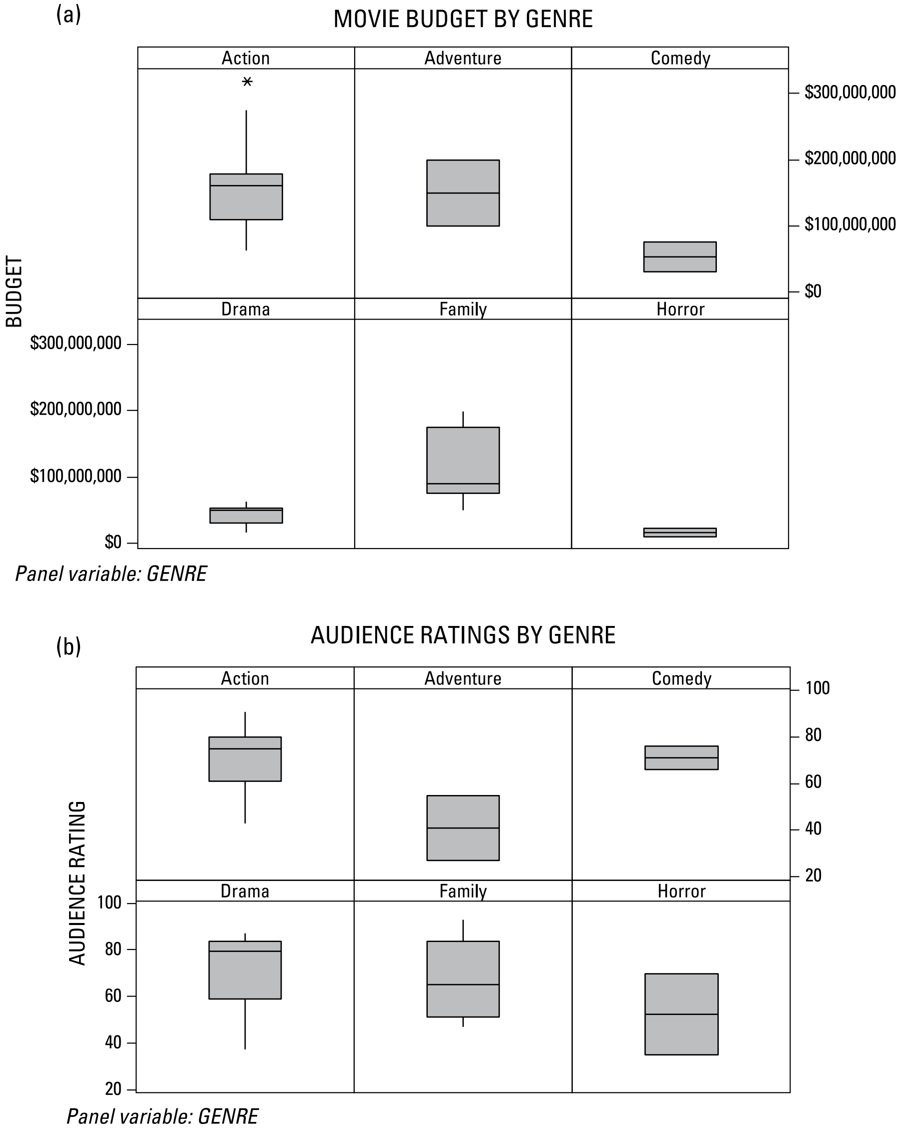


FIGURE 20-6: Side-by-side boxplots to look for relationships.

In Figure 20-6a, I made side-by-side boxplots to see which genre costs the most to make. This means you make one boxplot for each genre, with the variable being budget for each boxplot. You can see that horror movies, dramas, and comedies have the lowest budgets (mostly, people are talking in these movies, with horror movies containing some cheaper-style special effects). Family is the next lowest, with a median budget of around $100,000,000. Finally, action and adventure have the highest budgets, with medians in the range of $150,000,000 for adventure and

$200,000,000 for action. You can see an outlier in the action films denoted by an asterisk (*). This movie is *Avengers: Infinity War*, which had an extra-large budget of $321,000,000.

As for which genre gets the best audience ratings, check out Figure 20-6b. Starting at the bottom, the lowest ratings for top movies were for adventure movies (median 40%, IQR about 30%), followed by horror movies (median 50%, IQR about 40%), and then family movies (median 60%, IQR about 30%). Note that all of these boxplots show symmetric data for those three genres. As for the top three, you have action movies at an 80% median with 20% IQR; dramas, which also have a median of 80% but a larger IQR of 25%; and comedies close behind, with a 75% median but a small IQR of only 10%. So from these results, comedies get consistently high ratings from the audience. And, yes, a relationship appears to exist between audience ratings and the genre of top movies.

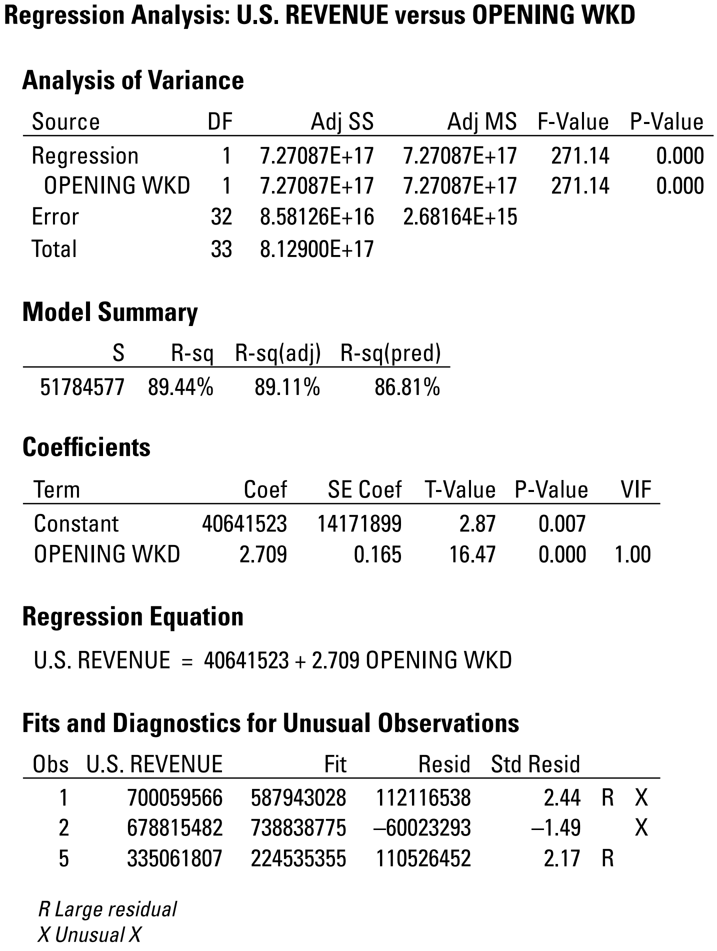
Building a Model for Predicting U.S. Revenue

Finally, you can see what all this work leads up to; you have some interesting theories to test about the preliminary results and some interesting relationships to try to model, starting with predicting a movie's U.S. box office revenue (relationship between two quantitative variables). You have a great candidate, opening weekend revenue, with a strong positive linear relationship according to the correlation and scatterplot.

The next step is to fit a simple (one-variable) linear regression model (see Chapter 5). Running the regression analysis using $X = \text{opening}$ weekend revenue and $Y = \text{U.S.}$ box office revenue, I get the following information, shown in Figure 20-7.

Figure 20-7 shows all the output that you get with a regression analysis in Minitab (it's the same for other software), starting with ANOVA output (see Chapter 13 for the relationship between ANOVA and linear regression). The Model Summary on the output shows R-squared at 89%, which is incredibly high, indicating that the model fits well. The Coefficients and Regression Equation are telling you the same thing, that the equation that best fits this relationship is as follows:

$$\text{U.S. Revenue} = \$40{,}641{,}523 + 2.709(\text{Opening Weekend})$$



```
Regression Analysis: U.S. REVENUE versus OPENING WKD

Analysis of Variance

Source        DF       Adj SS        Adj MS  F-Value  P-Value
Regression     1  7.27087E+17   7.27087E+17   271.14    0.000
  OPENING WKD  1  7.27087E+17   7.27087E+17   271.14    0.000
Error         32  8.58126E+16   2.68164E+15
Total         33  8.12900E+17

Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
51784577  89.44%     89.11%      86.81%

Coefficients

Term              Coef   SE Coef  T-Value  P-Value   VIF
Constant      40641523  14171899     2.87    0.007
OPENING WKD      2.709     0.165    16.47    0.000  1.00

Regression Equation

U.S. REVENUE  =  40641523 + 2.709 OPENING WKD

Fits and Diagnostics for Unusual Observations

Obs  U.S. REVENUE        Fit      Resid  Std Resid
  1     700059566  587943028  112116538       2.44  R  X
  2     678815482  738838775  –60023293      –1.49     X
  5     335061807  224535355  110526452       2.17  R

R Large residual
X Unusual X
```

FIGURE 20-7: Linear regression analysis using opening weekend revenue to predict U.S. box office revenue.

That is, if the opening weekend brings in $30,000,000, the movie is expected to make $\$40,641,523 + 2.709(\$30,000,000) = \$121,911,523$. Note that this is only for top movies, not all movies from 2018.

The Fits and Diagnostics for Unusual Observations at the bottom of the output identify three movies as being outliers. The observation number is actually the rank of the movie, so use that as your guide. The three movies listed as unusual are the movies with rank $1 =$ *Black Panther*; rank $2 =$ *Avengers : Infinity War*; and rank $5 =$ *Aquaman.* In the data set, these are the three points in Figure 20-8 with opening weekends above $150 million — crazy-high compared to the cloud of points to the left of them. However, the line fits these points quite nicely, so I don't see any reason to worry about the fit of this simple linear regression model.

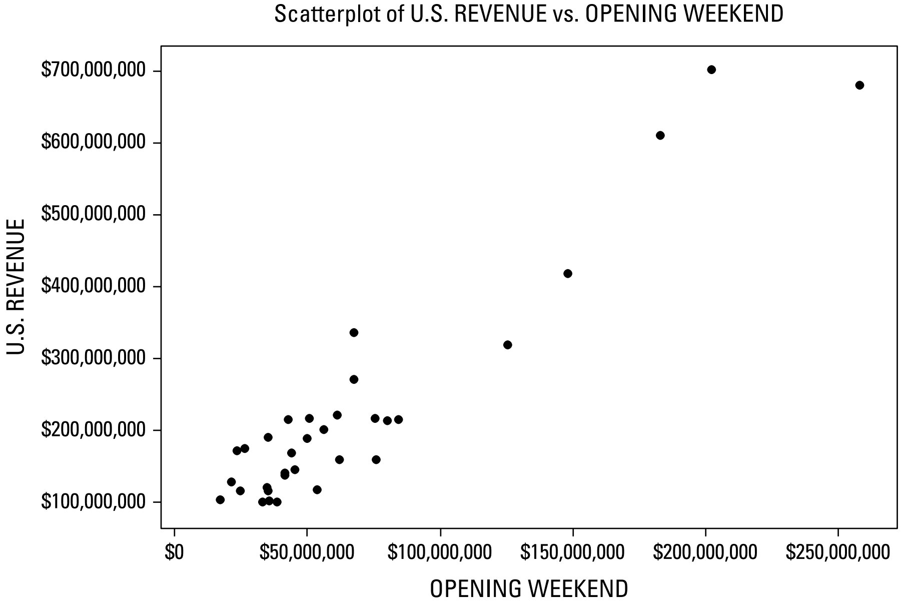
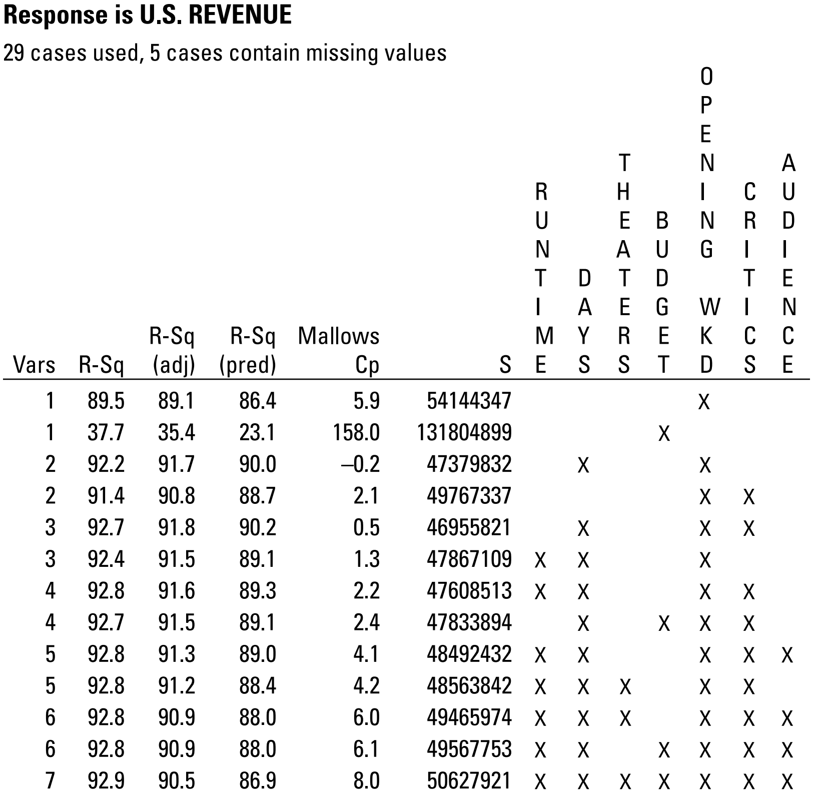


FIGURE 20-8: Final regression line on the scatterplot using opening weekend revenue to predict U.S. box office revenue.

TIP

This final model of using opening weekend revenue to predict U.S. box office revenue is also a good model because the *X* variable is something you can get ahead of time, not something you need to wait for until all the numbers come in — by then, the prediction wouldn't need to be made. You need a variable that fits well but that can also be obtained before the final numbers come out, and opening weekend revenue fits that bill.

Just for fun, I asked Minitab to choose the best model using the best subsets option (see Chapter 7), and it chose the same model I did (see the top line of Figure 20-9). Note that this model has a fairly low Mallow's C-p as well (see Chapter 7). You could add critics' ratings, and find that the model increases slightly in R-squared and lowers Mallow's C-p a little, but it may not be worth it to add that extra variable.



```
Response is U.S. REVENUE
29 cases used, 5 cases contain missing values
                                                      O
                                                      P
                                                      E
                                                  T   N   A
                                              R   H   I C U
                                              U   E B N R D
                                              N   A U G I I
                                              T D T D   T E
                                              I A E G W I N
                 R-Sq    R-Sq  Mallows        M Y R E K C C
Vars  R-Sq  (adj)  (pred)       Cp          S E S S T D S E
   1  89.5   89.1    86.4      5.9   54144347         X
   1  37.7   35.4    23.1    158.0  131804899       X
   2  92.2   91.7    90.0     –0.2   47379832   X     X
   2  91.4   90.8    88.7      2.1   49767337         X X
   3  92.7   91.8    90.2      0.5   46955821   X     X X
   3  92.4   91.5    89.1      1.3   47867109 X X     X
   4  92.8   91.6    89.3      2.2   47608513 X X     X X
   4  92.7   91.5    89.1      2.4   47833894   X   X X X
   5  92.8   91.3    89.0      4.1   48492432 X X     X X X
   5  92.8   91.2    88.4      4.2   48563842 X X X   X X
   6  92.8   90.9    88.0      6.0   49465974 X X X   X X X
   6  92.8   90.9    88.0      6.1   49567753 X X   X X X X
   7  92.9   90.5    86.9      8.0   50627921 X X X X X X X
```

FIGURE 20-9: Results of best subsets regression model selection for predicting U.S. box office revenue.

Writing It Up

Writing up your results may seem like an arduous chore, but it’s a very important part of the job. You’ve spent a great deal of time analyzing your data to this point, and you want to let your audience know what you found. What did you find out by analyzing the movie data? And what do you want to tell your audience, whoever they may be?

After taking all the information found in this chapter about the top U.S. money-making movies of 2018, including pertinent information on the data itself (such as variable names, and so on), and making an outline of the process I went through, keeping the highlight findings in mind, I wrote a short report. Your report may be longer if your audience needs more details; it may also have a different voice, depending on your audience, but the main ideas will be similar to mine. Note that I’m not including the graphs and tables again in my report because they are sprinkled throughout this chapter already; however, I would include them in a normal report and refer to them as I refer to tables and figures in this book.

Stats of Top Money-Making Movies Spark Interest

The top money-making movies of 2018 brought up some interesting points when studied as a group. (A movie was declared a top money-making movie if it made over $100,000,000 at the U.S. box office.) The top movie of 2018 was Black Panther, *making over $700 million in U.S. box office revenue, followed closely by* Avengers: Infinity War, *making over $678 million. Thirty-four movies in total made the cut.*

More of these movies were released in mid-summer and over the winter break; the content rating was over 55% for PG-13, with the remaining top movies split between PG and R. Most, by far, were action movies (15/34). Over 85% of them made a profit in the U.S.

The runtime of top movies varied from 84 minutes for Mama Mia! Here We Go Again *(a musical) to 149 minutes for* Avengers: Infinity War *(an action movie). Half of the movies ran between about 110 and 130 minutes (close to 2 hours, give or take 10 minutes). Days in theaters also varied, from 37 days for* Rampage *to about 37 weeks for* The Mule. *Half of the movies lasted between 100 and 120 days (17/34). Half of the movies were shown in between 3,800 and 4,250 theaters across the U.S., with a low of about 3,400 and a high of about 4,500.*

A relationship exists between content rating and genre. All the comedies and most of the action movies were rated PG-13, while almost all the family movies were rated PG. Adventure movies were split between PG-13 and R, and all the horror movies were rated R. Month of release was also related to genre; more family movies came out in summer, more horror movies in the fall, and more dramas in winter.

Budgets varied a great deal; the average budget of top movies was $113,676,471, with a minimum of $100,000,000 for Venom, *and a maximum of $321,000,000 for* Avengers: Infinity War. *U.S. box office revenue averaged $222,494,739, from $100,407,760 for the movie* Fifty Shades Freed *to $700,059,566 for* Black Panther. *Opening weekend revenue averaged $67,120,349 for the top movies of 2018, and peaked out at over $257 million for* Avengers: Infinity War.

Budget and genre were related, showing that not all types of movies had the same budgets. Horror movies, dramas, and comedies had the lowest budgets, with family movies coming in at the next budget-level, and action and adventure movies coming in at the top by a big margin.

Critic and audience ratings for the top money-makers of 2018 were both positive and skewed left (a few lower ratings, and mostly higher ratings). Critics' ratings averaged 69.29/100 and audience ratings averaged 67.09; critics' ratings varied a bit more than audience ratings (with standard deviations of 22.82 and 17.18, respectively). However, a strong correlation exists between critics' and audience ratings ($r = 0.712;\ p < 0.0005$).

Opening weekend revenue and U.S. box office revenue were highly correlated for the top money-making movies ($r = 0.946$; $p < 0.0005$), and it turns out that opening weekend was a good predictor of total U.S. box office revenue. You can predict total U.S. box office revenue using the equation $\$40{,}641{,}523 + 2.709$(opening weekend revenue).

IN THIS CHAPTER

- » Exploring a refrigerator data set
- » Analyzing the data
- » Writing it up

Chapter 21

Looking Inside the Refrigerator

Another data set I assembled myself is about refrigerators. I was buying a new refrigerator at the time and I figured, what better way to do my research on refrigerators than to put together a data set about some of the more common ones and analyze it? (Yeah, only a statistician would do something like that.) The data set was collected from refrigerators being sold at Home Depot (.com) and is available at `www.dummies.com/go/statisticsIIfd2e` if you want to examine it yourself or play along with me. (You don't *need* to do this, however. I've done all the work for you in this chapter.) Note that refrigerator models and prices are subject to change frequently.

This chapter is a boiled-down version of my findings upon creating, wrangling, and analyzing the data, sharing my graphs, and writing a short report.

Refrigerator Data — The Variables

The refrigerator data include information from 41 models of refrigerators appearing on the homedepot.com website in February, 2018. The 13 variables I chose to collect on each refrigerator model include the following:

Brand: Amana, GE, Whirlpool, Frigidaire, and Samsung.

Capacity: In cubic feet (a.k.a. cubic foot capacity).

Price: In dollars.

Type: Door arrangement — for example, bottom freezer, top freezer, French doors, and so on.

Depth: Inches at deepest part.

Width: Inches.

Height: Inches.

Icemaker: Yes or No.

Stars: Customer rating on a scale of 0 to 5 stars.

Colors: Number of colors offered.

Yearly electric use: Measured in kilowatt-hours (KwH).

Shelves: Number of shelves.

Model Number: The model number of the refrigerator.

In this data set, I was interested in comparing the refrigerator brands to see if anything jumped out at me.

Exploring the Data

Upon looking at several graphs that compared the refrigerator brands, here are some of the points I noticed.

Basic statistics were gathered on the overall group of 41 refrigerators. The results are shown in Table 21-1. Looking at depth, width, and height, width had the largest standard deviation (it's good to measure it before you try to install your fridge!). I didn't know that a refrigerator could have seven shelves, but the number of shelves ranged from 4 to 7. As far as the ratings go, it seems that customers were fairly happy with these brands of refrigerators overall, with the lowest rating being 3.8 and the highest being 4.6. Capacity had a fairly large standard deviation, with the minimum being as low as 10.100 cubic feet, and the maximum coming in at 28.000 cubic feet. Yearly electric use also had a large standard deviation, comparatively speaking, with the average being 575.6 KwH, and the maximum as high as 755.0 KwH.

TABLE 21-1 Descriptive Statistics for 41 Refrigerators

Variable	Mean	StDev	Minimum	Median	Maximum
Cubic ft. capacity	21.544	4.486	10.100	22.000	28.000
Price ($)	1,579	714	599	1,554	3,110
Depth (in.)	31.539	1.980	26.250	31.630	35.000
Width (in.)	32.724	3.337	23.750	32.813	36.000
Height (in.)	68.331	2.377	60.500	69.500	71.130
Stars (out of 5)	4.2805	0.2015	3.8000	4.3000	4.6000
Colors	3.195	1.054	1.000	3.000	5.000
Yearly Electric use (KwH)	575.6	137.3	297.0	633.0	755.0
Shelves	4.902	1.068	4.000	4.500	7.000

The prices were all over the board, with an average of $1,579, minimum of $599, and maximum of $3,110. The median price was $1,554, indicating some right-skewness in the prices. Figure 21-1 shows a histogram of the refrigerator prices.

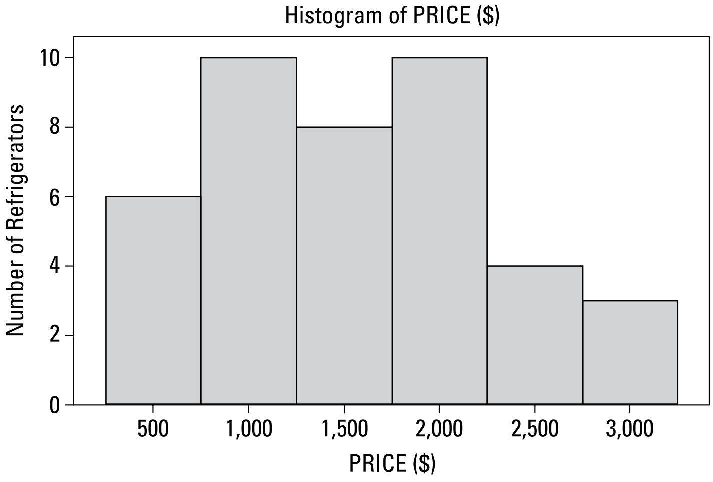


FIGURE 21-1: Histogram of refrigerator prices.

Regarding price of refrigerators broken down by brand, Figure 21-2 shows that Samsung appeared to be the most expensive overall. GE had the smallest IQR by far, but it also had two outliers, one on each end, in terms of refrigerator prices. Interesting marketing strategy!

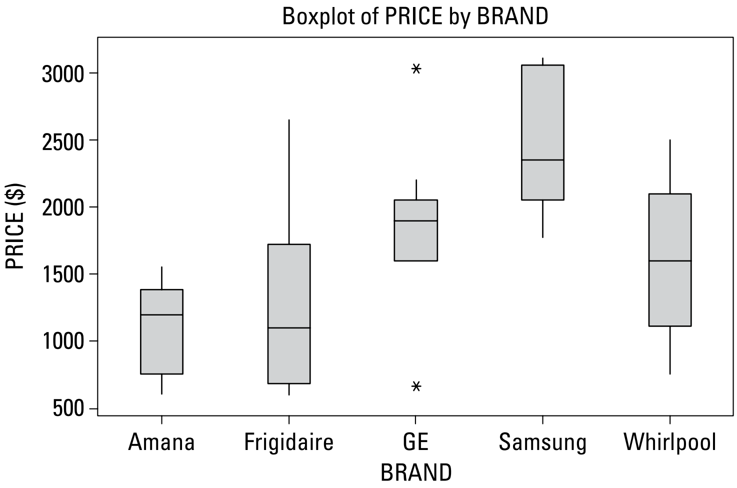
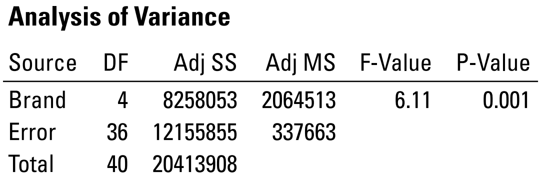


FIGURE 21-2: Boxplots of refrigerator price by brand.

Analyzing the Data

To see if there is a statistically significant difference in the mean prices of the five brands, I conducted an ANOVA (see Chapter 10). The conditions for ANOVA were robust, and I had a small sample of refrigerators to work with, but it is reasonable to assume normal distributions and independence in this example for demonstration purposes.

The results of the ANOVA are shown in Figure 21-3. The null hypothesis of the overall *F*-test is saying that the means are all equal, and the alternative hypothesis is saying they are not all equal.



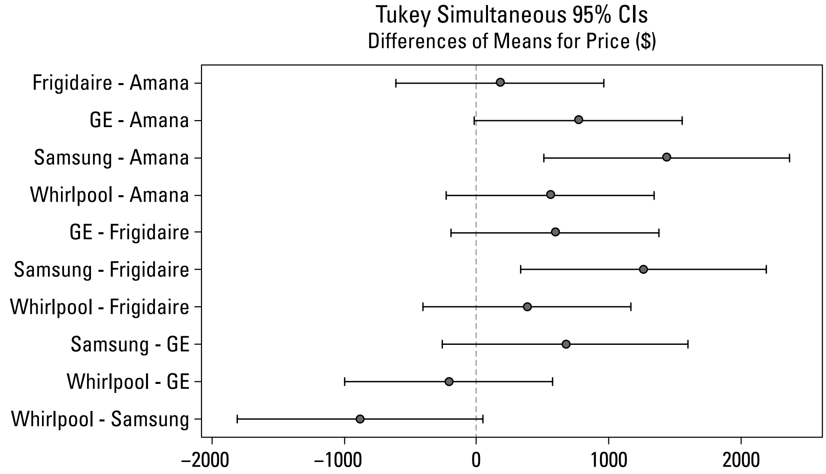
```
Analysis of Variance

Source  DF    Adj SS   Adj MS  F-Value  P-Value
Brand    4   8258053  2064513     6.11    0.001
Error   36  12155855   337663
Total   40  20413908
```

FIGURE 21-3: ANOVA for testing equality of mean prices across the refrigerator brands.

You see in Figure 21-3 that the p-value is 0.001, which is very small, leading you to reject H_0, and say that the mean prices were not all equal across brands (which I suspected). Now I could apply multiple comparisons to see which means were statistically higher than others and which were statistically equal.

Graphically, the results of Tukey's multiple comparisons are shown in Figure 21-4a (see Chapter 11), and the grouping information is shown in Figure 21-4b.

(a)



If an interval does not contain zero, the corresponding means are significantly different.

(b)

Grouping Information Using the Tukey Method and 95% Confidence

Brand	N	Mean	Grouping	
Samsung	5	2512	A	
GE	9	1842	A	B
Whirlpool	9	1631	A	B
Frigidaire	9	1251		B
Amana	9	1075		B

Means that do not share a letter are significantly different.

FIGURE 21-4: Tukey's multiple comparisons using confidence intervals and by brand.

The main idea is that if Tukey's confidence intervals overlap for two brands in Figure 21-4a, and they receive the same group letter in Figure 21-4b, then they are determined to be indistinguishable, and are put in the same group regarding price. The results of Figures 21-4a and b show, due to the small amounts of overlap here and there, that you have Samsung, GE, and Whirlpool in one group, and GE, Whirlpool, Frigidaire, and Amana in another group. Samsung is almost alone at the top pricewise, but not quite. Samsung's prices were higher than those of

Frigidaire and Amana. Not all output is squeaky clean; a friend of mine has a T-shirt that says "Messy Data Happen!" and they're right. But you get the general idea here.

If you want to build a model to predict the price of a certain brand of refrigerator, which variable (or variables) should be included? Looking at which variables might have a linear relationship with price, as you may have suspected, capacity (in cubic feet) is a pretty good predictor, with an R-squared of 0.771 (*p*-value <0.0001). See the fit of the line in the scatterplot in Figure 21-5.

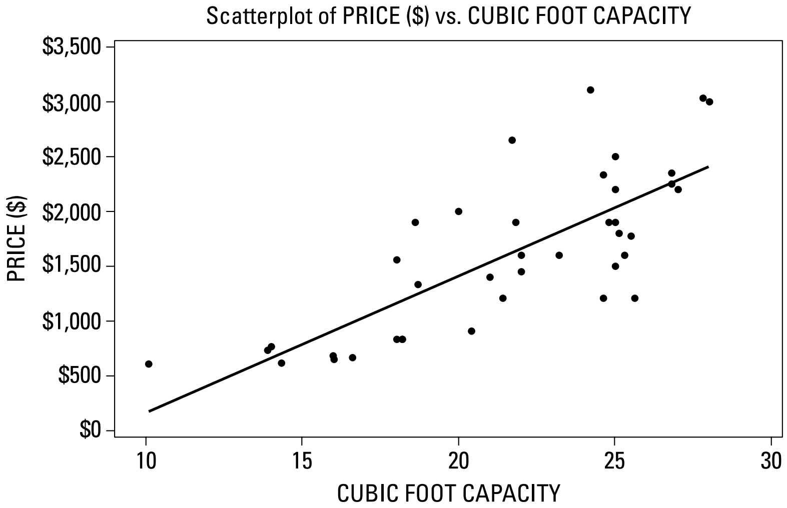


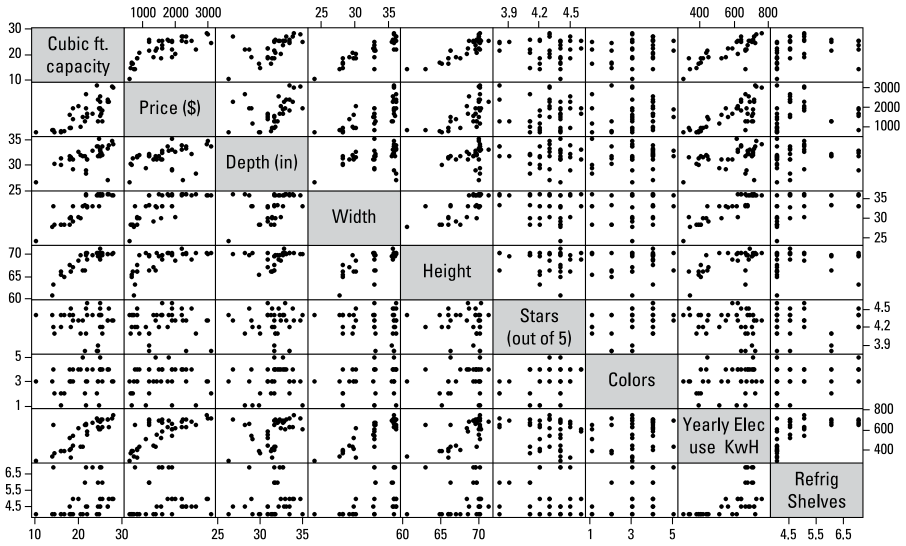
FIGURE 21-5: Cubic-foot capacity being used to predict refrigerator price.

Is there an even better model to use? To check this out, I made a matrix of all the correlations between the quantitative variables (price, cubic foot capacity, depth, width, height, stars, colors, yearly electric use, and number of shelves). I also had Minitab calculate all the correlations (and their *p*-values) between the pairs of variables. The results are shown in Figures 21-6a and b.

Taking the Bonferroni adjustment into account (see Chapter 20), I found the following pairs of relationships to be significant to a level that is less than 0.0001, and also practically significant (I only used *r* values of at least 0.50), so they are meaningful:

- Price and cubic foot capacity $(p = 0.771)$
- Price and width $(p = 0.750)$
- Price and height $(p = 0.652)$

(a) Matrix Plot of Cubic Foot Capacity, Price ($), Depth (in), Width, Height, …



(b) **Correlation: Cubic ft. capacity, Price ($), Depth (in), … H, Refrig Shelves**

```
Correlations
                  Cubic ft. capaci  Price ($)  Depth (in)  Width
Price ($)                    0.771
                             0.000
Depth (in)                   0.597      0.398
                             0.000      0.013
Width                        0.896      0.750       0.402
                             0.000      0.000       0.012
Height                       0.870      0.652       0.226  0.718
                             0.000      0.000       0.207  0.000
Stars (out of 5)            –0.065     –0.049      –0.049 –0.157
                             0.692      0.763       0.772  0.328
Colors                       0.281      0.113       0.172  0.266
                             0.083      0.482       0.303  0.092
Yearly Elec use              0.836      0.743       0.413  0.902
                             0.000      0.000       0.010  0.000
Refrig Shelves               0.324      0.100       0.142  0.491
                             0.044      0.534       0.396  0.001

                  Height  Stars (out of 5)  Colors  Yearly Elec use
Stars (out of 5)  –0.095
                   0.582
Colors             0.127             0.301
                   0.461             0.056
Yearly Elec use    0.673            –0.212   0.270
                   0.000             0.183   0.088
Refrig Shelves     0.188            –0.079   0.395            0.623
                   0.273             0.624   0.011            0.000

Cell Contents
   Pearson correlation
   P-Value
```

FIGURE 21-6: Matrix of scatterplots for refrigerator data and matrix of correlations and *p*-values for refrigerator data.

- Price and yearly electric use $(p = 0.743)$
- Cubic foot capacity and depth $(p = 0.597)$
- Cubic foot capacity and height $(p = 0.870)$
- Cubic foot capacity and width $(p = 0.896)$
- Cubic foot capacity and yearly electric use $(p = 0.836)$
- Height and width $(p = 0.718)$
- Yearly electric use and width $(p = 0.902)$
- Yearly electric use and number of shelves $(p = 0.623)$

The issue here is that I had a great deal of multicollinearity between the variables that could be used to predict price. Cubic foot capacity is strongly linearly related to price, but then, depth, height, width, and yearly electric use are strongly linearly related to cubic foot capacity, so they are knocked out of the running to help predict price (see Chapter 7).

Nothing else is strongly linearly related to price, so the best model has price being predicted by cubic foot capacity. The equation is as follows:

$$\text{Price} = -1{,}085 + 124.7 * \text{Capacity (in cubic feet)}$$

Interpreting the slope, each additional cubic square foot of capacity is associated with a price increase of $124.70. Note that you cannot interpret the *y*-intercept here, as it is negative.

WARNING

You cannot start making predictions until you've reached a capacity of at least 10.10 (the minimum value in the data set; see Figure 21-1). That price estimate is $-1{,}085 + 124.7(10.10) = \$174.47$. Note that the actual retail price (as *The Price is Right* show would say) is $599 from the data set, making this data point a big underestimate. However, it's an outlier, on the edge of the rest of the points (see Figure 21-5), so it should be looked at with caution. Starting with capacities around 15 will give you better estimates.

I also looked at yearly electric use. What can predict this? I wouldn't want to use price, because it makes more sense to use yearly electric use to predict price than the other way around. But what about number of shelves? What about capacity?

Turns out, any of the dimensions of a fridge (height, width, depth) are highly correlated with yearly electric use. And they are also highly correlated with each other, as you can see in Figures 21-6a and b. This means you should just choose one dimension as your predictor of yearly electric use if you are going to choose any.

With a correlation of 0.902 (see Figure 21-6b), width seems to be the best candidate. It's quite amazing that I found such a high correlation here; correlations like this are hard to come by in real life. And the nice thing is, width is a very easy variable to get a hold of. The resulting scatterplot and regression line are shown in Figure 21-7, using width to predict yearly electric use.

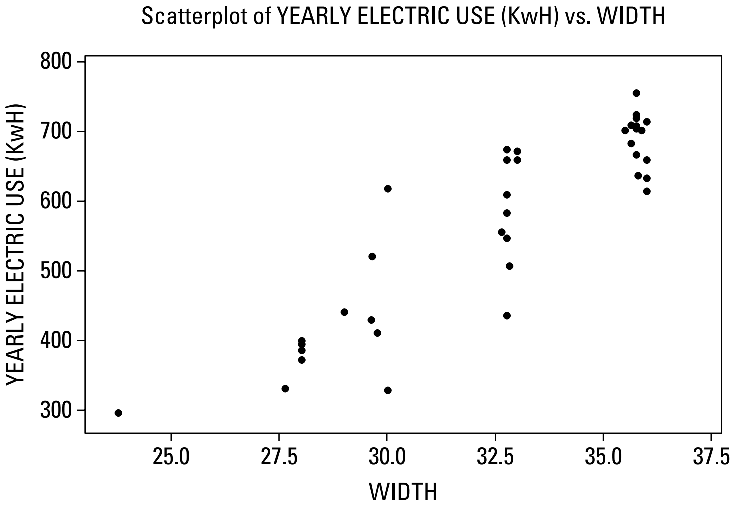


FIGURE 21-7: Scatterplot and regression line for predicting yearly electric use using width of the refrigerator.

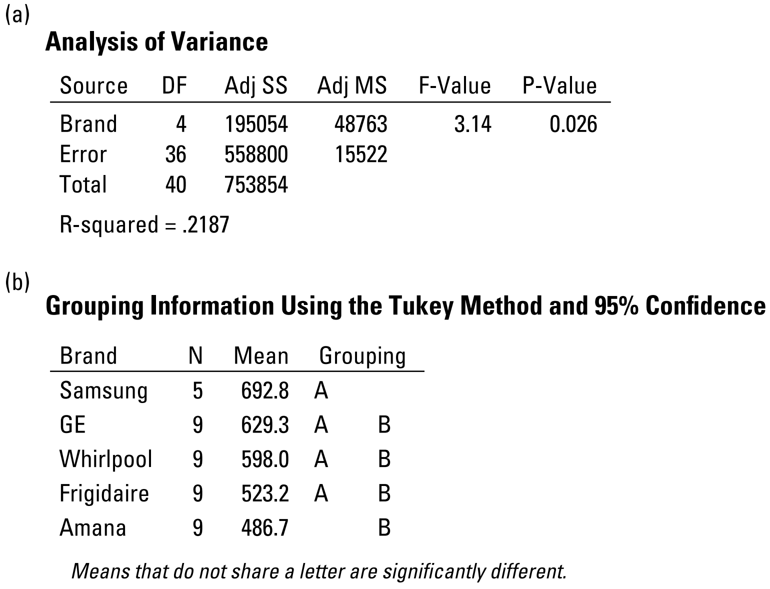
The equation of the regression line is

$$\text{Yearly Electric Use (in KwH)} = -638.4 + 37.10 * (\text{Width})$$

To interpret the slope, each additional inch of width on a refrigerator results in 37.10 additional kilowatt-hours of yearly electric use, on average, according to this model. And looking at Figure 21-7, you can see the range of widths for which I was comfortable making predictions. (The y-intercept is negative, and not interpretable here, but no data exists where $x = \text{width of a refrigerator} = 0$, either.)

You can also make comparisons of the average yearly electric use by brands, again using ANOVA. Are some brands more energy-conscious than others? Figures 21-8a and b show the results of the analysis of variance and Tukey's multiple comparisons in this case (see Chapters 10 and 11, respectively).

The groupings in Figure 21-8b tell you that Samsung (which, if you remember, was the most expensive from Figure 21-4b) is also the brand that uses the most energy. (I would have thought it would be the other way around, but perhaps some refrigerators have more bells and whistles that drive up the price and the electric use.) And Amana has the lowest energy use (it's also tied with Frigidaire as less expensive than Samsung in Figure 21-4b).



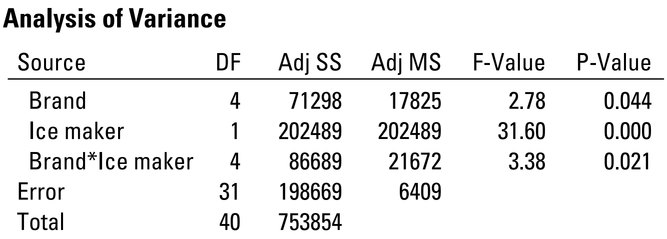
```
(a)
Analysis of Variance

Source   DF   Adj SS   Adj MS   F-Value   P-Value
Brand     4   195054    48763      3.14     0.026
Error    36   558800    15522
Total    40   753854

R-squared = .2187

(b)
Grouping Information Using the Tukey Method and 95% Confidence

Brand        N    Mean   Grouping
Samsung      5   692.8   A
GE           9   629.3   A     B
Whirlpool    9   598.0   A     B
Frigidaire   9   523.2   A     B
Amana        9   486.7         B

Means that do not share a letter are significantly different.
```

FIGURE 21-8: a) Analysis of Variance (ANOVA) for comparing yearly electric use by brand; b) Tukey's multiple comparisons of yearly electric use by brand.

The R-squared for this model is only 21.87% (see Figure 21-8a), which is okay, but not great. A lot of information remains out there to explain the variability in yearly electric use, although brand is a significant variable, as you can see in Figure 21-8a. Can you add another variable that will help break down the variance in yearly electric use and help raise the R-squared of this model? After some thinking and checking, I decided to include the icemaker variable.

Whether or not a refrigerator has an icemaker may be related to its electric use, and it may provide additional information to this model beyond brand, because each brand has some models that have icemakers, and some models that don't.

I created a two-factor ANOVA using the factors brand and icemaker (yes/no) to break down the response variable, yearly electric use. The results are shown in Figure 21-9. (See Chapter 12 for all the information on two-factor ANOVA.)



```
Analysis of Variance

Source              DF   Adj SS   Adj MS   F-Value   P-Value
  Brand              4    71298    17825      2.78     0.044
  Ice maker          1   202489   202489     31.60     0.000
  Brand*Ice maker    4    86689    21672      3.38     0.021
Error               31   198669     6409
Total               40   753854
```

FIGURE 21-9: Two-factor ANOVA with brand and icemaker as the factors and yearly electric use as the response.

As you can see in Figure 21-9, the *p*-value for the interaction term Brand*Icemaker is statistically significant $(p = 0.021)$. This means you can't explore the main effects of brand and icemaker by themselves. The interaction plot of Brand*Icemaker is shown in Figure 21-10.

You can see the lines are not parallel — the direction is the same for each brand. Moving from no icemaker to having an icemaker does increase the yearly electric use, but the amount of increase is very different, depending on the brand. For example, the yearly electric use going from no icemaker to icemaker hardly increases at all for Whirlpool (the line has a very small slope), while the difference for Frigidaire is huge (a very large slope).

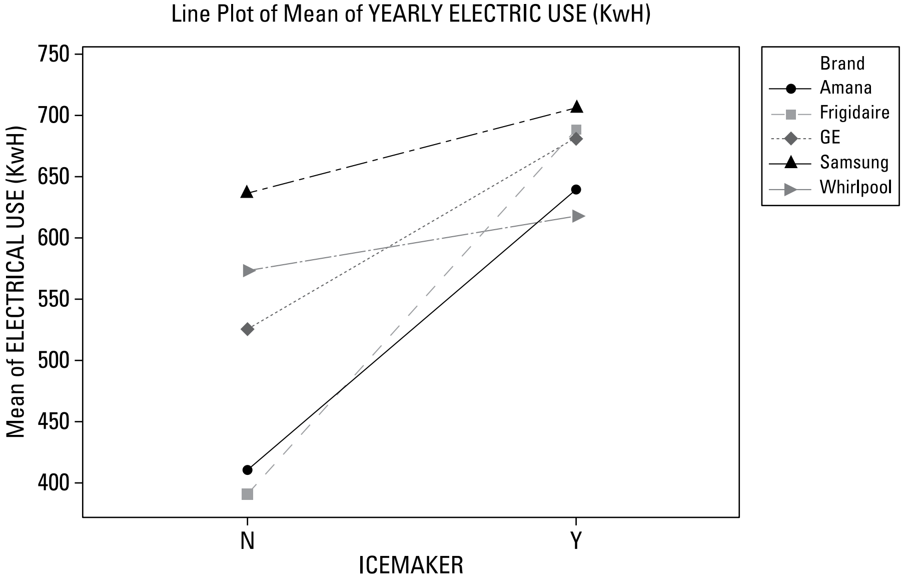


FIGURE 21-10: Interaction plot showing mean yearly electric use for each brand, broken down by whether it had an icemaker or not (Brand* Icemaker).

The next step would be to run ANOVA on the $5 * 2 = 10$ combinations of Brand*Icemaker (for example, Whirlpool and icemaker, Whirlpool and no icemaker, Amana and icemaker, Amana and no icemaker) and see where the differences lie. The results of the one-way ANOVA using Brand*Icemaker as the single factor show basically two groups as far as energy use is concerned: the icemaker group versus the no icemaker group. Two "outliers" from this pattern were found: one of the Samsung refrigerators without an icemaker had an extra-high amount of energy use, and one of the Whirlpool refrigerators with an icemaker had an extra-low amount of energy use.

Writing It Up

In this section, I provide a short report on what I was able to glean from the refrigerator data. Your report may be longer, shorter, or have a different voice, but the main ideas will be about the same.

REMEMBER

Your audience is very important to consider when writing a statistical report. What they know and what details they care about are paramount. Some care a lot, while some want to get to the nitty gritty. Also be aware of the amount of space you have available; you don't want to give too many details if you only have a small amount of space, but you don't want to cut it short if you have more space. No matter how much space you have, be sure to be true to the data — don't stretch the truth and don't sell it short.

I don't do it again here in my report because I did it throughout the chapter, but be sure to refer to any graphs that back up your work, and do include *p*-values (as I do) where appropriate. Readers need to see *p*-values so they can make their own decisions.

Examining Refrigerator Data — So Many Differences!

We are all used to staring into the refrigerator, wondering what will pop out at us and look good to eat. Now you've got the chance to look at many refrigerators at the same time, compare and contrast them, and figure out where their differences lie. (And also figure out what looks good to you!)

I compared 41 different models of refrigerators from five brands to see what they had to offer. The variables I examined included capacity, price, door arrangement (type), customer rating (stars), number of shelves, colors, yearly electric use, whether they had an icemaker or not, and the measurements of height, depth, and width.

Customer ratings were fairly positive overall; the average rating was 4.28 with a minimum of 3.80. The highest rating was 4.60, given to two Whirlpool refrigerators: one top freezer model and one side-by-side model. Capacity had a large standard deviation, with the smallest fridge at 10.1 cubic feet and the largest at 28.0 cubic feet. Yearly electric use varied as well, with an average of 575.6 KwH per year, and a maximum as high as 755.0 KwH. The minimum usage was 297.0KwH from a 10-cubic-foot capacity fridge.

Prices were also all over the board, with an average price of $1,579, a minimum of $599, and a maximum of $3,110. A few refrigerators cost way more than the rest. By brand, Samsung was the most expensive overall; GE had the smallest amount of deviation in price, except for two outliers on the high side and the low side. Overlap existed between GE and Whirlpool at the higher end, with Samsung and the other four brands in a lower price group overall.

If you want to predict the price of a refrigerator, check its capacity (in cubic feet). The best prediction line I could find was $price = -1{,}085 + 124.7$ *(capacity in cubic feet). Start at capacities of at least 15 cubic feet to use this equation.*

If you want to predict yearly electric use (in KwH), interestingly, the width of the refrigerator does a great job. Use the equation, $yearly\ electric\ use = -638.4 + 37.10$ *(width), starting at widths of around 25 inches. To interpret this equation, every additional 1 inch of width of a refrigerator is related to a 37.10 KwH increase in yearly electric use, on average. Some brands use more energy than others; Samsung was among those at the higher end, with Amana at the lower end. Brand and ice-maker combinations turned out to be a factor in determining KwH use. Some brands had a big difference in KwH when an icemaker was present; some brands had only a small increase.*

6
The Part of Tens

IN THIS PART . . .

Discover ten common errors in statistical conclusions.

Find out ten ways that knowing statistics can help you get ahead.

Check out ten cool jobs that use statistics.

IN THIS CHAPTER

- » Recognizing and avoiding mistakes when interpreting statistical results
- » Deciding whether or not someone's conclusions are credible

Chapter 22

Ten Common Errors in Statistical Conclusions

Stats II is all about building models and doing data analysis. It focuses on looking at data and figuring out the story behind it. It's about making sure that the story is told correctly, fairly, and comprehensively. In this chapter, I discuss some of the most common errors I've seen as a teacher and statistical consultant for many moons. You can use this "not to do" list to pull ideas together for homework and reports or as a quick review before a quiz or exam. Trust me — your professor will love you for it!

Claiming These Statistics Prove . . .

Be skeptical of anyone who uses *these statistics* and *prove* in the same sentence. The word *prove* is a definitive, end-all-be-all, case-closed, lead-pipe-lock sort of concept, and statistics by nature isn't definitive. Instead, statistics gives you evidence for or against someone's theory, model, or claim based on the data you collected; then it lets you come to your own conclusions. Because the evidence is based on data that changes from sample to sample, the results can change as well — that's the challenge, the beauty, and sometimes the frustration of statistics. The best you can say is that your statistics suggest, lead you to believe, or give you sufficient evidence to conclude — but never go as far as to say that your statistics prove anything.

It's Not Technically Statistically Significant, But . . .

REMEMBER

After you set up your model and test it with your data, you have to stand by the conclusions no matter how much you believe they're wrong. Statistics must lend objectivity to every process.

Suppose Barb, a researcher, has just collected and analyzed the heck out of her data, and she still can't find anything. However, she knows in her heart that her theory holds true, even if her data can't confirm it. Barb's theory is that dogs have ESP — in other words, a "sixth sense." She bases this theory on the fact that her dog seems to know when she's leaving the house, when he's going to the vet, and when a bath is imminent because he gets sad and finds a corner to hide in.

Barb tests her ESP theory by studying ten dogs, placing a piece of dog food under one of two bowls and asking each dog to find the food by pushing on a bowl. (Assume the bowl is thick enough that the dogs can't cheat by smelling the food.) She repeats this process ten times with each dog and records the number of correct responses. If the dogs don't have ESP, you would expect that they would be right 50 percent of the time because each dog has two bowls to choose from and each bowl has an equal chance of being selected.

As it turns out, the dogs in Barb's study are right 55 percent of the time. Now, this percentage is technically higher than the long-term expected value of 50 percent, but it's not enough (especially with so few dogs and so few trials) to warrant statistical significance. In other words, Barb doesn't have enough evidence for the ESP theory. But when Barb presents her results at the next conference she attends, she puts a spin on her results by saying, "The dogs were correct 55 percent of the time, which is more than 50 percent. These results are *technically* not enough to be statistically significant, but I believe they do show some evidence that dogs have ESP" (causing every statistician in the room to scream "NOT!").

Some researchers use this kind of conclusion all the time — skating around the statistics when they don't go their way. This game is very dangerous because the next time someone tries to replicate Barb's results (and believe me, someone always does), they find out what you knew from the beginning (through ESP?): When Barb starts packing to leave the house, her dog senses trouble coming and hides. That's all.

Concluding That x Causes y

Do you see the word that makes statisticians nervous? The first two seem pretty tame, and *x* and *y* are just letters of the alphabet, it's got to be that word *cause.* Of all the words used too loosely in statistics, *cause* tops the list.

Here's an example of what I mean. For your final report in stats class, you study which factors are related to a student's final exam score. You collect data on 500 statistics students, asking each one a variety of questions, such as "What was your grade on the midterm?", "How much sleep did you get the night before the final?", and "What's your GPA?" You conduct a multiple linear regression analysis (using techniques from Chapter 5) and conclude that study time and the amount of sleep the night before the test are the most important factors in determining exam scores. You write up all your analyses in a paper, and at the very end, you say, "These results demonstrate that more study time and a good sleep the night before cause a student's exam grade to increase."

I was with you until you said the word *cause.* You can't say that more sleep or more study time causes an increase in exam score. The data you collected shows that people who get a lot of sleep and study a lot do get good grades, and those who don't do those things don't get the good grades. But that result doesn't mean a flunky can just sleep and study more and all will be okay. This theory is like saying that because an increase in height is related to an increase in weight, you can get taller by gaining weight.

The problem is that you didn't take the same person, change their sleep time and study habits, and see what happened in terms of their exam performance (using two different exams of the same difficulty). That study requires a *designed experiment.* When you conduct a *survey,* you have no way of controlling other related factors going on, which can muddy the waters, like quality of studying, class attendance, grades on homework, and so on.

REMEMBER

The only way to control for other factors is to do a randomized experiment (complete with a treatment group, a control group, and controls for other factors that may ordinarily affect the outcome). Claiming causation without conducting a randomized experiment is a very common error some researchers make when they draw conclusions. Drawing causal conclusions from observational studies takes expertise beyond the scope of this book (for example, showing smoking causes lung cancer) and much more evidence to be mounted.

Assuming the Data Was Normal

The operative word here is *assuming.* To break it down simply, an assumption is something you believe without checking. Assumptions can lead to wrong analyses and incorrect results — all without the person doing the assuming even knowing it.

Many analyses have certain requirements. For example, data should come from a normal distribution (the classic distribution that has a bell shape to it). If someone says, “I assumed the data was normal,” they just assumed that the data came from a normal distribution. But is having a normal distribution an assumption you just make and then move on, or is more work involved? You guessed it — more work.

For example, in order to conduct a one-sample t-test (see Chapter 4), your data must come from a normal distribution unless your sample size is large, in which case you get an approximate normal distribution anyway by the Central Limit Theorem (remember those three words from Stats I?). Here, you aren’t making an assumption, but examining a *condition* (something you check before proceeding). You plot the data in a histogram (see Chapter 3), see if the data meets the condition, and if it does, you proceed. If not, you can use nonparametric methods instead (discussed in Chapters 17 and 18).

REMEMBER

Nearly every statistical technique for analyzing data has at least some conditions on the data in order for you to use it. Always find out what those conditions are, and check to see whether your data meets them (and if not, consider using nonparametric statistics; see Chapters 17 and 18). Be aware that many statistics textbooks wrongly use the word *assumption* when they actually mean *condition.* It’s a subtle but very important difference.

Only Reporting “Important” Results

WARNING

As a data analyst, you must not only avoid the pitfall of reporting only the significant, exciting, and meaningful results, but also be able to detect when someone else is doing so. Some number crunchers examine every possible option and look at their data in every possible way before settling on the analysis that gets them the desired result.

You can probably see the problem with that approach. Every technique carries the chance for error. If you’re doing a t-test, for example, and the α level is 0.05, over the long term 5 out every 100 t-tests you conduct will result in a false alarm (meaning you declare a statistically significant result when it wasn’t really there) just by chance. So, if an eager researcher conducts 20 hypothesis tests on the same data set, odds are that at least one of those tests could result in a false alarm just by chance, on average. As this researcher conducts more and more tests, they're unfairly increasing their odds of finding something that occurred just by chance and running the risk of a wrong conclusion in the process.

It's not all the eager researcher's fault, though. They are pressured by a results-driven system. It's a sad state of affairs when the only results that get broadcast on the news and appear in journal articles are the ones that show a statistically significant result (when H_0 is rejected). Perhaps it was a bad move when statisticians came up with the term *significance* to denote rejecting H_0— as if to say that rejecting H_0 is the only important conclusion you can come to. What about all the times when H_0 couldn't be rejected? For example, when doctors failed to conclude that drinking diet cola causes weight gain, or when pollsters didn't find that people were unhappy with the president? The public would be better served if researchers and the media were encouraged to report the statistically insignificant but still important results along with the statistically significant ones.

TIP

The bottom line is this: In order to find out whether a statistical conclusion is correct, you can't just look at the analysis the researcher is showing you. You also have to find out about the analyses and results they're *not* showing you and ask questions. Avoid the urge to rush to reject H_0.

Assuming a Bigger Sample Is Always Better

Bigger is better in some things, but not always with sample sizes. On the one hand, the bigger your sample is, the more precise the results are (if no bias is present). A bigger sample also increases the ability of your data analysis to detect smaller differences from a model or to deny some claim about a population (in other words, to reject H_0 when you're supposed to). This ability to detect a certain differences from H_0 and reject H_0 is called the *power* of a test (see Chapter 4). However, some researchers can (and often do) take the idea of power too far. They increase the sample size to the point that even the tiniest difference from H_0 sends them screaming to press that all-important "reject H_0" button.

WARNING

Sample sizes should be large enough to provide precision and repeatability of your results, but there's such a thing as being too large, believe it or not. You can always take sample sizes big enough to reject any null hypothesis, even when the actual deviation from it is embarrassingly small. What can you do about this? When you read or hear that a result was deemed statistically significant, ask what the sample mean actually was (before it was put into the *t*-formula) and judge its significance to you from a practical standpoint. Beware of someone who says, "These results are statistically significant, and the large sample size of 100,000 gives even stronger evidence for that."

Suppose research claims that the typical in-house dog watches an average of ten hours of TV per week. Bob thinks the true average is more, based on the fact that his dog Fido watches at least ten hours of cooking shows alone each week. Bob sets up the following hypothesis test: $H_o : \mu = 10$ versus $H_a : \mu > 10$. He takes a random sample of 100 dogs and has their owners record how much TV their dogs watch per week. The result turns out a sample mean of 10.1 hours, and the sample standard deviation is 0.8 hour. This result isn't what Bob hoped for because 10.1 is so close to 10. He calculates the test statistic for this test using the formula $t = \frac{(\bar{x} - \mu)}{\frac{s}{\sqrt{n}}}$ and comes up with a value of $t = \frac{(10.1 - 10.0)}{\frac{0.8}{\sqrt{100}}} = \frac{0.1}{0.08}$, which equals 1.25 for t. Because the test is a right-tailed test (> in H_a), Bob can reject H_o at level α if t is beyond 1.645, and his t-value of 1.25 is far short of that value. Note that because $n = 100$ here, you find the value of 1.645 by looking at the very last row of the t-distribution table (visit Table A-1 in the Appendix). The row is marked with the infinity sign to indicate a large sample. So Bob can't reject H_o. To add insult to injury, Bob's friend Joe conducts the same study and gets the same sample mean and standard deviation as Bob did, but Joe uses a random sample of 500 dogs rather than 100. Consequently, Joe's t-value is $t = \frac{(10.1 - 10.0)}{\frac{0.8}{\sqrt{500}}} = \frac{0.1}{0.036}$, which equals 2.78. Because 2.78 is greater than 1.645, Joe gets to reject H_o (to Bob's dismay).

REMEMBER

Why did Joe's test find a result that Bob's didn't? The only difference was the sample size. Joe's sample was bigger, and a bigger sample size always makes the standard error smaller (see Chapter 4). The standard error sits in the denominator of the t-formula, so as it gets smaller, the t-value gets larger. A larger t-value makes it easier to reject H_o. (See Chapter 4 for more on precisions and margin of error.)

Now, Joe could technically give a big press conference or write an article on his results (his mom would be so proud), but you know better. You know that Joe's results are technically *statistically* significant, but not *practically* significant — they don't mean squat to any person or dog. After all, who cares that he was able to show evidence that dogs watch just a tiny bit more than ten hours of TV per week versus exactly ten hours per week? This news isn't exactly earth-shattering.

It's Not Technically Random, But . . .

When you take a sample on which to build statistical results, the operative word is *random.* You want the sample to be randomly selected from the population. A problem is that people often collect a sample that they think is mostly random, sort of random, or random enough — and that doesn't cut it. A plan for taking a sample is either random or it isn't.

One day I gave each of the 50 students in my class a number from 1 to 50, and I drew two numbers randomly from a hat. The two students I picked were sitting in the first row, and not only that, they were right next to each other. My students immediately cried foul!

After this seemingly odd result, I took the opportunity to talk to my class about truly random samples. A *random sample* is chosen in such a way that every member of the original population has an equal chance of being selected. Sometimes people who sit next to each other are chosen. In fact, if these seemingly strange results never happen, you may worry about the process; in a truly random process, you're going to get results that may seem odd, weird, or even fixed. That's part of the game.

In my consulting experiences, I always ask how my clients chose or plan to choose their samples. They always say they'll make sure it's random. But when I ask them how they'll do this, I sometimes get less-than-stellar answers. For example, someone needed to get a random sample from a population of 500 free-range chickens in a farmyard. They needed five chickens and said that they'd select them randomly by choosing the five that came up to them first. The problem is, animals that come up to you may be friendlier, more docile, older, or perhaps more tame. These characteristics aren't present in every chicken in the yard, so choosing a sample this way isn't random. The results are likely biased in this case.

REMEMBER

Always ask the researcher how they selected a sample, and when you select your own samples, stay true to the definition of random. Don't use your own judgment to choose a random sample; use a computer to do it for you!

Assuming That 1,000 Responses Is 1,000 Responses

A newspaper article on the latest survey says that 50 percent of the respondents said blah blah blah. The fine print says the results are based on a survey of 1,000 adults in the United States. But wait — is 1,000 the actual number of people selected for the sample, or is it the final number of respondents? You may need to take a second look; those two numbers hardly ever match.

For example, Jenny wants to know the percentage of Americans who have ever knowingly cheated on their taxes. In her statistics class, she found out that if she gets a sample of 1,000 people, the margin of error for her survey is only ±3 percent, which she thinks is groovy. So she sets out to achieve the goal of 1,000 responses to her survey. She knows that these days, it's hard to get people to respond to a survey, and she's worried that she may lose a great deal of her sample that way, so she has an idea. Why not send out more surveys than she needs, so that she gets 1,000 surveys back?

Jenny looks at several survey results in the newspapers, magazines, and on the Internet, and she finds that the response rate (the percentage of people who actually responded to the surveys) is typically around 25 percent. (In terms of the real world, I'm being generous with this number, believe it or not. But think about it: How many surveys have you thrown away lately? Don't worry, I'm guilty of it too.) So, Jenny does the math and figures that if she sends out 4,000 surveys and gets 25 percent of them back, she has the 1,000 surveys she needs to do her analysis, answer her question, and have that small margin of error of ±3 percent.

Jenny conducts her survey, and just like clockwork, out of the 4,000 surveys she sends out, 1,000 come back. She goes ahead with her analysis and finds that 400 of those people reported cheating on their taxes (40 percent). She adds her margin of error, and reports, "Based on my survey data, 40 percent of Americans cheat on their taxes, ±3 percentage points."

Now hold the phone, Jenny. She only knows what those 1,000 people who returned the survey said. She has no idea what the other 3,000 people said. And here's the kicker: Whether or not someone responds to a survey is often related to the reason the survey is being done. It's not a random thing. Those nonrespondents (people who don't respond to a survey) carry a lot of weight in terms of what they're not taking time to tell you.

For the sake of argument, suppose that 2,000 of the people who originally received the survey were uncomfortable with the question because they *do* cheat on their taxes; they just didn't want anyone to know about it, so they threw the survey in the trash. Suppose that the other 1,000 people don't cheat on their taxes, so they didn't think it was an issue and didn't return the survey. If these two scenarios were true, the results would look like this:

$$\text{Cheaters} = 400(\text{respondents}) + 2,000(\text{nonrespondents}) = 2,400$$

These results raise the total percentage of cheaters to 2,400 divided by $4,000 = 60$ percent. That's a huge difference!

You could go completely the other way with the 3,000 nonrespondents. You could suppose that none of them cheat, but they just didn't take the time to say so. If you knew this information, you would get $600(\text{respondents}) + 3{,}000\,(\text{nonrespondents}) = 3{,}600$ noncheaters. Out of 4,000 surveyed, this would mean 90 percent didn't cheat, and only 10 percent did. The truth is likely to be somewhere between the two examples I just gave you, but nonrespondents make it too hard to tell.

And the worst part is that the formulas Jenny uses for margin of error don't take into account that the information she put into them is based on biased data, so her reported 3 percent margin of error is wrong. The formulas happily crank out results no matter what. It's up to you to make sure that you put good, clean information into the formulas.

TIP

Getting 1,000 results when you send out 4,000 surveys is nowhere near as good as getting 1,000 results when sending out 1,000 surveys (or even 100 results from 100 surveys). Plan your survey based on how much follow-up you can do with people to get the job done, and if it takes a smaller sample size, so be it. At least the results have a better chance of being statistically on target.

Of Course the Results Apply to the General Population

Making conclusions about a much broader population than your sample actually represents is one of the biggest no-no's in statistics. This kind of problem is called *generalization*, and it occurs more often than you may think. People want their results instantly; they don't want to wait for them, so well-planned surveys and experiments take a back seat to instant web surveys and convenient samples.

For example, a researcher wants to know how cable news channels have influenced the way Americans get their news. The researcher also happens to be a statistics professor at a large research institution and has 1,000 students in their class. They decide that instead of taking a random sample of Americans, which would be difficult, time-consuming, and expensive, they'll just put a question on their final exam to get their students' answers. Their data analysis shows that only 5 percent of their students read the newspaper and/or watch network news programs; the rest watch cable news. For their class, the ratio of students who exclusively watch cable news compared to those students who don't is 20 to 1. The professor reports this and sends out a press release about it. The cable news channels pick up on it and the next day are reporting, "Americans choose cable news channels over newspapers and network news by a 20-to-1 margin!"

Do you see what's wrong with this picture? The professor's conclusions go way beyond their study, which is wrong. They used the students in their statistics class to obtain the data that serves as the basis for their entire report and the resulting headline. Yet the professor reports that these results are true for all Americans. I think it's safe to say that a sample of 1,000 college students taking a statistics class at the same time at the same college doesn't represent a cross section of America.

If the professor wants to make conclusions about America, they have to select a random sample of Americans to take their survey. If the professor uses 1,000 students from their class, their conclusions can only be made about that class and no one else.

TIP

To avoid or detect generalization, identify the population that you're intending to make conclusions about and make sure the sample you selected represents that population. If the sample represents a smaller group within that population, you also have to downsize the scope of your conclusions.

Deciding Just to Leave It Out

It seems easier sometimes to just leave out information. I see this all too often when I read articles and reports based on statistics. But, this error isn't the fault of only one person or group. The guilty parties can include

- **The producers:** Some researchers may leave statistical details out of their reports for a variety of reasons, including time and space constraints. After all, you can't write about every element of the experiment from beginning to end. However, other items they leave out may be indicative of a bigger problem. For example, reports often say very little about how they collected the data or chose the sample. Or they may discuss the results of a survey but not show the actual questions they asked. Ten out of 100 people may have dropped out of an experiment, and the researchers don't tell you why. All these items are important to know before making a decision about the credibility of someone's results.

  Another way in which some data analysts leave information out is by removing data that doesn't fit the intended model (in other words, "fudging" the data). Suppose a researcher records the amount of time spent surfing the Internet and relates it to age. They fit a nice line to their data indicating that younger people surf the Internet much more than older people and that surf time decreases as age increases. All is good except for Claude the outlier, who's 80 years old and surfs the Internet day and night, leading his own bingo chat rooms and everything. What to do with Claude? If not for him, the relationship looks beautiful on the graph; what harm would it do to remove him? After all, he's only one person, right?

No way. Everything is wrong with this idea. Removing undesired data points from a data set is not only very wrong but also very risky. The only time it's okay to remove an observation from a data set is if you're certain beyond doubt that the observation is just plain wrong. For example, someone writes on a survey that they spend 30 hours a day surfing the Internet or that their IQ is 2,200.

- **The communicators:** When reporting statistical results, the media leave out important information all the time, which is often due to space limitations and tight deadlines. However, part of it is a result of the current, fast-paced society that feeds itself on sound bites. The best example is survey results in which the margin of error isn't communicated. You can't judge the precision of the results without it.
- **The consumers:** The general public also plays a role in the leave-things-out mindset. People hear a news story and instantly believe it to be true, ignoring any chance for error or bias in the results. For example, you need to make a decision about what car to buy, and you ask your neighbors and friends rather than examine the research and the resulting meticulous, comprehensive ratings. At one time or another, everyone neglects to ask questions as much as they should, which indirectly feeds the entire problem.

In the chain of statistical information, the producers (researchers) need to be comprehensive and forthcoming about the process they conducted and the results they got. The communicators of that information (the media) need to critically evaluate the accuracy of the information they're getting and report it fairly. The consumers of statistical information (the rest of us) need to stop taking results for granted and to rely on credible sources of statistical studies and analyses to help make important life decisions.

WARNING

In the end, if a data set looks too good, it probably is. If the model fits too perfectly, be suspicious. If it fits exactly right, run and don't look back! Sometimes what's left out speaks much louder than what's put in.

IN THIS CHAPTER

- Knowing what information to look for
- Being skeptical and confident
- Piecing together the statistics puzzle and checking your answers
- Knowing how best to present your findings

Chapter 23

Ten Ways to Get Ahead by Knowing Statistics

One of my personal goals of teaching statistics is to help people get very good at being able to say, "Wait a minute!" and stop a wrong analysis or a misleading graph in its tracks. I also want to help them become the stats gurus in their workplaces — those people who aren't afraid to work with statistics and do so correctly and confidently (and to also know when to consult a professional statistician). This chapter arms you with ten ways of trusting your statistics instincts and increasing your professional value through your understanding of the critical world of stats.

Asking the Right Questions

Every study, every experiment, and every survey is done because someone had a question they wanted answered. For example, "How long should this warranty last?"; "What's the chance of me developing complications during surgery?"; "What does the American public think about banning public smoking?" Only after a clear question has been defined can proper data collection begin.

Suppose a restaurant owner tells me that they want to conduct a survey to learn more about the clientele at their restaurant. I talk about various variables to look at, including the number of people in the party, how often they've been there before, the type of food ordered, the amount they pay, how long they stay, and so on. After I collect some data and go over the results, the owner suddenly has a major realization: What they really want to do is compare the clientele of their lunch crowd to their dinner crowd. Does the dinner crowd spend more money? Are they older? Do they stay longer? But sadly, they can't answer any of those questions because they didn't mention collecting data on whether the customers were there for lunch or dinner.

What happened here is a common mistake. The restaurant owner said they "just wanted to study" their clientele; they never mentioned comparisons because they hadn't thought that far ahead. If they had thought about it, they would have realized the real question was, "How does my lunch clientele compare with my dinner clientele?" Then, including a question on whether diners were there for lunch or dinner would have been a no-brainer. Always ask the right questions to get the answers you need.

REMEMBER

Testing the waters a bit before plunging into a full-blown study can be very helpful. One way to do this is to conduct what researchers call a pilot study. A *pilot study* is a small exploratory study that you use as a testing ground for the real thing. For example, you design a survey and try it out on a small group to see if they find any confusing questions, redundancies, spelling errors, and so on. Pilot studies are a quick and inexpensive way to help ensure that all goes well when the actual study takes place.

Being Skeptical

Being statistically skeptical is a good thing (within reason). Some folks have given up on statistics, thinking that people can say anything they want if they manipulate the data enough. So those who have a healthy degree of skepticism can get ahead of the game.

Colorful charts and graphs can catch your eye, especially if they have neat little captions, and long and detailed professional reports may show you more information than you want to know, all laid out in neat tables, page after page. What's most important, however, is not how nice-looking the information is, or how professionally sound or scientific it looks. What's most important is what's happened behind the scenes, statistically speaking, in order to produce results that are correct, fair, and clear.

REMEMBER

Many folks know only enough statistics to be dangerous. And many reported results are incorrect, either by mistake or by design (unfortunately). It's better to be skeptical than sorry!

Here's how to put your skepticism to good use:

- **Get a copy of survey questions asked.** If the questions are misleading, the survey results aren't credible.
- **Find out about the data-collection process.** When was the survey conducted? Who was selected to participate? How was the information collected? Surveys conducted on the Internet and those based on call-in polls are almost always biased, and their results should be thrown out the window.
- **Find out about the response rate of the survey.** How many people were initially contacted? How many responded? If many were contacted and few responded, the results are almost certainly biased because survey respondents typically have stronger feelings than those who don't respond.

Collecting and Analyzing Data Correctly

On the one hand, it's very important to think very critically and even be skeptical at times about statistical results that you come across in everyday life and in the workplace. You should always ask questions before you deem the results to be credible.

On the other hand, it's very important to remember that others are also thinking critically about your results, and you need to avoid the skepticism that you see others receiving. To avoid potential potshots that may be taken at your results, you need to make sure you've done everything right.

Because you're reading this book, by now you should have many tools to help you do data collection and analysis correctly. In each chapter you hear the same theme song: Using the wrong analysis or too many analyses isn't good. For each type of analysis I present, you see how to check to make sure that particular analysis is okay to use with the data you have. Chapters 1 and 2 serve as a reference to which techniques are needed, and where to find them in the book.

REMEMBER

Ninety percent of the work involved in a statistical analysis happens before the data even goes into the computer. Here's a basic to-do list of what to check for:

- Design your survey, your experiment, or your study to avoid bias and ensure precision.

- Make sure you conduct the study at the right time and select a truly random sample of individuals to participate.
- Follow through with those participants to make sure your final results have a high response rate.

This to-do list can be challenging, but in the end, you'll be safe in knowing that your results will stand up to criticism because you did everything right.

Calling for Help

One of the toughest things for nonstatisticians to get is that they don't have to do all the statistics themselves. In fact, it's not a good way to go in many instances. The six most important words for any nonstatistician are "Know when to consult a statistician." Know when to ask for help. And the best time to ask for help is *before* you collect any data.

So how can you tell when you're in a bit over your head and you need someone to throw you a statistical lifeline? Here are some examples to help give you an idea of when to call:

- If your boss wants no less than a 100-page marketing results report on their desk by Monday and you haven't collected data point #1, CALL.
- If you're reading *Cosmopolitan* on your lunch break and you want to analyze how you and your friends came out on the "Who's the Gossip Queen in Your Workplace?" quiz, DON'T CALL.
- If the list of questions on your survey becomes longer than you are tall, CALL.
- If you want to make a bar graph of how many of your Facebook friends are fans of the 1970s, 80s, or 90s, DON'T CALL.
- If a scatterplot of your data looks like it should be in a Rorschach inkblot test, CALL (and fast!).
- If you want to know the odds that someone you haven't seen since high school is on the same plane to Africa as you are, DON'T CALL.
- If you have an important job to do that involves statistics and you are unsure of how to begin or how you'll analyze your data once you get it, CALL. The sooner you call, the more the professionals can do to help you look good!

Retracing Someone Else's Steps

At some point in your work life, you'll take out a report, read it, and you'll have a question about it. You'll go to find the data, and after much searching, you'll bring up a spreadsheet with row upon row and column upon column of numbers and characters. Your eyes will glaze over; you'll have no idea what you're looking at. You'll tell yourself not to panic and just to find the person who entered all the data and find out what's going on.

But then comes the bad news. Someone named Bob collected the data and entered it a couple of years ago, and Bob doesn't work for the company anymore. Now what do you do? More than likely, you'll have to ditch the data and the report, start all over again from scratch, and lose valuable time and money in the process.

How could this disaster have been prevented? All the following issues should have been addressed before Bob passed on his report:

- The report should include a couple of paragraphs telling how and when the data were collected, the names of the variables in the data set, where they're located in the spreadsheet, and what their labels are.
- The report should include a note about missing data. Missing data are sometimes left blank, but they also can be written as a negative sign (–) or a decimal point. (Using zeroes for missing data is a special no-no because they will be confused with actual data values that equal zero.)
- The rows of the data set should be defined. For example, does each row represent one person? Do they have ID numbers?

REMEMBER

Unfortunately, many people create statistical reports and then disappear without a trace, leaving behind a data mess that often can't be fixed. It's common courtesy to take steps to avoid leaving other people in the lurch, the way Bob did. Always leave a trail for the next person to pick up where you left off. And on the flipside, always ask for the explanation and background of a data set before using it.

Putting the Pieces Together

You should never jump right into an analysis expecting to get a one-number answer and then walk away. Statistics requires much more work! You should view every statistical problem as a puzzle whose pieces need to be put together before you can see the big picture of what's really going on.

For example, suppose a coffee vendor wants to predict how much coffee they should have ready for an upcoming football game in Buffalo, New York. Their first step is to think about what variables may be related to coffee sales. Variables may be cost of the coffee, ease of carrying it, seat location (who wants to walk a mile for a cup of coffee?), and age of fans. The vendor also suspects that temperature at the game may affect coffee sales, with low temps translating into higher sales.

The vendor collects data on all these variables and explores the relationships. They find that coffee sales and temperature are somewhat related. But is there more to this story than temperature?

To find out, the vendor compares coffee sales for two games with the same temperature and notices a big difference. Looking deeper, they notice one game was on a Sunday and one was on a Monday. Attendance was higher on Monday, and that game had more adults in attendance. By analyzing the data, the vendor found that temperature is related to coffee sales, but so is attendance, day of the week that the game was played, and age of the fans. Knowing this information, the vendor was able to predict coffee sales more accurately with a lower chance of running out of coffee or wasting it. This example illustrates that putting the pieces together to keep an eye on the big picture can really pay off.

Checking Your Answers

After your data have been analyzed and you get your results, you need to take one more step before running giddily to your boss, saying, "Look at this!" You have to be sure that you have the right answers.

REMEMBER

By right answers, I don't mean that you need to have the results that your boss wants to hear (although that would be great, of course). Rather, you need to make sure your data analysis and calculations are correct and don't leave you high and dry when the questions start to come. Follow these basic steps:

1. **Double-check that you entered the data correctly, and weed out numbers that obviously make no sense (such as someone saying that they are 200 years old, or that they sold 500 billion light bulbs at their store last year).**

   Mistakes influence the data and the results, so catch them before it's too late.

2. **Make sure that your numbers add up when they're supposed to.**

   For example, if you collected data on number of employees for 100 companies and you don't list enough number groups to cover them all, you're in trouble! Also be on the lookout for data on an individual that have been entered twice. This error shows up if you sort the data by rows.

3. **If you intend to make conclusions, make sure you're using the right numbers to do so.**

   If you want to talk about how crime has increased in your area over the last five years, showing the number of crimes on a graph is incorrect. The number of crimes can increase simply because the population size increases. For correct statistical conclusions about crime, you need to report the crime rate, which is the number of crimes per person (per capita), or the number of crimes per 100,000 people. Just take the number of crimes divided by the population size, or divided by 100,000, respectively. This approach takes population size out of it.

Explaining the Output

Computers certainly play a major role in the process of collecting and analyzing data. Many different statistical software packages exist, including Microsoft Excel, Minitab, SAS, SPSS, and a host of others. Each type has its own style of printing out results. Understanding how to read, interpret, and explain computer output is an art and a science that not everyone possesses. With your statistical knowledge, though, you can be that person!

Computer output is the raw form of the results of doing any statistical summary or analysis. It can be graphs, charts, scatterplots, tables, regression analysis results, an analysis of variance table, or a set of descriptive statistics. Often the analysis is labeled by the computer; for example, ANOVA indicates an analysis of variance has been conducted (see Chapter 10). However, graphs, charts, and tables require the user to tell the computer what labels, titles, or legends (if any) to include so that the audience can quickly understand what's what.

Interpreting computer output involves sifting through what can seem like an intimidating amount of information. The trick is to know exactly what results you want and where the computer places them on the output. For example, in the output from a regression analysis, you find the equation of the regression line by looking in the Coef column of the output (see Chapter 5).

Most of the time there's information on a computer output that you don't need; sometimes there's also information that you don't understand. Before skipping everything, you may want to consult a statistician to make sure you aren't missing an important step, such as examining the correlation coefficient before doing a regression analysis (see Chapter 5).

REMEMBER

Most importantly, make sure the analysis is correct before explaining it to anyone. Sometimes it's easy when analyzing data to click on the wrong variable or to highlight the wrong column of data, which makes the analysis totally wrong.

Making Convincing Recommendations

As one moves up the corporate ladder, they have less time to read reports and carefully examine statistics. The best data analysis in the world won't mean squat if you can't communicate your results to someone who doesn't have the time or interest to get into the nitty-gritty. In this data-driven world, statistics can play a major role in good decision-making. The ability to use statistics to make an effective argument, make a strong case, or give solid recommendations is critical.

Put yourself in the following situation. You've done the work, you've collected marketing and sales data, and you've done the analyses and processed the results. Based on your study of product placement for your Sugar Surge Pop, you determined that the best strategy for placing this product on grocery store shelves is to put it in the checkout aisle at eye level so children can see it. (You never see nail clippers or hand sanitizers on the kids' eye-level shelves in the impulse aisle, do you?) Word is that your boss favors putting this product in the candy aisle of the store. (Of course, they have no data to support this, just their own experience people-watching in the candy aisle.) How do you convince your boss to follow your recommendation?

Probably the worst thing you can do is go into their office with a 100-page report loaded with everything from soup to nuts. Loads of complex information may impress your mom, but it won't impress your boss. Save that report in case they ask for it (or in case you need a doorstop). What you need is a short, succinct, and straightforward report that makes the point. Here's how to craft it:

1. **Start out with a statement of the problem.**

   "We want to determine which location has the most sales of the Sugar Surge Pop."

2. **Briefly outline your data-collection process.**

   "We chose 50 stores at random and placed the product in the checkout aisle at 25 stores and in the candy aisle in the other 25. We controlled for other factors such as number of products placed."

3. **Describe what data you collected.**

   "We tracked the sales of the product over a six-month period, calculating weekly sales totals for each store." At this point, show your charts and graphs of the sales over time for the two groups.

4. **Tell briefly how you analyzed the data, but spend the most time on your findings.**

   Don't show the output — your boss doesn't need to see that. You know the expression, "Never let them see you sweat"? That's important here. What you do want to say is, "I did a statistical analysis comparing average sales at these locations, and I found sales in the checkout aisle to be significantly higher than sales in the candy aisle." You can quantify the difference with percentages.

Follow up with your recommendation for product placement at kids' eyelevel in the checkout aisle, being sure to answer the original question you started with in Step 1. Then the most important point is to let your boss think the optimal placement was all their idea!

Establishing Yourself as the Statistics Go-To Person

Nothing is more valuable than someone in the workplace who isn't afraid to do statistics. Every office has one person with the courage to calculate, the confidence to make confidence intervals, the willingness to wrestle with the output, and the gumption to graph. This person is eventually everyone's friend and the first person to get to know when starting a new job.

What are the perks of being the statistics go-to person? It's the glory of knowing that you're saving the day, taking one for the team, and standing tall in the face of disaster. Your colleagues will say, "I owe you one," and you can take them up on it.

But seriously, the statistics go-to person has a more secure job because their boss knows that statistics is a staple of the workplace, and having someone to jump in when needed is invaluable.

REMEMBER

Statistics and statistical analyses can be intimidating, yet they're critical for the workplace. In most any career these days, you need to know how to select samples, write surveys, set up a process for collecting the data, and analyze it.

IN THIS CHAPTER

- » Using statistics in a wide range of jobs (yours might be next!)
- » Tracking data from birds to sports to crime
- » Taking stats into the professional world of medicine, law, and the stock market

Chapter 24 Ten Cool Jobs That Use Statistics

This book is meant to be a guide for folks who need to know statistics for their everyday life (which is all of us) as well as in their workplaces (which is most of us). If I think about it long enough, I can come up with some use of statistics in almost every job out there (except maybe a psychic advisor).

This chapter features a cross section of ten careers that all involve statistics in some way, shape, or form. You may be surprised at how often statistics turn up in the workplace! So don't burn this book when your stats course is over; you may find it to be useful in your job hunt or your job. (My accountant has a copy of this book on his shelf — what does that say? As long as he doesn't have a copy of *Accounting For Dummies* alongside it, I guess we'll be okay.)

One of my personal goals as a teacher of statistics is to help my students become the go-to folks in the workplace. You know, that person with a background in statistics who knows what they're doing, and when it's crunch time, they can do the statistics needed correctly and confidently. With experience and help from this book, you too can become that person. You'll become a hero, and your job will be all the more secure for it.

Pollster

Pollsters collect information on people from populations they're interested in. Some of the big names in professional polling include Gallup Inc., the Associated Press (AP), Zogby International, Harris Interactive, and the Pew Research Center. Major news organizations such as NBC, CBS, and CNN also conduct polls, as do many other agencies and organizations.

The purposes of polls vary, from the medical field trying to determine what's causing obesity, to political pollsters who want to keep up with the daily pulse of American opinion, to surveys that provide feedback and ideas to corporations.

REMEMBER

Knowledge of statistics is considered golden in the polling industry, because jobs can include designing surveys; selecting a proper sample of participants; carrying out a survey to collect data; and then recording, analyzing, and presenting the results.

All these tasks are part of statistics — the art and science of collecting and making sense of data. But don't just take my word for it; here's a quote from a job posting for the Gallup organization for a Research Analyst. I have to say it totally screams STATISTICS!

> If you have a strong academic record in the social sciences or economics, a familiarity with quantitative and categorical research and statistical tools in Market Research/Survey Research or Consulting, enjoy pulling together research data and abstract concepts to tell a meaningful story, while continually learning — this is the place to manage processes and projects that deliver perfect completion of client engagements.

And here's something you don't see every day. I found a job posting for a polling analyst with roughly the same requirements but a very different work setting. The job was for a company that provides security and intelligence for the United States government. For this job you need federal security clearance. You never know where your statistical background is going to take you!

Other positions related to polling that I've seen listed are quantitative research specialist and public polling research analyst.

TIP

A great website for finding out more about what pollsters do and what their work looks like is, appropriately, `www.pollster.com`.

Data Scientist

Data science is currently the hottest area out there, and people who can do data science are in high demand. Everyone talks about "BIG DATA." So if you wanted to get involved, what kind of job would you be looking for? Here's a description from a university that is building a data science program:

> The ability to transform a sea of data into actionable insights can have a profound impact — from predicting the best new diabetes treatment to identifying and thwarting national security threats. That's why businesses and government agencies are rushing to hire data science professionals who can help do just that.
>
> By extrapolating and sharing these insights, data scientists help organizations to solve vexing problems. Combining computer science, modeling, statistics, analytics, and math skills — along with sound business sense — data scientists uncover the answers to major questions that help organizations make objective decisions.

Sound like fun? If you have good math skills, enjoy statistics, and like to apply it to many different complex problems in a team setting, data science might be for you!

Ornithologist (Bird Watcher)

Everyone watches birds on occasion. I gladly admit to being a semi-serious bird-watcher, always trekking to Magee Marsh on Lake Erie in May for International Migratory Bird Day. But have you ever thought about getting paid to watch birds and other wildlife? Today's ever-increasing awareness of the environment includes a great deal of focus on identifying, studying, and protecting wildlife of all kinds.

Ornithology is the science of bird study. Ornithologists are always collecting data and finding and studying statistics on birds — often on a certain type of bird and its behavior. Some examples of common bird statistics include the following:

- Bird counts (number of birds per square unit of space on a particular day)
- Nest locations and territory maps
- Number of eggs laid and hatched
- Food preferences and foraging techniques
- Behaviors caught on tape and quantified

TIP

You can tap into a website totally dedicated to jobs that need birdwatchers and wildlife watchers and the use of their statistical skills. Here's one of the job postings supplied by the Ornithological Society:

> FLAMMULATED OWL SURVEY TECHNICIANS (2) needed for Idaho Bird Observatory study of Flammulated Owls and other forest birds in Idaho (approx. 2.5 months). Duties will consist mainly of standardized surveys and data entry. Qualifications of applicants should include: 1) good eyesight and hearing, 2) proficiency with standardized survey procedures, 3) ability to identify Western birds by sight and sound, and 4) willingness to give your all. Candidates must be physically fit and undaunted by the prospects of heat, humidity, bugs, and mud. (Indeed!)

With more experience and knowledge, you can eventually become a research wildlife biologist for the U.S. Department of the Interior U.S. Geological Survey. A job description for this position just found today actually requires 15 credit hours of statistics, proving that the government is onto the whole statistics thing.

Sportscaster or Sportswriter

Every good sportscaster or sportswriter knows that you're nothing without good juicy statistics that no one else knows. You do your homework by studying training camps and poring over printouts, spreadsheets, and historical data. You read newspapers, look at record books, and watch videos. There's no shortage of data out there, and your audience can't get enough.

Sports fans are statistics addicts! (Being an Ohio State Buckeye, I'm as rabid as the rest of 'em.) Here's just a sampling of the statistics recorded and presented in my favorite sport of college football:

- Points scored
- Points against
- Rushing yards
- Receiving yards
- Passing yards
- Interceptions
- Fumbles
- Punt and kick returns

- Number and distance of field goals attempted and made
- Kicker's career longest
- Number of first downs
- Third down conversions
- Fourth down conversions
- Penalties
- All-purpose yards
- Total offense
- Total defense
- Number of sacks
- Rushing defense
- Passing defense
- Turnover margins
- Passing efficiency
- Scoring offense
- Scoring defense
- Scoring by special teams
- Coaches' Poll standings
- AP Poll standings
- CFP standings (that's another book for another day!)
- The most 12+ win seasons
- Coaching records
- Single-game high scores
- Game attendance
- Toughness of schedule
- Winning and losing streaks
- Coin toss winners

It's obvious that we need a new saying: "Those who play sports play. Those who watch sports do statistics."

Journalist

Journalists of all types at some point or another have to work with data. They have a hard row to hoe, because the data comes to them from an infinite number of possible avenues and channels on an infinite number of topics, and they need to make sense of it, pick out what they feel is most important, boil it down, present the results, and write a story around it, all under a very strict deadline (sometimes only a few hours). That's a big job!

As a consumer of the media, I see many good uses of statistics that are clear, correct, and make interesting and important points. However, I also see many incorrect and misleading statistics in the media, and I cringe every time.

Some of the most common problems include making simple math errors, reporting percentages above 100 percent, assuming cause-and-effect relationships that aren't proven, using misleading graphs, and leaving out information (such as the number of people surveyed, the rate of nonresponse, and the margin of error). But for me, the biggest problem is reading a headline that sounds catchy, eye-opening, perhaps even shocking, only to find that it's not corroborated by the statistics in the article.

Having a couple of solid statistics courses under your belt puts you way ahead in that job interview for a journalist position. The statisticians around the world are counting on you to get out there on your white horse and do things right! (Don't forget to take this book with you in your saddle bag!)

To recognize the importance and appreciation of the difficult task that journalists have in using and reporting with statistics, the Royal Statistical Society has established an Award for Statistical Excellence in Journalism. Following is the description for the award, and I couldn't agree with it more!

> The Royal Statistical Society wishes to encourage excellence in journalists' use of statistics to question, analyze, and investigate the issues that affect society at large. Journalistic excellence in statistics helps to hold decision makers in all sectors to account — through accessible communication of complex information, highlighting of success, and exposure of important missing information.

Crime Fighter

Crime statistics help the nation's crime fighters, such as police officers, determine which kinds of crimes occur where, how often, to whom, and by whom. Crime statistics for the entire nation are compiled and analyzed by the

U.S. Department of Justice. National Crime Victimization Surveys are also conducted to help understand trends in crimes of various types.

Police officers record every incident they're involved in, forming large databases that city, county, and state officials can use to determine the number of police officers needed and which areas to focus most heavily on, and also to make changes in the policies and procedures of their police departments. The FBI can also use these huge databases to track criminals, look for patterns in crime types and occurrences, and keep track of overall trends in the number of crimes as well as the types of crimes that occur over time.

People looking for a new home or a new school can consult freely available information on crime statistics, and politicians use it to show that crime is going up or down, that money should or should not be spent on more police officers, and how safe their city or state has become with them in office.

Here's an overview of how the U.S. Department of Justice website uses data and statistics to help fight crime:

> All states have established a criminal record repository which maintains criminal records and identification data and responds to law enforcement inquiries and inquiries for other purposes such as background checks and national security. Criminal records include data provided by all components of the criminal justice system: law enforcement, prosecution, courts, and corrections . . . . Records developed for statistical purposes describe and classify each criminal incident and include data on offender characteristics, relationships between the offender and the victim, and offense impact. Statistical data are extracted from operational records using uniform criteria for classification and collection. Detailed statistical data permit localities to identify problem areas and to allocate manpower and limited financial resources in an efficient and effective manner.

Medical Professional

People who work in the medical field depend on statistics to do research and find new cures, therapies, medicines, and procedures to increase the health and well-being of all people. Medical researchers conduct clinical trials to measure every conceivable side effect of every drug that goes through the approval process. Comparative studies are done all the time to determine what factors influence weight, height, intelligence level, and ability to survive a certain disease. Statistics are a lifeline to being more confident that what works for a sample of individuals will also work for the population for which it was meant.

In the medical profession, the use of statistics starts as soon as a patient's name is called in the waiting room. Suppose you're a nurse. The first thing you ask the patient to do is to "go ahead and step on the scale, please" (eight dreaded words heard in doctors' offices round the world). From there, you check their vital signs (also known as vital statistics): temperature, blood pressure, pulse rate, and sometimes, respiration (breathing) rate. You record their numbers in your computer and compare them to what has been determined to be the normal range. Setting the normal ranges involves statistics as well, through analyzing historical data and medical research.

An anesthesiologist put it this way in one of their journal articles:

> The need for statistics in medicine extends beyond research — into daily clinical practice. Every patient-physician encounter is imbued with statistics — although we fail to recognize this. For a man presenting to an emergency room with chest pain, several diagnoses are possible. The brain of a trained physician combines several elements of history and clinical findings to arrive at one or a few diseases that are highly likely in him, a few that are possible but less likely and several that are highly unlikely. His brain then matches this list of diagnoses with performance characteristics — such as sensitivity, specificity, and predictive values — of various diagnostic tests, to select a few tests that are most likely to be helpful. And then follows the choice of treatment that is most likely to succeed. Each of these steps involves the use of statistical principles — such as the probability theory, and the Bayes' theorem. All this happens imperceptibly. However, a physician who understands the principles underlying this process can be expected to do better — just like an engineer who does not merely use a machine but also understands how it works.
>
> — Aggarwal, R., *Ann Card. Anaesth.* 2018 Oct-Dec; 21(4): 349–350.

Marketing Executive

Marketing is critical to any product's success. That's why companies spend millions of dollars for 30-second commercials during the Super Bowl. Researching who will buy your product, where, when, and for how much is a job that includes lots of statistics.

TECHNICAL STUFF

Some data is what statisticians call *quantitative,* such as surveys of existing, past, and potential customers, sales information and trends, economic and demographic information, and data regarding competitors. Other data is *qualitative* or *categorical,* including in-depth one-on-one interviews and focus groups to get a general picture of what consumers think, how they feel, what ideas they have, and what additional information they need about your product. (See Chapter 2 or your Stats I textbook for a review of quantitative and categorical data.)

Consider the example of Mars, Incorporated, which makes M&M'S candy. How has the company's product become a national icon for kids of all ages? The secrets have to be the company's innate ability to change with the times and its knowledge of what its customers want.

Statistics plays a huge role in this success through collecting data on sales, but most importantly, through getting direct feedback from customers using interviews, focus groups, and surveys. (I can't imagine the strain of trying out M&M'S and talking about what I think. "Oh wait, I need another sample before I can give you a good answer.") By analyzing this data, Mars is able to determine some of the most important and intricate details that spell success and longevity for any company.

For example, in 1995 Mars conducted a nationwide survey asking customers to choose the newest M&M'S color. That's when cyan blue came on the scene. Later, the survey went global and purple became the new addition to the M&M'S palette at M&M'S World. The Mars company continually uses statistics to find ways to be innovative in making new colors, flavors, styles, and even allowing for customized M&M'S, yet it still retains the classic essence of the M&M'S that started it all.

Lawyer

You've no doubt heard the phrase "beyond a reasonable doubt." It's the code that jurors use to make a decision of guilty or not guilty. The field of statistics plays a major role in determining whether laws are being followed or broken, whether a defendant is guilty or innocent, and whether laws need to be created or changed. Statistical information is very powerful evidence.

Lawsuits are often settled on the basis of statistical evidence collected in multiple situations over the course of years. Statistics also allow lawmakers to break down information in order to propose new laws. For example, using statistics to show that the first two hours are the most critical in terms of finding a missing child led to the Amber Alerts being broadcast on TV and radio, and posted on highway billboards when children go missing.

Prosecutors and defense attorneys often use probability and statistics to help make their cases, too. They also have to make these statistics understandable to a jury. (Maybe copies of this book should be a requirement for sequestered juries!) Statisticians are often brought onto the legal team to help attorneys gather information, decipher the results, and use the data in a jury trial situation.

A law career website says the following about useful skills in being an attorney:

> Although the LSAT does not include a math section and law schools don't teach math as part of their curricula, basic mathematical competence is useful to attorneys. Many lawyers feel that training in math improved their analytic skills, and there are some branches of legal practice that require lawyers to work with statistics, personal finance concepts, and accounting principles. Being comfortable with numbers can help attorneys practicing in these areas to serve their clients more effectively.

Attorneys may use correlation to show that certain variables have a linear relationship, such as skid distance and amount of a certain type of concrete in pavement, or the strength of a bridge beam related to the weight placed upon it.

Statistics also can help test claims. For example, suppose Shipping Company A claims its packages are delivered on average two days faster than Company B. If a random sample of packages takes longer than two days to arrive and the difference is large enough to have strong evidence against Company A's claim, it could get in trouble for false advertising. Of course, any decision based on statistics can be wrong, just by chance. In this case, if the random sample of packages just happened to take longer than usual and doesn't represent the typical average delivery time, Shipping Company A can fight back, saying they were unjustly accused of false advertising. It's a tight line to walk, and statisticians try their best to set up procedures to help the real truth come to light.

Appendix A
Reference Tables

This Appendix includes commonly used tables for five important distributions for Stats II: the *t*-distribution, the binomial distribution, the Chi-square distribution, the *F*-distribution, and the *Z*-distribution.

t-Table

Table A-1 shows right-tail probabilities for the *t*-distribution (refer to Chapter 3). To use Table A-1, you need four pieces of information from the problem you're working on:

- The sample size, *n*
- The mean of *x*, denoted μ
- The standard deviation of your data, *s*
- The value of *x* for which you want the right-tail probability

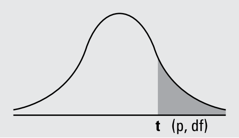
After you have this information, transform your value of *x* to a *t*-statistic (or *t*-value) by taking your value of *x*, subtracting the mean, and dividing by the standard error (see Chapter 3) by using the formula $t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$.

Then look up this value of *t* on Table A-1 by finding the row corresponding to the degrees of freedom for the *t*-statistic $(n-1)$. Go across that row until you find two values between which your *t*-statistic falls. Then go to the top of those columns and find the probabilities there. The probability that *t* is beyond your value of *x* (the right-tail probability) is somewhere between these two probabilities. Note that the last row of the *t*-table shows DF = ∞, which represents the values of the *Z*-distribution, because for large sample sizes, *t* and *Z* are close.

TABLE A-1

The *t*-Table

t-distribution showing area to the right



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
∞	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

Binomial Table

Table A-2 shows probabilities for the binomial distribution (refer to Chapter 17). To use Table A-2, you need three pieces of information from the particular problem you're working on:

- The sample size, *n*
- The probability of success, *p*
- The value of *x* for which you want the cumulative probability

TABLE A-2 **The Binomial Table**

Numbers in the table represent the probabilities for values of *x* from 0 to *n*.

Binomial probabilities: $\binom{n}{x} p^x(1-p)^{n-x}$

		p										
n	*x*	*0.1*	*0.2*	*0.25*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.75*	*0.8*	*0.9*
1	*0*	0.900	0.800	0.750	0.700	0.600	0.500	0.400	0.300	0.250	0.200	0.100
	1	0.100	0.200	0.250	0.300	0.400	0.500	0.600	0.700	0.750	0.800	0.900
2	*0*	0.810	0.640	0.563	0.490	0.360	0.250	0.160	0.090	0.063	0.040	0.010
	1	0.180	0.320	0.375	0.420	0.480	0.500	0.480	0.420	0.375	0.320	0.180
	2	0.010	0.040	0.063	0.090	0.160	0.250	0.360	0.490	0.563	0.640	0.810
3	*0*	0.729	0.512	0.422	0.343	0.216	0.125	0.064	0.027	0.016	0.008	0.001
	1	0.243	0.384	0.422	0.441	0.432	0.375	0.288	0.189	0.141	0.096	0.027
	2	0.027	0.096	0.141	0.189	0.288	0.375	0.432	0.441	0.422	0.384	0.243
	3	0.001	0.008	0.016	0.027	0.064	0.125	0.216	0.343	0.422	0.512	0.729
4	*0*	0.656	0.410	0.316	0.240	0.130	0.063	0.026	0.008	0.004	0.002	0.000
	1	0.292	0.410	0.422	0.412	0.346	0.250	0.154	0.076	0.047	0.026	0.004
	2	0.049	0.154	0.211	0.265	0.346	0.375	0.346	0.265	0.211	0.154	0.049
	3	0.004	0.026	0.047	0.076	0.154	0.250	0.346	0.412	0.422	0.410	0.292
	4	0.000	0.002	0.004	0.008	0.026	0.063	0.130	0.240	0.316	0.410	0.656
5	*0*	0.590	0.328	0.237	0.168	0.078	0.031	0.010	0.002	0.001	0.000	0.000
	1	0.328	0.410	0.396	0.360	0.259	0.156	0.077	0.028	0.015	0.006	0.000
	2	0.073	0.205	0.264	0.309	0.346	0.312	0.230	0.132	0.088	0.051	0.008
	3	0.008	0.051	0.088	0.132	0.230	0.312	0.346	0.309	0.264	0.205	0.073
	4	0.000	0.006	0.015	0.028	0.077	0.156	0.259	0.360	0.396	0.410	0.328
	5	0.000	0.000	0.001	0.002	0.010	0.031	0.078	0.168	0.237	0.328	0.590
6	*0*	0.531	0.262	0.178	0.118	0.047	0.016	0.004	0.001	0.000	0.000	0.000
	1	0.354	0.393	0.356	0.303	0.187	0.094	0.037	0.010	0.004	0.002	0.000
	2	0.098	0.246	0.297	0.324	0.311	0.234	0.138	0.060	0.033	0.015	0.001
	3	0.015	0.082	0.132	0.185	0.276	0.313	0.276	0.185	0.132	0.082	0.015
	4	0.001	0.015	0.033	0.060	0.138	0.234	0.311	0.324	0.297	0.246	0.098
	5	0.000	0.002	0.004	0.010	0.037	0.094	0.187	0.303	0.356	0.393	0.354
	6	0.000	0.000	0.000	0.001	0.004	0.016	0.047	0.118	0.178	0.262	0.531
7	*0*	0.478	0.210	0.133	0.082	0.028	0.008	0.002	0.000	0.000	0.000	0.000
	1	0.372	0.367	0.311	0.247	0.131	0.055	0.017	0.004	0.001	0.000	0.000
	2	0.124	0.275	0.311	0.318	0.261	0.164	0.077	0.025	0.012	0.004	0.000
	3	0.023	0.115	0.173	0.227	0.290	0.273	0.194	0.097	0.058	0.029	0.003
	4	0.003	0.029	0.058	0.097	0.194	0.273	0.290	0.227	0.173	0.115	0.023
	5	0.000	0.004	0.012	0.025	0.077	0.164	0.261	0.318	0.311	0.275	0.124
	6	0.000	0.000	0.001	0.004	0.017	0.055	0.131	0.247	0.311	0.367	0.372
	7	0.000	0.000	0.000	0.000	0.002	0.008	0.028	0.082	0.133	0.210	0.478

(continued)

TABLE A-2 ***(continued)***

Binomial probabilities:

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

						p						
n	*x*	*0.1*	*0.2*	*0.25*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.75*	*0.8*	*0.9*
8	*0*	0.430	0.168	0.100	0.058	0.017	0.004	0.001	0.000	0.000	0.000	0.000
	1	0.383	0.336	0.267	0.198	0.090	0.031	0.008	0.001	0.000	0.000	0.000
	2	0.149	0.294	0.311	0.296	0.209	0.109	0.041	0.010	0.004	0.001	0.000
	3	0.033	0.147	0.208	0.254	0.279	0.219	0.124	0.047	0.023	0.009	0.000
	4	0.005	0.046	0.087	0.136	0.232	0.273	0.232	0.136	0.087	0.046	0.005
	5	0.000	0.009	0.023	0.047	0.124	0.219	0.279	0.254	0.208	0.147	0.033
	6	0.000	0.001	0.004	0.010	0.041	0.109	0.209	0.296	0.311	0.294	0.149
	7	0.000	0.000	0.000	0.001	0.008	0.031	0.090	0.198	0.267	0.336	0.383
	8	0.000	0.000	0.000	0.000	0.001	0.004	0.017	0.058	0.100	0.168	0.430
9	*0*	0.387	0.134	0.075	0.040	0.010	0.002	0.000	0.000	0.000	0.000	0.000
	1	0.387	0.302	0.225	0.156	0.060	0.018	0.004	0.000	0.000	0.000	0.000
	2	0.172	0.302	0.300	0.267	0.161	0.070	0.021	0.004	0.001	0.000	0.000
	3	0.045	0.176	0.234	0.267	0.251	0.164	0.074	0.021	0.009	0.003	0.000
	4	0.007	0.066	0.117	0.172	0.251	0.246	0.167	0.074	0.039	0.017	0.001
	5	0.001	0.017	0.039	0.074	0.167	0.246	0.251	0.172	0.117	0.066	0.007
	6	0.000	0.003	0.009	0.021	0.074	0.164	0.251	0.267	0.234	0.176	0.045
	7	0.000	0.000	0.001	0.004	0.021	0.070	0.161	0.267	0.300	0.302	0.172
	8	0.000	0.000	0.000	0.000	0.004	0.018	0.060	0.156	0.225	0.302	0.387
	9	0.000	0.000	0.000	0.000	0.000	0.002	0.010	0,040	0.075	0.134	0.387
10	*0*	0.349	0.107	0.056	0.028	0.006	0.001	0.000	0.000	0.000	0.000	0.000
	1	0.387	0.268	0.188	0.121	0.040	0.010	0.002	0.000	0.000	0.000	0.000
	2	0.194	0.302	0.282	0.233	0.121	0.044	0.011	0.001	0.000	0.000	0.000
	3	0.057	0.201	0.250	0.267	0.215	0.117	0.042	0.009	0.003	0.001	0.000
	4	0.011	0.088	0.146	0.200	0.251	0.205	0.111	0.037	0.016	0.006	0.000
	5	0.001	0.026	0.058	0.103	0.201	0.246	0.201	0.103	0.058	0.026	0.001
	6	0.000	0.006	0.016	0.037	0.111	0.205	0.251	0.200	0.146	0.088	0.011
	7	0.000	0.001	0.003	0.009	0.042	0.117	0.215	0.267	0.250	0.201	0.057
	8	0.000	0.000	0.000	0.001	0.011	0.044	0.121	0.233	0.282	0.302	0.194
	9	0.000	0.000	0.000	0.000	0.002	0.010	0.040	0.121	0.188	0.268	0.387
	10	0.000	0.000	0.000	0.000	0.000	0.001	0.006	0.028	0.056	0.107	0.349
11	*0*	0.314	0.086	0.042	0.020	0.004	0.000	0.000	0.000	0.000	0.000	0.000
	1	0.384	0.236	0.155	0.093	0.027	0.005	0.001	0.000	0.000	0.000	0.000
	2	0.213	0.295	0.258	0.200	0.089	0.027	0.005	0.001	0.000	0.000	0.000
	3	0.071	0.221	0.258	0.257	0.177	0.081	0.023	0.004	0.001	0.000	0.000
	4	0.016	0.111	0.172	0.220	0.236	0.161	0.070	0.017	0.006	0.002	0.000
	5	0.002	0.039	0.080	0.132	0.221	0.226	0.147	0.057	0.027	0.010	0.000
	6	0.000	0.010	0.027	0.057	0.147	0.226	0.221	0.132	0.080	0.039	0.002
	7	0.000	0.002	0.006	0.017	0.070	0.161	0.236	0.220	0.172	0.111	0.016
	8	0.000	0.000	0.001	0.004	0.023	0.081	0.177	0.257	0.258	0.221	0.071
	9	0.000	0.000	0.000	0.001	0.005	0.027	0.089	0.200	0.258	0.295	0.213
	10	0.000	0.000	0.000	0.000	0.001	0.005	0.027	0.093	0.155	0.236	0.384
	11	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.020	0.042	0.086	0.314

(continued)

Binomial probabilities:

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

		p										
n	*x*	*0.1*	*0.2*	*0.25*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.75*	*0.8*	*0.9*
12	*0*	0.282	0.069	0.032	0.014	0.002	0.000	0.000	0.000	0.000	0.000	0.000
	1	0.377	0.206	0.127	0.071	0.017	0.003	0.000	0.000	0.000	0.000	0.000
	2	0.230	0.283	0.232	0.168	0.064	0.016	0.002	0.000	0.000	0.000	0.000
	3	0.085	0.236	0.258	0.240	0.142	0.054	0.012	0.001	0.000	0.000	0.000
	4	0.021	0.133	0.194	0.231	0.213	0.121	0.042	0.008	0.002	0.001	0.000
	5	0.004	0.053	0.103	0.158	0.227	0.193	0.101	0.029	0.011	0.003	0.000
	6	0.000	0.016	0.040	0.079	0.177	0.226	0.177	0.079	0.040	0.016	0.000
	7	0.000	0.003	0.011	0.029	0.101	0.193	0.227	0.158	0.103	0.053	0.004
	8	0.000	0.001	0.002	0.008	0.042	0.121	0.213	0.231	0.194	0.133	0.021
	9	0.000	0.000	0.000	0.001	0.012	0.054	0.142	0.240	0.258	0.236	0.085
	10	0.000	0.000	0.000	0.000	0.002	0.016	0.064	0.168	0.232	0.283	0.230
	11	0.000	0.000	0.000	0.000	0.000	0.003	0.017	0.071	0.127	0.206	0.377
	12	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.014	0.032	0.069	0.282
13	*0*	0.254	0.055	0.024	0.010	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	1	0.367	0.179	0.103	0.054	0.011	0.002	0.000	0.000	0.000	0.000	0.000
	2	0.245	0.268	0.206	0.139	0.045	0.010	0.001	0.000	0.000	0.000	0.000
	3	0.100	0.246	0.252	0.218	0.111	0.035	0.006	0.001	0.000	0.000	0.000
	4	0.028	0.154	0.210	0.234	0.184	0.087	0.024	0.003	0.001	0.000	0.000
	5	0.006	0.069	0.126	0.180	0.221	0.157	0.066	0.014	0.005	0.001	0.000
	6	0.001	0.023	0.056	0.103	0.197	0.209	0.131	0.044	0.019	0.006	0.000
	7	0.000	0.006	0.019	0.044	0.131	0.209	0.197	0.103	0.056	0.023	0.001
	8	0.000	0.001	0.005	0.014	0.066	0.157	0.221	0.180	0.126	0.069	0.006
	9	0.000	0.000	0.001	0.003	0.024	0.087	0.184	0.234	0.210	0.154	0.028
	10	0.000	0.000	0.000	0.001	0.006	0.035	0.111	0.218	0.252	0.246	0.100
	11	0.000	0.000	0.000	0.000	0.001	0.010	0.045	0.139	0.206	0.268	0.245
	12	0.000	0.000	0.000	0.000	0.000	0.002	0.011	0.054	0.103	0.179	0.367
	13	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.010	0.024	0.055	0.254
14	*0*	0.229	0.044	0.018	0.007	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	1	0.356	0.154	0.083	0.041	0.007	0.001	0.000	0.000	0.000	0.000	0.000
	2	0.257	0.250	0.180	0.113	0.032	0.006	0.001	0.000	0.000	0.000	0.000
	3	0.114	0.250	0.240	0.194	0.085	0.022	0.003	0.000	0.000	0.000	0.000
	4	0.035	0.172	0.220	0.229	0.155	0.061	0.014	0.001	0.000	0.000	0.000
	5	0.008	0.086	0.147	0.196	0.207	0.122	0.041	0.007	0.002	0.000	0.000
	6	0.001	0.032	0.073	0.126	0.207	0.183	0.092	0.023	0.008	0.002	0.000
	7	0.000	0.009	0.028	0.062	0.157	0.209	0.157	0.062	0.028	0.009	0.000
	8	0.000	0.002	0.008	0.023	0.092	0.183	0.207	0.126	0.073	0.032	0.001
	9	0.000	0.000	0.002	0.007	0.041	0.122	0.207	0.196	0.147	0.086	0.008
	10	0.000	0.000	0.000	0.001	0.014	0.061	0.155	0.229	0.220	0.172	0.035
	11	0.000	0.000	0.000	0.000	0.003	0.022	0.085	0.194	0.240	0.250	0.114
	12	0.000	0.000	0.000	0.000	0.001	0.006	0.032	0.113	0.180	0.250	0.257
	13	0.000	0.000	0.000	0.000	0.000	0.001	0.007	0.041	0.083	0.154	0.356
	14	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.007	0.018	0.044	0.229

(continued)

TABLE A-2 ***(continued)***

Binomial probabilities:

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

		p										
n	*x*	*0.1*	*0.2*	*0.25*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.75*	*0.8*	*0.9*
15	*0*	0.206	0.035	0.013	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1	0.343	0.132	0.067	0.031	0.005	0.000	0.000	0.000	0.000	0.000	0.000
	2	0.267	0.231	0.156	0.092	0.022	0.003	0.000	0.000	0.000	0.000	0.000
	3	0.129	0.250	0.225	0.170	0.063	0.014	0.002	0.000	0.000	0.000	0.000
	4	0.043	0.188	0.225	0.219	0.127	0.042	0.007	0.001	0.000	0.000	0.000
	5	0.010	0.103	0.165	0.206	0.186	0.092	0.024	0.003	0.001	0.000	0.000
	6	0.002	0.043	0.092	0.147	0.207	0.153	0.061	0.012	0.003	0.001	0.000
	7	0.000	0.014	0.039	0.081	0.177	0.196	0.118	0.035	0.013	0.003	0.000
	8	0.000	0.003	0.013	0.035	0.118	0.196	0.177	0.081	0.039	0.014	0.000
	9	0.000	0.001	0.003	0.012	0.061	0.153	0.207	0.147	0.092	0.043	0.002
	10	0.000	0.000	0.001	0.003	0.024	0.092	0.186	0.206	0.165	0.103	0.010
	11	0.000	0.000	0.000	0.001	0.007	0.042	0.127	0.219	0.225	0.188	0.043
	12	0.000	0.000	0.000	0.000	0.002	0.014	0.063	0.170	0.225	0.250	0.129
	13	0.000	0.000	0.000	0.000	0.000	0.003	0.022	0.092	0.156	0.231	0.267
	14	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.031	0.067	0.132	0.343
	15	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.013	0.035	0.206
20	*0*	0.122	0.012	0.003	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1	0.270	0.058	0.021	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	2	0.285	0.137	0.067	0.028	0.003	0.000	0.000	0.000	0.000	0.000	0.000
	3	0.190	0.205	0.134	0.072	0.012	0.001	0.000	0.000	0.000	0.000	0.000
	4	0.090	0.218	0.190	0.130	0.035	0.005	0.000	0.000	0.000	0.000	0.000
	5	0.032	0.175	0.202	0.179	0.075	0.015	0.001	0.000	0.000	0.000	0.000
	6	0.009	0.109	0.169	0.192	0.124	0.037	0.005	0.000	0.000	0.000	0.000
	7	0.002	0.055	0.112	0.164	0.166	0.074	0.015	0.001	0.000	0.000	0.000
	8	0.000	0.022	0.061	0.114	0.180	0.120	0.035	0.004	0.001	0.000	0.000
	9	0.000	0.007	0.027	0.065	0.160	0.160	0.071	0.012	0.003	0.000	0.000
	10	0.000	0.002	0.010	0.031	0.117	0.176	0.117	0.031	0.010	0.002	0.000
	11	0.000	0.000	0.003	0.012	0.071	0.160	0.160	0.065	0.027	0.007	0.007
	12	0.000	0.000	0.001	0.004	0.035	0.120	0.180	0.114	0.061	0.022	0.000
	13	0.000	0.000	0.000	0.001	0.015	0.074	0.166	0.164	0.112	0.055	0.002
	14	0.000	0.000	0.000	0.000	0.005	0.037	0.124	0.192	0.169	0.109	0.009
	15	0.000	0.000	0.000	0.000	0.001	0.015	0.075	0.179	0.202	0.175	0.032
	16	0.000	0.000	0.000	0.000	0.000	0.005	0.035	0.130	0.190	0.218	0.090
	17	0.000	0.000	0.000	0.000	0.000	0.001	0.012	0.072	0.134	0.205	0.190
	18	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.028	0.067	0.137	0.285
	19	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.007	0.021	0.058	0.270
	20	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.003	0.012	0.122

(continued)

Find the portion of Table A-2 that's devoted to your *n*, and look at the row for your *x* and the column for your *p*. Intersect that row and column, and you can see the probability for *x*. To get the probability of being strictly less than, greater than, greater than or equal to, or between two values of *x*, you sum the appropriate values of Table A-2.

Chi-Square Table

Table A-3 shows right-tail probabilities for the Chi-square distribution (you can use Chapter 15 as a reference for the Chi-square test). To use Table A-3, you need three pieces of information from the particular problem you're working on:

- The sample size, *n*.
- The value of Chi-squared for which you want the right-tail probability.
- If you're working with a two-way table, you need *r* = number of rows and *c* = number of columns. If you're working with a goodness-of-fit test, you need $k-1$, where *k* is the number of categories.

The degrees of freedom for the Chi-square test statistic is $(r-1)*(c-1)$ if you're testing for an association between two variables, where *r* and *c* are the number of rows and columns in the two-way table, respectively. Or, the degrees of freedom is $k-1$ in a goodness-of-fit test, where *k* is the number of categories; see Chapter 16.

Go across the row for your degrees of freedom until you find the value in that row closest to your Chi-square test statistic. Look up at the number at the top of that column. That value is the area to the right of (beyond) that particular Chi-square statistic.

TABLE A-3

The Chi-Square Table

Numbers in the table represent Chi-square values whose area to the right equals *p*.

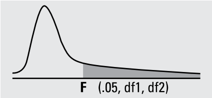
df/p	0.10	0.05	0.025	0.01	0.005
1	2.71	3.84	5.02	6.64	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.82	9.35	11.35	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.65	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.54	20.09	21.96
9	14.68	16.92	19.02	21.67	23.59
10	15.99	18.31	20.48	23.21	25.19
11	17.28	19.68	21.92	24.73	26.76
12	18.55	21.03	23.34	26.22	28.30
13	19.81	22.36	24.74	27.69	29.819
14	21.06	23.69	26.12	29.14	31.32
15	22.31	25.00	27.49	30.58	32.80
16	23.54	26.30	28.85	32.00	34.27
17	24.77	27.59	30.19	33.41	35.72
18	25.99	28.87	31.53	34.81	37.16
19	27.20	30.14	32.85	36.19	38.58
20	28.41	31.41	34.17	37.57	40.00
21	29.62	32.67	35.48	38.93	41.40
22	30.81	33.92	36.78	40.29	42.80
23	32.01	35.17	38.08	41.64	44.18
24	33.20	36.42	39.36	42.98	45.56
25	34.38	37.65	40.65	44.31	46.93
26	35.56	38.89	41.92	45.64	48.29
27	36.74	40.11	43.20	46.96	49.65
28	37.92	41.34	44.46	48.28	50.99
29	39.09	42.56	45.72	49.59	52.34
30	40.26	43.77	46.98	50.89	53.67
40	51.81	55.76	59.34	63.69	66.77
50	63.17	67.51	71.42	76.15	79.49

F-Table

Table A-4 shows the critical values on the F-distribution where ⟨ is equal to 0.05. (*Critical values* are those values that represent the boundary between rejecting H_0 and not rejecting H_0; refer to Chapter 10.) To use Table A-4, you need three pieces of information from the particular problem you're working on:

- The sample size, *n*
- The number of populations (or treatments being compared), *k*
- The value of *F* for which you want the cumulative probability

To find the critical value for your F-test statistic using Table A-5, go to the column representing the degrees of freedom you need ($k-1$ and $n-k$). Intersect the column degrees of freedom ($k-1$) with the row degrees of freedom ($n-k$), and you find the critical value on the F-distribution.



df2/df1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817	243.9060	245.9499	248.0131	249.0518	250.0951	251.1432	252.1957	253.252
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959	19.4125	19.4291	19.4458	19.4541	19.4624	19.4707	19.4791	19.487
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	8.7446	8.7029	8.6602	8.6385	8.6166	8.5944	8.5720	8.549
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117	5.8578	5.8025	5.7744	5.7459	5.7170	5.6877	5.658
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6777	4.6188	4.5581	4.5272	4.4957	4.4638	4.4314	4.398
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600	3.9999	3.9381	3.8742	3.8415	3.8082	3.7743	3.7398	3.704
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747	3.5107	3.4445	3.4105	3.3758	3.3404	3.3043	3.267
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2839	3.2184	3.1503	3.1152	3.0794	3.0428	3.0053	2.966
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729	3.0061	2.9365	2.9005	2.8637	2.8259	2.7872	2.747
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130	2.8450	2.7740	2.7372	2.6996	2.6609	2.6211	2.580
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876	2.7186	2.6464	2.6090	2.5705	2.5309	2.4901	2.448
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866	2.6169	2.5436	2.5055	2.4663	2.4259	2.3842	2.341
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.6037	2.5331	2.4589	2.4202	2.3803	2.3392	2.2966	2.252
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022	2.5342	2.4630	2.3879	2.3487	2.3082	2.2664	2.2229	2.177
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753	2.4034	2.3275	2.2878	2.2468	2.2043	2.1601	2.114
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935	2.4247	2.3522	2.2756	2.2354	2.1938	2.1507	2.1058	2.058
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499	2.3807	2.3077	2.2304	2.1898	2.1477	2.1040	2.0584	2.010
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117	2.3421	2.2686	2.1906	2.1497	2.1071	2.0629	2.0166	1.968
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779	2.3080	2.2341	2.1555	2.1141	2.0712	2.0264	1.9795	1.930
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	2.3479	2.2776	2.2033	2.1242	2.0825	2.0391	1.9938	1.9464	1.896
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660	2.3210	2.2504	2.1757	2.0960	2.0540	2.0102	1.9645	1.9165	1.865
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967	2.2258	2.1508	2.0707	2.0283	1.9842	1.9380	1.8894	1.838
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201	2.2747	2.2036	2.1282	2.0476	2.0050	1.9605	1.9139	1.8648	1.812
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547	2.1834	2.1077	2.0267	1.9838	1.9390	1.8920	1.8424	1.789
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365	2.1649	2.0889	2.0075	1.9643	1.9192	1.8718	1.8217	1.768
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197	2.1479	2.0716	1.9898	1.9464	1.9010	1.8533	1.8027	1.748
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	2.2043	2.1323	2.0558	1.9736	1.9299	1.8842	1.8361	1.7851	1.730
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	2.1900	2.1179	2.0411	1.9586	1.9147	1.8687	1.8203	1.7689	1.713
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229	2.1768	2.1045	2.0275	1.9446	1.9005	1.8543	1.8055	1.7537	1.698
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646	2.0921	2.0148	1.9317	1.8874	1.8409	1.7918	1.7396	1.683
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772	2.0035	1.9245	1.8389	1.7929	1.7444	1.6928	1.6373	1.576
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401	1.9926	1.9174	1.8364	1.7480	1.7001	1.6491	1.5943	1.5343	1.467
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588	1.9105	1.8337	1.7505	1.6587	1.6084	1.5543	1.4952	1.4290	1.351

TABLE A-4 **The *F*-Table** $(\alpha = 0.05)$

Z-Table

Table A-5 shows less-than-or-equal-to probabilities for the Z-distribution — that is, $p(Z \le z)$ for a given Z-value. To use Table A-5, do the following:

1. **Determine the *Z*-value for your particular problem.**

   The *Z*-value should have one leading digit before the decimal point (positive, negative, or zero) and two digits after the decimal point; for example, $z = 1.28, -2.69$, or 0.13.

2. **Find the row of the table corresponding to the leading digit and the first digit after the decimal point.**

   For example, if your *Z*-value is 1.28, look in the "1.2" row; if $z = -1.28$, look in the "–1.2" row.

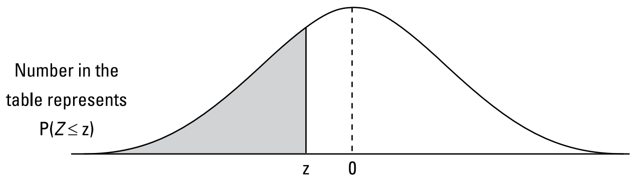
3. **Find the column corresponding to the second digit after the decimal point.**

   For example, if your *Z*-value is 1.28 or –1.28, look in the ".08" column.

4. **Intersect the row and column from Steps 2 and 3.**

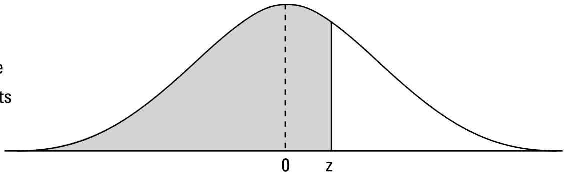
   This number is the probability that *Z* is less than or equal to your *Z*-value. In other words, you've found $p(Z \le z)$. For example, if $z = 1.28$, you see $p\left(Z \le 1.28\right) = 0.8997$. For $z = -1.28$, you see $p\left(Z \le -1.28\right) = 0.1003$.

TABLE A-5 The *Z*-Table



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
–3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
–3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0003	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
–1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
–1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
–1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
–1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
–1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
–1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
–0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
–0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
–0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
–0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
–0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
–0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
–0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
–0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
–0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
–0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Number in the
table represents
P($Z \leq$ z)



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

Index

About the Author

Deborah Rumsey has a PhD in Statistics from Ohio State University, where she's an associated professor in the Department of Statistics. Dr. Rumsey has the distinction of being named a Fellow of the American Statistical Association. She has also won the Presidential Teaching Award from Kansas State University. She's the author of *Statistics For Dummies,* 2nd Edition, *Statistics Workbook For Dummies,* 2nd Edition, and *Probability For Dummies,* and she has published numerous papers and given many professional presentations on the subject of statistics education. Her passions include being with her family, bird watching, working and playing on her goat, hay, and cattle ranch, and cheering the Ohio State Buckeyes on to another National Championship.

Dedication

To my husband, Eric: My sun rises and sets with you. To my son, Clint: I love you up to the moon and back.

Author’s Acknowledgments

Thanks to Lindsay Lefevere for giving me the opportunity to write this book; to Chrissy Guthrie for her unwavering support and perfect chiseling and molding of my words and ideas; to Sara Conroy, PhD, The Ohio State University, for a thorough technical review; and to Marylouise Wiack for great copy editing. Special thanks to Elizabeth Stasny for guidance and support from day one, and to Joan Garfield for her constant inspiration and encouragement.

Publisher's Acknowledgments

Executive Editor: Lindsay Sandman Lefevere

Acquisitions Editor: Elizabeth Stilwell

Editorial Project Manager and Development Editor: Christina N. Guthrie

Managing Editor: Michelle Hacker

Copy Editor: Marylouise Wiack

Technical Editor: Sara Conroy, PhD

Production Editor: Mohammed Zafar Ali

Cover Image: © mstanley/Shutterstock

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

20 MILLION PAGE VIEWS **EVERY SINGLE MONTH**

15 MILLION **UNIQUE** **VISITORS PER MONTH**

43% OF ALL VISITORS ACCESS THE SITE **VIA THEIR MOBILE DEVICES**

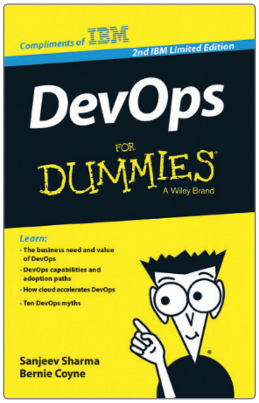
700,000 NEWSLETTER **SUBSCRIPTIONS** **TO THE INBOXES OF** *300,000* UNIQUE **INDIVIDUALS EVERY WEEK**

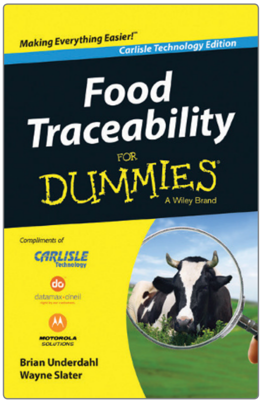
# of dummies

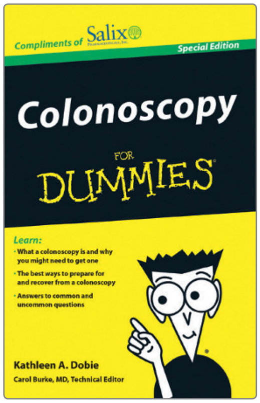
Custom Publishing

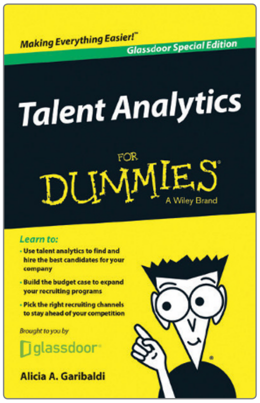
Reach a global audience in any language by creating a solution that will differentiate you from competitors, amplify your message, and encourage customers to make a buying decision.

- Apps
- Books
- eBooks
- Video
- Audio
- Webinars









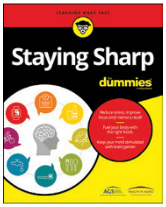
Brand Licensing & Content

Leverage the strength of the world's most popular reference brand to reach new audiences and channels of distribution.

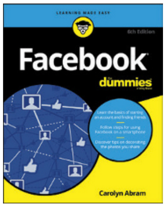
For more information, visit dummies.com/biz

dummies®
A Wiley Brand

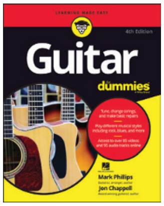
PERSONAL ENRICHMENT



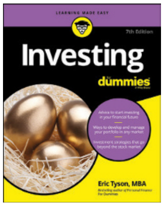
9781119187790
USA $26.00
CAN $31.99
UK £19.99



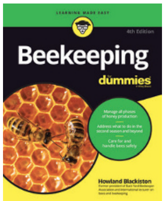
9781119179030
USA $21.99
CAN $25.99
UK £16.99



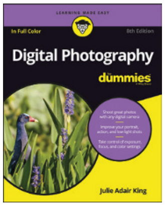
9781119293354
USA $24.99
CAN $29.99
UK £17.99



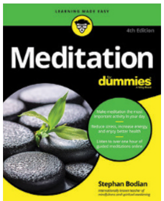
9781119293347
USA $22.99
CAN $27.99
UK £16.99



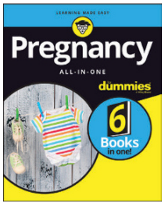
9781119310068
USA $22.99
CAN $27.99
UK £16.99



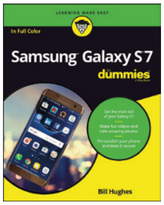
9781119235606
USA $24.99
CAN $29.99
UK £17.99



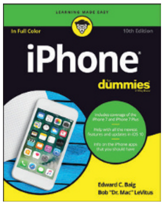
9781119251163
USA $24.99
CAN $29.99
UK £17.99



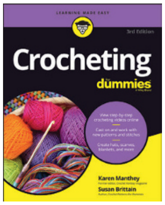
9781119235491
USA $26.99
CAN $31.99
UK £19.99



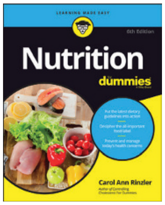
9781119279952
USA $24.99
CAN $29.99
UK £17.99



9781119283133
USA $24.99
CAN $29.99
UK £17.99

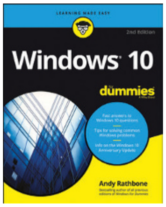


9781119287117
USA $24.99
CAN $29.99
UK £16.99

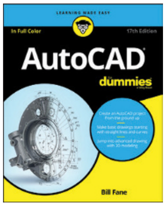


9781119130246
USA $22.99
CAN $27.99
UK £16.99

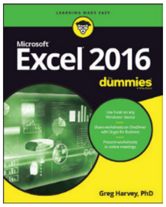
PROFESSIONAL DEVELOPMENT



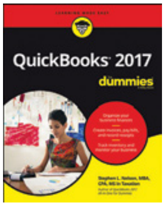
9781119311041
USA $24.99
CAN $29.99
UK £17.99



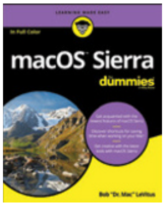
9781119255796
USA $39.99
CAN $47.99
UK £27.99



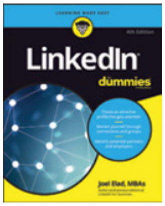
9781119293439
USA $26.99
CAN $31.99
UK £19.99



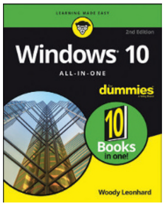
9781119281467
USA $26.99
CAN $31.99
UK £19.99



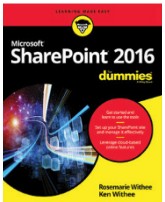
9781119280651
USA $29.99
CAN $35.99
UK £21.99



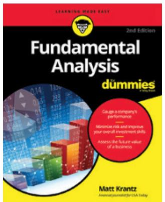
9781119251132
USA $24.99
CAN $29.99
UK £17.99



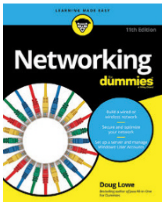
9781119310563
USA $34.00
CAN $41.99
UK £24.99



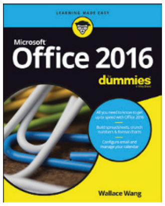
9781119181705
USA $29.99
CAN $35.99
UK £21.99



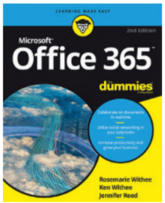
9781119263593
USA $26.99
CAN $31.99
UK £19.99



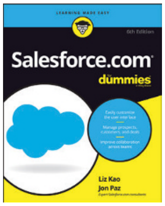
9781119257769
USA $29.99
CAN $35.99
UK £21.99



9781119293477
USA $26.99
CAN $31.99
UK £19.99



9781119265313
USA $24.99
CAN $29.99
UK £17.99



9781119239314
USA $29.99
CAN $35.99
UK £21.99

9781119293323
USA $29.99
CAN $35.99
UK £21.99

dummies.com

dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.