Supplementary Chapters for Essential Statistics, Regression, and Econometrics

Gary Smith

Pomona College

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1800, San Diego, California 92101-4495, USA
The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, UK

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

For information on all Academic Press publications visit our website: w*ww.elsevierdirect.com*

Contents

Chapter 12

The Binomial Model

Chapter Outline

You can observe a lot by just watching.

—Yogi Berra

There are two kinds of data: *quantitative* data, like heights and weights, which have natural numerical values, and *categorical* data, like whether a person is male or female, which do not have natural numerical values. This section will describe a simple, yet powerful, model for analyzing categorical data.

The outcome of a coin flip—heads or tails—is an example of categorical data. When an evenly balanced coin is flipped, there is an 0.5 probability of heads and an 0.5 probability of tails. What about ten flips? What is the probability of exactly five heads? What are the chances of six, seven, or even ten heads? And what about that hypothetical long run? Are we guaranteed exactly 50 percent heads, or is there some chance of 51 percent, or even 60 percent, heads? In this chapter we will see how to answer these questions, and many more.

Coin flips are an example of many phenomena that were analyzed by James Bernoulli (1654–

1705), one of an eminent family of Swiss mathematicians, and can be described as follows. In the *binomial model*, there are n trials and

1. Each trial has two possible categories of outcomes, not necessarily equally likely, which are labeled *success* and *failure*.
2. The probability p of success is the same for each trial.
3. The trials are independent, in that the outcome of any trial is not affected by the outcomes of other trials.

Coin flips satisfy these binomial-model conditions. So do dice rolls if we label some outcomes successes, for example, odd numbers or the number 5. So do the following situations. Each item in a group of electronic components has the same probability of failure (and failures are independent). A person is given independent medical tests, each with the same probability of a particular diagnosis. We conduct a political poll and, for each independent interview, have a constant probability of selecting someone who favors a particular position or candidate. The probability that stock prices will increase on any given day is constant, no matter what happened to stock prices on other days.

An enormous variety of situations can be described by the binomial model. By the time you finish this chapter, you may be seeing the binomial model all around you, and indeed it is.

12.1 Sampling Without Replacement

If we draw cards, one after another, from a standard deck of 52 cards, the binomial model is not exactly correct because probabilities change as cards are removed from the deck. Similarly, if we interview 1,000 randomly selected households, each selection changes the pool of remaining households and so alters the probability of choosing households with certain characteristics. When a selected item is removed and cannot be selected again, this is called *sampling without replacement*. The binomial model would be appropriate if each selection of cards (or households) were replaced and the "deck" thoroughly reshuffled before the next selection is made. If each selected item is replaced and can be selected again, this procedure is called *sampling with replacement*.

More complicated models can handle the messy details of sampling without replacement. Sometimes, however, the "deck" is so large that the binomial model is a reasonable and convenient approximation. Imagine that we have a million decks of cards thoroughly shuffled to form one enormous deck with 52 million cards. Perhaps success is drawing a heart. Because there are 13 million hearts scattered among the 52 million cards, the initial probability of a heart is exactly 0.25. If the first card drawn is a heart and it is not replaced, the probability of a heart on the second draw falls to

$$\frac{12{,}999{,}999}{51{,}999{,}999} = 0.2499999856$$

If the first card is not a heart, the probability of a heart on the second draw rises to

$$\frac{13{,}000{,}000}{51{,}999{,}999} = 0.2500000048$$

For most purposes, it would be perfectly acceptable to assume that the probability of a heart is still 0.25 and that the binomial model is applicable. There is undoubtedly more error introduced by inadequately shuffling 52 million cards than by neglecting to replace a few cards taken from

the deck. Of course, drawing 26 million cards without replacement from a 52-million card deck could change the probability of a heart substantially. A rule of thumb is that the binomial model is a good approximation when the sample is not more than ten percent of the size of the population.

In this same spirit, the binomial model is commonly applied to public opinion surveys, where a few thousand people are selected from a population of millions, even though the selection is done without replacement (we don't interview the same person twice). The slight variation in probabilities is trivial compared to the more worrisome danger that the "deck" is not adequately shuffled, giving every person an equal chance of being interviewed.

12.2 Binomial Probabilities

A random variable x is a variable whose numerical value is determined by chance, the outcome of a random phenomenon. In the binomial model, we can let x be the number of successes in n trials. The binomial distribution states that in n binomial trials with probability p of success on each trial, the probability of exactly x successes is

$$P[x \text{ successes}] = \binom{n}{x} p^x (1-p)^{n-x} \tag{12.1}$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} = \frac{n(n-1)(n-2)\ldots(n-x-1)}{x(x-1)(x-2)\ldots 1}$$

The term $p^x(1-p)^{n-x}$ is the probability of any particular sequence of x successes and $n-x$ failures, for example, x successes followed by $n-x$ failures or $n-x$ failures followed by x successes. The term $\binom{n}{x}$ (spoken "n, choose x") is the number of possible sequences of x successes and $n-x$ failures. For example, if a coin is flipped 5 times and a success is heads, there are

$$\begin{aligned}\binom{n}{x} &= \binom{5}{3} \\ &= \frac{5!}{(3!)(2!)} \\ &= \frac{(5)(4)(3)(2)(1)}{(3)(2)(1)(2)(1)} \\ &= 10\end{aligned}$$

possible sequences of 3 heads and 2 tails: HHHTT, HHTHT, HHTTH, HTHHT, HTHTH, HTTHH, THHHT, THHTH, THTHH, TTHHH.

Whenever possible, we avoid using Equation 12.1 to calculate binomial probabilities by hand. It is easier, faster, and more accurate to use statistical software.

It is important to recognize that the binomial formula in Equation 12.1 gives the probability of *exactly x* successes. For example, the probability of exactly 3 heads in 5 coin tosses is 0.3125:

$$P[3 \text{ successes}] = \binom{5}{3} 0.5^{x}(1-0.5)^{n-x}$$
$$= 0.3125$$

If we want to know the probability of 3 or more heads, we add together the probability of 3 heads, the probability of 4 heads, and the probability of 5 heads.

Table 12.1 shows the binomial probabilities for the six possible values of x (0, 1, 2, 3, 4, and 5) when five coins are tossed fairly.

Table 12.1 The Probability of x Heads in 5 Coin Flips

Number of Heads	Probability
0	0.0313
1	0.1563
2	0.3125
3	0.3125
4	0.1563
5	0.1313

Notice that because heads and tails are equally likely, this probability distribution is symmetrical: three tails are as likely as three heads, four tails are as likely as four heads, and five tails are as likely as five heads.

If success and failure are not equally likely, the binomial probabilities are not symmetrical. Suppose that we have a six-sided die, with the sides numbered 1 through 6, and consider success to be the number 1. The probability of success on each roll is $p = 1/6$ and the probability of failure is $1 - p = 5/6$. The roll of six dice (or one die six times) can be described by the binomial model, and Table 12.2 shows the binomial probabilities derived from Equation 12.1. The most likely outcome is one success. Two successes are slightly less likely than none, and there are rapidly declining probabilities of three, four, five, or six successes.

Table 12.2 The Probability of x 1s in 6 Dice Rolls

Number of 1s	Probability
0	0.33490
1	0.40188
2	0.20094
3	0.05358
4	0.00804
5	0.00064
6	0.00002

The Population Mean and Standard Deviation

The mean and standard deviation of the number of successes x in the binomial model are as follows.

$$\begin{aligned} &\text{Mean of } x: && np \\ &\text{Standard deviation of } x: && \sqrt{np(1-p)} \end{aligned} \tag{12.2}$$

We can also specify the mean and standard deviation for the success proportion x/n, the fraction of the trials that are successes:

$$\begin{aligned} &\text{Mean of } \frac{x}{n}: && p \\ &\text{Standard deviation of } \frac{x}{n}: && \sqrt{\frac{p(1-p)}{n}} \end{aligned} \tag{12.3}$$

The formulas for the standard deviation of x and x/n are difficult to explain intuitively, so we should just accept them as being correct. However, the formulas for the means of x and x/n can be explained easily. Suppose that we were to flip five coins, record the number of heads x, and then repeat this experiment over and over again until we have a billion experiments, with each experiment consisting of five coin flips. In some experiments, there will be five heads; in others zero heads; and in others one, two, three, or four heads. What do you predict will be the average number of heads in these five-flip experiments? What do you predict will be the average proportion of the five flips that are heads? If your answers are an average of 2.5 heads, which is fifty percent of five flips, then Equations 12.2 and 12.3 confirm that your intuition is correct.

12.3 The Fallacious Law of Averages

The binomial distribution was used by Bernoulli and other mathematicians to show that, in the long run, the success proportion x/n is almost certain to be close to the probability of success p. For example, in a large number of trials the fraction of coin flips that land heads will almost certainly be close to 0.5. Far too many people have misinterpreted this argument as implying that in the long run the number of heads and tails must be exactly equal and, therefore, any short-run deficit of heads or tails must soon be balanced out by an offsetting surplus. In this section we will look at the correct logic and see where intuition has so often failed us.

The Law of Large Numbers

Figures 12.1 and 12.2 show the binomial probabilities for 10 and 100 coin flips. Notice that each probability distribution is centered at $x/n = 0.5$ and that the probability distribution narrows around $x/n = 0.5$ as n increases. This is Bernoulli's law of large numbers: as n increases, the probability that x/n will be close to the success probability p approaches 1. In our coin flip example, the probability that there will be between 49 and 51 percent heads is 0.236 with one hundred flips, 0.494 with one thousand flips, and 0.956 with ten thousand flips.

A Common Fallacy

The law of large numbers is mathematically correct and intuitively obvious, but it is frequently misinterpreted as a fallacious *law of averages*, which states that, in the long run, x/n must be exactly equal to p. For example, some people think that in 1,000 coin flips there must be exactly 500 heads and 500 tails; thus if tails come up more often than heads in the first 10, 50, or 100 flips, heads must subsequently come up more often than tails in order to "average" or "balance"

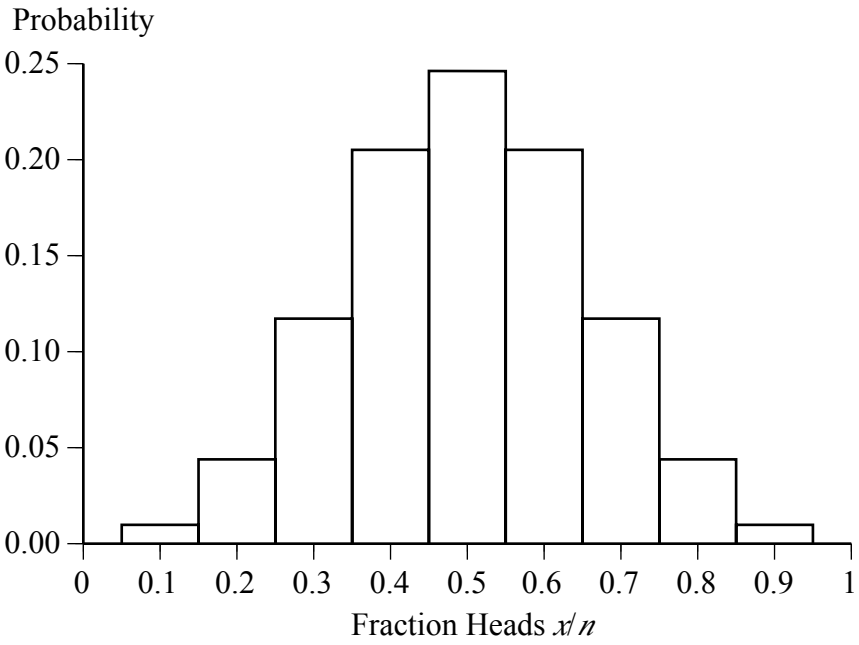


Figure 12.1 Probability distribution for 10 coin flips

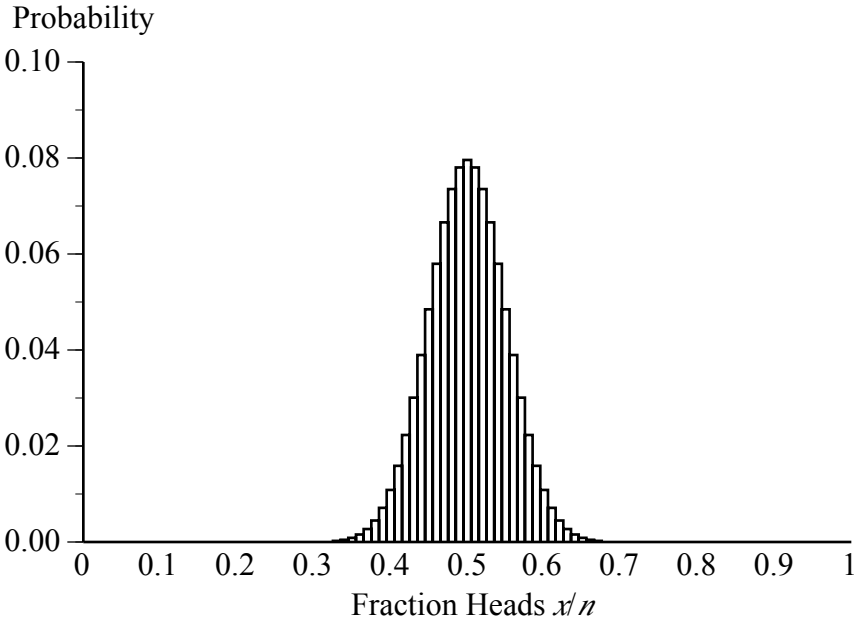


Figure 12.2 Probability distribution for 100 coin flips

things out. This erroneous belief is also known as "the gambler's fallacy." This belief is wrong, but widespread. For example, one gambler wrote:

> Regardless of mathematics and the theory of probability, if I'm in a craps game and the shooter makes ten consecutive [wins], I'm going to bet against him on his eleventh throw. And if he wins, I'll bet against him again on the twelfth. Maybe it's true that the mathematical odds on every throw of the dice remain the same, regardless of what's gone before—but how often do you see a crapshooter make 13 or 14 straight [wins]?

Thirteen consecutive wins is rare, but it is a very different question to ask the probability of 13 straight wins, *given that you have already won 12*. Dice, coins, and other inanimate objects do not remember what happened in the past and do not care what happens in the future. The outcome of a fair game of chance does not depend on what happened in the past.

The Mathematical Error

The law of large numbers does not say that the *number* of heads must equal the number of tails. Instead, it says that, if we begin making a large number of coin flips, the probability is close to one that the *fraction* of these forthcoming throws that are heads will be close to 0.5.

Notice that the law of large numbers says nothing about the last 12 throws or the last 12 million throws—which have already happened, cannot be changed, and have no effect on future throws. It says, looking ahead to predict the next 1,000 throws, the probability is high that heads will come up close to half the time.

We can use the binomial distribution to make an even stronger statement. While we are almost certain that in the long run, the fraction of the flips that are heads will be very close to 0.5, we are also almost certain that the number of heads will not equal the number of tails! To illustrate this crucial distinction, Table 12.3 shows, for different values of n, the probability that x/n will differ from 0.5 by less than 0.001 and also shows the probability that the number of heads will differ from the number of tails by more than 100. For larger values of n, we are more certain that the fraction that are heads will be very close to 0.50. At the same time, we are increasingly certain that there will be at least 100 more heads than tails or at least 100 more tails than heads. The law of large numbers does not require the numbers of heads and tails to be exactly equal!

Table 12.3 We Are Almost Certain that for Large n, x/n will be Very Close to 0.5 and That There Will Be a Large Difference Between the Numbers of Heads and Tails

Number of Tosses n	Probability that x/n will be between 0.499 and 0.501	Probability that there will be at least 100 more heads than tails or at least 100 more tails than heads
1,000	0.08	0.0017
10,000	0.17	0.3222
100,000	0.48	0.7542
1,000,000	0.95	0.9211
1,000,000,000	≈ 1.00	0.9970

Some Examples

The fallacious law of averages is all around us. It is commonly appealed to by gamblers hoping to find a profitable pattern in the chaos created by random chance. The sports pages are another fertile source of law-of-averages fallacies. Penn State's football team beat West Virginia in 1956 and 1957, tied them in 1958 and then won the next 13 games in a row. On the eve of the 1972 game, newspapers reported that West Virginia had the law of averages on its side. Penn State won that game and eleven more after that—a long wait for those counting on the law of averages to give West Virginia a victory. In 1984, West Virginia ended 28 years of frustration by finally beating Penn State, 17–14. It wasn't the law of averages that won the game. Going into the 1984 game, West Virginia had six wins and one loss and was ranked 14th and 18th in the AP and UPI polls; Penn State was 5–2 and ranked 19th and 20th. After the game, the West Virginia coach said that, "This is the greatest win I've been associated with since I've been here. Beating Oklahoma and Florida and Pitt twice is great, but to beat a team with the class and talent of Penn State ranks above them all." This win was earned by the players, not some fallacious law of averages.

Admittedly, sports do differ from games of chance in at least two ways. First, we may not have much information about the relative strengths and weaknesses of players or teams. If Penn State and West Virginia had never played each other or common opponents, we might be very unsure about the outcome; but when Penn State beats West Virginia year after year, we become convinced that Penn State has the stronger football program. As the victories pile up, it does not become more probable that Penn State will lose the next game. If anything, we become more convinced that Penn State will win yet again.

The second difference between sports and mechanical games of chance is that players do have memories and emotions. They do care about wins and losses and they do remember the last game. Sometimes a team will be beaten badly and play harder the next time, seeking revenge. Other times, teams that lose lots of games also seem to lose self-confidence and grow accustomed to losing. If West Virginia loses to Penn State year after year, West Virginia players may get fired up and Penn State may be nonchalant, but it is also possible that Penn State will play with confidence and West Virginia will be anxious. Destructive self-doubts are seen in almost all sports, and athletes have turned to sports psychologists and even hypnotists to help them relax and forget past failures.

Mary Sue Younger, a statistics professor at the University of Tennessee, told me about another misuse of the law of averages: "I was told by my automobile insurance salesman that some insurance companies will raise your premiums after so many years without a claim, figuring you are coming due for an accident!" If any conclusion can be drawn from many years without an accident, it is probably the opposite: you are a safe driver and should have your premiums reduced.

Hot Streaks in Sports

Sports fans often observe "hot" or "cold" streaks by athletes—the basketball player who makes (or misses) five shots in a row, the football quarterback who completes five passes in a row or throws five straight incomplete passes, the baseball player who gets four hits in a row or makes four outs in a row. What most fans do not appreciate is that even if each outcome is independent of previous outcomes, streaks will happen.

To illustrate this point, I flipped a coin 20 times, which we can interpret as 20 shots by a basketball player or 20 passes by a quarterback, with a heads counting as a successful shot or completed pass, and a tails counting as a missed shot or an incomplete pass. A string of consecutive heads is a hot streak, a string of tails is a cold streak. Here are my results:

H H T T T T H H T H
H H H T H H T T H H

Beginning with the third flip, a streak of four tails in a row occurs. Beginning with the tenth flip, there is a streak of four heads in a row. I had a cold streak and then, a little later, a hot streak, even though every coin flip was independent!

I did this experiment ten times. In seven of my ten experiments, I had a hot or cold streak of four or longer. I had two streaks of four, four streaks of five, and one streak of ten in a row! Table 12.4 shows the theoretical probability of observing streaks of various lengths in a sequence of 20 coin flips. There is a 0.768 probability of a streak of four or more, and seven of my ten experiments yielded such streaks. My streak of ten is a row is unusual, but has about an 11 percent chance if, as here, the experiment is repeated ten times.

Such experiments do not prove that human hot or cold streaks are not due to physical or emotional factors. However, they do show that the occurrence of a hot or cold streak does not, by itself, prove that physical or emotional forces are at work. Streaks occur even with independent coin tosses. The important implication for players, coaches, and fans is that a hot streak does not guarantee continued success and a cold streak does not guarantee continued failure, any more than four heads in a row guarantees heads on the next coin flip.

Table 12.4 Streak Probabilities for 20 Coin Flips

Length of Streak	Probability
< 3	0.021
3	0.211
4	0.310
5	0.222
6	0.121
7	0.061
8	0.029
9	0.013
> 9	0.012

12.4 Approximating the Binomial Distribution

The *central limit theorem* states that if Z is the sum of n independent, identically distributed random variables with a finite, nonzero standard deviation, then the probability distribution of Z approaches the normal distribution as n increases. The number of successes in the binomial model satisfies these conditions. Therefore, the binomial distribution approaches the normal distribution as n increases. Figure 12.2 is an example for the specific case $p = 0.5$, but the central

limit theorem applies for any value of the success probability.

We can consequently use the normal distribution to calculate approximate values of binomial probabilities. Figure 12.2 shows the familiar bell shape for a binomial probability distribution with success probability $p = 0.50$ and $n = 100$ trials. To find the exact binomial probability that x is greater than or equal to 60, we would have to compute and add up the 41 separate (and complicated) binomial probabilities for $x = 60, 61, \ldots, 100$. Instead, we can make a quick approximation by finding the area to the right of $x = 60$ under a normal curve that has been fit to this binomial probability distribution.

Computer software shows that the exact binomial probability is 0.0284:

$$P[x \geq 60] = \sum_{x=60}^{100} 0.5^{x} 0.5^{100-x} = 0.0284$$

To calculate a normal approximation to a binomial probability, we use Equations 12.2 and 12.3, which give the mean and standard deviation of the number of successes x for a binomial distribution, to convert to a standardized Z value:

$$\begin{aligned} Z &= \frac{x - np}{\sqrt{\frac{p(1-p)}{n}}} \\ &= \frac{60 - 100(0.5)}{\sqrt{\frac{0.5(1-0.5)}{100}}} \\ &= 2.0 \end{aligned}$$

Appendix Table A.1 in the textbook shows that the probability of a Z value larger than 2 is 0.0228, which is close to the exact binomial probability.

In general, a satisfactory normal approximation requires a large enough value of n to put the mean np safely away from the extreme values $x = 0$ and $x = n$. To ensure that the mean is at least 5 units away from these extremes, we can use the rule of thumb that a normal approximation is reasonably accurate if np and $n(1 - p)$ are both larger than 5.

A more fundamental question is why we would use a normal approximation. Decades ago, when binomial probabilities had to be calculated by hand, normal approximations were a welcome alternative to excruciatingly tedious computations. Today, researchers have an even better alternative—statistical software that will calculate exact binomial probabilities. This, too, is what you should do.

The purposes of this section are somewhat different. First, we provided hard evidence of the power of the central limit theorem by showing that the normal distribution is a reasonable approximation to the binomial distribution. Second, when we gauge the accuracy of political polls and other empirical data, we will rely heavily on the fact that the binomial distribution is approximately normal with a mean and standard deviation given by Equations 12.2 and 12.3. The calculations in this section justify that reliance.

12.5 Estimating a Success Probability

We can now apply the estimation procedures explained in the textbook to categorical data, for example, the probability that a manufactured product will be defective, the probability that a

patient will benefit from a medication, the probability that a randomly selected person will favor a certain candidate.

To structure our discussion, we focus on a political poll. Among the population of all voters, a fraction p would vote for Candidate A if the election were held today. If a small random sample of size n is drawn from a presumably large population, each draw has a probability p of picking a person who prefers A and a $1 - p$ probability of picking someone who prefers another candidate. Our random sample can be considered a sequence of n independent observations with a constant probability p of "success."

In probability analysis, our reasoning runs from an assumed probability to possible outcomes: we use the binomial model to determine the probability of x successes in n observations if the probability of success on each observation is p. For example, if 40 percent of all voters prefer A, the binomial distribution tells us the probability that a random sample of 1,000 voters will find that more than half prefer A. In statistical analysis, our logic runs in the opposite direction, from outcomes to probability. In a random sample with categorical data that yield x successes in n observations, the success proportion x/n is an estimate of p, the probability of success for each observation.

Our general approach is the same as when we used the sample mean to estimate the population mean μ. It is easy to see why this is so. For each observation of categorical data, we could assign a value 1 if there is a success and 0 if there is a failure. The success proportion x/n, giving the total number of successes divided by the number of observations, is really just a sample mean—the average number of successes—and can consequently be analyzed by the very same procedures that you have already learned: the sample mean is an estimator and the standard deviation of its sampling distribution can be used to determine a confidence interval.

The Mean and Standard Deviation

The only quirk is that the mean and standard deviation for a binomial distribution are related. For normally distributed random variables, the mean μ and standard deviation σ are specified separately. For the binomial distribution, Equations 12.2 and 12.3 show that the mean and standard deviation of x/n depend on the same parameter, the success probability p:

$$\text{Mean of } \frac{x}{n} : p$$

$$\text{Standard deviation of } \frac{x}{n} : \sqrt{\frac{p(1-p)}{n}}$$

Instead of specifying the mean and standard deviation separately, we specify p, which simultaneously determines the mean and standard deviation.

If a fraction $p = 0.40$ of the population prefers candidate A, then the mean and standard deviation of the sampling distribution for the sample proportion x/n for a sample of size $n = 25$ are

$$\text{Mean of } \frac{x}{n} : p = 0.40$$

$$\text{Standard deviation of } \frac{x}{n} : \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(1-0.4)}{25}} = 0.098$$

Here are the means and standard deviations of the sample proportion for various sample sizes:

Sample Size n	Mean of x/n	Standard Deviation of x/n
25	0.40	0.098
100	0.40	0.049
400	0.40	0.024
1600	0.40	0.012

Notice that the mean of the sampling distribution for the x/n is 0.40, no matter the size of the sample, but that the standard deviation declines as the sample increases. (Because the standard deviation depends on the square root of the sample size, we must quadruple the sample in order to halve the standard deviation.)

Since the sample proportion x/n can be interpreted as a sample mean, the central limit theorem tells us that, like the mean of any sufficiently large sample, the sampling distribution for x/n is approximately normal:

$$\frac{x}{n} \sim N\left[p, \sqrt{\frac{p(1-p)}{n}}\right] \tag{12.4}$$

A normal approximation to a binomial distribution is generally satisfactory if both np and $n(1-p)$ are larger than 5. For our polling example with $p = 0.40$, even a sample as small as $n = 25$ easily satisfies these criteria: $np = 25(0.40) = 10$ and $n(1-p) = 25(0.60) = 15$.

Confidence Intervals

Because there is a 0.95 probability that x/n will be within 1.96 standard deviations of p, a 95 percent confidence interval for p is

$$\begin{aligned}95\%\text{ confidence interval for } p &: \frac{x}{n} \pm 1.96\left(\text{standard deviation of } \frac{x}{n}\right)\\ &= \frac{x}{n} \pm 1.96\sqrt{\frac{p(1-p)}{n}}\end{aligned}$$

More generally, for a given confidence level, such as 95 percent or 99 percent, Appendix Table A.1 in the textbook gives the appropriate Z value z^* for the confidence interval:

$$\begin{aligned}\text{confidence interval for } p &: \frac{x}{n} \pm z^*\left(\text{standard deviation of } \frac{x}{n}\right)\\ &= \frac{x}{n} \pm z^*\sqrt{\frac{p(1-p)}{n}}\end{aligned}$$

Notice that the calculation of a confidence interval requires a value for p, which is the very thing we are trying to estimate. A logical solution is to use our estimate x/n.

$$\text{confidence interval for } p : \frac{x}{n} \pm z^*\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} \tag{12.5}$$

For our polling example, suppose that a random sample of 1,000 voters finds that 407 prefer candidate A, a 95 percent confidence interval is

$$\text{confidence interval for } p: \frac{x}{n} \pm z^* \sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

$$= \frac{407}{1000} \pm 1.96 \sqrt{\frac{\frac{407}{1000}\left(1-\frac{407}{1000}\right)}{1000}}$$

$$= 0.407 \pm 0.030$$

To reinforce our logic, consider this very different example. Suppose that historically only forty percent of the people receiving the standard treatment for a certain type of cancer live more than three years. However, of a random sample of 30 patients who were given a new genetically engineered product, 20 lived more than three years. We want to use these sample data to estimate the probability p that a cancer patient receiving this treatment will live more than three years.

First, we calculate the success proportion in the sample; here,

$$\frac{x}{n} = \frac{20}{30}$$

$$= 0.667$$

To see if we can use a normal approximation to construct a confidence interval, we check that np and $n(1-p)$ are both larger than 5. Using $x/n = 0.667$ as an estimate of p: $np = 30(0.667) = 20$ and $n(1-p) = 30(1-0.667) = 10$. Okay. For a 95 percent confidence interval, we use Equation 12.5 with $z^* = 1.96$:

$$\text{confidence interval f}or\ p: \frac{x}{n} \pm z^* \sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

$$= 0.667 \pm 1.96 \sqrt{\frac{0.667(1-0.667)}{30}}$$

$$= 0.667 \pm 0.169$$

These data indicate that a 95 percent confidence interval for the probability that a cancer patient who receives this new treatment will live more than three years is 0.667 ± 0.169. For a more precise estimate (with a narrower margin of error), we need to test more than 30 patients.

Choosing a Sample Size

One way of deciding how large a random sample to use is to specify a desired margin m for sampling error, and then set m equal to the margin for sampling error given in Equation 12.4:

$$m = z^* \sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

Rearranging, the requisite sample size is

$$n = \left(\frac{z^*}{m}\right)^2 \left(\frac{x}{n}\right)\left(1-\frac{x}{n}\right)$$

We don't know what x/n will turn out to be. However, trial and error confirms that the maximum value of $(x/n)(1 - x/n)$ is at $x/n = 0.5$. Therefore, if we set $x/n = 0.5$, we obtain a conservative value for the necessary sample size—conservative in the sense that if our estimate of p does not equal 0.5, our margin for error will be even smaller than our target margin.

$$\begin{aligned} n &= \left(\frac{z^*}{m}\right)^2 (0.5)(1-0.5) \\ &= \left(\frac{z^*}{2m}\right)^2 \end{aligned} \tag{12.6}$$

To illustrate the procedure, suppose that we want a 95 percent confidence interval for a political poll to have a 2 percent margin for sampling error. Equation 12.6 tells us that we need a sample size of 2,401:

$$\begin{aligned} n &= \left(\frac{z^*}{2m}\right)^2 \\ &= \left(\frac{1.96}{2(0.02)}\right)^2 \\ &= 2{,}401 \end{aligned}$$

By comparing alternatives, we can make an informed choice based on the tradeoff between a smaller margin of error and a larger (and more expensive sample). These illustrative calculations for a 95 percent confidence interval show how a larger sample size gives a smaller margin for sampling error:

Sample Size	Margin for Sampling Error
10	± 0.310
100	± 0.098
500	± 0.044
1,000	± 0.031
2,500	± 0.020
5,000	± 0.014

A sample of size 10 would almost always be rendered useless by the 31 percent margin for sampling error. Usually, a better choice is a sample in the range 1,000 to 2,500, with a margin of error of 3 percent to 2 percent.

A Gallup Poll

We can use the final Gallup Poll in the 1996 presidential election poll to illustrate these procedures. During the weekend before the election, the Gallup organization conducted telephone interviews with 2,400 people, of whom 1,573 were identified as likely voters, based on the strength of their feelings and the reported frequency with which they had voted in the past. We assume that the 1,573 people used for their election prediction were a random sample from the population of election-day voters.

We can interpret the population parameter p as the fraction of the voting population who were going to vote for Bob Dole, the Republican candidate in 1996. If the Gallup poll is a random sample from a large population, then each voter had an equal chance of being selected

and, with every selection, the probability of choosing a Dole voter is p. Of the sample of n = 1,573 likely voters, Gallup identified x = 645 who indicated they would vote for Dole. Thus the sample proportion is 0.41:

$$\frac{x}{n} = \frac{645}{1{,}573} = 0.410$$

Recognizing that this is but one of many samples that might have been selected, we gauge the reliability of this estimate by calculating a 95 percent confidence interval:

$$\text{confidence interval for } p: \frac{x}{n} \pm z^* \sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

$$= 0.041 \pm 1.96\sqrt{\frac{0.041(1-0.041)}{1{,}573}}$$

$$= 0.041 \pm 0.024$$

This calculation predicted that Dole would receive 41.0 percent of the vote, with a 2.4 percent margin of error—meaning that we anticipate that, on average, 19 of 20 such confidence intervals will include the actual percentage the candidate receives in the election. As it turned out, Dole received 40.7 percent of the vote, which is inside this confidence interval.

Table 12.5 shows several Gallup presidential election polls. Given that these are not simple random samples, that pollsters must predict who will vote, that some voters do not tell the truth, and that some voters change their minds during the time between the poll and the election, this is an impressive record. Surveying roughly 2,000 likely voters, Gallup is generally able to predict within a few percentage points the outcome of elections involving more 100 million voters.

Table 12.5 The Gallup Presidential Poll's Forecasting Record

Year	Candidate	Final Gallup Survey	Election Result	Error
2008	Obama	55.0	53.0	2.0
	McCain	44.0	46.0	-2.0
2004	Bush	49.0	50.7	-1.7
	Kerry	49.0	48.3	0.7
2000	Bush	48.0	47.9	0.1
	Gore	46.0	48.4	-2.4
	Nader	4.0	2.7	1.3
1996	Clinton	52.0	49.2	2.8
	Dole	41.0	40.7	0.3
	Perot	7.0	8.4	-1.4
1992	Clinton	49.0	43.3	5.7
	Bush	37.0	37.7	-0.7
	Perot	14.0	19.0	-5.0

1988	Bush	56.0	53.0	3.0
	Dukakis	44.0	46.1	-2.1
1984	Reagan	59.0	59.2	-0.2
	Mondale	41.0	40.8	0.2
1980	Reagan	47.0	50.8	-3.8
	Carter	44.0	41.0	3.0
	Anderson	8.0	6.6	1.4
	Other	1.0	1.6	-0.6
1976	Carter	48.0	50.1	-2.1
	Ford	49.0	48.1	0.9
	McCarthy	2.0	0.9	1.1
	Other	1.0	0.9	0.1
1972	Nixon	62.0	61.8	0.2
	McGovern	38.0	38.2	-0.2
1968	Nixon	43.0	43.5	-0.5
	Humphrey	42.0	42.9	-0.9
	Wallace	15.0	13.6	1.4
1964	Johnson	64.0	61.3	2.7
	Goldwater	36.0	38.7	-2.7
1960	Kennedy	50.5	50.1	0.4
	Nixon	49.5	49.9	-0.4

source: Gallup.com

12.6 Test for a Success Probability

We often do things "instinctively" or "by nature." Where do instincts come from? If they are truly instinctive, how do we acquire them? Perhaps by natural selection. For instance, how do seed-bearing plants know how and when they need to release their seeds in order to ensure future plants? They didn't learn this during their lifetimes. Through countless generations of natural selection, plants that didn't release seeds at the right time or in the right manner didn't have offspring to carry on these ineffectual survival traits. Natural selection means that traits or behavior that are conducive to survival and reproduction are more likely to survive and be passed on to future generations, either in genes or in genetic programs.

Why do so many animals instinctively bark or roar when they confront strangers? This barking is intended to warn the stranger that the animal is not weak and is willing to fight. Similar animals that didn't issue effective warnings got into too many fights and were thinned out of the gene pool. Why do so many animals have fierce teeth, horns, or other built-in weapons? Because their ancestors were able to compete effectively for food and reproductive opportunities, and so perpetuate their genes. Why are gazelles, which are poor fighters, able to run fast? Those that weren't fast enough didn't pass on their genes. Why do some creatures that are neither strong nor swift blend so well into their natural surroundings? You know the answer.

This theory has many intriguing implications. Why, for example, are modern males, on average, larger and more muscular than women? Natural selection suggests that because young

men fought with each other for innumerable generations when the weapons were hands, rocks, and clubs, the largest and strongest lived while the smallest and weakest died (and with them, their genes).

Now, consider this question carefully. If I were to hand you a baby, would you hold it against your right shoulder or your left shoulder? There is a pronounced preference. What is it and why? We are about to learn the surprising answer.

If we were to conduct this experiment with 100 randomly selected people, our data would be categorical, as right and left do not have natural numerical values, but are instead two categories of possible outcomes. To analyze such categorical data, we can use the binomial model by thinking of each trial as a random draw from a population from which there is a probability p of selecting a person who will hold the baby against, say, the left shoulder.

The principles of hypothesis testing developed in the textbook can easily be extended to tests of the binomial success probability p. First, we specify the null and alternative hypotheses about the value of p. Then we use a random sample to calculate the success proportion x/n. The P value is the probability that the sample success proportion x/n would be so far (or farther) from the value of p if the null hypothesis were true. The P value can either be calculated exactly using the binomial distribution or (if the sample is large) approximated by the normal distribution. If the alternative hypothesis is two-sided, we double this probability to obtain the two-sided P value. That's all there is to it! We will use several examples to illustrate the application of these procedures, and then apply them to the baby on the shoulder.

Exact Binomial Probabilities

Sixteen pregnant women were asked to read a children's story aloud three times a day during the last six and a half weeks of pregnancy. These readings were tape recorded and, shortly after birth, each baby was allowed to choose (by varying the rhythm of its sucking on a special bottle) between a tape recording of the mother reading the familiar story and a tape recording of the mother reading an unfamiliar story. In 13 of 16 cases, the babies chose the familiar story. To assess the statistical significance of these results, the natural null hypothesis is that each baby has a 0.5 probability of choosing the familiar story. The alternative hypothesis should be two-sided, because we cannot rule out the possibility that a baby would prefer to listen to something new.

$$H_0: \quad p = 0.5$$
$$H_1: \quad p \neq 0.5$$

Using the binomial distribution

$$P[x \geq 13] = \sum_{x=13}^{16} \binom{16}{x} 0.5^x (1-0.5)^{16-x} = 0.0105$$

The two-sided P value is 2(0.0105) = 0.0210, which is sufficiently low to reject the null hypothesis at the 5 percent level, but not at the 1 percent level.

Normal Approximation to a Binomial

Earlier in this chapter, we saw that if the sample is sufficiently large (np and $n(1-p)$ both larger than 5), then the sample success proportion x/n is approximately normally distributed (Equation 12.4). Therefore, we can use this Z statistic to test a null hypothesis about the value of p:

$$Z = \frac{\text{observed statistic} - \text{null hypothesis}}{\text{standard deviation of statistic}}$$
$$= \frac{\frac{x}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

As an illustration, consider a 1958 study of the racial awareness of black children. The sample consisted of 253 black children, 134 from Arkansas and 119 from Massachusetts, whose ages ranged from three to seven years. Each child was offered four dolls, two white and two black, and told to, "Give me the doll that you would like to play with." Overall, 83 children chose a black doll and 169 chose a white doll. (One child refused.) Can the fact that two-thirds of these children chose a white doll be explained plausibly by chance—the coincidental result of color-blind choices—or does it indicate that black children in 1958 were well-aware of race?

For a statistical test, the null hypothesis is that the choices are color blind, so that each child has a probability $p = 0.5$ of selecting a white doll. Because we have no reason beforehand to rule out a preference for either white or black dolls, the alternative hypothesis is two-sided:

$$H_0: \quad p = 0.5$$
$$H_1: \quad p \neq 0.5$$

Evidence against the null hypothesis is provided by an observed sample proportion x/n that is far from 0.5. The exact probability that 169 or more of the 252 children who participated would choose a white doll is given by the binomial distribution. Using statistical software, the exact P value is

$$P[x \geq 169] = \sum_{x=169}^{253} \binom{253}{x} 0.5^x (1-0.5)^{253-x} = 0.000000032$$

For a two-sided P value, we double this probability: $2(0.000000032) = 0.000000064$.

Because this is a reasonably large sample, we can also calculate an approximate P value by using a normal approximation to the binomial distribution. Checking our rule of thumb, and using the value of p specified by the null hypothesis, $np = 252(0.5) = 126$ and $n(1-p) = 252(0.5) = 126$ are both (much) larger than 5. Using $x = 169$, $n = 252$, and $p = 0.5$ under the null hypothesis, the Z value is 5.42:

$$Z = \frac{\frac{x}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}$$
$$= \frac{\frac{169}{252} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{252}}}$$
$$= 5.42$$

We find the approximate probability that x/n is larger than 169/252 by determining the

probability that a standardized Z value is larger than 5.42:

$$P\left[\frac{x}{n} > \frac{169}{252}\right] = P[Z > 5.42]$$
$$= 0.000000031$$

The two-sided P value is 2(0.000000031) = 0.000000062, which is very close to the exact value.

How Do We Hold Babies?

We are now ready to answer the question posed at the beginning of this section. While observing a rhesus monkey at a zoo, Lee Salk, a psychologist, noticed that on 40 of 42 occasions the monkey held her baby on the left side of her body, close to her heart. He then observed 287 human mothers shortly after they had given birth. Of 255 right-handed mothers, 83 percent held their baby on the left side of their body; of 32 left-handed mothers, 78 percent held their babies on the left side. In contrast, when 438 shoppers who were carrying packages the approximate size of a baby left a supermarket with automatic doors, exactly half held the package on their left side.

In a controlled experiment, he compared the weight gains and amount of crying for babies in a nursery who listened to a tape recording of a human heartbeat 24 hours a day for four days with a control group that did not listen to the tape. The babies who listened to the heartbeats had a median weight gain of 409 grams and cried 38 percent of the time, compared to the control group, which had a median weight loss of 20 grams and cried 60 percent of the time. Because the babies consumed the same amount of food, he concluded that the weight gain was evidently due to decreased crying.

Salk speculated that "it is not in the nature of nature to provide living organisms with biological tendencies unless such tendencies have survival value....[W]hen a baby is held on the mother's left side, not only does the baby receive soothing sensations from the mother's heartbeat but also the mother has the sensation of her heartbeat being reflected back from the baby."

Evidently, mothers instinctively know that their infants are soothed by the sound of a heartbeat. We can use the data for the $n = 287$ human mothers to test the null hypothesis that a baby is equally likely to be held on the left or right side: H_0: $p = 0.5$. The observed sample proportion $x/n = 0.82$ gives this Z value:

$$Z = \frac{\frac{x}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}$$
$$= \frac{0.82 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{287}}}$$
$$= 10.84$$

The observed 0.82 sample proportion is 10.84 standard deviations away from the 0.5 value of p (were the null hypothesis true). The P value is virtually zero and the null hypothesis is decisively rejected.

Are Mendel's Data Too Good to Be True?

Gregor Mendel's probabilistic model of how genes pass from one generation to the next is one of our most important and elegant scientific theories. Mendel reported many experimental results supporting his theory, and other researchers have conducted numerous experiments confirming his insights. One of Mendel's original experiments involved the self-fertilization of hybrid yellow-seeded sweet peas. His theory implied that the offspring have a 0.75 probability of being yellow-seeded and a 0.25 probability of being green-seeded. Mendel reported the following data:

Offspring	Number
Yellow-seeded	6,021
Green-seeded	2,002
Total	8,023

To determine if these data support or reject Mendel's theory, we can use the binomial model with p equal to the probability of a yellow-seeded offspring. The null and alternative hypotheses are

$$H_0: \quad p = 0.75$$
$$H_1: \quad p \neq 0.75$$

The observed sample proportion, $x/n = 6{,}021/8{,}023 = 0.7505$, is almost exactly to 0.75, the value of p under the null hypothesis. Because the sample of $n = 8{,}023$ plants is certainly large enough to use a normal approximation, we convert to a standardized Z value:

$$Z = \frac{\frac{x}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.7505 - 0.75}{\sqrt{\frac{0.75(1-0.75)}{8{,}023}}} = 0.097$$

Since $P[Z > 0.097] = 0.4615$, the two-sided P value is $2(0.4615) = 0.923$, which is very far from the 0.05 required to reject the null hypothesis at the 5 percent level. Therefore, these data are consistent with Mendel's theory.

Under the null hypothesis, the population mean for the number of yellow-seeded plants $np = 8{,}023(0.75) = 6{,}017.25$. Mendel's reported number of 6,021 is off by fewer than 4 plants in an experiment involving 8,023 plants! Mendel's data support his theory amazingly well, perhaps too well. Some 70 years later, R.A. Fisher argued that Mendel's data were literally too good to be true.

Fisher reversed the usual hypothesis-test question by asking, if the null hypothesis is correct, what is the probability of observing a value of x/n that is so close to p? This is equal to one minus the two-sided P value: $1 - 0.923 = 0.077$. If there is a 0.923 probability that a normally distributed random variable will be more than 0.097 standard deviations (in either direction) from its population mean, then there is a 0.077 probability that a normally distributed variable will be within 0.097 standard deviations (in either direction) of its mean. Such a close correspondence between theory and data requires a bit of luck, or something less innocent.

Mendel studied seven inherited characteristics of pea plants and, in every single case, reported a close correspondence between his predictions and his results. The probability of such a phenomenal run of luck is about 0.000004. It is difficult to believe that Mendel was so lucky. Even worse, in one of his studies, he miscalculated the theoretical probability and then reported data that fit this erroneous probability almost exactly (and that are inconsistent with the correct probability). It is hard to disagree with Fisher's conclusion that Mendel's data are too good to be true. However, Fisher was so impressed by Mendel's scientific contributions and the fact that Mendel was a monk that he placed the blame elsewhere: "I have no doubt that Mendel was deceived by a gardening assistant, who knew only too well what his principal expected from each trial."

Exercises

12.1 Identify which of the following can be described by the binomial model and explain your reasoning:
   a. head or tail when a bent coin is flipped.
   b. safe or out when a baseball player bats.
   c. black or not when a bird lands on a telephone wire.
   d. Republican or not when a governor is elected in Oregon.

12.2 After an Iraqi missile hit the U.S. frigate *Stark*, novelist Tom Clancy wrote that

> It's quite possible that the *Stark*'s radar did not see the weapons separate from the launching aircraft. I have seen radar miss an aircraft carrier sitting in plain view on the horizon. The same radar will "paint" the blip nine times out of 10, but the laws of statistics mean that occasionally it will miss on two of the three consecutive passes.

Assuming that the binomial model applies with $p = 0.9$, what is the probability of two or more misses in three trials?

12.3 A traditional Cayuga game uses six peach pits that have been rubbed smooth and blackened on one side with fire. The six peach pits are placed in a cup, shaken thoroughly, and then viewed. The player is credited with five points if all six pits have a similar side face up (either all blackened or all not blackened) and is credited with one point if five of the six pits have a similar side face up, with one pit different. Otherwise, the player receives no points. Assuming that each pit is equally likely to land blackened or not-blackened, what is the probability of five points? What is the probability of one point?

12.4 In 1984 a sports columnist for the *Dallas Morning News* had a particularly bad week picking the winners of NFL football games—he got one right and twelve wrong, with one tie. Afterwards, he wrote that, "Theoretically, a baboon at the Dallas Zoo can look at a schedule of 14 NFL games, point to 1 team for each game and come out with at least 7 winners." The next week, Kanda the Great, a gorilla at the Dallas Zoo, made his predictions by selecting pieces of paper from his trainer. Kanda got nine right and four wrong, better than all six *Morning News* sportswriters. Assuming no ties, what is the probability that a baboon will select at least nine winners? Will select fewer than two winners?

12.5 In 1980 the United States tried a daring rescue of American hostages being held by Iranians in the American Embassy in Teheran. The American plan was to bring in soldiers in eight helicopters traveling 800 miles across the desert at low altitudes at night. The military commanders believed that at least six of the eight helicopters would have to reach Teheran for the mission to have a chance of succeeding. The rescue attempt was canceled when three of the helicopters became disabled. Use the binomial distribution to calculate the probability that at least six of eight helicopters would reach Teheran if each has a 0.75 probability of success. If each has a 0.90 probability of success.

12.6 Explain why you either agree or disagree with the following:

> This man has kept records of play on Blackjack, craps, and Roulette, and the house percentage on all three games works out inexorably, within a fraction of a percentage point—but there are times shown by his records when there are lengthy streaks of steady house wins along with other streaks of house losses.
>
> His records show that such streaks exist. But no matter how long he studies his figures, he hasn't been able to arrive at any pattern in them. He can see that the streaks are there, but he hasn't anything even resembling an explanation of why they occur when they do or why they last for a certain period.
>
> "Nonetheless," he insists, "if I'm scoring red and black in a Roulette game and red has come up 500 times in 900 spins of the wheel, I'm going to put my money on black the next 100 spins. What's more, I'll bet you that I come out ahead of the game."

12.7 In 1989, *Sports Illustrated* reported that Kansas City's Pat Tabler, a career 0.290 hitter, had 38 hits in 66 times at bat with the bases loaded, an astonishing 0.576 batting average in these key situations. Use a normal approximation to the binomial distribution to calculate the probability that an event with a 0.290 probability of success would happen 38 or more times in 66 trials. Why is this data grubbing?

12.8 The U.S. Postal Service has on-time standards for mail delivery; for instance, letters from Washington, D.C. to Manhattan should arrive within two days. The Postal Service claims that it delivers 90 percent of the mail to Manhattan on time. In the summer of 1988, *New York* magazine tested this claim by mailing 144 letters from around the country to Manhattan and found that 50 percent did not arrive on time. If the Postal Service's claims are correct what is the probability that 72 of more of these 144 letters would not arrive on time? Assume that the binomial model applies and use a normal approximation.

12.9 Mendel postulated that the self-fertilization of hybrid yellow-seeded sweet peas would yield offspring with a 0.75 probability of being yellow-seeded and a 0.25 probability of being green-seeded. In 1865, he reported that 8,023 such experiments yielded 6,021/8,023 = 0.7505 yellow-seeded plants and 2,002/8,023 = 0.2495 green-seeded plants. It has been suggested that these results are too good to be true, in that Mendel must have reported what he wanted rather than what he observed. If these were honest trials that could be described by the binomial model, what is the exact probability that the number of yellow-seeded plants would be between 6,013 and 6,021 (that is, no more than four plants from the expected value, 6,017.25)? Do not use a normal approximation.

12.10 Karl Pearson reported flipping a coin 24,000 times and obtaining 12,012 heads and 11,988 tails. If heads and tails are equally likely, what is the probability that the number of heads would be at least 11,988 and no more than 12,012 (that is, no more than 12 from the mean)? Do not use a normal approximation.

12.11 The U.S. Supreme Court compared 6-person and 12-person juries and declared "there is no discernible difference between the results reached by the two different sized juries." Since jurors are not allowed to discuss the case with anyone until all the testimony has been heard, we will look at an initial secret ballot conducted by the jury immediately after the testimony has been completed. Assume that jurors are a random sample from a population in which a fraction $p = 0.9$ would vote for conviction in this initial secret ballot and a fraction $1 - p$ would vote for acquittal. Compare the probabilities that all members of a six- and twelve-person jury will vote for conviction. What if $p = 0.8$?

12.12 During the January 1984 NFC championship game between Washington and San Francisco, it appeared that the game might be decided by a field goal. One of the CBS commentators, Jack Buck, pointed out ominously that Mark Mosely had already missed four field goals in the game. Hank Stram, the other CBS commentator, responded, "The percentages are in his favor." Explain why you either agree or disagree.

12.13 In a 1979 court case, a military supplier argued that the U.S. government had tested too small a sample when it rejected a shipment of 20,000 nose fuse adaptors, a component of artillery shells. Military procurement standards specified that a shipment of 10,001 to 35,000 adaptors would be tested by sampling 315 pieces, yet the government rejected this shipment of 20,000 pieces after testing 20 adapters and finding them all to be defective. What is the probability that every item in a random sample of 20 items from an infinite population will be defective if the probability of selecting a defective item is $p = 0.10$?

12.14 A *Los Angeles Times* basketball article in was titled, "Wright Defies the Percentages." The article explained that, "Oregon's strategy was obvious in the closing minutes—foul Gerry Wright, a 37.5 percent free throw shooter. But Wright, a reserve center, didn't fold under the pressure. He made 3 of his 6 foul shots in the final two minutes."

a. Assuming the binomial model to be appropriate, what is the probability that Wright would make more than 2 of 6 free throws?
b. Is the fact that he did so an unlikely defiance of the percentages?
c. What assumptions are needed for the binomial model to be appropriate?

12.15 In his final newspaper column, Melvin Durslag reminisced about some of the advice he had received in his fifty-one years as a sports columnist, including this suggestion from a famous gambler: "Nick the Greek tipped his secret. He trained himself so that he could stand at the table eight hours at a time without going to the washroom. It was Nick's theory that one in action shouldn't lose the continuity of the dice." Explain why you either agree or disagree with Nick's secret.

12.16 "Five out of ten people who have this disease die from it. It is lucky you came to me; my last five patients all died." Would you be comforted by such report? If there is a 0.5 probability of dying from this disease, what is the probability that five of five patients will die from it?

12.17 A study compared the ZIP code residences of 4,121 women who had immigrated to California from another country with the ZIP code residences of their adult daughters at comparable periods of their lives. Of the 4,121adult daughters, 581had the same ZIP code as their mother and 3,540 had different ZIP codes.
   a. Estimate a 95% confidence interval for the probability that the adult daughter will have the same ZIP code as her mother.
   b. Of the 3,540 daughters who lived in different ZIP codes, 1,986 lived in more affluent ZIP codes and 1,554 lived in less affluent ZIP codes. Calculate the exact two-sided *P* value for a test of the null hypothesis that a daughter who lives in a different ZIP code is equally likely to live in a more affluent or less affluent ZIP code.

12.18 Two baseball teams will play 5 games against each other. In each game, Team A has a 0.6 probability of winning and Team B has a 0.4 probability of winning. What is the probability that Team B will win at least 3 of these 5 games?

12.19 Fifty-six college students participated in an ESP test in which a professor flipped a coin ten times and each student attempted to identify each flip as a head or tail.
   a. If a student's guess is equally likely to be a head or tail, regardless of the professor's flip or the student's previous guesses, what is the probability that a student will end up with 10 guesses consisting of exactly 5 heads and 5 tails?
   b. Let *p* be the probability you found in question a. It turned out that 27 of these 56 students ended up with guesses consisting of exactly 5 heads and 5 tails. Use these data to test the null hypothesis that a student's guess is equally likely to be a head or tail, regardless of the professor's flip or the student's previous guesses.

12.20 A sociology professor argued that the stress associated with the number of 4 might cause Chinese and Japanese persons to suffer an unusually large number of fatal heart attacks on the fourth day of the month. Mortality data for Chinese and Japanese Californians for the years 1989-1998 show that 472 of the 1391 fatal heart attacks that occurred on days 3, 4, or 5 of a month occurred on day 4. Use these data to test the null hypothesis that of those fatal heart attacks occurring on days 3, 4, or 5 of a month, there is a 1/3 probability that the death will be on day 4. (Use a normal approximation.)

12.21 In the gambling game Chuck-A-Luck, a player can bet $1 on any number from 1 to 6. Three dice are thrown and the payoff depends on the number of times the selected number appears. (For example, if you pick the number 2, your payoff is $4 if all three dice have the number 2.) What is the expected value of the payoff?

Number of dice with number	0	1	2	3
Payoff (dollars)	0	2	3	4

12.22 When a National Football League (NFL) game is tied at the end of regulation, the teams play a 15-minute sudden death overtime with a coin flip used to decide which team kicks off. The first team to score wins. During the 2001–2007 seasons, 113 games were decided by sudden death overtime. The team that received the ball first won 66 games and the team that kicked off won 47 games. Use these data to calculate the exact two-sided *P* value for a test of the null hypothesis that of those games decided by sudden death overtime, the kicking and receiving teams have an equal chance of winning.

12.23 In the NCAA, tied football games are decided by giving each team the ball on the other team's 25 yard line and seeing which team is able to score the most points in their turn. If the teams are still tied after each team has had one chance to score, they do it again with the teams reversing the order. During the 2003-2007 seasons, there were 160 overtime games, with the team going first winning 71 games and the team going second winning 89 games. Use these data to calculate the exact two-sided *P* value for a test of the null hypothesis that the teams going first and second have an equal chance of winning.

12.24 In basketball, defensive field goal percentage (DFG) is the percentage of shots that the opposing team makes (not counting free throws). Thus, if the opposing team attempts 100 shots and makes 45, the DFG is 45 percent. In the 37 National Basketball Association (NBA) seasons from 1970-1971 through 2006-2007, 35 of the 37 NBA champions had a DFG that was lower than the average DFG for all teams that season. (The exceptions were the Chicago Bulls, led by Michael Jordan, in 1990-1991 and 1992-1993.) Calculate the exact two-sided *P* value for a test of the null hypothesis that the NBA champion is equally likely to have a DFG that is above or below the league average.

12.25 The pointspread set by bookmakers is a prediction of the margin of victory in a game. For example, in its first game of the 2000 National Football League season, Minnesota was a 4-point favorite to beat Chicago. Bettors who picked Minnesota would win their bets if Minnesota won by more than 4 points; those who picked Chicago would win their bets if Chicago won or if Minnesota won by fewer than 4 points. It turned out that Minnesota won 30-27; those who picked Chicago won their bets.

A researcher found that over the years 1993-2000 a strategy of betting against teams that beat their pointspreads the previous week would have won 902 wagers and lost 831 wagers. Test the null hypothesis that this strategy is equivalent to flipping a coin to pick winners.

12.26 William is running out of time on a multiple-choice exam. There are 20 questions left, each with 5 possible answers. He marks 20 answers without even reading the questions.

a. What is the probability that he will get more than half of these questions correct?
b. Is he more likely to get more than half his guesses correct if he guesses the answers to 10 questions or 20 questions? Use logic, not calculations, to answer this question.

12.27 A certain college has found that half of the people they accept for admission decide to enroll and half do not. Assuming that these enrollment decisions are independent, with each student having a 1/2 probability of enrolling, how many students should they admit if they want the expected value of the number of students who enroll to be 400? If they follow this admission policy, what is the probability that more than 420 will enroll?

12.28 Five polls completed on November 5, 2000, gave these results:

	Number Surveyed	Bush (%)	Gore (%)
CNN/USA Today/Gallup	2386	47	45
IBD/CSM/TIPP	989	48	42
ICR	1000	46	44
Reuters/MSNBC	1200	47	46
VOTER.COM	1000	46	37

Combine these polls to estimate the percentage of the vote that Bush would have received if the election had been held that day. Give a 95% confidence interval for this prediction.

12.29 In the 1940s, the Federal Drug Administration seized a shipment of 22,464 rubber prophylactics when tests of a sample of 408 prophylactics found 30 to be defective, with holes that would allow the transmission of disease. In federal court, the manufacturer argued that these 30 defective prophylactics represented only one-tenth of one-percent of the shipment, and that this was not sufficient grounds for seizure. If you were the judge in this case, would you find this argument persuasive? Estimate a 99 percent confidence interval for the fraction of the prophylactics in the shipment that were defective.

12.30 A 1995 survey for Fodor's, a publisher of travel guidebooks, asked 600 U.S. travelers to name their least desirable vacation destination. The winner, with 9 percent of the votes was Iraq/Iran; in second place, with 7 percent, was New York City. A spokesman for New York's mayor scoffed, "They only interviewed 600 people? That's the size of an apartment building." Assuming this to be a random sample, give a 95 percent confidence interval for the percentage of all U.S. travelers who would choose New York City as their least desirable vacation destination.

12.31 In 1988 the PGA Tour made a special study of the putting success of professional golfers. At 15 tournaments, they selected one green with a relatively flat and smooth surface and measured and recorded the result of each putt on that green. Here are some of their results:

Length of Putt	Number of Putts	Fraction Made
5	353	0.589
15	167	0.168

For each of these two distances, calculate a 95 percent confidence interval for the probability of making a successful putt, assuming that the binomial model applies and using a normal approximation to the binomial distribution. Why might the binomial model be inappropriate?

12.32 In the game Roshambo (rock-scissors-paper), two players simultaneously move their fists up and down three times and then show a fist (rock), two fingers (scissors), or an open hand (paper). Rock beats scissors, scissors beats paper, and paper beats rock. If there is a tie, the players repeat the game. A student played fifty matches against randomly selected opponents and recorded the frequency of the initial moves. Considering these data to be a random sample, estimate this student's probability of playing rock on the initial move and use a normal approximation to the binomial distribution to give a 99 percent confidence interval. Do the same for the opponents' probability of playing rock on the initial move.

	Student	Opponents
Rock	26	16
Scissors	9	17
Paper	15	17

12.33 Explain how you, as a statistician, would evaluate this advice: "Look for a sample of at least 600 interviews. If fewer than 400 people were polled, you can skip the survey."

12.34 Each of twelve college intramural basketball players was asked to take 30 shots 15 feet from the basket—ten shots from the free-throw line, ten shots from a 45-degree angle to the basket, and ten shots from the baseline. Assuming that the binomial model is appropriate, use the results shown below to give a 95 percent confidence interval for the probability of making each of these three shots.

	Attempts	Successes
Free-throw line	120	84
45-degree angle	120	62
Baseline	120	53

12.35 A study of a sample of 286 NBA basketball games during the 1989–1990 season found that 6,841 personal fouls were called against visiting teams and 6,574 against home teams. Use a normal approximation to the binomial distribution to calculate a 95 percent confidence interval for the probability that a randomly selected personal foul is against the visiting team.

12.36 A 37-car train carrying 2,500 kegs of beer was derailed during a snowstorm. The shipper's insurance policy covered the complete cost of this accident if more than 10 percent of the kegs were damaged. The shipper examined a sample of 50 kegs and found that 12 (24 percent) were damaged. The insurance company disputed the use of a sample in place of an examination of all 2,500 kegs; however, a court awarded damages based on the calculation of a 95 percent confidence interval. Determine a 95 percent confidence interval for p, the fraction of kegs damaged, using a normal approximation to the binomial distribution and assuming an infinite population. Does this confidence intervals include $p = 0.10$? If 10 percent of an infinite population of kegs are damaged, what is the probability that 12 or more kegs will be found damaged in a random sample of 50 kegs? (Use a normal approximation.)

12.37 A marketing survey will be used to estimate the fraction of the population that prefer a new spaghetti formula to the old formula. The pollsters advertise that there is a 0.95 probability that their sample estimate will be within two percentage points of the actual fraction of the population that prefer the new formula. How large a sample are they using?

12.38 A student mailed a short questionnaire to the 495 professors at her school. One of the questions was "Overall, would you consider your views more consistently in agreement with Democrats or Republicans?" Of the 248 professors who responded, 193 answered Democrats and 55 Republicans. Assuming (unrealistically) these responses to be a random sample from an infinite population, calculate a 95 percent confidence interval for the fraction of all the professors at her school who would answer Democrat. If we were to take into account the fact that the population was finite, would this shrink or widen the confidence interval? Give one good reason why these data may not be a random sample.

12.39 A marine biologist hypothesized that when there is a long interval between calving, female whales are more likely to have sons than daughters. Data collected by the Center for Coastal Studies found that whales with an interval of 1–2 years between calving had twenty-two sons and twenty daughters. For mothers calving at intervals of three, four, or five years, there were sixteen sons and only four daughters. If these are independent observations with male and female calves equally likely, what is the probability that of twenty calves, sixteen or more will be male? (Do not use a normal approximation.)

12.40 In February of 1995, *Sports Illustrated* reported that the most serious kinds of knee injuries "are virtually epidemic in women's college basketball." They found that among currently active basketball players in the six major Division I conferences, 83 women and 26 men had suffered tears in their anterior cruciate ligaments (ACL), one of the two main ligaments that support the knee. If we consider each ACL injury a binomial trial with a 0.5 probability of involving a male and a 0.5 probability of involving a female, what is the exact probability that of 109 injuries, more than 82 would be to females?

12.41 As of 2011, U.S. presidents had fathered 91 boys and 63 girls. Assuming these to be independent events, use a normal approximation to test the null hypothesis that the children of presidents are equally likely to be male or female.

12.42 A high school basketball coach who was named "coach of the decade" by the *Los Angeles Times* told his players that there is a 60 percent probability that a missed shot will bounce "long," to the opposite side of the basket from which the shot was taken. To test this claim, an examination of video tapes of three men's Division III college basketball games found that 84 of 169 missed shots bounced short and 85 bounced long. Calculate the Z value and use the two-sided P value to test the following null hypotheses:

a. A missed shot has a 0.6 probability of bouncing long.
b. A missed shot has a 0.5 probability of bouncing long.

12.43 A high school basketball coach said that a missed free throw by a right-handed shooter is more likely to bounce to the right, while the reverse if true of a left-hander. To investigate this theory, a student asked two college basketball players to shoot 50 free throws apiece. These players missed 21 of 100 free throws, of which 15 landed on the same side as their shooting hand and 6 bounced to the opposite side. Assuming the binomial model to be applicable, calculate the two-sided P value for testing the null hypothesis that missed free throws are equally likely to bounce to either side. Do not use use a normal approximation.

12.44 High doses of anti-cancer drugs can be effective in killing malignant cells in women with breast cancer, but also destroy vital bone marrow. In 1990, Professor Karen Antman of the Harvard Medical School reported that 150 of 259 patients with advanced breast cancer who received a combination of bone marrow transplants and high doses of anti-cancer drugs were free of cancer after 12 months. In contrast, of those patients with advanced breast cancer who receive conventional treatment, thirty percent were free of cancer after 12 months. Letting $p = 0.30$ be the null hypothesis, are Antman's results statistically significant, using a one-tail test at the 1 percent level and a normal approximation?

12.45 A soccer player organized a penalty-kick tournament for ten right-footed Division III college soccer players. A total of 152 penalty kicks were taken, of which 88 were shots to the right side of the goal (from the kicker's perspective) and 64 were shots to the left side.
   a. Use these data to calculate the two-sided *P* value for a test of the null hypothesis that shots to the right and left side of the goal are equally likely. Do not use a normal approximation.
   b. Now calculate the two-sided *P* value using a normal approximation.
   c. Overall, 110 of 152 shots were successful. Using $p = 0.5$ when estimating the standard deviation, calculate a 95 percent confidence interval for the probability that a penalty kick is successful.

12.46 In 1977, the U.S. Supreme Court (Castaneda v. Partida) observed that persons with Spanish surnames constituted 79 percent of the population of Hidalgo County, Texas, but only 339 of 870 recent jurors in this county. The Court calculated a *Z* value of 29 and commented that "as a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the jury drawing was random would be suspect."
   a. What parameter value constitutes the null hypothesis?
   b. Show how the Court calculated this *Z* value.
   c. Explain the reference to "large samples."
   d. Explain the reference to "two or three standard deviations."

12.47 Young children who play ice hockey are separated by age. In 1991, for example, children born in 1984 were placed in the 7-year-old league and children born in 1983 were placed in the 8-year-old league. A student with a December 11 birthday observed that someone born in January 1984 is eleven months older than someone born in December 1984. Because coaches give more attention and playing time to better players, this student suspected that children with early birth dates have an advantage when they are young that might cumulate over the years. To test this theory, he looked at the birth dates of 1,487 National Hockey League (NHL) players in 1991 and found that 934 of these players had birth dates during the first six months of the year. Assuming these players to be a random sample from the population of NHL players for all years, use a normal approximation to calculate the *Z* value and two-sided *P* value for a test of the null hypothesis that there is a 0.5 probability that an NHL player has a birth date in the first six months of the year. Is this null hypothesis rejected at the 5 percent level?

12.48 In Hazelwood School District v. the United States (1977), the federal government argued that this district in a rural area in St. Louis County had followed a "pattern or practice" of employment discrimination as evidenced by the fact that only 15 of 405 (3.7 percent) of the Hazelwood teachers hired during the preceding two years were black, while 15.4 percent of the teachers in St. Louis County, including the City of St. Louis, were black. The school district argued that the 3.7 percent figure in Hazelwood should be compared either with data showing that only 2 percent of the Hazelwood students were black, or with the fact that blacks constituted 5.7 percent of the teachers in St. Louis County, excluding the City of St. Louis (which was attempting to maintain a 50-percent-black

teaching staff). The Supreme Court ruled that student body was irrelevant, but instructed the lower courts to gather more information in order to determine the appropriate labor market. Use a normal approximation to calculate the Z values and two-sided P values for these two different values of p (0.154 and 0.057) under the null hypothesis that, with regard to race, hiring is a binomial process with a success probability p.

12.49 In December of 1984, *Sports Illustrated* reported that, "Mindy, 6, a dolphin at the Minnesota Zoo, predicts the outcomes of NFL games by choosing among pieces of Plexiglas, each bearing a different team's name, that are dropped into the pool. This season, Mindy has a 32-21 record." Is this record sufficient to reject at the 5 percent level the null hypothesis that Mindy has a 0.5 probability of picking a winner? Use a two-tailed test and calculate the exact P value. Explain why this report might be considered data grubbing.

12.50 A 1991 experiment at the emergency room at the University of California at San Diego Medical Center compared the accuracy of doctors and a computer program in quickly diagnosing patients who complained of chest pains and may have experienced a heart attack. The computer program used the same patient information available to the doctors —electrocardiogram test results, other physical symptoms, and health history—and used a data base that weighed the importance of this information. The accuracy of the diagnoses made by the doctors and the computer program was determined by the results of blood tests of enzyme levels. Because these enzyme tests take a few hours to process, an early diagnoses can be the basis for the application of life-saving procedures.

In this experiment, 36 of the 331 patients who complained of chest pains had actually experienced a heart attack. The computer program made the correct diagnoses for 35 of these 36 heart-attack victims, while the doctors were right 28 of 36 times. What additional data do we need before concluding that the computer program outperformed the doctors?

Chapter 13

Two-Sample Tests

Chapter Outline

Nothing is good or bad but by comparison.

—Thomas Fuller

In the 1890s, two newspaper reporters, Lincoln Steffens and Jacob Riis, created the illusion of major crime wave in New York by writing increasingly sensational stories that were intended to boost circulation. Even though there had been no increase in criminal activity, frightened readers demanded that the city leaders do something about the crime problem. Theodore Roosevelt, then president of the city's Board of Police Commissioners, ended the fictitious crime wave simply by asking Steffens and Riis to write about something else.

Reporters everywhere tend to focus on bad news—robberies, bankruptcies, and wars. If someone has a pleasant, carefree day, this is not a newsworthy event that will sell newspapers or attract television viewers. Evidently, people do not want to open the newspaper in the morning or turn on the television in the evening and see stories about people who had nice days.

How might we move beyond this anecdotal evidence and investigate whether when there are two stories of equal importance, one good and one bad, the news media really does devote more coverage to the bad?

Chapter 7 in the textbook explains how a random sample can be used to test a null hypothesis about a specified value of the population mean μ. Often, instead of a specific value for μ, we are interested in comparing two populations: whether nonsmokers live longer than cigarette smokers; whether children who watch television frequently are more violent than children who seldom watch television; whether people who take large doses of vitamin C have fewer colds than those who take less vitamin C. Educators compare the performance of students taught in different ways; engineers compare the reliability of different designs; pollsters compare the opinions of different demographic groups.

In this chapter, we extend hypothesis testing to encompass comparisons of two population means and comparisons of two success probabilities.

13.1 The Difference Between Two Means

Exercise 7.6 in the textbook compared the mean of 50 body-temperature readings to a benchmark of 98.6, the widely known figure set forth by Carl Wunderlich in the 1860s. Often, however, we have no benchmark, because we want to compare two groups and do not know the population mean for either. We may want to compare the durability of two kinds of thermometers. Or we may want to compare the effectiveness of two medications for dissolving blood clots in the arteries of heart-attack victims. When we compare two sample means to each other, rather than comparing one sample mean to an assumed population mean, we must modify the details of our test, but not the logic.

A General Framework

To illustrate the general methodology, consider the question posed at the beginning of this chapter. How might we investigate whether the news media give equal coverage to good news and bad news? The challenge is to find good and bad news of comparable importance. One economist had the ingenious idea of looking at television reporting of changes in the unemployment rate. Every month, the federal government releases its latest estimate of the unemployment rate and this figure is invariably reported on television news programs. A decrease in the unemployment rate is just as important as an increase. But if the media are preoccupied with bad news, newscasts may spend more time on this story when the unemployment rate increases than when it decreases.

Writing in 1985, he examined data for 1973 (the earliest available) through 1984 on the amount of time that network newscasts devoted to reporting increases or decreases in the unemployment rate. As shown in Table 13.1, this happened to be a period in which the number of increases in the unemployment rate was almost exactly equal to the number of decreases. The average amount of time devoted to bad news (increasing unemployment) was indeed larger than the average time devoted to good news (declining unemployment), 161.8 seconds versus 123.6 seconds. The statistical question is whether this observed difference in the sample means can be explained by the large standard deviations and limited number of observations. If not, then these data provide persuasive evidence of the tendency of network news programs to spend different amounts of time on good and bad economic news.

Table 13.1 Television Network Reporting of Changes in the Unemployment Rate

	Increase in Unemployment	Decrease in Unemployment
Number of observations	171	170
Average news time (seconds)	161.8	123.6
Standard deviation	110.8	103.9

To answer this question statistically, we need to assume that the amount of time that a network news program spends reporting an increase in the unemployment rate can be described by a probability distribution with a population mean μ_1 and standard deviation σ_1, from which these 171 observations are a random sample, and that the amount of time spent reporting a

decrease in the unemployment rate can be described by a probability distribution with a population mean μ_2 and standard deviation σ_2, from which these 170 observations are a random sample. (The sampling assumption might be justified by arguing that these newscasts are a sample from a much larger population of newscasts.)

Our null hypothesis is that the two population means are equal, so that any observed difference in the sample means is due to sampling error:

$$H_0: \quad \mu_1 = \mu_2$$
$$H_1: \quad \mu_1 \neq \mu_2$$

The alternative hypothesis is two-sided because, before looking at the data, we cannot rule out the possibility that network news programs spend more time on good news than bad, or vice versa. It is useful to rewrite these hypotheses in terms of the difference between the population means:

$$H_0: \quad \mu_1 - \mu_2 = 0$$
$$H_1: \quad \mu_1 - \mu_2 \neq 0$$

After specifying the null and alternative hypotheses, we need to choose an appropriate estimator to use as a test statistic. It is natural to use the difference between the two sample means to estimate the size of the difference between the population means. If the underlying populations are normal (or if both samples are sufficiently large), the difference in the sample means is normally distributed with a population mean equal to the difference in the population means and a standard deviation equal to the square root of the *sum* of the population variances:

$$\text{Mean of } \bar{X}_1 - \bar{X}_2 = \mu_1 - \mu_2$$
$$\text{Standard deviation of } \bar{X}_1 - \bar{X}_2 = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

We can summarize these relations in this shorthand notation:

$$\bar{X}_1 - \bar{X}_2 \sim N\left[\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] \tag{13.1}$$

Chapter 7 in the textbook showed how to convert a normally distributed test statistic into a standardized *Z* statistic:

$$Z = \frac{\text{observed statistic} - \text{null hypothesis parameter value}}{\text{standard deviation of statistic}}$$

Here, the observed statistic is the difference between the two sample means; the population value under the null hypothesis is the specified value of $\mu_1 - \mu_2$ (which is 0); and the standard deviation of the statistic is shown in Equation 13.1. Thus, our *Z* statistic is

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - 0}{\text{standard deviation of } \left(\bar{X}_1 - \bar{X}_2\right)}$$
$$= \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{13.2}$$

We can also determine a confidence interval for the difference in the population means by using the probability distribution of the difference in the sample means given in Equation 13.1. However, the Z statistic and confidence interval in Equations 13.2 and 13.3 require values for the population standard deviations, σ_1 and σ_2. We seldom know these values, but the next two sections explain how the sample standard deviations can be used in their place.

Unequal Standard Deviations

Our first option is to use the sample standard deviations to estimate the population standard deviations, and then use a t distribution in place of the normal distribution to determine the P value. The second option, discussed later, is to assume that both samples are from distributions with the *same* standard deviation σ.

If we use the standard deviations from each sample, s_1 and s_2, the Z statistic in Equation 13.2 is replaced by this t statistic:

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - 0}{\text{standard error of } \left(\bar{X}_1 - \bar{X}_2\right)} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{13.3}$$

Although this statistic does not have an exact t distribution, a t distribution generally gives an accurate approximation, using either of two methods for calculating the degrees of freedom. The recommended approach is a value (usually not a whole number) estimated from the data, which will be at least as large as the smaller of $n_1 - 1$ and $n_2 - 1$ and never larger than $n_1 + n_2 - 2$. The calculation is quite complicated and should be done by statistical software. If you don't have such software, a more conservative, approach is to use the smaller of $n_1 - 1$ and $n_2 - 1$.

Because of the power of the central limit theorem, the t distribution is an excellent approximation if each sample has at least 5 observations from distributions that are generally similar and not extremely asymmetric, or at least 20 observations from dissimilar or very asymmetric distributions.

For a confidence interval, the appropriate t value t^* is determined by using statistical software that will estimate the degrees of freedom, or, in the absence of such software, letting the degrees of freedom equal the smaller of $n_1 - 1$ and $n_2 - 1$:

$$\begin{aligned}\text{Confidence interval for } \left(\mu_1 - \mu_2\right) &= \bar{X}_1 - \bar{X}_2 \pm t^* \left(\text{standard error of } \left(\bar{X}_1 - \bar{X}_2\right)\right) \\ &= \bar{X}_1 - \bar{X}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\end{aligned} \tag{13.4}$$

We will first use the study of reported changes in the unemployment rate to illustrate these calculations. Substituting the appropriate data from Table 13.1 into Equation 13.3, the t value is

$$t = \frac{161.8 - 123.6}{\sqrt{\frac{110.8^2}{171} + \frac{103.9^2}{170}}} = 3.284$$

Statistical software shows the estimated degrees of freedom to be 337.9, and the corresponding two-sided *P* value to be 0.0011. (If we use 170 – 1 = 169 as a conservative estimate of the degrees of freedom, the two-sided *P* value is 0.0012.)

If these data are random samples from populations with the same mean, there is only roughly a one-in-a-thousand chance of observing such a large disparity in the average time devoted to good and bad news about the unemployment rate. This researcher concluded that there is a statistically significant difference in the television reporting of increases and decreases in the unemployment rate, a difference in the direction that he had suspected beforehand: bad news is emphasized and good news is slighted. The difference is substantial. Thirty-eight seconds is a lot of time on a television news program, and represents 30 percent more time spent on bad news than on good.

For a 95 percent confidence interval with 337.8 degrees of freedom, the appropriate *t* value is approximately 1.97 and our confidence interval is

$$\text{Confidence interval for } (\mu_1 - \mu_2) = \bar{X}_1 - \bar{X}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= 161.8 - 123.6 \pm 1.97 \sqrt{\frac{110.8^2}{171} + \frac{103.9^2}{170}}$$

$$= 38.2 \pm 22.9$$

Notice that, consistent with rejection of the null hypothesis $\mu_1 = \mu_2$ at the 5 percent level, 0 is not inside this 95 percent confidence interval.

For an example with much smaller samples, consider an experimental test of a possible cure for cancer. In this laboratory test, cancerous tumors were implanted in nine mice, three of which were given a promising anticancer drug. Each tumor was later removed and weighed to see if the drug had slowed its growth. Here are the tumor weights (in grams):

	Control	Treatment
	1.29	0.96
	1.60	1.59
	2.27	1.14
	1.31	
	1.88	
	2.21	
Mean	1.76	1.23
Standard deviation	0.4303	0.3245

Using Equation 13.4, here is the *t* value for testing the null hypothesis that these sample data are from populations with the same means:

$$t = \frac{1.76 - 1.23}{\sqrt{\frac{0.4303^2}{6} + \frac{0.3245^2}{3}}}$$

$$= 2.064$$

Statistical software estimates the degrees of freedom as 5.4 and the two-sided *P* value as 0.088.

These samples are too small and the standard deviations too large to reject decisively the possibility that the observed difference in mean tumor weights may be due to sampling error.

On the other hand, if we think of *P* values as a continuum, 0.088 is not much larger than 0.05, and while the drug didn't stop the growth of these cancerous tumors, it did slow their growth substantially. The samples were very small, but the experiment had promising results in the neighborhood of statistical significance at the 5 percent level. Additional tests can provide supplementary data that will replicate and reinforce these results or show them to be a fluke.

Equal Standard Deviations

In contrast to these approximate probability calculations, an exact *t* distribution can be used if we are willing to assume that both random samples are from normal distributions with the *same* standard deviation σ. In this case, we use the *pooled variance* s_P^2 as an estimator of the common variance σ^2:

$$s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 1} \tag{13.5}$$

The pooled variance lies between the two sample variances, depending on the relative size of the samples. If the samples are the same size, the pooled variance is exactly halfway between the two sample variances. Otherwise, it is closer to the variance of the larger sample.

We replace the separate estimates of the two sample standard deviations with the common, pooled estimate to obtain this *t* statistic with $n_1 + n_2 - 2$ degrees of freedom:

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{\sqrt{\frac{s_P^2}{n_1} + \frac{s_P^2}{n_2}}} \tag{13.6}$$

For a confidence interval, we chose a confidence level, such as 95 or 99 percent, and determine the *t* value *t** that corresponds to this probability:

$$\text{Confidence interval for } \left(\mu_1 - \mu_2\right) = \bar{X}_1 - \bar{X}_2 \pm t^* \sqrt{\frac{s_P^2}{n_1} + \frac{s_P^2}{n_2}}$$

Returning to our network news example, the pooled variance is

$$s_P^2 = \frac{(171-1)110.8^2 + (170-1)103.9^2}{171+170-1} = 11{,}538.11$$

Because of the nearly equal sample sizes, the *t* value is equal to the value obtained earlier:

$$t = \frac{161.8 - 123.6}{\sqrt{\frac{11{,}538.11}{171} + \frac{11{,}538.11}{170}}} = 3.284$$

With 171 + 170 – 2 = 339 degrees of freedom, statistical software shows the two-sided *P* value to be 0.0011, as before.

For the laboratory test of the potential anti-cancer drug, the pooled sample variance is

$$s_P^2 = \frac{(6-1)0.4303^2 + (3-1)0.3245^2}{6+3-2}$$
$$= 0.162$$

and the t value is

$$t = \frac{1.76 - 1.23}{\sqrt{\frac{0.162}{6} + \frac{0.162}{3}}} = 1.860$$

The two-sided P value is 0.103.

A comparison of this example with the previous unemployment news example illustrates the general rule that the t statistics in Equations 13.3 and 13.6 are equal when the samples are the same size and will be close if the samples are approximately the same size.

A Victim of Incorrect Statistical Advice

A small data-processing firm tried to persuade a large grocery chain that it could reduce the chain's cost of handling workers' compensation invoices. This small firm and the firm that had been handling the invoices based their charges on complex formulas, depending on the number of lines in the doctors' bills and other factors. The only way to tell which firm was less expensive was to compare their charges for actual invoices. As a test, the current firm processed a random sample of 3,800 bills and charged an average of $9.26 per bill, with a standard deviation of $9.70; the new firm processed a random sample of 3,500 bills and charged an average of $7.35, with a standard deviation of $5.26. A consultant advised the grocery chain's management that it should stick with its current data-processing firm, because the large standard deviations meant that the observed difference could easily be explained by sampling error. Do you agree?

The consultant evidently compared the observed $1.91 difference in the sample means to the $9.70 and $5.26 sample standard deviations, forgetting that the standard deviation of a sample mean is equal to the sample standard deviation divided by the square root of the sample size. Equation 13.3 shows that, because these are very large samples, the standard error of the difference in the sample means is a mere 18 cents:

$$\text{standard error of } \left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$
$$= \sqrt{\frac{9.70^2}{3800} + \frac{5.26^2}{3500}}$$
$$= 0.18$$

The $1.91 difference in the sample means is more than ten standard deviations from 0, and provides overwhelming evidence that the new firm is, on average, less expensive.

More formally, the null hypothesis is that the handling costs come from populations with the same mean: H_0: $\mu_1 = \mu_2$. Using Equation 13.3, the t value for testing the null hypothesis is

$$t = \frac{9.26 - 7.35}{\sqrt{\frac{9.70^2}{3800} + \frac{5.26^2}{3500}}} = 10.6$$

The *P* value is minuscule, providing decisive evidence against the null hypothesis that the population means are equal. This $1.91 difference is also substantial, representing more than a 20 percent saving for the grocery chain, a potential saving that was lost because the chain's management was given incorrect statistical advice.

Matched-Pair Samples

A possible sample design for a company that wants to test consumer reaction to two colas is to select two random samples and have the people in the first sample taste and rate one cola, while the people in the second sample taste and rate the other cola. Perhaps the ratings turn out as shown in Table 13.3, with the first cola rated higher among those surveyed. A statistical test needs to take into account the possible sampling error when we choose two random samples from a population with diverse cola preferences. The second cola may have received lower ratings because, by the luck of the draw, it was tasted by a disproportionate number of people who who don't like cola or like to give low ratings.

Table 13.3 Cola Ratings for Two Independent Samples

	First Sample's Rating of First Cola	Second Sample's Rating of Second Cola
	8	6
	2	3
	6	7
	10	7
	9	6
	2	1
	8	6
	10	8
	1	2
	6	4
Mean	6.20	5.00
Standard deviation	3.4254	2.3570

To gauge this potential sampling error, our statistical test takes into account the size of the samples and the standard deviations of the ratings. Allowing for the population standard deviations to be unequal and using Equation 13.3, the *t* value is

$$t = \frac{6.2 - 5.0}{\sqrt{\frac{3.4254^2}{10} + \frac{2.3570^2}{10}}} = 0.913$$

Statistical software shows that this statistic has 16.0 degrees of freedom and the two-sided *P* value is 0.375. The standard deviations are too large and the samples too small to rule out

sampling error as an explanation for the observed difference in the average ratings.

One way to reduce the sampling error inherent in the selection of two independent random samples is to use a single random sample, having everyone taste and rate both colas. What if the data in Table 13.3 were a comparison sample, with each pair of observations representing two ratings by a single person? Table 13.4 shows the revised statistical calculations.

Table 13.4 Cola Ratings by a Single Sample

Person	Rating of First Cola	Rating of Second Cola	Difference
1	8	6	2
2	2	3	–1
3	6	7	–1
4	10	7	3
5	9	6	3
6	2	1	1
7	8	6	2
8	10	8	2
9	1	2	–1
10	6	4	2
Mean			1.20
Standard deviation			1.6193

Each difference in an individual's ratings of these two colas can be treated as a single observation X. If the underlying ratings are normally distributed or if the sample is large, then the sample mean of X is approximately normally distributed and we can test the null hypothesis that the population mean of X is 0 by using the single-sample t statistic given by Equation 7.3 in Chapter 7 in the textbook:

$$t = \frac{\bar{X} - 0}{s / \sqrt{n}}$$
$$= \frac{1.2 - 0}{1.6193 / \sqrt{10}}$$
$$= 2.34$$

With 10 – 1 = 9 degrees of freedom, the two-sided P value is 0.042.

The observed 1.2 average difference in the sample ratings is statistically significant at the 5 percent level if the data are from a single comparison sample, but not if they are from two independent samples.

The possibility that one sample may include harsher judges implies that the relative merits of the two colas may be confounded with the effects of harsher judging. By using a single sample, we control for this confounding effect. If we select a person who gives relatively low ratings, this person will be included in both samples, and we can focus our attention on which cola this person prefers, without being concerned with the level of the person's ratings.

A single, comparison sample can be used not only in taste tests, but in any situation where a direct comparison is appropriate; for instance, the performance of ten machines doing two different tasks, the speed of ten runners on two different track surfaces, or the accuracy of two forecasters predicting ten different events. Unfortunately, a comparison test using a single sample is often unfair or impractical. In a taste test, the first item may leave a lingering taste that distorts the rating of the second item. In medical experiments, we cannot compare two drugs by giving both to the same patient. We can't compare two textbooks or two teaching styles by making students take the same course twice. In comparing the opinions of males and females, whites or blacks, Republicans or Democrats, two samples are required by the very nature of the poll.

If a comparison sample is inappropriate, the product tester, medical researcher, or pollster can attempt to assemble two samples that contain matched pairs that are virtually identical with respect to influences that might confound the results of two independent samples. If truly identical twins can be found, it is as if a single person is tested twice, and the difference in each matched-pair result can be treated as a single observation.

To illustrate the procedure, consider a test to see if a cholesterol-reducing drug affects the risk of heart attack. Two independent samples might be inconclusive because the effects of the drug are confounded with other factors—including age, gender, smoking habits, and blood pressure. In small samples, there is a non negligible chance, for example, that the drug will be given mostly to nonsmokers, while the control group consists mostly of smokers. A lower incidence of heart problems might be attributed to the drug, when it is at least partly due to the differences in smoking habits. Statistically persuasive results will consequently require large samples, which reduce the probability of the results being tainted by this kind of sampling error—it is unlikely that a large, randomly selected sample will consist mostly of smokers, while another consists mostly of nonsmokers. A heart-attack study needs large samples to begin with, because relatively few people suffer heart attacks. If we add to this burden the need to guard against confounding effects, our samples will have to be enormous to be statistically persuasive.

A single-sample comparison would protect against the confounding effects, but we can't put the same person in both the treatment group and the control group. The next best thing is matched pairs, giving the drug to one person while using a very similar person as a control. This is, in fact, just what researchers at the National Heart, Lung, and Blood Institute did. A sample of 3,806 middle-aged males who had high cholesterol levels but no history of heart disease were divided into two samples, with each person in the treatment group matched with someone in the control group according to age, cholesterol level, smoking habits, blood pressure, and other factors that are believed to affect one's chances of having a heart attack. Each matched pair served as a virtual twin, as if a single person were simultaneously in the treatment and control groups.

The patients were selected by doctors at twelve medical centers throughout the United States. Those in the treatment group were given a cholesterol-reducing drug over a ten-year period, while those in the control group were given a placebo. Neither the patients nor their doctors knew who was in the control group and who was in the treatment group. At the end of this ten-year period, the researchers found that those in the treatment group had 19 percent fewer heart attacks, 20 percent fewer chest-pain attacks, and 21 percent fewer coronary-bypass operations.

The most important advantage of matched-pair sampling is that it allows for a smaller sample by controlling for the influence of confounding factors. The researcher doesn't need to rely on sample size to reduce sampling error. In addition, as a practical matter, it defuses criticism that the samples may have slanted the results. Critics cannot raise doubts by speculating that the medication might have been given mostly to people who don't smoke or have low blood pressure if, in fact, the people taking the medication were similar to those in the control group.

As in this cholesterol study, the accurate identification of matching pairs is a time-consuming process. It requires knowledge of the important factors that need to be controlled and a careful attention to the matching process. Matching is subjective and subject to error. Some factors may be inadvertently neglected, so that the matched pairs aren't really virtual twins. However, when done carefully, matched-pair samples offer an attractive sample design.

Testing Automobile Emissions Using New Fuels

The Environmental Protection Agency (EPA) tests new automobile fuels and fuel additives to see whether they have a substantial adverse impact on automotive emissions. In a 1981 test of a new fuel called Petrocal, the emission levels of several noxious gases were recorded as sixteen automobiles were driven with a standard fuel and then driven with Petrocal. The nitrogen oxide emissions were as follows:

Petrocal	Standard	Difference	Petrocal	Standard	Difference
1.385	1.195	0.190	0.875	0.687	0.188
1.230	1.185	0.045	0.541	0.498	0.043
0.755	0.755	0.000	2.186	1.843	0.343
0.775	0.715	0.060	0.809	0.838	–0.029
2.024	1.805	0.219	0.900	0.720	0.180
1.792	1.807	–0.015	0.600	0.580	0.020
2.387	2.207	0.180	0.720	0.630	0.090
0.532	0.301	0.231	1.040	1.440	–0.400

These are matched-pair samples since each car used both fuels. The 16 observed differences have a mean of 0.0841, with a standard deviation of 0.1672. For testing the null hypothesis that there is no difference in the mean emissions in the population, the t value is

$$t = \frac{\bar{X} - 0}{s / \sqrt{n}}$$
$$= \frac{0.0841 - 0}{0.1672 / \sqrt{16}}$$
$$= 2.01$$

With 16 – 1 = 15 degrees of freedom, the probability of such a large t value is 0.030, which is statistically significant at the 5 percent level using a one-tail test, but not with a two-tail test. Relative to the 1.0754 average level of emissions with the standard fuel, the observed 0.0841 difference represents about an 8 percent increase in nitrogen oxide emissions. Nonetheless, the EPA concluded that the environmental impact was small and, because the fuel had "narrowly missed" passing other standards, they approved Petrocal. A lawsuit by the Motor Vehicles Manufacturers Association challenged this conclusion, arguing that the EPA was ignoring its own standards. An appellate court agreed and reversed the EPA's decision.

14.2 The Difference Between Two Success Probabilities

So far in this chapter, we have developed test procedures for the case where we have quantitative data that can be averaged. An analogous procedure can be used to handle categorical data.

To illustrate, we will first analyze data from the 1971–1972 Toronto tests of Linus Pauling's claim that large doses of vitamin C help prevent colds. Eight hundred volunteers were randomly divided into a group that received placebos and a second group that received one gram of vitamin C each day and four grams-a-day if cold symptoms appeared. At the end of the experiment, 72 (18 percent) of those taking the placebo had not had a cold, while 104 (26 percent) of those taking megadoses of vitamin C had been cold-free. Data on the number of colds or the number of sick days would be quantitative data that can be averaged and analyzed by the statistical procedures already developed in this chapter. But information on whether or not a person made it through the winter cold-free is categorical data. The person either had a cold or didn't. We need slightly different statistical procedures to analyze these data.

Chapter 12 ("The Binomial Model") explains how categorical data from a single sample can be used to calculate the sample success proportion x/n, which can then be used to test a hypothesized value of a success probability p. In one example, 16 pregnant women read a children's story aloud during the last six and a half weeks of pregnancy and, after birth, their babies were allowed to choose between a tape recordings of their mother reading the familiar story or an unfamiliar story. In 13 of 16 cases, the babies chose the familiar story. The sample success proportion $x/n = 13/16$ was used to test the null hypothesis that each baby has a 0.5 probability of choosing the familiar story. This is a one-sample test.

Our vitamin C example is a two-sample test. We can let p_1 be the probability that a randomly selected person taking the placebo would make it through this Toronto winter cold-free, and let let p_2 be the probability that a randomly selected person taking megadoses of vitamin C would be cold-free. The natural null hypothesis is that these two probabilities are equal:

$$H_0: \ p_1 = p_2$$
$$H_1: \ p_1 \neq p_2$$

If the data reject this null hypothesis, then we will have shown that megadoses of vitamin C affect the probability of being cold-free. The alternative hypothesis is two-sided if we cannot rule out beforehand the possibility that p_1 is either larger or smaller than p_2. We can rewrite these hypotheses in terms of the difference between the probabilities:

$$H_0: \ p_1 - p_2 = 0$$
$$H_1: \ p_1 - p_2 \neq 0$$

A natural test statistic for the difference in probabilities is the difference between the success proportions in random samples drawn from each population: $x_1/n_1 - x_2/n_2$. This test statistic is unbiased, in that the population mean of the difference in the sample success proportions is equal to the difference in the population probabilities:

$$\text{Mean of } \frac{x_1}{n_1} - \frac{x_2}{n_2} = p_1 - p_2$$

The standard deviation is given by this formula:

$$\text{Standard deviation of } \frac{x_1}{n_1} - \frac{x_2}{n_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Finally, if the samples are reasonably sized, the central limit theorem implies that the difference between the two sample success proportions is approximately normally distributed. Therefore, a shorthand summary is:

$$\frac{x_1}{n_1} - \frac{x_2}{n_2} \sim N\left[p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right] \tag{13.7}$$

Again, we use the following general formula to convert our normally distributed test statistic to a standardized Z statistic:

$$Z = \frac{\text{observed statistic} - \text{null hypothesis parameter value}}{\text{standard deviation of statistic}}$$

Here, the observed statistic is the difference between the two sample success proportions and the null hypothesis is the assumed population value of $p_1 - p_2$ (which is 0). Thus, our Z statistic is

$$Z = \frac{\left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right) - 0}{\text{standard deviation of } \left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)} \tag{13.8}$$

Equation 13.7 shows that the standard deviation depends on the unknown values of the success probabilities, p_1 and p_2. For a test of the null hypothesis that $p_1 = p_2$, the best estimate of these two equal success probabilities is provided by the pooled estimator:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \tag{13.9}$$

Using Equations 13.7 and 13.8, the Z statistic is

$$Z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \tag{13.10}$$

However, for a confidence interval for $p_1 - p_2$, we don't assume that p_1 is equal to p_2 and it is better to use the two separate sample proportions when estimating the standard deviation:

$$\text{Confidence interval for } p_1 - p_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2} \pm z^* \sqrt{\frac{\left(\frac{x_1}{n_1}\right)\left(1 - \frac{x_1}{n_1}\right)}{n_1} + \frac{\left(\frac{x_2}{n_2}\right)\left(1 - \frac{x_2}{n_2}\right)}{n_2}} \tag{13.11}$$

where z^* is the Z value corresponding to our desired confidence level; for example $z^* = 1.96$ for a 95 percent confidence interval.

Let's apply these admittedly intimidating formulas to our Vitamin C example. The pooled

sample success proportion is the total number of people who were cold-free divided by the total number of people participating in the study:

$$\begin{aligned}\hat{p} &= \frac{x_1 + x_2}{n_1 + n_2} \\ &= \frac{72 + 104}{400 + 400} \\ &= \frac{176}{800} \\ &= 0.22\end{aligned}$$

Using Equation 13.10, the Z statistic is

$$\begin{aligned}Z &= \frac{\frac{72}{400} - \frac{104}{400}}{\sqrt{\frac{0.22(1-0.22)}{400} + \frac{0.22(1-0.22)}{400}}} \\ &= -2.73\end{aligned}$$

which has a two-sided P value of 2(0.0032) = 0.0064. The Toronto study showed a statistically significant effect at the 1 percent level.

We can use Equation 13.11 to calculate a 95 percent confidence interval for the difference in the population probabilities of not getting a cold:

$$\frac{72}{400} - \frac{104}{400} \pm z^* \sqrt{\frac{\left(\frac{72}{400}\right)\left(1 - \frac{72}{400}\right)}{400} + \frac{\left(\frac{104}{400}\right)\left(1 - \frac{104}{400}\right)}{400}} = -0.080 \pm 0.057$$

Whether the 0.08 difference between 0.26 and 0.18 is substantial (something to sneeze about?) is a matter of opinion. For some, the cost of vitamin C (and possible side effects) is not worth the modest improvement in the chances of being cold-free. For others, it is worth the cost.

Do Headstart Programs Help?

In the early 1960s, 123 black children of ages 3 to 4 from low-income families in Ypsilanti, Michigan, participated a study of the effectiveness of preschool programs for disadvantaged children. Coin flips were used to assign the children randomly to either the experimental group (which received intensive high-quality preschooling) or the control group (which received no preschooling). Because siblings were kept together and coin flips are random, the two groups were slightly different sizes (58 in the experimental group and 65 in the control group). Here are some data collected when these youngsters were 19 years of age (two members of the control group could not be located).

	Preschool (n = 58)	No Preschool (n = 63)
Completed high school	39	31
Currently on welfare	10	20
Arrested at some time	18	32
Currently employed	29	20

To evaluate whether these observed differences are substantial and statistically significant, we can calculate the success proportions and then test the null hypothesis that an intensive preschool program does not affect the probability of each outcome described here.

For example, a fraction 39/58 = 0.67 of the preschoolers completed high school, as compared to 31/63 = 0.49 of the control group. The difference between a 67-percent and 49-percent graduation rate is substantial. The pooled sample success proportion is

$$\begin{aligned}\hat{p} &= \frac{39+31}{58+63} \\ &= \frac{70}{121} \\ &= 0.579\end{aligned}$$

For testing the null hypothesis that the probability of completing high school is not affected by participation in an intensive preschool program, the Z statistic is

$$\begin{aligned}Z &= \frac{\frac{39}{58}-\frac{31}{63}}{\sqrt{\frac{0.579(1-0.579)}{58}+\frac{0.579(1-0.579)}{63}}} \\ &= 2.01\end{aligned}$$

which is barely larger than the 1.96 cutoff for statistical significance at the 5 percent level. Using Appendix Table A.1 in the textbook, we see that the two-sided P value is 2(0.0222) = 0.0444.

Table 13.5 shows the sample success proportions, Z values, and two-sided P values for all four outcomes. In each case, the outcome proportions are distressing for both groups, but substantially less so for those who participated in the intensive preschool program. All results are in the neighborhood of statistical significance at the 5 percent level, with three of the four two-sided P values a little less than 0.05 and one a little above. The cumulative force of the four tests strengthens the statistical case that this program has long-term beneficial effects.

One of the reasons that intensive preschool programs are supported by politicians from all parts of the spectrum is that these expenditures are an investment with long-term financial benefits. This study estimated that a $4,000 preschool investment would return the government $29,000 in reduced welfare payments, diminished jail expenses, and increased tax revenue.

Table 13.5 The Ypsilanti Data

	Sample Success Proportion				
	Preschool	No Preschool	Overall	Z value	Two-Sided P
Completed high school	0.67	0.49	0.59	2.01	0.044
Currently on welfare	0.18	0.32	0.25	–1.85	0.065
Arrested at some time	0.31	0.51	0.41	–2.21	0.027
Currently employed	0.50	0.32	0.41	2.04	0.041

Job Discrimination: Statistical Significance and the 80-Percent Rule

Statistical tests have frequently been accepted by U.S. courts as evidence of employment discrimination that violates Title VII of the Civil Rights Act of 1964. In Griggs v. Duke Power Company, 1971, the Supreme Court ruled that it is illegal to use a job qualification test that has a "disparate impact" on minority applicants unless the employer can show a business necessity for the test (for example, an eye examination for pilots). In Hazelwood School District v. United States, 1977, the Supreme Court concluded that "gross statistical disparities" can be sufficient to prove discrimination. However, courts have also recognized that a statistically significant difference may not be substantial. For instance, in Moore v. Southwestern Bell, 1979, a promotion test for clerks was passed by 453 of 469 whites (96.6 percent) and by 248 of 277 blacks (89.5 percent). Using the pooled success proportion (94.0 percent), the Z value is 3.91, which has a two-sided P value of 0.000092:

$$Z = \frac{0.966 - 0.895}{\sqrt{\frac{0.940(1-0.940)}{469} + \frac{0.940(1-0.940)}{277}}} = 3.91$$

The district court ruled that although the difference between the 96.6 percent and 89.5 percent pass rates was statistically significant, it was not substantial enough to prove discrimination.

The U.S. Equal Employment Opportunity Commission (EEOC) has argued, and the Supreme Court has endorsed (in Connecticut v. Teal, 1982), the use of an "80-percent rule" to supplement tests of statistical significance in evaluating evidence of employment discrimination against members of a specified ethnic, racial, or sexual group. The 80-percent rule considers whether the selection rate for this group is less than 80 percent of the selection rate for the most favored group.

For instance, in the case of Moore v. Southwestern Bell cited above, the ratio of the selection rate for blacks to the selection rate for whites is 0.895/0.966 = 0.927. Although the difference is statistically significant, it is not, by the 80-percent rule, substantial. The opposite situation occurred in Chicano Police Officers Association v. Stover, 1975, where 3 of 26 Spanish surnamed officers passed a promotion test, a selection rate of 3/26 = 0.115, while 14 of 64 other officers passed, a selection rate of 14/64 = 0.219. The ratio 0.115/0.219 = 0.53 is less than 0.80, indicating a substantial difference in selection rates. However, because the samples are so small, this difference is not statistically significant at the 5 percent level. Using the pooled-sample success proportion (3 + 14)/(26 + 64) = 0.189, the Z value is 1.14 and the two-sided P value is 0.256:

$$Z = \frac{0.219 - 0.115}{\sqrt{\frac{0.189(1-0.189)}{64} + \frac{0.189(1-0.189)}{26}}} = 1.14$$

Courts have found statistical evidence of discrimination to be the most persuasive when the observed differences are both statistically significant and substantial.

Exercises

13.1 U.S. soldiers in World War II were taller, heavier, and stronger than those in World War I. To see if the same was true of U.S. women during theis period, a statistics class at Mount Holyoke College analyzed the data below. Although these first-year students were not randomly selected, we can consider them representative of students in similar years at similar colleges. Do these data show substantial and statistically significant changes at the 1 percent level in the mean heights, weights, and grips of 1918 and 1943 female students? Assuming these to be random samples from populations with possibly unequal standard deviations, calculate the *t* value, two-sided *P* value, and 99 percent confidence interval for each of these characteristics.

	1918 (250 students)		1943 (308 students)	
	Mean	Standard Deviation	Mean	Standard Deviation
Height (cm)	161.7	6.3	164.8	5.9
Weight (lb)	118.4	16.6	128.0	17.2
Grip (kg)	30.9	4.6	32.3	5.4

13.2 In the 1800s, sailors near the coast of Peru labeled the unusual currents and warm sea temperatures that sometimes appear around Christmas El Niño, the (Christ) boy. El Niño conditions have been linked to flooding in Central and South America, mild winters in Alaska, Canada, and the Northern United States, and heavy rainfall in the Southwestern United States. A researcher examined annual Los Angeles rainfall (in inches) during 61 rain years (July 1 to June 30) from 1929–1989. Here are the data in the ten years identified by meteorologists and climatologists as moderate or extreme El Niño years:

Rain Year	Precipitation	Rain Year	Precipitation
1931–1932	16.95	1969–1970	7.77
1940–1941	32.76	1972–1973	17.45
1953–1954	11.99	1976–1977	14.97
1957–1958	21.13	1982–1983	32.19
1965–1966	20.44	1986–1987	9.11

In the other 51 years, Los Angeles precipitation averaged 13.78 inches with a standard deviation of 6.23. Calculate the two-sided *P* value for a difference-in-means test comparing precipitation in El Niño years with other years, assuming these data to be random samples from populations with unequal standard deviations. Are these really random samples?

13.3 A cold vaccine was tested with 548 college students who believed themselves especially susceptible to colds. Half were given the experimental vaccine and half, the control group, were given plain water. The test results on the number of colds per person are shown below. Assuming possibly unequal standard deviations, calculate the *t* value and two-sided *P* value for a difference-in-means test. Explain why the researchers concluded that "the effect is too small to be of practical importance."

	Vaccine ($n = 272$)	Control Group ($n = 276$)
Mean	1.6	2.1
Standard deviation	0.8	1.0

13.4 Telephone calls were made to 20 randomly selected college students, who were told that they were going to be surveyed about environmental issues. Students who agreed to participate were asked to, "Hold, please." If the student agreed to hold, the researcher timed the number of seconds that elapsed before the student eventually hung up:

Males	32	66	58	136	9	83	247	11	47	93
Females	258	91	132	51	183	86	49	106	319	213

Use two box plots to compare these data. Assuming the population standard deviations to be equal, is the observed difference between male and female holding times statistically significant at the 5 percent level?

13.5 A Stanford University team examined changes in the triglyceride level in 30 blood samples that had been in frozen storage for eight months. Use a matched-pair test to see whether the observed differences are statistically significant at the 5 percent level.

Before	After	Before	After	Before	After	Before	After	Before	After
74	66	177	185	126	133	83	81	131	127
80	85	88	96	72	69	79	74	228	227
75	71	85	76	301	302	194	192	115	129
136	132	267	273	99	106	124	129	83	81
104	103	71	73	97	94	42	48	211	212
102	103	174	172	71	67	145	148	169	182

13.6 In 1973, the U.S. Public Health Service studied the effects of vitamin C on the incidence of colds at a Navajo boarding school in Arizona during a 14-week period, from February to May. Daily doses of from 1 to 2 grams of vitamin C were given to 321 children, while 320 children received placebos that were identical in taste and appearance to the vitamin C tablets. They found that 143 of those who took vitamin C and 92 of those taking placebos had no sick days during this 14-week period. Is this difference substantial and statistically significant at the 1 percent level?

13.7 In 1983, 22,000 male doctors volunteered to participate in an experiment testing the effects of aspirin on the incidence of heart attacks. Half the doctors were given a single aspirin tablet every other day; the other half took a placebo. Neither the volunteers or the researchers knew which volunteers were taking aspirin. After five years of what was intended to be a seven-year test, the experiments were stopped by an outside committee monitoring the results. During this period, fatal heart attacks were suffered by 18 of those taking placebos and by 5 of those taking aspirin. Nonfatal heart attacks were experienced by 171 of those taking placebos and by 99 of those taking aspirin. Use the Z value and the two-sided P value to determine if the difference in fatal heart attacks is statistically significant at the 1 percent level. What about the difference in nonfatal heart attacks? Why do you suppose that outside committee stopped the experiment two years early?

13.8 A 1983 federal court case (Elison v. City of Knoxville) concerned a candidate at the Knoxville Police Academy who claimed that the physical qualification tests used by the academy were discriminatory because they had a disparate impact on the pass rates of males and females. In her class, 6 of 9 females passed and 34 of 37 males passed. The judge accepted Knoxville's argument that the court should consider data for all classes,

not just the plaintiff's class. For all classes, 16 of 19 females passed and 64 of 67 males passed. Using each set of data, is the 80-percent rule satisfied and is the difference in pass rates statistically significant at the 5 percent level?

13.9 In a penalty-kick tournament for ten right-footed Division III college soccer players, a total of 152 penalty kicks were taken, of which 88 were shot to the right side of the goal (from the kicker's perspective) and 64 were shot to the left side. Of 88 shots to the right, 59 were successful; of the 64 shots to the left, 51 were successful. Use these data to calculate the Z value and two-sided P value for testing the null hypothesis that a shot to the left or the right is equally likely to be successful.

13.10 A 1980 poll asked women, "If you had enough money to live as comfortably as you would like, would you prefer to work full time, work part time, do volunteer work, or work at home caring for the family?" Overall, 39 percent wanted to work at home. Among women working outside the home, 28 percent wanted to work at home. Assume 1,000 women were surveyed, 500 of whom work outside the home. Use a two-sided P value to test the null hypothesis that a preference for working outside the home is equally prevalent among women who do and do not work outside the home.

13.11 A 1986 newspaper headline read "Barnstable Auto Fatalities Drop 600 Percent," referring to the fact that there had been 12 fatalities in 1980 and 2 in 1985. How large a percentage drop would you report? The accompanying article also noted that, "Of particular significance is that last year neither of the two fatalities involved alcohol, [the selectman] said. That compares with alcohol being a factor in three of the 12 deaths in 1980." The selectman's daughter died in an alcohol-related auto accident in 1982 and there was subsequently a community-wide campaign to reduce drunken driving. The article used these data to show the effects of this campaign:

	Before Campaign			After Campaign		
	1980	1981	1982	1983	1984	1985
Fatalities	12	12	6	4	5	2
Alcohol-Related	3	4	2	0	1	0

Assuming the population standard deviations to be equal, is the observed difference between the average annual number of fatalities in 1980–1982 and 1983–1985 statistically significant at the 5 percent level? What about the difference in alcohol-related fatalities?

13.12 One hundred randomly selected students were asked to rate the qualifications of two job candidates. Fifty were given a resume for a mythical John Benson and fifty were given another resume, identical in all respects except that the name was Mary Benson. Assuming the population standard deviations to be equal, use a two-sided P value to determine if the difference in the average ratings is statistically significant at the 5 percent level.

	John Benson	Mary Benson
Average rating	8.23	8.02
Standard deviation	1.15	2.18

13.13 To see whether persons with glaucoma have abnormally thick corneas, the corneas were measured in eight persons with glaucoma in one eye but not in the other. Here are the thicknesses, in microns:

Glaucomatous eye	488	478	480	426	440	410	458	460
Other eye	484	478	492	444	436	398	464	476

In what sense are these a matched-pair sample? Are the observed differences in corneal thickness statistically significant at the 1 percent level?

13.14 A random sample of students at Pomona College obtained these data on grade point average (GPA) by race (other races are not shown):

	Number Surveyed	GPA (A = 4, B = 3, ...) Mean	Standard Deviation
Caucasian/White	127	3.391	0.339
Pacific/Asian American	46	3.303	0.331

a. Calculate the overall average GPA when the two categories are combined, giving a single sample of 173 students.
b. Explain why a difference-in-means test comparing the average GPA of the 127 Caucasian/White students to the average GPA of all 173 students is inappropriate.
c. Assuming possibly unequal population standard deviations, use an appropriate difference-in-means test for statistical significance at the 5 percent level. Is the difference substantial?

13.15 A study of changes over a 100-year period in the lead content of people's hair, measured in micrograms of lead per gram of hair, analyzed the sample data shown below. Allowing the sample standard deviations to be unequal, are there statistically significant differences at the 5 percent level in the average lead content between children 1871–1923 and adults 1871–1923? Children 1924–1971 and adults 1924–1971? Children 1871–1923 and children 1924–1971? Adults 1871–1923 and adults 1924–1971?

	Sample Size	Mean	Standard Deviation
Children (1871–1923)	36	164.2	124.0
Adults (1871–1923)	20	93.4	72.9
Children (1924–1971)	119	16.2	10.6
Adults (1924–1971)	28	6.6	6.2

13.16 Before coming to college, a student watched a videotape that claimed to reveal proven strategies for guaranteeing A grades. One strategy is to sit in the front of the class. To test this theory, 50 randomly selected college students were asked to write down their grade point average (GPA) and to indicate where they typically sit in large classrooms. Using the data below, he calculated the average GPA of the 26 students who sit either in the front or towards the front and used a difference-in-means test to compare this average to the average GPA of all 50 students. What crucial assumption is violated by this procedure?

The 26 students who sit in the very front or towards the front had an average GPA of 9.68 with a standard deviation of 1.1468; the thirteen students who sit in the very back or towards the back had an average GPA of 8.79 with a standard deviation of 1.1445.

Assuming these data are from normal distributions with the same standard deviation, determine if the difference in the sample means is substantial and statistically significant at the 5 percent level. Why, even if the difference is substantial and statistically significant, do these data not ensure that students can raise their GPAs by sitting in the front of the class?

	Number of Students	Average GPA (12-Point Scale)
In the very front	5	10.94
Towards the front	21	9.38
In the middle	11	9.38
Towards the back	10	8.37
In the very back	3	10.20

13.17 Data were obtained for the finals of the women's 800-meter run at the Southern California Intercollegiate Athletic Conference Championships. In three of the years studied (1983, 1986, and 1990), the meet was held on a dirt track; in three other years (1987, 1988, and 1989), an artificial surface was used. Although these data are not a random sample, we can plausibly consider them to be representative of similar female athletes in similar years. Allowing the population standard deviations to be unequal, is the difference in the average times statistically significant at the 1 percent level?

	Dirt Track	Artificial Track
Number of runners	18	18
Average time (seconds)	144.90	143.11
Standard deviation	4.83	4.33

13.18 A survey of single and double majors obtained these grade-point-average (GPA) data:

	Number Surveyed	Mean	Standard Deviation
One major	20	3.257	0.354
Two majors	16	3.525	0.230

a. Not assuming a common standard deviation, calculate the t value and two-sided P value for testing the null hypothesis that the average GPAs of all single and double majors are equal.
b. Assuming normality and a common standard deviation, calculate the pooled sample variance and determine the t value and two-sided P value for testing the null hypothesis that the average GPAs of all single and double majors are equal.
c. Does the difference in GPAs seem substantial to you?
d. Provide an interpretation of the observed difference in GPAs other than, "The way to raise your GPA is to major in two subjects, instead of one."

13.19 In a September 18, 1983 article, "The Honeymoon Might be Over for Steve Garvey," *The Los Angeles Times* quoted the manager of the San Diego Padres baseball team as saying, "I'm not taking anything away from Steve [Garvey], but we've played better won-lost percentage ball without him." At the time, the Padres had won 50 and lost 52 with Garvey and won 24 and lost 22 without him. Is this difference substantial or statistically significant at the 5 percent level?

13.20 A study of 546 women and 1,590 men who underwent coronary angioplasty—a surgical procedure in which small balloons are inserted into arteries in order to unclog them—found that 14 women and 4 men died during the operation. Is this observed difference in gender mortality statistically significant at the 1 percent level? Is it substantial? Is a one-tail or two-tail test appropriate?

Women tend to develop heart disease later than men, and older patients are more at risk during surgeries. In this study, the average age of the 546 women was 4 1/2 years older than the average age of the 1,590 men, and 12 of the 14 women who died were over 65. How might this age difference bias the results? How could you correct for it?

13.21 In the Pacific Crest Outward Bound program, students travel through the wilderness for seven to nine days carrying food and equipment in backpacks. A study compared the ratio of backpack weight to student weight for a random sample of 50 students who did not suffer injuries and 40 students who suffered knee injuries. Is the difference substantial and statistically significant at the 5 percent level? (Use a pooled variance.)

	Injuries	No Injuries
Mean	0.383	0.370
Standard deviation	0.029	0.030

13.22 In 1991 the National Education Association (NEA) surveyed 2,000 of the 1,750,000 kindergarten-through-6th-grade teachers in the United States and reported that only 12 percent were male, down from 13.8 percent in 1986 and 17.7 percent in 1981.

a. If half of all 1,750,000 teachers are male, what is the probability that a random sample of 2,000 teachers would include so few males? Use a normal approximation.
b. If the 1981 and 1991 surveys each included 2,000 teachers, is the reported decline from 17.7 percent to 12 percent statistically significant at the 5 percent level?

13.23 Connecticut v. Teal, 1982, was concerned with a promotion test that was passed by 206 of 259 white employees and by 26 of 48 black employees. Is the observed difference statistically significant at the 1 percent level? Substantial by the 80 percent rule?

13.24 A study was made of 10,590 men who had vasectomies, each matched to a man of approximately the same age, race, and martial status who had not been vasectomized. Over the 5 to 14 years covered by the study, 212 of the vasectomized men and 326 of the nonvasectomized men died. If these are assumed to be independent random samples, is observed difference statistically significant at the 5 percent level? If we took into account that these were matched pairs, do you think that the calculated Z value would be higher or lower? The researchers caution that, "Exposure to vasectomy cannot be considered random; the men chose to be vasectomized." Explain why this is a reason for caution.

13.25 A three-year study in the 1960s tested the effectiveness of sodium monofluorophosphate (MFP) as a decay-fighting agent in toothpaste. The 208 children in the MFP group had an average of 19.98 cavities and a standard deviation of 10.61; the 201 children in the control group had an average of 22.39 and a standard deviation of 11.96. Is this difference substantial and statistically significant at the 5 percent level?

13.26 A compilation of over 12,000 times at bat by major league baseball players in 1951 and 1952 separated the data into four categories, depending on whether the pitcher and batter were right or left handed. Calculate the batting averages for each of these four groups. Then combine the data into two groups, depending on whether the pitcher and batter are same-handed or opposite-handed. Assuming these data to be independent observations, test the null hypothesis that the probability of getting a hit does not depend on whether the pitcher and batter are same-handed or opposite-handed.

Batter	Pitcher	Times at Bat	Hits
right	right	5,197	1,201
left	left	1,164	270
left	right	4,002	1,055
right	left	2,245	590

13.27 The first five columns in Exercise 6.30 in Chapter 6 in the textbook are female temperatures; the last five columns are male temperatures. Use these data to find the two-sided *P* value for a test of the null hypothesis that the average body temperatures of females and males are the same. Do not assume that the population standard deviations are equal, but do use two histograms to check for any evidence that these data came from dissimilar or very asymmetrical distributions.

13.28 To compare the age at death of left handed and right handed persons, researchers examined data for major league baseball players who threw and batted with the same hand. Of 1,472 such strongly right handed players, the average age at death was 64.64 years, with a standard deviation of 15.5 years. Of 236 strongly left handed players, the average age at death was 63.97 years, with a standard deviation of 15.4 years. Assuming these data to be a random sample for the purposes of estimating death ages, is there a statistically significant difference at the 1 percent level in the average age at death of right handers and left handers? Do not assume the population standard deviations are equal.

13.29 A survey of 33 female and 31 male randomly selected college students obtained these data on the number of biological children they expected to have during their lifetimes:

	Females	Males
mean	2.1818	2.0968
standard deviation	1.0141	1.2742

Allowing the population standard deviations to differ, calculate the two-sided *P* value for testing the null hypothesis that the population means are equal.

13.30 A study of four years of accident reports by the Dallas Fire Department found that fire trucks painted red made 153,348 runs and had 20 accidents, while identical fire trucks that had been painted yellow made 135,035 runs and had 4 accidents. Assuming these to be independent random samples, is this difference substantial and statistically significant at the 1 percent level?

13.31 To investigate the relationship between plaque, gingivitis, and periodontis, dental assessments were made of 747 undergraduate students (average age 21.4 years) at the University of Indonesia, a group with high socioeconomic status and substantial access to dental services, and of 592 young-adult Indonesian tea pickers (average age 24.2 years)

who have no non-emergency access to dental care. The dentists compiled plaque and gingivitis indexes for each person and also noted whether the person had periodontis:

	Plaque (scale 0–6)		Gingivitis (scale 0–2)		Cases of
	Mean	Std. Dev.	Mean	Std. Dev.	Periodontis
students	1.49	0.8	0.30	0.28	29
tea pickers	2.48	1.01	0.55	0.38	25

For each of these three sets of dental data, see if there is a statistically significant difference at the 5 percent level between the students and the tea pickers.

13.32 A study of the Brown University corpus of one million words of written American English and the Lancaster-Oslo-Bergen (LOB) corpus of one million words of British English found that 37 percent of the Brown corpus were nouns and 36 percent of the LOB corpus were nouns. Assuming the binomial model to be appropriate, are these differences substantial and statistically significant at the 1 percent level?

13.33 A random sample of 59 Pomona College students were asked their grade point averages (GPAs) and whether they had been admitted to the college through the early decision program or as part of the regular admissions process. Here are the results:

	Early (12)	Regular (47)
Mean	3.57	3.37
Standard deviation	0.22	0.33

Do these data show a substantial and statistically persuasive difference in the GPAs of these two groups of students? Do not assume that the population standard deviations are equal.

13.34 To see whether major league baseball players bat better during day or night games, a researcher randomly selected 75 players (not including pitchers) who batted at least 100 times during the 1994 baseball season. The mean batting average for the 75 players was 0.2841 during day games (with a standard deviation of 0.0503) and 0.2761 during night games (with a standard deviation of 0.0383). Assuming these to be independent samples from populations with possibly unequal variances, test the null hypothesis that the population mean batting averages are the same for day and night games. Why might a matched-pair test be appropriate for these data? Make a matched-pair test using the sample mean of 0.0080 and standard deviation of 0.0572.

13.35 A price comparison was made of 40 everyday items at L grocery stores and R grocery stores. (The items were chosen beforehand at a different grocery chain.) Here are the results, in dollars:

	L	R
Mean	2.63	3.01
Standard deviation	0.61	0.66

Are the observed differences statistically significant at the 5 percent level? Assume that the population standard deviations are equal.

13.36 On average, men are taller than women and therefore take longer strides. One way to adjust for this height advantage in comparing male and female track records is to divide the distance run by the runner's height, giving a proxy for the number of strides it would

take this runner to cover the distance. The division of this figure by the runner's time gives an estimate of the runner's strides per second. For instance, in 1994 Leroy Burrell set the world record of 9.85 seconds for the 100 meters. Dividing 100 meters by his height (1.828 meters) gives 54.70, showing that 100 meters represents approximately 54.70 strides for him. For his world record time of 9.85 seconds, this converts to 54.70/9.85 = 5.55 strides per second. In comparison, Florence Griffith Joyner, the women's record holder in 1994 was 1.7018 meters tall, so that 100 meters is 100/1.7018 = 58.76 strides for her, and her world record time of 10.49 seconds converts to 58.76/10.49 = 5.60 strides per second. A study of the top three runners in several events during seven recent years yielded the data below on strides per second. In every event, there were 21 observations for each gender.

	Men		Women	
	Mean	Standard Deviation	Mean	Standard Deviation
100 meters	5.52	0.20	5.38	0.18
200 meters	5.49	0.12	5.30	0.13
400 meters	4.90	0.13	4.68	0.17
800 meters	4.24	0.12	4.11	0.10
1500 meters	4.04	0.15	3.77	0.11
5000 meters	3.71	0.15	3.33	0.15
10,000 meters	3.57	0.14	3.26	0.13

Treating these data as random samples from a population of top runners in comparable years, see if any of these seven differences in male and female strides per second is statistically significant at the 5 percent level. Do not assume that the population standard deviations are the same for males and females.

13.37 A 1976 study of residents of a Florida retirement community (all of whom were over the age of 65) obtained the data below on cholesterol level. Use these data and a pooled variance to test the null hypothesis that elderly females and males have the same average cholesterol level.

	Females (n = 538)	Males (n = 351)
Age (years)		
Mean	73.7	74.1
Standard deviation	5.3	5.3
Cholesterol (mg/dl)		
Mean	228.9	202.4
Standard deviation	37.1	36.1

13.38 In the 1980s Kahneman and Tversky reported that when 200 people were asked the following question, 92 answered yes and 108 answered no: "Imagine that you have decided to see a play and paid the admission price of $10 per ticket. As you enter the theater, you discover that you have lost the ticket. The seat was not marked and the ticket cannot be recovered. Would you pay $10 for another ticket?"

When 183 people were asked the following question, 161 answered yes and 22 answered no: "Imagine that you have decided to see a play where admission is $10 per

ticket. As you enter the theater, you discover that you have lost a $10 bill. Would you still pay $10 for a ticket for the play?"

Assuming these to be independent random samples, are the observed differences statistically significant at the 1% level?

13.39 A study of survey response rates mailed 357 surveys containing a $2 bill and mailed a control group 368 surveys with no financial incentive. Responses were received from 225 of the first group and 162 of the latter. Are these observed differences substantial and statistically significant at the 1 percent level?

13.40 The 794 students in Pomona College's freshmen classes of 1987–1988 and 1988–1989 were separated into two groups: 552 who had attended public high schools and 242 who had attended private schools. Of those from public schools, 48 (8.7%) were Pomona Scholars their freshman year (an A– grade point average); of those from private schools, 13 (5.4%) were Pomona Scholars. Does the observed difference seem substantial? The author calculated the Z value to be 1.64 and concluded that although "two classes of Pomona freshmen proved the null hypothesis to be true,...the sample size is minuscule in comparison to the population of college and university students." Did she prove the null hypothesis to be true? Are her results invalidated by the small sample size?

13.41 A study found that male derelicts are less often balding than are male college professors:

	Number Studied	Median Age	Number Balding
Derelicts	141	47.5	51
Professors	49	47.5	35

Assuming these to be random samples, test the null hypothesis that male derelicts and professors are equally likely to be balding. Are the differences substantial?

13.42 Writer's Workbench is a computerized writing analysis program developed at Bell Labs to assist their technical writers. It has been used extensively at Colorado State University (CSU) to help students in first-year composition classes and in English as a Second Language (ESL) classes. The program has also been used by CSU instructors to analyze student writing styles. For example, an analysis was made of two different kinds of 30-minute essays written by ESL students. One essay asked the students to describe a bar chart or pie chart. A second essay asked them to discuss a compare/contrast topic, such as the two ways to spend leisure time.

	Describe Chart	Compare/Contrast
Number of essays	64	69
Number of words in essay		
Mean	151.05	198.23
Standard Deviation	51.25	67.11
Average sentence length		
Mean	22.18	19.78
Standard Deviation	6.43	5.48

Are the observed differences between the chart description and compare/contrast essay statistically significant at the 5 percent level? Do not assume that the population standard deviations are equal. Summarize your findings.

13.43 In a 1990 study, half of 20,800 patients in thirteen countries who had been hospitalized for heart attacks were given streptokinase to dissolve clots in their arteries and half were given TPA. By the end of the average two-week hospitalization, 8.5 percent of the streptokinase group had died, compared to 8.9 percent of the TPA group. Is this difference is statistically significant at the 1% level?

13.44 Seven psychologists each tried two different approaches to reducing the cigarette consumption of a matched-pair of heavy smokers who professed a desire to quit smoking. One person was given verbal encouragement and supportive counseling while the other person was given electrical shocks while he smoked. Here are the post-treatment cigarette consumption data, expressed as a percentage of pre-treatment consumption:

Psychologist	1	2	3	4	5	6	7
Supportive	125.0	59.0	43.0	0.0	50.0	18.5	66.0
Punitive	59.5	14.0	72.0	27.5	80.5	17.5	83.0

Which approach reduced cigarette consumption the most, on average? Use a matched-pair test to see if the observed difference is statistically significant at the 1 percent level.

13.45 To see whether runners tend to improve during their high school years, a researcher looked at the season-best times (in seconds) for the 3000 meters by the top ten sophomore females, senior females, sophomore males, and senior males in Oregon in 1990:

Female Sophomores	Female Seniors	Male Sophomores	Male Seniors
635	617	542	516
636	631	548	516
637	640	552	520
645	645	555	525
653	646	559	528
657	654	562	529
663	657	563	529
665	658	565	530
666	663	567	530
667	674	568	534

Assume that these data are random samples from populations consisting of the nation's best high school 3000-meter runners and that the population standard deviations are not necessarily equal. Is there a substantial and statistically significant difference (at the 5 percent level) in the sophomore and senior female times? In the sophomore and senior male times?

13.46 In horseshoes, each player pitches two shoes in each inning. If the player pitches two ringers, this is called a double. A study of the 2000 World Horseshoe Pitching championships looked at whether players were more likely or less likely to throw a double after throwing a double in the previous inning. For example, female pitchers threw 1641 doubles and 1691 nondoubles after a nondouble inning, and threw 2142 doubles and 1666 nondoubles after a doubles inning. Is this observed difference substantial and statistically persuasive? What is your null hypothesis?

13.47 A French hospital put 412 patients who had suffered one heart attack on a traditional Mediterranean diet (including olive oil, fruit, and bread); the control group consisted of 358 heart-attack patients who were given a recommended low-fat diet. Four of the patients on the Mediterranean diet and 17 of the patients on the low-fat diet suffered a second heart attack during the two years of the study. Is this observed difference substantial and statistically persuasive?

13.48 From the 30 stocks in the Dow Jones Industrial Average, students identified the 10 stocks with the highest dividend-price (D/P) ratio and the 10 stocks with the lowest D/P ratio. The percentage returns on each stock were then recorded over the next year:

High D/P	Low D/P
39.77	–37.74
32.01	46.83
28.93	–5.81
32.30	48.04
31.28	20.91
12.17	20.24
40.68	1.05
14.33	–28.10
40.39	17.03
37.74	–1.55

a. Display these data in two side-by-side box plots. Do the data appear to have similar or dissimilar means and standard deviations?

b. Considering these data to be two independent random samples of size 10, calculate the two-sided *P* value for a test of the null hypothesis that the population means are equal. Do not assume that the population standard deviations are equal.

13.49 A study compared the economic status of 4,121 women who had immigrated to California from another country with the economic status of their adult daughters at comparable periods of their lives. Of the 3,540 daughters who lived outside their mother's ZIP code, 1,986 lived in a more affluent ZIP code and 1,554 lived in a less affluent ZIP code. For comparison, the study looked at 5,510 adult daughters of California-born mothers, with 2,755 moving to a more affluent ZIP code and 2,755 to a less affluent ZIP code. Calculate the two-sided *P* value for a test of the null hypothesis that, among daughters who more to different ZIP codes, the probability of living in a more affluent ZIP code does not depend on whether the mother is foreign-born or California born.

13.50 A study comparing the grades given by female and male students to an essay written by a male found that female graders gave an average grade of 7.125 and male graders gave an average grade of 7.500. Explain why you either agree or disagree with this interpretation of the results: "The *t* value was 0.5800, with a *P* value of 0.5721. This can be interpreted as there being a 57.21% chance that the means of 7.500 for male graders and 7.125 for female graders were 0.5800 standard deviations away from each other."

Chapter 14

Using ANOVA to Compare Several Means

Chapter Outline

They call it stormy Monday, but Tuesday's just as bad.

—T-bone Walker

Our statistical tests to this point have been been limited to a single mean or success probability, or to comparisons of two means or two success probabilities. We tested whether the average body temperature of humans is 98.6 degrees Fahrenheit and whether there is a 50 percent chance that a baby will choose a familiar or unfamiliar story from its mother. We compared the average amount of time devoted to good and bad economic news, and we compared the percentage of those taking large amounts of vitamin C who caught colds with the percentage for those who took placebos. Often, however, we are interested in the fine detail that an average or percentage neglects—we don't want to suffer the fate of the statistician who drowned crossing a river that had an average depth of three feet.

If we suspect that a certain casino is using loaded dice, we are less interested in the average value of the numbers rolled than in the frequencies with which the individual numbers occur. If we look at the months in which people are born, marry, or die, we want to look at all twelve months, not just one month or a pair of months. In this chapter, we will develop a very useful statistical procedure developed by the great British statistician Sir Ronald A. Fisher (1890–1962) for handling multiple samples of quantitative data. In Chapter 15, we look at a test that can be used for categorical data.

14.1 The Logic Behind an ANOVA *F* Test

The logic behind Fisher's approach can be explained as follows. There are many situations where we suspect that a certain medication, diet, equipment, or procedure may make a difference. The ill may get well sooner with certain medicines; people may be more productive with certain computer software, students may learn more from certain teachers. However, variations in the outcomes can be caused by other factors too. Some patients get well faster than others, even if they receive identical treatment. Some workers are more productive than others, even if they use

identical software. Some students learn more than others, even if taught by the same teacher. How do we distinguish differences in the effects of the medication, software, or teacher from those caused by variations among the people participating in the experiment? That's what you will learn in this chapter.

For concreteness, we look at daily returns in the U.S. stock market. Mondays have a reputation on Wall Street for often being bad days for the stock market. To investigate whether this reputation is well-deserved, an empirical study looked at the daily percentage returns over the period July 3, 1963, to December 18, 1978, with 826 observations for each day of the week. The means, variances, and standard deviations are shown in Table 14.1. The –0.134 return is 134 thousandths of a percent, not 13.4 percent, and so on. Only Monday had a negative average return, and it works out to be a staggering –33.5 percent on an annual basis.

Table 14.1 S&P 500 Daily Percentage Returns

	Monday	Tuesday	Wednesday	Thursday	Friday
Mean	–0.134	0.002	0.096	0.028	0.084
Variance	0.670	0.551	0.644	0.483	0.479
Standard Deviation	0.819	0.742	0.802	0.695	0.692

We want to investigate the question of whether these observed differences in average returns by days of the week are statistically persuasive or can be explained by the inevitable variation in returns. It would be astonishing if the average returns were exactly the same for each day of the week. The statistical issue is whether the observed differences are too large to be explained by chance. The differences among the five means are suggestive, but the substantial standard deviations show that there is considerable variation in daily returns.

Figures 14.1 and 14.2 illustrate two very different explanations for observed differences among the sample means. (To unclutter the figure, we only look at two sample means.) In Figure 14.1, the two sample means come from a single population with a relatively large standard deviation. In Figure 14.2, the two sample means come from two populations with different means and relatively small standard deviations.

How can we choose between these competing explanations for the observed differences among the sample means? The key is to use the sample standard deviations to gauge whether the observed differences among the sample means are large or small relative to the variation within each sample. If the observed differences among the sample means are small relative to the observed variation in the data within each sample, then the samples may well have come from the same population. If, however, the observed differences among the sample means are large relative to the observed variation in the data within each sample, then it is unlikely that the samples came from the same population.

If we only had two sample means to compare, we could use the difference-in-means *t* test described in Chapter 13. It might therefore be tempting to use a sequence of *t* tests on the 10 possible pairings of the stock return data in Table 14.1: Monday with Tuesday, Monday with Wednesday, Tuesday with Wednesday, and so on. However, as mentioned in the textbook

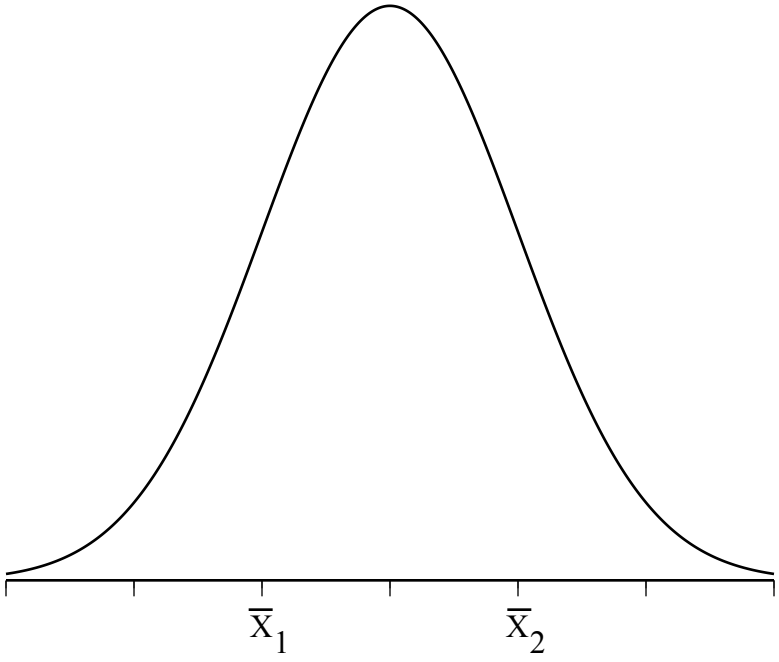


Figure 14.1 Two Sample Means from the Same Distribution

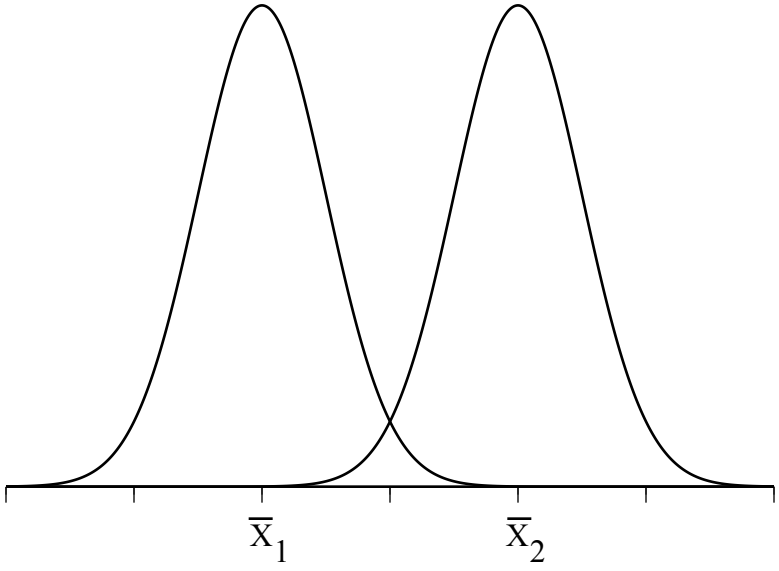


Figure 14.2 Two Sample Means from Different Distributions

discussion of data grubbing, as we increase the number of tests, we increase the probability of incorrectly rejecting at least one of the many null hypothesis we are testing. The *F* statistic described in the next section provides an attractive alternative: a single test of the null hypothesis that all of the population means are equal.

14.2 The *F* Statistic

To describe our statistical procedure in general terms, suppose we have 12 random samples. The null hypothesis is that all *k* of the population means are equal; the alternative hypothesis is that at least one of these population means is not equal to the others:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

H_1: at least one population mean differs from the others

Although we compare the sample means in order to test the equality of population means, our procedure is called analysis of variance (ANOVA) because our analysis is based on a comparison of two kinds of variation: the variation among the sample means and the variation within each sample. In honor of Sir Ronald A. Fisher's pioneering work, the test statistic is called the *F statistic*:

$$F = \frac{\text{variation among sample means}}{\text{variation within samples}} \tag{14.1}$$

The numerator of the ANOVA *F* statistic measures the variation in the samples means about the overall mean of all the data, $\bar{X}$:

$$\text{Variation among sample means} = \frac{n_1\left(\bar{X}_1 - \bar{X}\right)^2 + n_2\left(\bar{X}_2 - \bar{X}\right)^2 + \ldots + n_k\left(\bar{X}_k - \bar{X}\right)^2}{k-1} \tag{14.2}$$

If all the sample means turn out to be the same, so that the variation among the sample means is zero, the data provide no evidence whatsoever against the null hypothesis. In contrast, a large value for the numerator (and the *F* statistic) is evidence against the null hypothesis.

The denominator of the ANOVA *F* statistic is a weighted average of the *k* sample variances:

$$\text{Variation within samples} = s_P^2 = \frac{\left(n_1-1\right)s_1^2 + \left(n_2-1\right)s_2^2 + \ldots + \left(n_k-1\right)s_k^2}{n_1 + n_2 + \ldots + n_k - k} \tag{14.3}$$

This measure of the variation within the samples is a generalization of the pooled variance that was introduced in Chapter 13 for a difference-in-means test when two samples are taken from populations with the same standard deviation:

$$s_P^2 = \frac{\left(n_1-1\right)s_1^2 + \left(n_2-1\right)s_2^2}{n_1 + n_2 - 2}$$

Equation 14.3 is the pooled variance for *k* samples from populations with the same standard deviation. In fact, when there are two samples, the square of the *t* value for a difference-in-means test using the pooled variance is exactly equal to the *F* value for an ANOVA test, and the *P* values are identical! The advantage of an *F* test is that it can be used when we have more than two samples.

The observed differences among the sample means are more telling evidence against the null hypothesis when there is little variation within the samples (when the pooled variance is small).

If the data hardly vary at all within each sample, then large differences among the sample means cannot be explained persuasively by sampling error. If, on the other hand, there is a lot of variation within the samples, sampling error can plausibly explain why the sample means happen to differ somewhat from each other. Thus a small value for the denominator of Equation 14.1 (and a large for the F statistic) is evidence against the null hypothesis.

Figures 14.3 and 14.4 use box plots to illustrate this logic, this time with three samples. The data in each sample are roughly symmetrical with sample means that are close to the medians indicated by the vertical line in the middle of each box. Look first at Figure 14.3. Because there is substantial variation within each sample relative to the observed differences among the sample means, it is conceivable that the data came from populations with identical means, and that the proverbial luck of the draw caused the observed differences among the sample means. The box plots in Figure 14.4 tell a very different story. The means are exactly the same as in Figure 14.3, but because there is little variation within each sample, the observed differences among the sample means suggest that these data did not come from populations with the same mean.

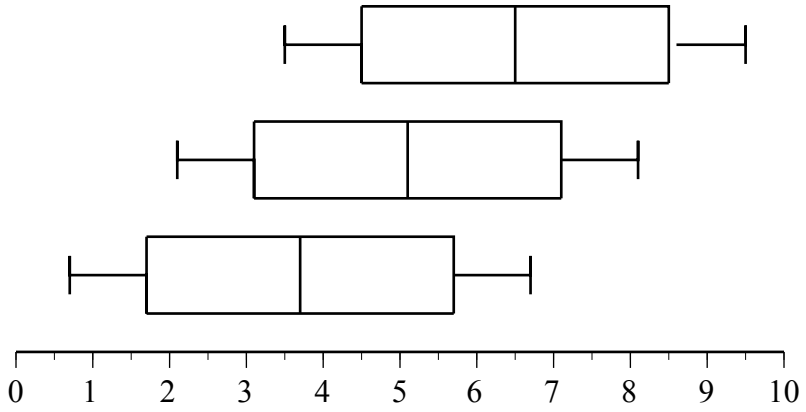


Figure 14.3 The means do not differ substantially relative to the variation within each group

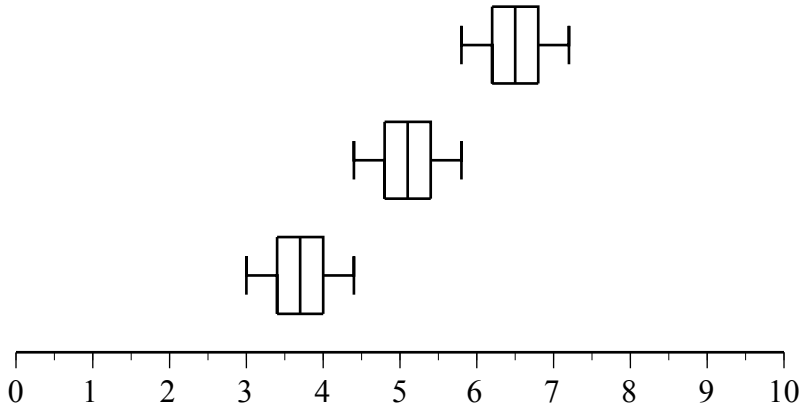


Figure 14.4 The means differ substantially relative to the variation within each group

Matters are actually more complicated than this, because we need to take the sample sizes into account in order to judge the amount of variation to be expected in the sample means. The *F* statistic does this. It considers the differences among the group means, the variation within each group, and the sample sizes.

The calculation of the *F* statistic from Equations 14.1, 14.2, and 14.3 is obviously complicated (and tedious) and best done with statistical software. Nonetheless, we will show one calculation to give you a feel for the messy details. The overall mean for the daily stock returns in Table 14.1 works out to be 0.015. The numerator of the *F* statistic is

$$\begin{aligned}\text{Variation among sample means} &= \frac{n_1\left(\bar{X}_1-\bar{X}\right)^2+n_2\left(\bar{X}_2-\bar{X}\right)^2+\ldots+n_k\left(\bar{X}_k-\bar{X}\right)^2}{k-1}\\ &= \frac{826(-0.134-0.152)^2+\ldots+826(-0.084-0.0152)^2}{5-1}\\ &= 6.992\end{aligned}$$

In order to calculate the denominator of the *F* statistic, we need to use the sample variances given in Table 14.1:

$$\begin{aligned}\text{Variation within samples} &= \frac{(n_1-1)s_1^2+(n_2-1)s_2^2+\ldots+(n_k-1)s_k^2}{n_1+n_2+\ldots+n_k-k}\\ &= \frac{825(0.670)+\ldots+825(0.479)}{826+826+826+826+826-5}\\ &= 0.5654\end{aligned}$$

Finally, the value of the *F* statistic is

$$\begin{aligned}F &= \frac{\text{variation among sample means}}{\text{variation within samples}}\\ &= \frac{6.992}{0.5654}\\ &= 12.37\end{aligned}$$

These calculations are obviously complex and taxing; statistical software provides a welcome alternative.

14.3 The *F* Distribution

In order to assess the value of the *F* statistic, we need to know its probability distribution so that, as with any hypothesis test, we can determine whether the observed *F* value is sufficiently unlikely to be persuasive evidence against the null hypothesis.

If all k samples come from normal distributions with the same standard deviation and if the null hypothesis is true, so that all the population means are equal, then the ANOVA *F* statistic has the *F distribution* with the following degrees of freedom:

Numerator: $k-1$

Denominator: $n_1+n_2+n_3+n_4-k$

The *F* distribution assumes that the populations have normal distributions. However, because of the power of the central limit theorem, the ANOVA *F* test is still reliable even if the underlying populations are not exactly normal. What is important is that the sample means are

approximately normally distributed. If the distributions are generally symmetrical without extreme outliers, the ANOVA *F* test works well with as few as 5 observations in each sample. Even with asymmetrical distributions and outliers, the test is acceptable with reasonably large samples.

The assumption that all the populations have the same standard deviation does not have to be exactly true either. The ANOVA *F* test still works well if the samples are roughly the same size and not too small, and the largest sample standard deviation is not more than twice the size of the smallest sample standard deviation. Our daily stock return data pass this test, in that the largest standard deviation is only 1.18 the size of the smallest: 0.819/0.692 = 1.18.

The *F* distribution has two separate degrees of freedom—one for the numerator and one for the denominator. Figure 14.5 shows two *F* distributions. (The *F* distribution becomes bell-shaped as both degrees of freedom increase.) Because the *F* statistic is the ratio of two sums of squares, it cannot be negative. As indicated in our earlier discussion, a large value of the ANOVA *F* statistic is evidence against the null hypothesis because it indicates that there are large differences among the sample means relative to the variation within each sample.

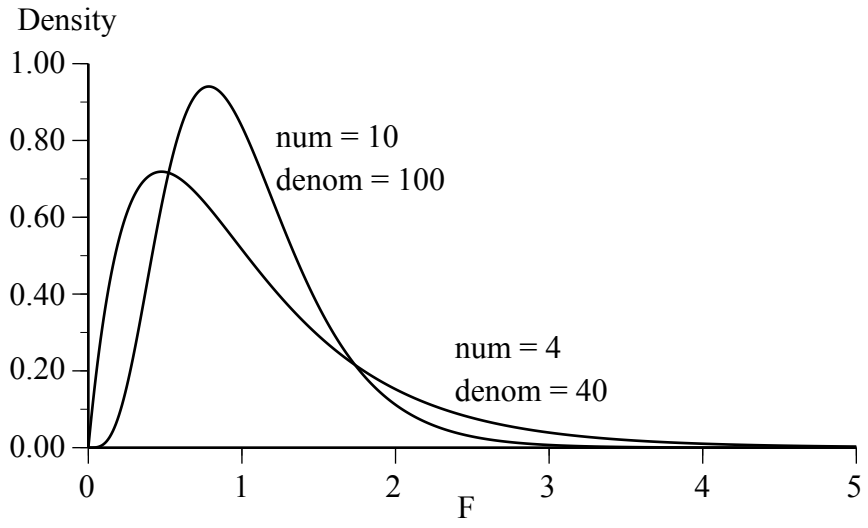


Figure 14.5 Two *F* Distributions

In our daily stock return example, there are $k = 5$ samples and the sample sizes all 826, giving these degrees of freedom:

Numerator: $k - 1 = 5 - 1 = 4$

Denominator: $n_1 + n_2 + n_3 + n_4 + n_5 - k = 5(826) - 5 = 4{,}125$

Figure 14.6 shows this *F* distribution. The *P* value is a minuscule 0.00000081. The observed 12.37 *F* value decisively rejects the null hypothesis that the days of the week do not matter. If the null hypothesis is correct, there is only a 0.00000081 probability that five samples of size 826 will yield an *F* value larger than 12.37. In these stock return data, the variation within each sample is too small to explain the observed differences among the daily mean returns.

Why is Monday so stormy? The researchers who compiled these data tried to find a reasonable explanation and came up empty, concluding that it is "an obvious and challenging empirical anomaly."

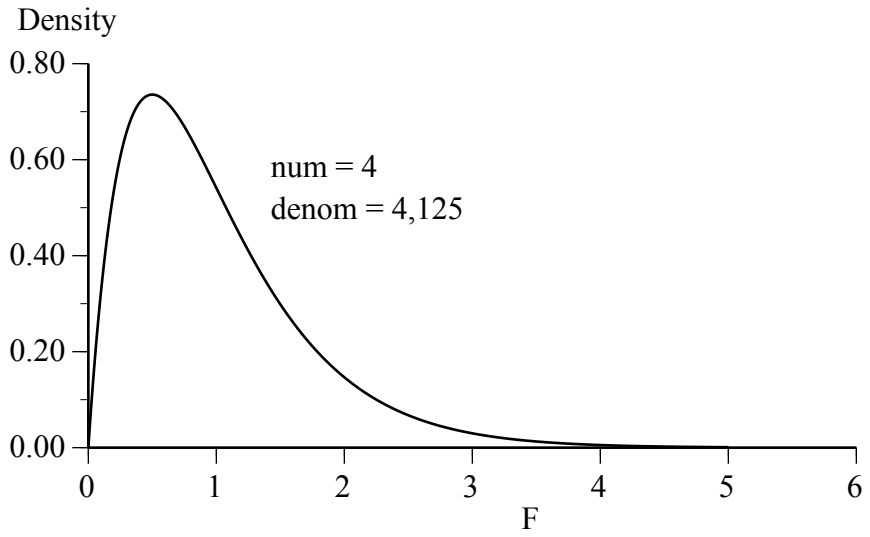


Figure 14.6 The *F* Distribution for Stock Returns by Day of the Week

The ANOVA Table

The results of an ANOVA *F* test are often summarized in an ANOVA table, as in Table 14.2. Software packages might use somewhat different labels for the rows and columns and arrange the columns in slightly different order, but all contain the same information in Table 14.2.

Table 14.2 An ANOVA Table

ANALYSIS OF VARIANCE					
SOURCE	DF	SS	MS	F	P
FACTOR	4	27.97	6.99	12.37	0.000
ERROR	4,125	2,332.27	0.57		
TOTAL	4,129	2,360.24			

The first column ("source") uses the labels "factor," "error," and total" to show that we are separating the total variation in the data into two parts—the variation among the sample means and the variation within each sample. The factor row (or "treatments" or "groups" in other software) refers to the variation among the samples, as measured by the variation among the sample means. The variation among the sample means is said to be explained by our model in that it is attributable to the different groups. The error row in the table refers to the remaining variation, the variation within each sample; it is not an error in the sense that we have made a

mistake, but rather refers to the fact that some of the variation in the data (the variation within each group) is not explained by the existence of different groups.

The second column in Table 14.2 shows the degrees of freedom. The third column (SS, for "sum of squares") shows the separation of the total sum of squared deviations of the data about the overall mean into two sums of squares: the variation among the sample means and the variation within each sample. The fourth column (MS, for "mean sum of squares") divides each sum of squares in the third column by the degrees of freedom in the second column: 27.97/4 = 6.99 and 2,332.27/4,125 = 0.57. The MS values shown in the fourth column are the numerator and denominator of the ANOVA F statistic, $F = 6.99/0.57 = 12.37$. The fifth column shows the value of this statistic and the sixth column gives the P value, the probability of obtaining an F value so far from zero if the null hypothesis were true.

Taste Tests of a New Food Product

In developing a new "easy-to-prepare, nutritious, on-the-run" food product, General Foods compared the taste of two versions of this product—one liquid and the other solid—to a successful competitive product. For this taste test, General Foods paid 75 females and 75 males who were divided randomly into three groups of 50; each with 25 females and 25 males. One group tasted the new liquid product, the second group tasted the new solid product, and the third group tasted the established product. Each person was asked to show her or his reaction to the product by marking a secret ballot showing seven faces ranging from delight to disgust.

For an ANOVA test, we assume that the seven faces correspond to a numerical scale with the difference between any two boxes reflecting a constant difference in taste. General Foods assigned numerical scores ranging from +3 for delight to –3 for disgust, with the results summarized in Table 14.3. General Foods was naturally pleased that its new product was more highly rated, on average, than the established competitor. The statistical question is whether the observed differences among the average ratings can be explained by taste variations from one person to the next. Perhaps, by the luck of the draw, the established product happened to be tasted mostly by people who tend to give low ratings. This sort of luck is always possible, but we can use an ANOVA F test to determine the probability of such bad luck.

Table 14.3 Ratings of Three Food Products

Score	New Liquid	New Solid	Established
+3	7	5	1
+2	13	11	5
+1	16	19	15
0	11	13	17
–1	2	1	8
–2	1	1	3
–3	0	0	1
Mean	1.18	1.06	0.22
Variance	1.38	1.16	1.44

The null hypothesis is that the three samples come from populations with the same mean:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

The overall mean of the data is 0.82. Fisher's F statistic can be calculated from

$$\text{Variation among sample means} = \frac{n_1\left(\bar{X}_1 - \bar{X}\right)^2 + n_2\left(\bar{X}_2 - \bar{X}\right)^2 + n_3\left(\bar{X}_3 - \bar{X}\right)^2}{k-1}$$

$$= \frac{50(1.18-0.82)^2 + 50(1.06-0.82)^2 + 50(0.22-0.82)^2}{3-1}$$

$$= 13.68$$

and

$$\text{Variation within samples} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2}{n_1+n_2+n_3-k}$$

$$= \frac{49(1.38) + 49(1.16) + \ldots + 49(1.44)}{50+50+50-3}$$

$$= 1.33$$

so that

$$F = \frac{\text{variation among sample means}}{\text{variation within samples}}$$

$$= \frac{13.68}{1.33}$$

$$= 10.32$$

Here are the results in ANOVA format:

ANALYSIS OF VARIANCE

SOURCE	DF	SS	MS	F	P
Explained	2	27.36	13.68	10.32	0.000
Unexplained	147	194.78	1.33		
Total	149	222.14			

With $k = 3$ samples and each sample of size 50, the degrees of freedom are $k - 1 = 3 - 1 = 2$ and $n_1 + n_2 + n_3 - k = 50 + 50 + 50 - 3 = 147$, and the probability of an F value larger than 10.3 is 0.0002. Bad luck is always possible, but here that explanation for the observed variation among the ratings is not very plausible. The differences among the sample means are too large to be credibly explained by sampling error. Evidently, the population does, on average, give higher ratings to General Foods' new product.

Simultaneous Confidence Intervals and *t* values

If we are persuaded to reject the null hypothesis that the sample data come from populations with identical means, then it is natural to look more closely at the differences among the means. Are all the sample means quite dissimilar? Or is one sample mean very different from the rest, which are themselves very similar? And is there a logical, plausible explanation for why the means differ in the way that they do?

One way to proceed is to use our sample means and the standard errors for these means to calculate confidence intervals for each of the population means. The general formula (Equation

6.7) from Chapter 6 of the textbook is

$$\text{confidence interval for } \mu = \overline{X} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the appropriate t value for a given confidence level, such as 95 percent or 99 percent, and depends on the degrees of freedom.

When we have a single sample, as in Chapter 6, we use that sample's standard deviation in our calculation of the standard error of the sample mean. Here, using ANOVA, we have several samples, and our F test is based on the assumption that the population standard deviations are all equal. The pooled standard deviation s_P defined in Equation 14.3 uses data from all k of the samples, and is consequently an appropriate estimator of the common population standard deviation σ. Thus, a confidence interval for a population mean μ_i is based on the mean of the sample from that population and the pooled standard deviation from all the samples:

$$\text{confidence interval for } \mu_i = \overline{X}_i \pm t^* \frac{s_P}{\sqrt{n_i}} \qquad (14.4)$$

The appropriate t^* value can be determined from the t distribution with $n_1 + n_2 + \ldots + n_k - k$ degrees of freedom.

Instead of calculating the pooled standard deviation directly from Equation 14.3, we can let our software do the calculations for us. The pooled standard deviation is the square root of the denominator of the ANOVA F statistic, which is shown as the mean sum of squares in an ANOVA table.

In our daily stock return data, the pooled standard deviation is the square root of the 0.57 value given in the MS column and ERROR row of Table 14.2:

$$s_P = \sqrt{0.57} = 0.755$$

Some software packages show the pooled standard deviation separately.

With five samples of size 826, there are $n_1 + n_2 + n_3 + n_4 + n_5 - k = 5(826) - 5 = 4{,}125$ degrees of freedom and the appropriate t value for a 95 percent confidence interval is $t^* = 1.96$. Substituting $t^* = 1.96$, $s_P = 0.755$, the sample means in Table 14.1, and the common sample sizes $n_i = 826$ into Equation 14.4, here are 95 percent confidence intervals:

$$\text{Monday}: -0.134 \pm 1.96\left(\frac{0.755}{\sqrt{826}}\right) = -0.134 \pm 0.051$$

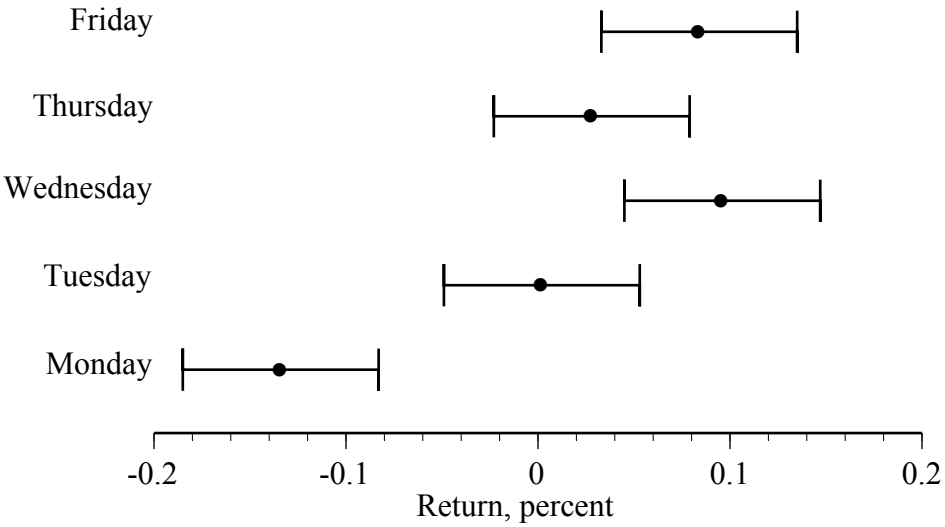
$$\text{Tuesday}: 0.002 \pm 1.96\left(\frac{0.755}{\sqrt{826}}\right) = 0.002 \pm 0.051$$

$$\text{Wednesday}: 0.096 \pm 1.96\left(\frac{0.755}{\sqrt{826}}\right) = 0.096 \pm 0.051$$

$$\text{Thursday}: 0.028 \pm 1.96\left(\frac{0.755}{\sqrt{826}}\right) = 0.028 \pm 0.051$$

$$\text{Friday}: 0.084 \pm 1.96\left(\frac{0.755}{\sqrt{826}}\right) = 0.084 \pm 0.051$$

In this example, the confidence intervals are equally wide because the samples are all the same size. A graph clarifies the fact that the Monday confidence interval does not overlap any other day of the week, and the confidence intervals for the other four days of the week all overlap each other:



In those cases where we reject the null hypothesis, we can use a sequence of *t* tests to compare all possible pairings of the sample means. Equation 13.6 in Chapter 13 gives the formula for calculating the *t* statistic for a difference-in-means test when the standard deviations are assumed to be equal:

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{\sqrt{\frac{s_P^2}{n_1} + \frac{s_P^2}{n_2}}} \tag{13.3}$$

In analysis of variance, we have more than two samples, all of which are assumed to come from populations with identical standard deviations. As with simultaneous confidence intervals, we should use the pooled sample standard deviation based on all *k* samples, and there are $n_1 + n_2 + \ldots + n_k - k$ degrees of freedom.

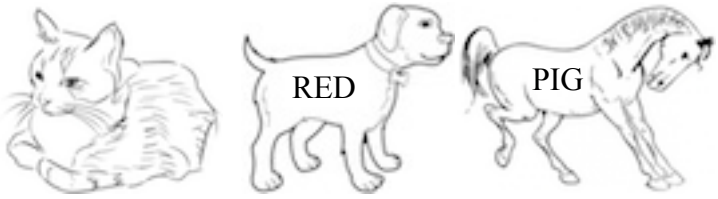
In ourdaily stock return example, there are 10 possible pairs of comparisons: Monday with Tuesday, Monday with Wednesday, Monday with Thursday, and so on. This procedure has the weakness described at the beginning of this chapter. By doing a large number of paired comparisons, we increase the chances that at least one *t* test will falsely reject the null hypothesis. Even more unfortunately, there is no general agreement about exactly how to take into account the multiplicity of tests.

It is clear that to reduce to probability of at least one false rejection of the null hypothesis, we need to raise the *t* value cutoff for rejecting the null hypothesis. But by how much? A number of competing alternatives (including the Bonferroni and Tukey methods) have been proposed and none have won general acceptance. One widely used statistical program offers more than a dozen choices. More advanced textbooks describe these alternatives in detail and offer suggestions for choosing among them.

My recommendation is to begin with the ANOVA *F* test in order to test the null hypothesis that all of the population means are equal. If the data do not persuade you to reject this hypothesis, then do not scrutinize the simultaneous confidence intervals and *t* tests. If you do reject the null hypothesis, then use the simultaneous confidence intervals described here to get a general idea of which specific differences among the sample means are responsible for the rejection of the null hypothesis.

The Stroop Effect

Name these three pictures as quickly as you can:



It is easiest to name the cat and most difficult to name the horse, because we have trained our minds to read words when we see them and this ingrained habit interferes with our efforts to name pictures. When a picture and word send competing signals to the brain, it is difficult to respond quickly and accurately, particularly if the competing word is closely associated with the picture. In the above figure, it is much easier to identify the dog than to identify the horse. Furthermore, because we can read words faster than we can name pictures, it is easier to read the word and ignore the picture than it is to name the picture and ignore the word.

The mental interference created by our trained inclination to read words was demonstrated in a series of classic experiments in the 1930s by a psychologist named J. R. Stroop. He found that it is hard for people to name the color that letters are written in if the letters spell the name of another color; for example, it is difficult to state that letters printed in yellow are yellow if the letters spell the word GREEN. Although we try to ignore the word and just identify the ink color, we cannot help reading the word—it happens automatically and interferes with our response.

This well known phenomenon—the Stroop effect—demonstrates how highly practiced psychological processes can occur without conscious control. A familiar example is when we are trying to ignore background conversation in a noisy room, and find that it is much harder to ignore someone who says our name than it is to ignore less familiar words.

In one test of the Stroop effect, 30 college students were timed for how fast they could name the color of a string of letters. The different conditions can be illustrated by the case of letters written in the color yellow, with the subject timed in the number of seconds it takes to respond "yellow":

Neutral: A string of O's: OOOOO
Unrelated: A word unrelated to any color: SHOOT
Related: A word related to a non-yellow color: GRASS
Color nonresponse: A word that is a color not used anywhere in the experiment: BROWN
Color Response: A word that is a color used elsewhere in the experiment: BLACK

Each student was tested several times, with the student's mean response time for each condition considered to be a single observation. Here are the sample means and standard deviations, in milliseconds:

Condition	Example	Mean	Standard Deviation
Neutral	OOOOO	654.09	61.05
Unrelated	SHOOT	673.50	59.44
Related	GRASS	696.38	67.88
Color Nonresponse	BROWN	755.06	89.54
Color Response	BLACK	793.65	100.09

The response time is slowed by the extent to which the letters interfere with recognizing the color yellow. To determine if the differences are statistically persuasive, we use an ANOVA test:

ANALYSIS OF VARIANCE

SOURCE	DF	SS	MS	F	P
FACTOR	4	407,045	101,764	17.02	0.000
ERROR	145	867,197	5,981		
TOTAL	149	1,274,242			

The null hypothesis that the population mean response times are equal is rejected decisively. If this null hypothesis were true, there is be less than a 0.001 probability that the F value would be 17.02 or larger. (The P value is, in fact, less than 0.0000000001.)

Using the pooled variance, here are 95 percent confidence intervals for the population means:

$$\text{neutral (OOOOO)} : 654.09 \pm 1.98\left(\frac{77.33}{\sqrt{30}}\right) = 654.09 \pm 27.96$$

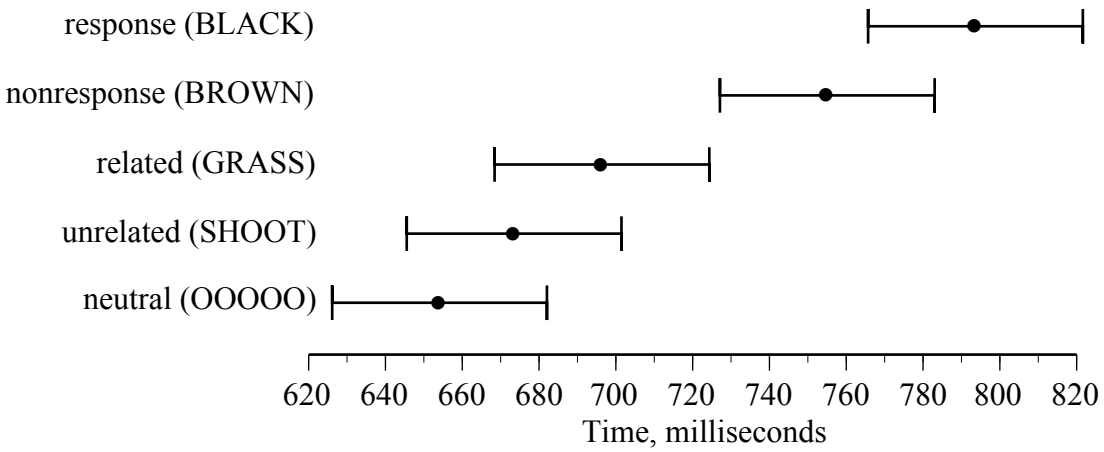
$$\text{unrelated (SHOOT)} : 673.50 \pm 1.98\left(\frac{77.33}{\sqrt{30}}\right) = 673.50 \pm 27.96$$

$$\text{related (GRASS)} : 696.38 \pm 1.98\left(\frac{77.33}{\sqrt{30}}\right) = 696.38 \pm 27.96$$

$$\text{color nonresponse (BROWN)} : 755.06 \pm 1.98\left(\frac{77.33}{\sqrt{30}}\right) = 755.06 \pm 27.96$$

$$\text{color response (BLACK)} : 793.65 \pm 1.98\left(\frac{77.33}{\sqrt{30}}\right) = 793.65 \pm 27.96$$

Here is a graphical representation of these confidence intervals:



While many of the adjacent confidence intervals overlap, there are clear differences between the extremes. The lower limit for the confidence interval for a color-response word (such as BLACK) is 84 milliseconds above the upper limit for neutral letters (OOOOO) and 64 milliseconds above the upper limit for unrelated letters (SHOOT). This experiment confirmed the Stroop effect.

Casual Sleep in Not Good For You

The final examination for one of my introductory statistics courses was at 9:00 in the morning. Just before passing out the exam, I asked the 30 students in this course to write on slips of paper their names and the number of hours they had slept during the past 24 hours. I told them that these slips of paper would be placed in a sealed envelope and not opened until the beginning of the following semester, when the data would be analyzed by that semester's statistics class.

After opening the envelope the following semester, the class analyzed these data in a variety of ways, including drawing a histogram and a box plot and calculating the mean, median, and standard deviation. I also looked up each student's grade on the final examination and matched that to the number of hours of sleep. These exam scores are shown in Table 14.4, with the data grouped into three categories based on the number of hours of sleep.

Table 14.4 Final Exam Scores Grouped by Hours of Sleep

less than 4 hours (n = 7)	4 to 8 hours (n = 17)		more than 8 hours (n = 6)
69	83	72	97
73	91	90	93
84	69	80	92
75	79	88	78
56	85	92	94
78	84	94	89
85	79	79	
	92	79	
	82		

The null hypothesis for an ANOVA F test is that the population means of the exam scores are the same for all sleep groups. Here are the sample means and standard deviations for each group:

	Less than 4 hours	4 to 8 hours	More than 8 hours
Mean	74.286	83.412	90.500
Standard deviation	9.895	7.133	6.656

Students who slept more got better grades on average. The ANOVA F test should be appropriate since the samples are reasonably sized and no standard deviation is twice the size of another::

ANALYSIS OF VARIANCE

SOURCE	DF	SS	MS	F	P
FACTOR	2	869.25	434.63	7.23	0.003
ERROR	27	1623.05	60.11		
TOTAL	29	2492.30			

If the null hypothesis were true, there is only a 0.003 probability that the observed differences among the sample means would be so large relative to the variation within each sample.

For determining confidence intervals for the population means, the pooled standard deviation is the square root of the error mean sum of squares (MS):

$$S_P = \sqrt{60.11} = 7.753$$

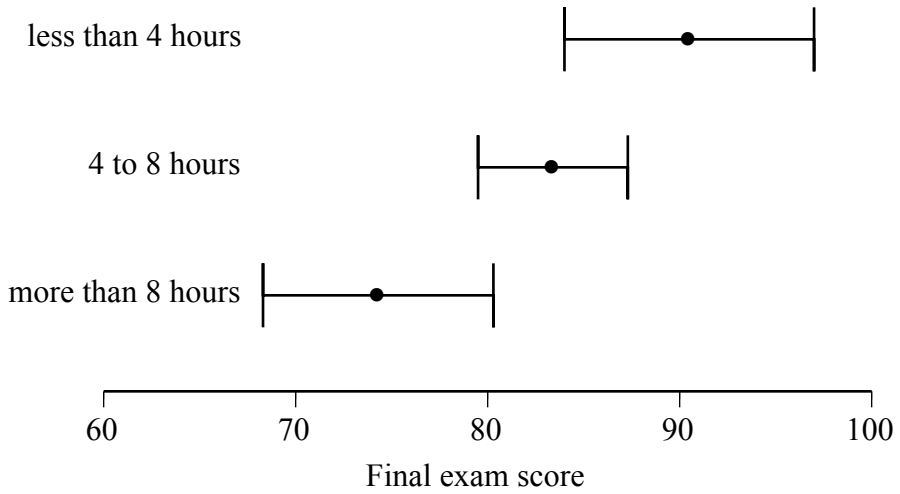
There are $n_1 + n_2 + n_3 - k = 7 + 17 + 6 - 3 = 27$ degrees of freedom and Table A.2 in the textbook shows that the appropriate t value for a 95 percent confidence interval is $t^* = 2.056$. Substituting these values and the sample means and standard deviations above into Equation 14.4, here are 95 percent confidence intervals:

$$\text{Less than 4 hours}: \quad 74.286 \pm 2.056\left(\frac{7.753}{\sqrt{7}}\right) = 74.3 \pm 6.0$$

$$\text{4 to 8 hours}: \quad 83.412 \pm 2.056\left(\frac{7.753}{\sqrt{17}}\right) = 83.4 \pm 3.9$$

$$\text{More than 8 hours}: \quad 90.500 \pm 2.056\left(\frac{7.753}{\sqrt{6}}\right) = 90.5 \pm 6.5$$

Here is a graphical representation:



The confidence interval for the middle category (4 to 8 hours) overlaps the intervals for the two extreme categories, but the 95 percent confidence intervals for the two extreme categories are distinct. The evidence is not decisive that those who get 4 to 8 hours of sleep do better, on average, than those who get less than 4 hours, or do worse than those who get more than 8 hours. However, it is clear that those who get more than 8 hours of sleep do a lot better, on average, than those who get less than 4 hours.

These results do not prove that additional sleep causes students to get better grades. It may be that good students who have worked conscientiously all semester get a full night's sleep because they do not need to cram the night before the final. My advice is to study hard during the semester and sleep well before the final exam, not during it.

Exercises

14.1 To investigate the growth of women between the ages of 18 and 21, we might take four random samples of 100 women of ages 18, 19, 20, and 21. Why might the sample means turns out to be different even if the populations from which these heights came all have the same mean and standard deviation? Use a graph of a sampling distribution to illustrate your reasoning.

14.2 Without doing any calculations, explain why you believe the *P* value for an ANOVA *F* test to be either larger or smaller than 0.05 for these data:

Group A	5.1	5.6	4.9	5.5	5.2	5.4	4.4	4.9	4.8
Group B	4.0	4.8	3.4	3.5	4.1	4.6	4.0	4.4	3.9
Group C	7.2	6.7	5.8	6.5	6.7	5.9	5.8	5.8	6.0

14.3 Without doing any calculations, explain why you believe the *P* value for an ANOVA *F* test to be either larger or smaller than 0.05 for these data:

Group A	5.0	5.6	4.3	3.7	4.6	5.2	4.6	4.9	5.1
Group B	5.6	5.8	5.5	4.9	5.0	5.2	5.6	4.1	5.5
Group C	5.6	4.9	6.1	4.9	5.1	4.9	5.8	5.2	6.4

14.4 Telephone calls were made to 20 randomly selected college students, who were told that they were going to be surveyed about environmental issues. Students who agreed to participate were asked to, "Hold, please." If the student agreed to hold, the researcher timed the number of seconds that elapsed before the student eventually hung up:

Males	32	66	58	136	9	83	247	11	47	93
Females	258	91	132	51	183	86	49	106	319	213

Calculate the *t* value and two-sided *P* value for a difference-in-means *t* test using a pooled variance. Now determine the *F* value and *P* value using an ANOVA test. Is there any evident relationship between the *t* value and the *F* value? Between the two *P* values?

14.5 Here are the annual stock returns for the S&P 500, over the 20-year period 1961–1980:

Year	Percent Return	Year	Percent Return
1961	26.89	1971	14.31
1962	–8.73	1972	18.98
1963	22.80	1973	–14.66
1964	16.48	1974	–26.47
1965	12.45	1975	37.20
1966	–10.06	1976	23.84
1967	23.98	1977	–7.18
1968	11.06	1978	6.56
1969	–8.50	1979	18.44
1970	4.01	1980	32.42

Separate these stock-return data into two groups: the odd-numbered years (1961, 1963, ...) and the even-numbered years (1962, 1964, ...). Now calculate the *t* value and two-sided *P* value for a difference-in-means *t* test using a pooled variance. What is your null hypothesis? Now determine the *F* value and *P* value using an ANOVA test. Is there any relationship between the *t* value and the *F* value? Which test has the larger *P* value?

14.6 Here are 132 S&P 500 monthly returns over the period 1970 (first row) through 1980 (last row):

Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
–7.43	5.86	0.30	–8.89	–5.47	–4.82	7.52	5.09	3.47	–0.97	5.36	5.84
4.19	1.41	3.82	3.77	–3.67	0.21	–3.99	4.12	–0.56	–4.04	0.27	8.77
1.94	2.99	0.72	0.57	2.19	–2.05	0.36	3.91	–0.36	1.07	5.05	1.31
–1.59	–3.33	–0.02	–3.95	–1.39	–0.51	3.94	–3.18	4.15	0.03	–10.82	1.83
–0.85	0.19	–2.17	–3.73	–2.72	–1.28	–7.59	–8.28	–11.70	16.57	–4.48	–1.77
12.51	6.74	2.37	4.93	5.09	4.62	–6.59	–1.44	–3.28	6.37	3.13	–0.96
11.99	–0.58	3.26	–0.99	–0.73	4.27	–0.68	0.14	2.47	–2.06	–0.09	5.40
–4.89	–1.51	–1.19	0.14	–1.50	4.75	–1.51	–1.33	0.00	–4.15	3.70	0.48
–5.96	–1.61	2.76	8.70	1.36	–1.52	5.60	3.40	–0.48	–8.91	2.60	1.72
4.21	–2.84	5.75	0.36	–1.68	4.10	1.10	6.11	0.25	–6.56	5.14	1.92
6.10	0.31	–9.87	4.29	5.62	2.96	6.76	1.31	2.81	1.87	10.95	–3.15

Determine the F value and P value using an ANOVA test of the null hypothesis that the expected value of the return does not depend on the month.

14.7 The Educational Testing Service claims that its SAT tests are a useful predictor of college performance. To test this theory, data were collected on 30 randomly selected first-year students enrolled at the same college, with ten students from each of the following categories of predicted first-year grade-point average (GPA): 2.33–2.66, 2.67–2.99, and 3.00–3.13. Their actual first-year GPAs are as shown:

2.33–2.66:	2.53	2.90	2.43	2.37	2.60	2.16	2.03	2.80	3.17	2.70
2.67–2.99:	2.43	3.20	2.57	2.80	2.10	3.50	3.00	2.70	3.30	2.90
3.00–3.13:	2.40	3.07	3.40	2.76	3.27	3.83	2.86	3.60	3.17	3.50

What are the average first-year GPAs for each of these groups? Calculate the F value and P value for a statistical test using these data. Identify the null hypothesis that you are testing and interpret your results.

14.8 Forty-five college randomly selected college students were each given five minutes to study a two-page article from a professional journal. Fifteen of these students studied while listening to hard rock/alternative music (Nirvana's "Smells Like Teen Spirit"); fifteen studied while listening to classical music (Beethoven's Symphony No. 9, Opera 125); and the last fifteen studied in quiet. After the five-minute study period, the music was turned off and the researcher attempted to remove the article from each subject's short-term memory by having them fill out a form identifying their gender, age, class, and major. Each student was then asked five questions about the article and the number of right answers was recorded:

Rock	3	3	4	4	4	3	4	3	2	2	3	2	4	3	2
Classical	3	3	2	3	4	3	4	4	4	2	3	4	3	2	4
Quiet	3	3	3	5	4	3	3	3	4	5	3	4	4	5	3

Assuming that these data are from populations with normal distributions and identical standard deviations, test the null hypothesis that the population means for the rock, classical, and quiet environments are equal.

14.9 Use the data below to investigate whether cigarette brands with high nicotine content also tend to have high carbon monoxide content (both data are in milligrams). To do this, separate the cigarette brands into three categories based on the nicotine content: less than 0.7, 0.7 to 1.0, and more than 1.0. Now record the carbon monoxide content of the cigarettes in each group, and use three box plots to display these carbon monoxide data. Finally use an ANOVA *F* test at the 5 percent level to compare the carbon monoxide means for these three groups.

	Nicotine	Carbon		Nicotine	Carbon
Alpine	0.86	13.6	MultiFilter	0.78	10.2
Benson&Hedges	1.06	16.6	NewportLights	0.74	9.5
CamelLights	0.67	10.2	Now	0.13	1.5
Carlton	0.40	5.4	OldGold	1.26	18.5
Chesterfield	1.04	15.0	PallMallLight	1.08	12.6
GoldenLights	0.76	9.0	Raleigh	0.96	17.5
Kent	0.95	12.3	SalemUltra	0.42	4.9
Kool	1.12	16.3	Tareyton	1.01	15.9
L&M	1.02	15.4	True	0.61	8.5
LarkLights	1.01	13.0	ViceroyRichLight	0.69	10.6
Marlboro	0.90	14.4	VirginiaSlims	1.02	13.9
Merit	0.57	10.0	WinstonLights	0.82	14.9

14.10 Use the data in Exercise 14.9 to estimate the simple regression model where carbon monoxide is the dependent variable and nicotine content is the explanatory variable. What is the two-sided *P* value for testing the null hypothesis that the slope is zero?

14.11 A statistics student made a study of the office hours posted by professors in each of three divisions (natural sciences, social sciences, and humanities) at his college:

	Natural Sciences	Social Sciences	Humanities
Mean	3.66	3.12	2.91
Standard deviation	1.42	1.21	1.68

Write a paragraph interpreting this computer output of the results:

ANALYSIS OF VARIANCE

SOURCE	DF	SS	MS	F	P
FACTOR	2	13.0	6.50	2.92	0.056
ERROR	148	329.75	2.23		
TOTAL	150	342.75			

14.12 At a large university, students in an introductory statistics course are randomly assigned to sections taught by different persons. Each section uses the same textbook, homework, and tests. Interpret these results of a study of the final exam scores for four sections:

ANALYSIS OF VARIANCE

SOURCE	DF	SS	MS	F	P
FACTOR	3	1543.69	514.56	5.08	0.003
ERROR	116	11748.90	101.28		
TOTAL	119	13292.59			

14.13 Here are the number of persons treated for hornet stings at the emergency room of a Cape Cod hospital before, during, and after Hurricane Bob, grouped into three categories.

0–14 days before:	1	2	3	3	3	2	3	3	1	5	2	2	6	2
0–14 days after:	24	23	31	23	25	26	10	10	20	11	14	13	3	5
15–28 days after:	4	5	6	2	8	20	7	4	2	2	3	1	0	3

Use three box plots to compare these data, and then apply an ANOVA F test. What is your null hypothesis and what is your conclusion? If you reject the null hypothesis at the 5 percent level, calculate 95 percent confidence intervals for the three population means.

14.14 A study of the language proficiency of international students attending college in the United States asked professors to rate on a scale of 1 to 5 (with 1 being most proficient and 5 least proficient) the grammatical abilities of international students in their classes whose native language was not English. The following data were obtained for 34 students in these three language groups: 16 Asian (native speakers of Chinese, Japanese or Korean), 7 Austronesian (native speakers of Malay or Tagalog), and 11 Indo-European (native speakers of French, German, Italian, or Spanish).

Asian	2.8	3.9	3.9	2.1	2.2	3.0	1.0	3.0	3.1	2.7	2.4
	2.1	2.7	2.6	2.5	2.7						
Austronesian	3.0	3.5	1.0	3.9	4.4	2.3	3.4				
Indo-European	2.1	2.1	1.6	1.3	3.0	1.6	1.1	3.1	3.2	2.4	1.0

Use box plots to display the data and then use an ANOVA F test to see if the differences among the sample means are statistically significant at the 5 percent level. If the P value is less than 0.05, calculate 95 percent confidence intervals for the population means.

14.15 A highway department conducted an experiment with three different brands of paint for traffic lanes. Each paint was used on parts of five different streets and then measured for brightness 12 months later. Use Fisher's F test to see if there are statistically significant differences at the 1 percent level among these three brands. If there are, calculate 99 percent confidence intervals for the population means.

Brand 1	Brand 2	Brand 3
72	80	65
67	71	59
58	65	58
72	71	63
65	66	58

14.16 A chemical firm tested the durability of four kinds of rubber-covered fabric. The following data show the weight loss from rubbing tests on four samples of each fabric:

Test	Fabric 1	Fabric 2	Fabric 3	Fabric 4
1	268	218	235	236
2	251	241	229	227
3	274	226	273	234
4	270	195	230	225

Use Fisher's ANOVA F test to see if the observed differences are statistically significant at the 1 percent level.

14.17 Twenty-one industrialized countries were divided into three groups, based on the percentage of their citizens' 1988 diets that consisted of vegetables. Shown below for each of these countries is the 1988 age-adjusted mortality rate (per 100,000 people) from coronary heart disease for persons of age 35–74:

Percent of Diet Consisting of Vegetables					
0.8 to 1.4		1.5 to 1.8		1.9 to 2.8	
Australia	210.9	Canada	190.6	Austria	166.6
Belgium	131.2	Denmark	220.4	Finland	296.8
France	71.3	Ireland	299.8	Iceland	211.4
Italy	107.3	Japan	35.6	Israel	182.9
Spain	86.4	Netherlands	166.5	Norway	226.5
UK	284.7	Switzerland	115.1	Sweden	207.1
New Zealand	265.6	West Germany	171.9	United States	198.6

Summarize these data with three box plots, and then use an ANOVA F test to see if the differences among the three group means are statistically significant at the 5 percent level. If so, calculate 95 percent confidence intervals for the three population means.

14.18 Three girls scout troops operating in very similar areas recorded the following annual sales of cookie boxes. Use box plots to display these data. Then test the null hypothesis that these data were drawn from the same normal distribution.

	Troop 287	Troop 319	Troop 403
1986	1238	815	1436
1987	1159	1478	1152
1988	906	1253	1282
1989	1246	1208	1414
1990	1648	1314	1234
1991	1280	1337	1543
1992	1113	948	1438
1993	1114	1215	1282
1994	852	1207	1548
1995	1197	992	1502

14.19 Three long jumpers at a track meet were each given five jumps. Do these data indicate a statistically significant difference at the 5 percent level in their long-jumping ability? What is the null hypothesis being tested here?

A	23'6"	22'10"	24'2"	24'4"	23'11"
B	24'9"	24'7"	23'5"	25'6"	25'0"
C	25'3"	26'8"	25'8"	24'9"	26'8"

14.20 The first and fifth columns in Exercise 6.26 in the textbook are data for first-year students; the second and sixth columns are second-year students; the third and seventh columns are third-year students; and the fourth and eighth columns are fourth-year students. Use an ANOVA F test to see if the differences among the hours of sleep for these four groups are statistically significant at the 5 percent level. If they are, calculate 95 percent confidence intervals for the population means.

14.21 A weather researcher suspected that the heat produced by workers in Melbourne during the workweek might affect the city's temperature. For each Melbourne winter (May through October) from 1964 to 1990, he calculated the average daily maximum temperature for each of the seven days of the week. Interpret this output (the temperatures are in degrees Celsius):

ANALYSIS OF VARIANCE

SOURCE	DF	SS	MS	F	P
FACTOR	6	5.405	0.901	1.90	0.082
ERROR	182	86.100	0.473		
TOTAL	188	91.505			

INDIVIDUAL 95 PCT CI'S FOR MEAN BASED ON POOLED STDEV

LEVEL	N	MEAN	STDEV	----------+---------+---------+------
Mon	27	16.350	0.655	(--------*--------)
Tue	27	16.280	0.686	(--------*--------)
Wed	27	16.234	0.553	(--------*--------)
Thu	27	16.130	0.684	(--------*--------)
Fri	27	15.992	0.786	(--------*--------)
Sat	27	15.871	0.712	(--------*--------)
Sun	27	15.945	0.717	(--------*--------)
				----------+---------+---------+------
POOLED STDEV = 0.688				15.90 16.20 16.50

14.22 Here are the self-reported heights (in inches) of singers in the New York Choral Society in 1979. The sopranos and altos are female; the tenors and basses are male. Use the means, medians, standard deviations, and box plots to describe these data.

Soprano					Alto			Tenor				Bass	
64	67	66	65	65	62	70	63	69	70	72	66	71	71
62	65	62	61	62	61	65	64	72	65	70	68	70	68
66	62	65	65	68	66	64	67	71	72	72	71	74	70
65	65	63	66	67	64	63	66	66	70	69	73	70	75
60	68	65	65	67	60	65	68	76	68	73	73	75	72
61	65	66	62	63	61	69		74	73	71	70	75	66
65	63	65		67	66	61		71	66	72	68	69	72
66	65	62		66	66	66		66	68	68	70	72	70
65	62	65		63	66	65		68	67	68	75	71	69
63	65	66		72	62	61		67	64	71	68	70	

Use an ANOVA *F* test to see if the differences among the average heights for these four groups are statistically significant at the 5 percent level. If they are, calculate 95 percent confidence intervals for the population means.

14.23 Below are data from a grocery store for 75 breakfast cereals on grams of sugars and the shelf location (1 = bottom, 2 = middle, 3 = top). Group the data by shelf location and use three box plots to compare the sugar content by shelf location. Now apply an ANOVA *F* test. What is your null hypothesis and what is your conclusion?

Sugar	Shelf	Sugar	Shelf	Sugar	Shelf	Sugar	Shelf	Sugar	Shelf
6	3	11	1	2	3	3	1	3	2
8	3	7	2	10	3	2	1	6	2
5	3	10	3	14	3	12	2	12	2
0	3	12	3	3	3	13	2	3	2
8	3	12	2	0	3	7	3	11	3
10	1	15	1	0	3	0	2	11	3
14	2	9	2	6	3	3	3	13	3
8	3	5	3	12	2	10	3	6	1
6	1	3	3	8	3	5	3	9	2
5	3	4	3	6	3	13	2	7	3
12	2	11	2	2	1	3	1	3	3
1	1	10	1	3	1	5	2	12	2
9	2	11	1	0	1	3	3	3	1
7	3	6	3	0	1	14	3	3	1
13	2	9	3	15	2	3	3	8	1

14.24 An atomic absorption spectrophotometry analysis of samples of Romano-British pottery at three sites in the United Kingdom yielded the data below on the percentage of aluminum oxide in each sample. Use box plots to compare these sites and apply an ANOVA test. If the differences among the sample means are statistically significant at the 5 percent level, calculate 95 percent confidence intervals for the three population means.

Llanederyn		Island Thorns	Ashley Rails
14.4	11.6	18.3	17.7
13.8	11.1	15.8	18.3
14.6	13.4	18.0	16.7
11.5	12.4	18.0	14.8
13.8	13.1	20.8	19.1
10.9	12.7		
10.1	12.5		

14.25 A standardized test of logical reasoning was given on the first day of class to students enrolled in introductory mathematics courses at a California State University campus. The test results were not made available to the professors teaching these courses until after the courses were completed and grades had been assigned. An ANOVA test was then used to see if the logical reasoning scores were related to the course grades. The F value was 12.85, with a P value of 0.000000001. Summarize these additional results:

Grade	Sample Size	Mean Logical Reasoning Score	Standard Deviation	95% Confidence Interval
A	85	16.90	2.53	16.90 ± 0.68
B	95	15.65	3.26	15.65 ± 0.65
C	85	14.01	3.27	14.01 ± 0.68
D	18	12.83	4.33	12.83 ± 1.48
F	14	13.57	4.38	13.57 ± 1.68

14.26 Below are measurements of the maximum skull breadth of samples of male Egyptians in three different time periods ranging from 4000 B.C. to 150 A.D.15 (The researchers theorized that change in skull size over time would be evidence of the interbreeding of the Egyptians with immigrant populations.) Compare these three samples with box plots and an ANOVA *F* test. If the *P* value is less than 0.05, calculate 95 percent confidence intervals for the three population means.

4000 B.C.			1850 B.C.			150 A.D.		
131	129	126	137	137	132	137	143	138
125	134	135	129	137	133	136	141	131
131	126	134	132	136	138	128	135	143
119	132	128	130	137	130	130	137	134
136	141	130	134	129	136	138	142	132
138	131	138	140	135	134	126	139	137
139	135	128	138	129	136	136	138	129
125	132	127	136	134	133	126	137	140
131	139	131	136	138	138	132	133	147
134	132	124	126	136	138	139	145	136

14.27 Thirty randomly selected college students tastee three brands of chocolate chip cookies (Pepperidge Farms, SnackWell, and Chips Ahoy) and rated each on a scale of 1 to 10, with 1 the worst possible rating and 10 the best. An ANOVA test of the data yielded the results below. Write a brief summary of the most important conclusions.

ANALYSIS OF VARIANCE

SOURCE	DF	SS	MS	F	P
FACTOR	2	44.60	22.30	8.65	0.0004
ERROR	87	224.30	2.58		
TOTAL	89	268.90			

VARIABLE	N	MEAN	STDEV	95 PCT CI'S FOR MEAN
Pepperidge	30	5.4667	1.7953	5.4667 ± 0.5830
SnackWell	30	6.0667	1.4606	6.0667 ± 0.5830
Chips Ahoy	30	4.3667	1.5421	4.3667 ± 0.5830

14.28 Shown below are the annual growth rates of real per capita income in 41 developing countries classified by the degree to which they restrict international trade. Summarize these data with three box plots, and then use an ANOVA *F* test to see if the differences among the three group means are statistically significant at the 1 percent level. Summarize your findings.

Free-Trade			Moderate Restrictions			Strong Restrictions	
6.5	1.5	5.6	1.3	–1.0	2.0	–1.1	–3.2
6.3	1.4	4.0	1.1	–1.2	2.0	–1.6	–3.4
5.4	0.4	3.3	0.3	–3.5	1.2	–2.0	
4.1	0.4	3.1	–0.1	–3.9	0.5	–2.3	
3.8	0.1	2.7	–0.8		0.5	–2.5	
2.9		1.8	–1.0		–0.4	–3.1	

14.29 To see whether runners tend to improve during their high school years, a researcher looked at the season-best times (in seconds) for the 3000 meters by the top ten sophomore females, senior females, sophomore males, and senior males in Oregon in 1990:

Female Sophomores	Female Seniors	Male Sophomores	Male Seniors
635	617	542	516
636	631	548	516
637	640	552	520
645	645	555	525
653	646	559	528
657	654	562	529
663	657	563	529
665	658	565	530
666	663	567	530
667	674	568	534

Assume that these data are random samples from populations consisting of the nation's best high school 3000-meter runners. Use an ANOVA *F* test to see if the differences among the average times for these four groups are statistically significant at the 5 percent level. If they are, calculate 95 percent confidence intervals for the population means.

14.30 Samples from a can of well-mixed dried eggs were sent to six commercial laboratories to be analyzed for the percentage fat content. The laboratories did not know that all of the samples came from the same can and consequently had the same fat content. Use an ANOVA *F* test to see if the differences among the laboratory measurements are statistically significant at the 1 percent level. If they are, calculate 99 percent confidence intervals for the population means, and display these intervals in a graph.

Lab1	Lab2	Lab3	Lab4	Lab5	Lab6
.62	.30	.46	.18	.35	.37
.55	.40	.38	.47	.39	.43
.34	.33	.27	.53	.37	.28
.24	.43	.37	.32	.33	.36
.80	.39	.37	.40	.42	.18
.68	.40	.42	.37	.36	.20
.76	.29	.45	.31	.20	.26
.65	.18	.54	.43	.41	.06

14.31 Explain why you either agree or disagree with this statement: "An ANOVA *F* test and a difference-in-means *t* test give the same *P* value if there are two samples."

14.32 Explain why you either agree or disagree with this statement: "The *F* value will be negative if all the sample means are negative."

14.33 Explain why you either agree or disagree with this statement: "A significance test that is significant at the 1% level is also significant at the 5% level."

14.34 Explain why you either agree or disagree with this statement: "If the *F* value is 1, then the *P* value is 0."

14.35 Explain the error(s) in this reasoning: "I used a normal approximation to the *F* distribution because the central limit theorem tells us that for large sample sizes, the distribution of the populations means is approximately normal."

14.36 Explain the error(s) in this reasoning: "The *P* value and *F* statistic are both very small, which demonstrates a lack of statistical significance and proves the null hypothesis."

14.37 Explain the error(s) in this reasoning: "The ANOVA *F* statistic can be used to test the null hypothesis that the sample means are equal."

14.38 *Consumer Sports* tested the gasoline mileage of three subcompact cars by driving 10 of each model from Claremont, California, to Brewster, Massachusetts. Do not do an ANOVA *F* test, but do explain why you believe that such a test either would or would not have a *P* value less than 0.05. What is the null hypothesis?

Model A:	32	31	30	27	31	33	31	34	28	31
Model B:	36	41	43	41	41	42	37	40	39	42
Model C:	39	37	33	35	32	36	36	34	37	33

14.39 A student looked at the 1997, 1998, and 1999 batting averages of the 10 major league baseball players who had the highest batting averages in 1998. Identify the error(s) in this conclusion: "The *F* value is 2.74 and the *P* value is 0.0808. Since the *P* value is larger than 0.05, these three samples are not statistically significant at the 5% level. Consequently the data reject the null hypothesis that all population means are equal."

14.40 An ANOVA test obtained an *F* value of 0.0; what is the *P* value?

14.41 For which of the following tests do we need equal sample sizes: difference-in-means, difference-in-proportions, ANOVA, regression?

14.42 Without doing any calculations, explain why you believe that an ANOVA *F* test using these three data sets either would or would not have a *P* value less than 0.05:

	Sample1	Sample2	Sample3
Sample size	30	30	30
Mean	71.33	84.73	81.77
Standard deviation	13.66	6.93	10.84

14.43 Temperature data were compiled for 1970–1989 for these four seasons: winter (December–February), spring (March–May), summer (June–August), and fall (September–November). These data are annual averages of temperatures from more than 1200 weather stations for the contiguous United States (Alaska and Hawaii are omitted). The data below have been adjusted relative to the base period, 1900–1969. Each winter temperature shows how much warmer or colder winter was that year compared to the average winter temperature from 1900 through 1969; each summer temperature shows how much warmer or colder summer was that year compared to 1900–1969; and so on.

	Winter	Spring	Summer	Fall
1970	–0.24	–0.60	0.70	–0.07
1971	–0.08	–1.59	0.14	1.06
1972	0.43	0.66	–0.22	0.15
1973	–0.35	1.04	0.52	1.67
1974	1.00	1.52	–0.46	–0.65

1975	1.20	–1.23	0.17	–0.30
1976	1.65	0.03	–0.85	–3.04
1977	–3.44	2.12	1.20	1.28
1978	–2.29	0.61	0.28	0.43
1979	–5.19	0.35	–0.26	–0.07
1980	1.36	–0.26	0.89	–0.02
1981	1.69	1.63	1.15	0.70
1982	–1.48	0.25	–0.08	0.16
1983	4.26	–0.46	1.16	1.75
1984	–1.34	–0.35	0.68	0.18
1985	–1.34	2.42	–0.29	0.15
1986	0.46	2.39	0.93	0.94
1987	2.80	1.94	0.71	–0.25
1988	–0.40	0.08	1.15	0.13
1989	–0.44	0.84	0.34	–0.43

Draw four side-by-side box plots for these seasonal temperatures and then use an ANOVA F test to see if the differences among the four group means are statistically significant at the 5 percent level.

14.44 Table 13.3 in the textbook shows cola ratings for two samples. Treating these as independent samples, the t value is 0.913 and the two-sided P value is 0.375. Calculate the F value and P value for an ANOVA test.

14.45 Eighteen college-educated adults between the ages of 20 and 40 were asked to rank the taste of samples of duck liver mousse, Spam, and Newman's Own canned dog food, with 1 best, 2 second best and, 3 third best.

Duck	1	1	3	1	1	1	1	1	1	2	2	2	1	1	1	3	1	1
Spam	3	2	2	2	2	3	2	2	2	1	1	1	2	3	2	2	2	2
Dog food	2	3	1	3	3	2	3	3	3	3	3	3	3	2	3	1	3	3

Assume these scores are numerical data and calculate the P value for an ANOVA F test.

14.46 From the 30 stocks in the Dow Jones Industrial Average, students identified the 10 stocks with the highest dividend-price (D/P) ratio and the 10 stocks with the lowest D/P ratio. The percentage returns on each stock were then recorded over the next year:

High D/P	Low D/P
39.77	–37.74
32.01	46.83
28.93	–5.81
32.30	48.04
31.28	20.91
12.17	20.24
40.68	1.05
14.33	–28.10
40.39	17.03
37.74	–1.55

a. Display these data in two side-by-side box plots. Do the data appear to have similar or dissimilar means and standard deviations?
b. Considering these data to be two independent random samples of size 10, calculate the *P* value for an ANOVA test of the null hypothesis that the population means are equal.

14.47 The deans of two small engineering schools compared a random sample of the starting salaries of their 1991 graduates who majored in engineering:

Eastern Institute of Technology	Western Institute of Technology
$45,000	$42,000
42,500	42,000
40,000	41,000
39,000	38,500
38,000	36,000
37,500	35,000
37,000	33,000
36,000	33,000
35,000	32,000
34,000	30,000
32,500	
31,000	
30,000	
30,000	

Determine the *P* value for an *F* test of the null hypothesis that the population means are equal.

14.48 The first five columns in Exercise 6.30 in Chapter 6 in the textbook are female temperatures; the last five columns are male temperatures. Use these data to determine the *P* value for an ANOVA *F* test of the null hypothesis that the average body temperatures of females and males are the same.

14.49 A researcher looked at the number of points scored by the Los Angeles Lakers in each quarter. Explain the error(s) in this description: "The null hypothesis for my *F* test will be H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$, meaning that the average points scored in each quarter should equal all other quarters. My alternative hypothesis is that $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$,, meaning that the average difference in points in each quarter is significant with respect to each quarter."

14.50 Survey data were collected from 120 college students on the average time they went to bed each night during the previous semester and their GPA that semester. Bedtimes were divided into 10–12, 12–1, 1–2, and 2–4 with respective GPAs of 3.68, 3.53, 3.56, and 3.38. An ANOVA test gave a P value of 0.0263. The author concluded that, "The null hypothesis is rejected at the 5 percent level, which indicates we are 95% sure that bedtime affects GPA." Explain the error in this interpretation of the results.

Chapter 15

Chi-Square Tests for Categorical Data

Chapter Outline

Statistics is the grammar of science.

—Karl Pearson

Following psychiatrist Alfred Adler's seminal work in the 1920s, there have been thousands of studies of behavioral differences among first-born and later-born children. Many are contradictory and unpersuasive; some are rigorous and compelling.

The Harvard Preschool Project, a major study lasting more than a decade, found that parents spend twice as much time with a first-born child than with a later-born, and that first-born children score higher at three years of age on standardized tests of language and intelligence. Other studies emphasize how personality might be influenced by the dynamics among children of different ages interacting with each other and with their parents. First-born and only children often seem to have more self-control and self-esteem, and to be driven to succeed. Middle children seek attention and do not respond well to punishment. The last-born are said to be carefree, charming, and manipulative.

A 25-year study by Frank Sulloway, supported in part by a "genius grant" from the MacArthur Foundation, argues that genetic differences evolved during eons of favoritism toward oldest children, who were the biggest, strongest, and most likely to survive. An extreme example is that the practice of infanticide when resources are scarce never involves the first-born; less extreme is the labeling of young children in ancient Japan as "cold rice," because they were fed after the oldest son. Sulloway reasons that the first-born carry genes primordially programmed to respect parents and the established order, while the later born are biologically inclined to take risks and challenge authority. John Adams, Calvin Coolidge, Ayn Rand, and Rush Limbaugh were first-borns; Charles Darwin, Karl Marx, Lenin, Fidel Castro, and Bill Gates were later-borns. By his calculations, in the 1850s, later-borns were five times as likely as first-borns to accept Darwin's revolutionary theories of evolution.

All of these theories—time with parents, family dynamics, and genetics—suggest that first-borns are more likely than later-borns to be law-abiding. Is it true? In this chapter, we will use a statistical test to answer this question.

The ANOVA *F* test described in Chapter 14 can be used with numerical data that can be averaged, for example, a comparison of average stock returns on different days of the week. This chapter will introduce a test that can be used with categorical data: a person is first born or not, a prediction is correct or incorrect.

This test can be used for two very different situations.We begin with the case of data that have been separated into several categories; for, instance, separating the first letter of politicians' last names into three alphabetical groups: A–G, H–S, T–Z. The null hypothesis might be that the alphabetical order of last names is politically unimportant, so that if 30 percent of the population has a last name beginning with a letter in the category A–G, then there is a 30 percent chance that a politician selected at random will have a last name in this category.

The second situation that we consider involves data that have been categorized in two different ways; for example, by political views and gender, or by juvenile delinquency and birth order. Here, our statistical procedure enables us to test the null hypothesis that the two categorizations are independent—political views are independent of gender, juvenile delinquency is unrelated to birth order.

15.1 Tests for Several Categories

We begin with an example from a student's statistics paper. While mulling over the fallacious law of averages, this student speculated that the common belief that the average outcome is the most likely outcome might explain a tendency of people to avoid extremes when they predict a randomly selected number. If asked to choose a number from 1 to 4, few choose 1 or 4; if asked to choose a number from 1 to 10, very few choose 1 or 10. Perhaps people mistakenly believe that because the average is in the middle, a middle number is more likely to occur.

To test this theory, he set up mock a ESP experiment in which he thoroughly shuffled four cards numbered 1 to 4, selected a card, and concentrated on the number written on the card while the participant tried to identify the card he had selected. His real interest was whether they would avoid the extreme numbers 1 and 4. He repeated this charade for forty randomly selected people and obtained these results:

Card Number	Times Selected
1	2
2	13
3	21
4	4

The experiment turned out the way he had anticipated, with very few people choosing the extreme values. But are the tendencies shown in this sample of 40 people persuasive, or might they be explained reasonably well by chance?

Our null hypothesis is that people are equally likely to choose any of the four numbers, just as if their selections were made by rolling a four-sided die. If so, what is the probability that the number 1 will appear just twice, or that the number 3 will appear 21 times? We might be tempted to do a series of binomial tests for some (or all) of the individual numbers, but there are several reasons we shouldn't. First, the outcomes are related, in that if there are few 1s, there must be a large number of 2s, 3s, or 4s. Second, if we make several tests, it is more likely that that we will incorrectly reject at least one null hypothesis. Third, if we look at the data and pick out

anomalies, such as the large number of 3s, this is data grubbing. We should specify the null hypothesis before we see the data, not after. All these reasons argue for a statistical test that looks at the number of 1s, 2s, 3s, and 4s simultaneously.

The Chi-Square Statistic

In 1900, Karl Pearson developed a test statistic that simultaneously compares all the observed and expected numbers when the possible outcomes are divided into mutually exclusive categories. The *chi-square statistic* is

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \tag{15.1}$$

where χ is the lower-case Greek letter chi, and the summation denoted by the upper-case Greek letter sigma Σ is over the categories of possible outcomes.

The calculation is in Table 15.1. First, we specify the expected value in each category if the null hypothesis is true. In our example, the null hypothesis is that each of the four possible numbers has a 0.25 probability of being selected by a participant, so that the expected value of the number of times it will appear in 40 observations is $np = 40(0.25) = 10$. Second, we record the number of observations. Third, we measure the extent to which the null hypothesis is contradicted by the data by squaring the deviation between the observed and expected numbers. This squaring keeps positive and negative deviations from canceling each other and, in addition, emphasizes large deviations because the improbability of large deviations makes their presence more convincing evidence against the null hypothesis. Fourth, we divide each squared deviation by the expected number for that category. This scaling takes into account the fact that a deviation of 10 when 10 are expected provides more evidence against the null hypothesis than does a deviation of 10 when 1,000 are expected. Finally, we add the scaled squared deviations. Because the chi-square statistic is the sum of squared deviations, it cannot be negative. A chi-square value close to zero shows that the observed values are close to what we expect if the null hypothesis is true. A large chi-square value is evidence against the null hypothesis.

Table 15.1 shows that the chi-square value is 23.0 for our card-picking example. Is this value large enough to discredit the null hypothesis? To answer this question, we need to know the probability that a chi-square value would be so far from zero if the null hypothesis were true. To answer this question, we need to know the probability distribution for the chi-square statistic.

Table 15.1 The Calculation of a Chi-Square Value

Card	Observed	Expected	(Observed – Expected)^2	$\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$
1	2	10	$(-8)^2 = 64$	6.4
2	13	10	$(3)^2 = 9$	0.9
3	21	10	$(11)^2 = 121$	12.1
4	4	10	$(-6)^2 = 36$	3.6
Total	40	40	0	23.0

The Chi-Square Distribution

The calculation of the exact probability is very difficult. Not having a computer, Pearson figured out an approximation, based on the ubiquitous central limit theorem. Because each deviation in Equation 15.1 is approximately normally distributed, the probability distribution for the chi-square statistic can be approximated by the chi-square distribution, which is the distribution for the sum of the squared values of normally distributed variables.

Like the *t* distribution and the *F* distribution, the chi-square distribution depends on the degrees of freedom. Figure 15.1 shows three chi-square distributions. Notice that negative values are not possible and that as the degrees of freedom increases, the chi-square distribution approaches a normal distribution. A rule of thumb is that the chi-square distribution is a good approximation as long as no expected value is smaller than 5. If some categories have very small expected numbers, categories can be combined. For example, if 50 participants chose a number from 1 to 100, we might group the responses into ten categories: 1–11, 11–20, and so on.

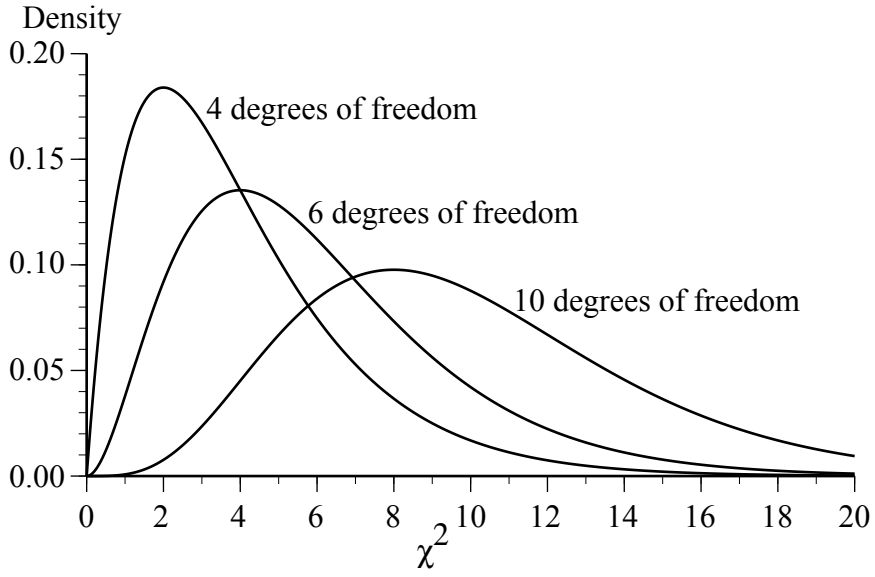


Figure 15.1 Three chi-square distributions

When there is a single list of categories, the degrees of freedom equals the number of categories minus 1. This can be explained by the fact that if we know the values in all but one category, then we know the value in the last category too. For our card example, there are 4 categories and consequently 4 – 1 = 3 degrees of freedom. Figure 15.2 shows this chi-square distribution. With 3 degrees of freedom, a chi-square value above 7.81 is statistically significant at the 5 percent level. For the observed chi-square value of 23.0, the *P* value is a minuscule 0.00014. The tendency to choose the middle numbers (2 and 3) and avoid the extremes (1 and 4) yielded deviations between the observed and expected numbers under the null hypothesis that are too large to be explained plausibly by sampling error. (I now repeat this experiment every semester, and there is invariably a statistically significant aversion to the numbers 1 and 4.)

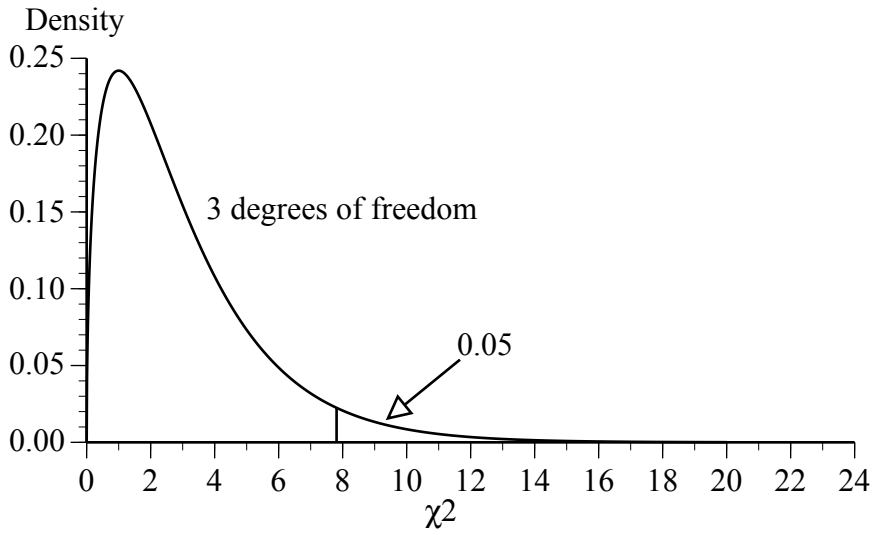


Figure 15.2 A chi-square test

Alphabetical Listing of Political Candidates

Many political observers believe that election results are influenced by the order in which candidates are listed on the ballot. Some indecisive citizens reputedly vote for the first name that sounds reasonable, or at least familiar, thus giving candidates whose names appear early on a list an advantage over those listed later. In cases where names are listed alphabetically, candidates whose surname (last name) begins with a letter early in the alphabet presumably have an advantage. A political party wanting to exploit this advantage could nominate candidates with early-letter names. (This is the same logic that causes some firms that advertise in telephone directories to name their company AA Plumbing or AAAAA Appliance Repair.) Concerned that order may matter, many governments now draw names out of a hat, or use some other random process to determine the order in which candidates' names are listed on ballots.

Researchers investigated the effects of alphabetical listing of candidates in the 1973 general elections in the Republic of Ireland by analyzing the data in Table 15.2.

Table 15.2 Irish Surnames and Elections

First Letter of Surname	Fraction of Irish Voters	Number of Candidates	Number Elected
A–C	0.203	91	47
D–G	0.179	63	32
H–L	0.172	59	23
M–O	0.253	75	22
P–Z	0.194	47	20
Total	1.001	335	144

The surnames have been divided into five categories that encompass an approximately equal number of registered voters. (Last names that begin with M are especially common in Ireland.) First, we use the chi-square statistic to test the null hypothesis that the nominated candidates are selected without regard for surname. If so, we expect roughly 20.3 percent of the 335 candidates to have last names that begin with the letter A, B , or C; 17.9 percent to begin with D, E, F, or G; and so on. Thus we have the following data on the observed and expected numbers under the null hypothesis:

Surname	Observed	Expected
A–C	91	0.203(335) = 68.0
D–G	63	0.179(335) = 59.9
H–L	59	0.172(335) = 57.6
M–O	75	0.253(335) = 84.7
P–Z	47	0.194(335) = 65.0
Total	335	335.2

There are many more candidates with A–C surnames and many fewer with P–Z surnames than would be expected if candidates were chosen without regard for the alphabetical ordering of names. Using Equation 15.1, the value of the chi-square statistic is 14.1:

$$\chi^2 = \frac{(91 - 68.0)^2}{68.0} + \frac{(63 - 59.9)^2}{59.9} + \frac{(59 - 57.6)^2}{57.6} + \frac{(75 - 84.7)^2}{84.7} + \frac{(47 - 65.0)^2}{65.0}$$
$$= 14.1$$

With 5 – 1 = 4 degrees of freedom, statistical software shows the *P* value to be 0.007. These data reject at the 1 percent level the null hypothesis that candidates were selected without regard for surname.

We can also ask whether, given the candidates presented to them, Irish voters are influenced by the alphabetical listing of names on the ballot. The null hypothesis is that they are not. Because 91 of 335 candidates had A–C surnames, the null hypothesis implies that we expect a fraction 91/335 of the 144 people elected to have A–C surnames. Using similar logic for the other four categories,

Surname	Observed	Expected
A–C	47	(91/335)144 = 39.1
D–G	32	(63/335)144 = 27.1
H–L	23	(59/335)144 = 25.4
M–O	22	(75/335)144 = 32.2
P–Z	20	(47/335)144 = 20.2
Total	144	144.0

These results are much less clear cut. The differences are noticeably smaller, and the P–Z surnames won almost exactly as many elections as expected if the null hypothesis were true. The chi-square value works out to be 5.94 and the *P* value is 0.20. While these data reject at the 1 percent level the null hypothesis that the candidates are selected without regard for surname, they do not reject the null hypothesis that those elected are selected from the candidates without regard for surname. Perhaps the voters care less about alphabetical listing than do the politicians.

15.2 Contingency Tables

Pearson's chi-square test is a remarkably versatile way of gauging how closely the data agree with the detailed implications of a specified null hypothesis—a procedure so general that it is often called a *goodness-of-fit test*. We divide the possible outcomes into mutually exclusive categories and then compare the observed number of outcomes in each category with the expected number if the hypothesis were true. One application, as we have seen, is to a single list of possible outcomes, such as four numbered cards or five categories of surnames. Another common application is where the data are labeled in more than one way.

Consider a study that investigated whether male juvenile delinquency is related to birth-order. Based on school truancy, theft, and gang fights, 1,137 boys were separated into two categories—most delinquent and least delinquent—and then separated into four categories based on birth order, with the results displayed in a *contingency table* (or *two-way table*):

	Oldest	In-Between	Youngest	Only Child	Total
Most delinquent	127	123	93	17	360
Least delinquent	345	209	158	65	777
Total	472	332	251	82	1,137

What patterns are in these data? The column percentages help describe the data:

	Oldest	In-Between	Youngest	Only Child	Total
Most delinquent	127	123	93	17	360
	(26.9%)	(37.0%)	(37.1%)	(20.7%)	(31.7%)
Least delinquent	345	209	158	65	777
	(73.1%)	(63.0%)	(62.9%)	(79.3%)	(68.3%)
Total	472	332	251	82	1,137

Overall, 31.7 percent of these boys were classified as most delinquent; the percentages are somewhat smaller for the oldest and only children, and somewhat larger for the in-between and youngest children. Are these observed differences statistically significant, or might they be explained by sampling error—the inevitable variation that occurs in samples?

Our null hypothesis is that there is no relationship between birth order and delinquency. If so, we expect the percentage of delinquent children in each birth-order category to be equal to the overall percentage: 31.7 percent. Thus, among the 472 oldest children, the expected number of "most delinquent" is 0.317(472) = 149.4. Similar calculations give these expected values:

	Oldest	In-Between	Youngest	Only Child	Total
Most delinquent	149.4	105.1	79.5	26.0	360
Least delinquent	322.6	226.9	171.5	56.0	777
Total	472	332	251	82	1,137

For the chi-square distribution to give a good approximation to the exact probabilities, the total number of observations should be at least five times the number of cells and no expected value should be less than 1 (or less than 5 if there are only two rows and two columns).

A comparison of the observed and expected numbers shows the same pattern revealed by a comparison of the column percentages. There are fewer oldest-child and only-child delinquents than expected if the null hypothesis were true and more in-between and youngest-child delinquents than expected. To test the null hypothesis statistically, we use the chi-square statistic

to measure how far the observed values are from the expected numbers:

$$\chi^2 = \frac{(127 - 149.4)^2}{149.4} + \frac{(123 - 105.1)^2}{105.1} + \ldots + \frac{(158 - 171.5)^2}{171.5} + \frac{(65 - 56.0)^2}{56.0} = 17.3$$

We saw earlier that when data are labeled in one direction, the degrees of freedom equals the number of categories minus 1. When, as here, the data are labeled in two directions, the degrees of freedom is

(number of rows – 1)(number of columns – 1)

In our birth-order example, there are (4 – 1)(2 – 1) = 3 degrees of freedom. (Given the row and column totals, three numbers in the first (or second) row will determine all the other entries.) Statistical software shows that the *P* value is 0.00093. These deviations between the observed and expected numbers are sufficiently large to reject the null hypothesis at the 1 percent level. These data indicate that oldest and only children are less likely to be delinquent.

For the simplest possible contingency table, with two columns and two rows, the chi-square test is exactly equivalent to the *Z* test of the difference between two sample proportions discussed in Chapter 12. The advantage of the chi-square test is that it can handle more than two categories.

The Unshuffled Draft Lottery

A national lottery was held on December 1, 1969 to determine the order in which U.S. males born in 1952 would be drafted in 1970 to fight in the Vietnam War. The federal government prepared 366 capsules, each containing a birthday: January 1, January 2, and so on. Each month's capsules were carried to the drawing in a separate box and emptied into a large drum, one after another—first January, then February, and on through the months, with December's capsules poured in last. After rotating the drum a few times, the capsules were selected, with those drawn first drafted first and those drawn last most likely not drafted at all. The results are summarized in Table 15.3.

Table 15.3 The 1969 Draft Lottery

	1–122 (first drafted)	123–244 (middle)	245–366 (last drafted)	Total
January	9	12	10	31
February	7	12	10	29
March	5	10	16	31
April	8	8	14	30
May	9	7	15	31
June	11	7	12	30
July	12	7	12	31
August	13	7	11	31
September	10	15	5	30
October	9	15	7	31
November	12	12	6	30
December	17	10	4	31
Total	122	122	122	366

A casual inspection suggests that the late months, August through December, got more than their share of the "first-drafted" numbers, while the early months, January through May, got a lot of the "last-drafted" numbers. A plausible explanation is that a few rotations of the drum did not do a very good jobs of mixing up the capsules. The early months mostly stayed at the bottom, while the later months stayed mostly stayed on top, ready to be picked first. Is this suspicion just sour grapes from a statistician with a December birthday? After all, even with a perfectly random selection, some days have to be picked first and these may turn out to be December days. Are the apparent patterns in Table 15.3 so strong that they cannot be explained by the luck of the draw?

This is the sort of question that the chi-square test is designed to answer. The null hypothesis is that the capsules were well-shuffled, so that each day is equally likely to be selected. If true, we expect each month to have its days equally divided among the three categories 1–122, 123–244, and 245–366. Months with 30 days would average 30/3 = 10 days per category in the long run, over many such drawings. For months with 31 days, the expected number in each category is 31/3 = 10.33. February, with 29 days in 1952, would average 29/3 = 9.67 days. To measure how severely the observed numbers differ from these expected numbers, we calculate the chi-square value:

$$\chi^2 = \frac{(9 - 10.33)^2}{10.33} + \frac{(12 - 10.33)^2}{10.33} + \ldots + \frac{(10 - 10.33)^2}{10.33} + \frac{(4 - 10.33)^2}{10.33}$$
$$= 37.2$$

Because Table 15.3 has 12 rows and 3 columns, there are (12 – 1)(3 – 1) = 22 degrees of freedom. The *P* value is 0.02. These data reject at the 5 percent level, but not the 1 percent level, the null hypothesis that the capsules were well-shuffled. In response to this statistical evidence, the government was more careful the following year, when they conducted a draft lottery for men born in 1953. In this lottery, two drums were used—one with dates and one with numbers—and each drum was loaded in a random order.

Nicotine Replacement Therapies

Several types of nicotine replacement therapy (NRT) have been developed to help people stop smoking cigarettes. Table 15.4 shows data from an aggregation of the results of fifty independent tests of the efficacy of four different therapies (gum, inhalation, intranasal spray, and transdermal patches) and the control groups for these studies.

Table 15.4 Four Nicotine Replace Therapies and the Control Groups

	Gum	Inhaler	Nasal Spray	Patches	Control	Total
Quit	1,149	22	30	255	1,016	2,472
	(18.2%)	(15.2%)	(25.9%)	(20.5%)	(10.6%)	(14.2%)
Didn't	5,179	123	86	990	8,584	14,962
	(81.8%)	(84.8%)	(74.1%)	(79.5%)	(89.4%)	(85.8%)
Total	6,328	145	116	1,245	9,600	17,434

Overall, 14.2 percent were able to quit smoking. For those in the control groups, 10.6 percent

quit; for those using one of the four therapies, 15.2 to 25.9 percent quit. For each therapy, the differences between the NRT group and control group seem substantial. Among the four therapies, the nasal spray and inhaler seem to have considerably different success rates. Whether the other differences are substantial is perhaps debatable.

We can use a chi-square test to determine if the differences among the five categories are statistically persuasive. Here are the expected numbers if there were no relationship between the five categories and the results (quitting or not quitting):

	Gum	Inhaler	Nasal Spray	Patches	Control	Total
Quit	897.3	20.6	16.4	176.5	1,361.2	2,472
Didn't	5,430.7	124.4	99.6	1,068.5	8,238.8	14,962
Total	6,328	145	116	1,245	9,600	17,434

A comparison of the observed with the expected numbers confirms our interpretation of the column percentages. The number of people in the control group who quit smoking was considerably smaller than would be expected if the therapies were ineffective. Those using a nasal spray or patches quit much more often than would be expected; those using gum or an inhaler quit somewhat more often than would be expected.

For a formal statistical test, the chi-square value is 238.1. With $(2 - 1)(5 - 1) = 4$ degrees of freedom, the *P* value is virtually zero. Thus these studies provide overwhelming evidence that smokers using nicotine replacement therapies are more likely to quit smoking. It is far from a sure thing that someone who uses an NRT will quit smoking, but this empirical evidence indicates a substantial and statistically persuasive relationship.

Exercises

15.1 Two hundred randomly selected college students were asked to pick a number from 1 to 10. Use the data shown below to test the null hypothesis that each number is equally likely to be chosen. What pattern do you see in the responses?

Number	1	2	3	4	5	6	7	8	9	10
Responses	4	8	31	37	13	16	57	24	6	4

15.2 A study of the relationship between auto weight and accident frequency used the data below. Test the null hypothesis that the probability that an accident will be in a weight class is equal to its registration frequency.

Auto Weight (Pounds)	Registration Frequency (Percent)	Number of Accidents
Less than 3,000	21.04	162
3,000 – 3,999	46.13	318
4,000 – 4,999	31.13	689
5,000 or more	1.70	35

15.3 A student counted the number of candies of each color in a one-pound bag of plain M&Ms. Assume that these data are a random sample from the population of all M&Ms.

Color	Brown	Green	Orange	Red	Blue	Yellow
Number	154	39	48	86	61	134

Test at the 1 percent level the following null hypotheses about the population:

a. All colors are equally likely.

b. The manufacturer's claim that 30 percent of all plain M&Ms are brown, 10 percent are green, 10 percent are orange, 20 percent are red, 10 percent are blue, and 20 percent are yellow.

15.4 A study compared the birth and death dates of 193 deceased people listed in the book *Four Hundred Notable Americans* who died within 90 days of their birthday, to see whether famous people might be able to postpone death until after the celebration of a birthday. These 193 deaths were divided into six 30-day categories:

Days	–90 to –61	–60 to –31	–30 to –1	0 to 29	30 to 59	60 to 89
Number	26	23	30	32	41	41

where the –90 to –61 means that the person's death was 90-to-61 days before the birth date, and so on. Test the null hypothesis that a death is equally likely to occur in any of the 6 categories.

15.5 The random walk hypothesis states that changes in stock prices cannot be predicted from previous changes. To test this theory, a student randomly selected 128 5-day periods and, for each 5-day period, calculated the number of days that the Dow Jones Industrial Average went up. Use her data to test the null hypothesis that daily changes in the Dow Jones Industrial Average are independent binomial observations with $p = 0.5$.

Days market went up	5	4	3	2	1	0
Number of occurrences	3	27	48	31	17	2

15.6 A study investigated whether a mouse raised by a foster mother is more likely to fight with other mice than a mouse raised by its natural mother. Show the column percents and see if these data show a statistically significant relationship at the 1 percent level.

	Fighters	Nonfighters
Natural	27	140
Foster	47	93

15.7 When the Titanic sank on April 14,1912, it was carrying 1,315 passengers: 402 adult females, 805 adult males, and 108 children. Of the 498 passengers who were saved, 296 were adult females, 146 were adult males, and 56 were children. In which of these three groups was the number of people saved disproportionately large and in which was it disproportionately small? Are these patterns statistically significant at the 1 percent level?

15.8 Flu vaccines are usually effective and safe in children and adults, but less effective with infants under six months, the most vulnerable group. After a nose-drip vaccine was shown to be effective in a sample of 9,000 children and adults, doctors tested this vaccine for thirty infants aged two to five months. Six babies got a low dose, 15 got a stronger dose, and a control group of 9 infants got drops with no virus. Two-thirds of the first group and nearly ninety percent of the second group developed antibodies that protect against the flu. The control group developed no antibodies. Use a chi-square test to determine if these results are significant at the 5 percent level.

	Developed Antibodies	No Antibodies
Low Dose	4	2
Stronger Dose	13	2
Control	0	9

15.9 A University of Southern California study examined the relationship between a woman's shoe size and whether the woman frequently experienced foot pain. Do these data show a statistically significant relationship at the 5 percent level?

Shoe Size	Number Surveyed	Number With Foot Pain
Smaller than 8	150	87
Between 8 and 10	130	104
10 or larger	120	101

15.10 A sample of live births in the United States between 1969 and 1971 obtained the data below on the gender of babies in relation to their birth order. (Thus, a birth order of 1 signifies a mother's first child, while a birth order of 2 indicates her second child.) Is there a statistically significant relationship at the 1 percent level between birth order and a baby's gender?

Birth Order	Males	Females	Total
1	849,710	798,600	1,648,310
2	620,978	586,935	1,207,913
3	365,000	345,317	710,317
4	201,484	191,524	393,008
5	110,339	105,085	215,424
6	60,579	58,137	118,716
7	34,322	32,876	67,198
8 or higher	50,459	48,611	99,070

15.11 As part of an investigation of the inheritance of various traits, the athletic skills of brothers were evaluated. Test the null hypothesis that the athletic skills of brothers are independent.

	First Brother		
Second Brother	Athletic	Betwixt	Nonathletic
Athletic	906	20	140
Betwixt	20	76	9
Nonathletic	140	9	370

15.12 Yale University admitted its first women undergraduates in 1969. For the first few years, male and female students were admitted separately, to preserve Yale's tradition of "graduating 1,000 fine young men each year." However, a separate admission policy can cause disparities between the caliber of the male and female students. A University Committee on Coeducation compared the grades of male and female Yale undergraduates:

	Males	Females
Honors	6,591	1,593
High pass	7,573	1,655
Pass	3,108	539
Other	590	117

What patterns do you see in these data? Is a there a statistically significant relationship at the 5 percent level between grades and gender?

15.13 In a 1984 lawsuit (Bazemore v. Friday), racial discrimination was alleged in the merit pay raises awarded by the North Carolina Agricultural Extension Service. At the trial, the data were divided into six geographic districts:

	Pay Raise		No Pay Raise	
	White	Black	White	Black
Northcentral	47	24	12	9
Northeast	35	10	8	3
Northwest	57	5	9	4
Southeast	54	16	10	7
Southwest	59	7	12	4
West	47	0	13	0

The court analyzed six 2-by-2 contingency tables (one for each geographic district in which blacks were employed) and found that only in northwestern region was there close to statistical significance at the 5 percent level. Combine the six geographic areas to give a single 2-by-2 contingency table relating pay raises to race. Is there a statistically significant relationship at the 5 percent level?

15.14 After attempting to imitate a popular television character, a young man concluded that whether one is right-handed or left-handed affects how far apart the two middle fingers on each hand can be spread. If a wider "V" is made with the right hand, the person is probably left-handed; if a wider V is made with the left hand or there is no difference, the person is probably right-handed. After an article explaining this theory appeared in *Current Science*, 3,225 readers reported the following results:

	Test Predictions	
Readers	Right-Handed	Left Handed
Right-Handed	2,057	783
Left-Handed	56	289

Use a chi-square test to determine if there is a statistically significant relationship at the 1 percent level between handedness and the test predictions.

15.15 A student asked a random sample of 54 of her classmates whether they were nearsighted and whether their grade point average (GPA) was above or below B+:

	Nearsighted	Not Nearsighted
GPA < B+	14	15
GPA ≥ B+	17	8

Do these data reject (at the 1 percent level) the null hypothesis that GPA and nearsightedness are independent? This student then multiplied all her data by 3:

	Nearsighted	Not Nearsighted
GPA < B+	42	45
GPA ≥ B+	51	24

"In doing this, I assumed that I originally picked a perfectly random sample of students, and that if I were to have polled three times as many people, my data would have been greater in magnitude, but still distributed on the same normal distribution." Does this procedure make sense to you? How do you think it affected her chi-square value?

15.16 Part way through the 1994 National Basketball Association season, the publicity department for the San Antonio Spurs released these data on the relationship between the hair color of one of the players (Dennis Rodman) and the team's performance:

	Win	Lose
Blonde	22	3
Purple	9	2
Red or blue	6	7

Is there a statistically significant relationship at the 5 percent level? Why, even, if there is a statistically significant relationship, should we treat these results cautiously?

15.17 A study of video tapes of four of Monica Seles' Grand-Slam tennis matches categorized Seles' first serves as being to the opponent's forehand, body, or backhand and as to whether the serve was effective or ineffective. A serve was considered effective if the opponent was unable to return it or returned it so weakly that Seles' immediately hit a winning shot. Test the null hypothesis that the effectiveness of Seles' first serve is independent of whether she serves to the forehand, body, or backhand.

	Forehand	Body	Backhand
Effective	16	9	22
Ineffective	26	12	136

15.18 Twelve college basketball players took 30 shots 15 feet from the basket—ten shots from the free-throw line, ten from a 45-degree angle to the basket, and ten from the baseline. Use these data to test the null hypothesis that hits and misses are independent of location:

	Free-Throw Line	45-Degree Angle	Baseline
Hit	84	62	53
Miss	36	58	67

15.19 A study of 2,000 major league baseball games during the 1959 season tabulated the number of runs scored by inning. The data below are for the first inning of all 2,000 games and for the ninth inning of 1,576 games in which both teams batted. (The game ends if the home team is ahead after the visiting team has batted in the ninth inning.) Use these data to make a chi-square test of the null hypothesis that run scoring does not depend on whether it is the first or ninth inning of a baseball game.

	Runs Scored					
	0	1	2	3	4	> 4
First inning	1400	318	162	62	36	22
Ninth inning	1213	203	95	43	16	6

15.20 Two surveys, conducted 20 years apart, asked five questions to gauge people's trust in government (for example, "Are quite a few of the people running the government a little crooked?") Based on their responses, those surveyed were labeled as follows:

	Cynical	Middle	Trusting
1958	200	456	1,057
1978	1,198	599	438

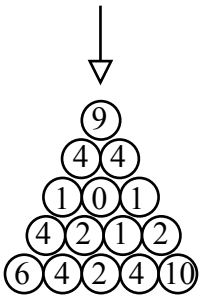
What patterns do you see in these data? Is there a statistically significant difference at the 5 percent level between the 1958 and 1978 responses?

15.21 To study social mobility in Great Britain, a 1954 sample compared the occupations of 3,497 male workers and their fathers:

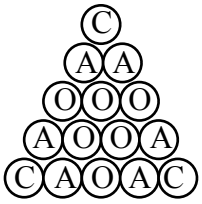
	Son's Status		
Father's Status	Upper	Middle	Lower
Upper	588	395	159
Middle	349	714	447
Lower	114	320	411

If there was no social mobility, what would the data look like? If there were unrestricted mobility, what would the data look like? Compute the column percentages and then test the null hypothesis that the occupational status of fathers and sons are independent.

15.22 Ninety-nine break shots on a full rack of billiard balls resulted in 54 sunk balls, allocated among the 15 balls as follows:

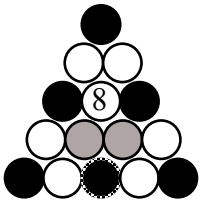


Thus the balls in the three corners of the rack were sunk 9, 6, and 10 times. Dividing the rack into three categories (corners, adjacent, and other),



is there a statistically significant relationship between a ball's position in the rack and its likelihood of being sunk on the break?

15.23 Using the data in Exercise 15.22, divide the rack according to the placement of the 8-ball, the seven striped balls (white in the diagram), and the seven solid balls (black):



The two gray balls indicate that the seventh striped ball and seventh solid ball can be placed in either position. Assume that of the three balls sunk from these positions, 1.5 was solid and 1.5 was striped. Omit the 8-ball entirely from your analysis. Is there a statistically significant relationship between a ball being solid or striped and its likelihood of being sunk on the break?

15.24 A study of severe heart attacks suffered by German workers (either on or off the job) grouped the data according to the day of the week when the heart attack occurred:

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
145	105	111	120	97	115	98

Is there a statistically significant relationship between day of the week and the incidence of severe heart attacks?

15.25 Table 15.4 shows some data on the effectiveness of four different kinds of nicotine replacement therapy (NRT). Omit the control group data so that we can concentrate on differences among the four therapies. Are the differences in effectiveness statistically significant at the 1 percent level?

15.26 A one-year study was made of injuries at a large British engineering company that used a weekly rotation of 8-hour shifts beginning at 6:00 a.m., 2:00 p.m., and 10:00 p.m. so that every worker worked the same number of morning, afternoon, and night shifts. Do the following data on the number of injuries that occurred on each shift reject at the 5 percent level the null hypothesis that the incidence of injuries does not depend on the shift?

Shift	Number of Injuries
morning	1372
afternoon	1578
night	1686

15.27 An animal-behavior specialist studied John McEnroe's behavior as he won the 1983 Wimbledon tennis tournament. Among the data gathered were whether McEnroe grunted audibly on his serve and whether the serve was an ace (winner), error, or other:

	Grunt	Silent
Ace	61	35
Error	32	8
Other	144	53

Is there a statistically significant relationship (at the 5 percent level) between his grunting and the outcomes of his serves?

15.28 Suspecting that a disproportionate number of people might be born in November, which is nine months after Valentine's Day, a survey of 148 college students obtained the following birth months:

January	7	May	7	September	11
February	15	June	12	October	12
March	16	July	8	November	21
April	6	August	16	December	17

Determine the *P* value for a test of the null hypothesis that all birth months are equally likely. Summarize your findings.

15.29 Basketball players are often told that a missed shot usually bounces long, to the opposite side of the basket from which the shot is taken; for example, a missed shot from the right side of the basket usually bounces to the left of the basket. A college player suspected that this tendency might be reversed at the end of a game, when the missed shots of tired players bounce off the front of the rim. Video tapes of four Division III college women's

basketball games were studied to see whether missed shots bounced short (towards the shooter) or bounced long (away from the shooter). The rebounds were further characterized by whether they occurred during the first or last ten minutes of the game:

	Short	Long
First ten minutes	16	25
Last ten minutes	20	10

A similar study of videotapes of three different Division III men's basketball games yielded the data below from the first and last ten minutes of each half:

	Short	Long
First ten minutes	49	39
Last ten minutes	35	46

For each set of data, determine the *P* value for a chi-square test of the null hypothesis that short and long rebounds are independent of when the shot is taken.

15.30 A 1986 study by the National Transportation Safety Board of 30 frontal automobile crashes involving 139 people used the following data:

	Killed	Injured	Uninjured
No restraint	4	49	4
Lap belt	13	36	1
Lap and shoulder belt	1	29	2

Calculate the *P* value for a test of the null hypothesis that the injury outcomes are unrelated to the restraint categories. Summarize your conclusions.

15.31 One hundred and twenty college students (62 male and 58 female) participated in a study in which they were given a sheet of paper with the following instructions:

> Please make up a story (a few sentences) creating a fictional character who fits the following theme (Please do not write about yourself): *In a large coeducational institution the average student will feel isolated in his introductory courses.*

The students were then asked to write down the name of their fictional character. In half of the cases, the instructions replaced "his introductory courses" with "his or her introductory courses." The researcher was interested in whether this change would affect the subject's choice of a male or female character. Here are the results:

	Male Subjects (62)		Female Subjects (58)	
	Male Character	Female Character	Male Character	Female Character
His	27	4	23	6
His or her	17	14	13	16

a. Why did the study ask the subjects not to write about themselves?
b. Combine the male and female subjects and make a chi-square test of the null hypothesis that the choice of a male or female character is independent of whether "his" or "his or her" are used in the instructions.
c. What patterns do you see in the results?

15.32 Explain why this conclusion is incorrect: "The chi-squared value is 5.3 with a probability of 0.22 that the values are independent of their categories. This is a low probability that does not pass the 95% confidence interval test that the data are independent."

15.33 A researcher looked at male and female choices of Coke and Diet Coke and obtained a chi-square value of 0.0383: "This means that there is only a 3.8% chance that men and women do purchase Diet Coke at the same frequency." Explain the error in this conclusion.

15.34 Infants typically begin crawling at 6-to-8 months of age. For those born in the summer and fall in cities with cold winters, crawling may be restricted and delayed by layers of clothing. To investigate this theory, the Denver Infant Study Center compiled data for 425 infants on the birth month and the age at which the infant was first able to crawl four feet in one minute. The median age at which these infants first began to crawl was 30.89 weeks. For each season of birth, they then calculated the number of infants who began crawling before 30.89 weeks and the number who began crawling later:

Season of Birth	Before 30.89 Weeks	After 30.89 Weeks
Winter (December–February)	64	51
Spring (March–May)	43	36
Summer (June–August)	44	53
Fall (September–November)	50	84

Use these data to determine the *P* value for a statistical test of the null hypothesis that there is no relationship between birth month and crawling age. Summarize your results.

15.35 Fifty college students were asked whether the amount of time that elapsed between making their beds was more than or less than 15 days. What patterns do you see in these data? Use the chi-square distribution to see if there is a statistically significant relationship at the 5 percent level between gender and bed making:

	More Than 15 Days	Less Than 15 Days
Females	18	7
Males	12	13

15.36 Use the data in the preceding exercise to make a difference-in-means test of the null hypothesis that male and female students are equally likely to allow more than 15 days to elapse before making their beds. Are your results consistent with those obtained in the preceding exercise? Are you surprised?

15.37 Twenty-three-thousand pregnant New England women were asked whether they had used a hot tub during the first two months of pregnancy. When these women gave birth, the researchers recorded whether their babies had brain and spinal cord defects caused by the failure of the neural tube to close:

	Hot Tub	No Hot Tub
Neural tube defects	7	41
No neural tube defects	1,247	21,404

Use the chi-square distribution to calculate the *P* value for a test of the null hypothesis that the incidence of neural tube defects is unrelated to hot tub use.

15.38 Seventy two randomly selected college students tasted three unlabeled chocolate chip cookies from a local grocery and identified the one they liked best. Thirteen students chose Chips Ahoy; 12 chose Lady Lee, and 47 chose supermarket bakery cookies. Are these results statistically significant at the 1 percent level? What is your null hypothesis?

15.39 The Mandarin, Cantonese, and Japanese pronunciation of "four" and "death" are almost identical, and many Chinese and Japanese people consider the number 4 to be unlucky. Californians have some say in the last four digits of their telephone numbers. A study of the last four digits of the telephone numbers of 1984 Chinese and Japanese restaurants in California counted the number of times the digits 0 – 9 appeared:

Digit	0	1	2	3	4	5	6	7	8	9
Occurrences	960	834	789	725	562	650	726	657	1332	701

Calculate the *P* value for a test of the null hypothesis that all 10 digits are equally likely.

15.40 One hundred and twenty college students who had been in a serious romantic relationship in the past two years were asked to name the month in which their most recent relationship began. The chi-square value was 18.6 and, with 11 degrees of freedom, the *P* value was 0.0686. Explain why you either agree or disagree with this interpretation of the results: "This study concludes that the data are statistically substantial because there are more than 5 degrees of freedom, and statistically significant because the *P* value is not below 0.05."

15.41 It has been argued that people may be able to postpone their deaths until after the celebration of an important event. Data were collected for 773 deceased Jewish members of Sinai Memorial Chapel who died within 4 weeks of Passover between January 1, 1987 and December 31, 1995. What is your null hypothesis, statistical test, and *P* value?

Weeks before (–) or after (+) Passover	–4	–3	–2	–1	1	2	3	4
Number of deaths	119	102	90	99	100	97	78	88

15.42 A study of injuries suffered by Pomona College varsity athletes analyzed these data:

	Females	Males
Basketball	214	344
Soccer	155	236
Swimming	161	34
Tennis	82	82
Track	151	233

Identify the patterns in these data and determine whether they are statistically persuasive.

15.43 A study of major league baseball (MLB) managers from 1871 through 2007 found that 540 managers played professional baseball before becoming a manager, with the primary positions grouped as follows: pitcher (51), catcher (116), infielder (262) and outfielder (111). Use these data to text the null hypothesis that the respective probabilities are 1/9 pitcher, 1/9 catcher, 4/9 infielder, and 3/9 outfielder. What patterns do you see in your data? Are these patterns statistically persuasive?

15.44 Eighteen college-educated adults between the ages of 20 and 40 were asked to rank the taste of samples of duck liver mousse, Spam, and Newman's Own canned dog food, with 1 best, 2 second best and, 3 third best.

Duck	1	1	3	1	1	1	1	1	1	2	2	2	1	1	1	3	1	1
Spam	3	2	2	2	2	3	2	2	2	1	1	1	2	3	2	2	2	2
Dog food	2	3	1	3	3	2	3	3	3	3	3	3	3	2	3	1	3	3

Use these data to test the null hypothesis that the ratings are random.

15.45 The manufacturer of plain M&M's claims that the overall percentages of the candies of different colors are as follows: 30 percent brown, 20 percent red, 20 percent yellow, 10 percent blue, 10 percent green, and 10 percent orange. A randomly selected 17.6 ounce bag of plain M&M's yielded the following data: 34 percent brown, 18 percent red, 21 percent yellow, 12 percent blue, 7 percent green, and 8 percent orange. The researchers used this chi-square statistic:

$$\chi^2 = \frac{(34-30)^2}{30} + \frac{(18-20)^2}{20} + \ldots + \frac{(8-10)^2}{10} = 2.48$$

With 5 degrees of freedom $P[\chi 2 > 2.48] = 0.78$. Their conclusion: "The probability of every color matching the null hypothesis was fairly high, and could not be rejected at the 5% level."

a. What is the correct null hypothesis?
b. Identify the error in their calculation of the chi-square statistic. Do not just say, "They should have used this formula." Explain why the correct formula is more sensible than the formula they used.

15.46 Explain the error in this conclusion:"The chi-squared value is 1; therefore, the probability that the data occurred by chance is 1."

15.47 A researcher believed that students whose last names begin earlier in the alphabet are given higher grades than students whose names come later in the alphabet. To test this theory, last names were divided into three groups: early (A - H), middle (I - P), and late (Q - Z). The following data were collected for students receiving A grades in a large introductory college course:

Last Name	Students	A Grades
A - H	40.5%	43.2%
I - P	35.7%	29.2%
Q - Z	23.8%	27.6%
Total	100.0%	100.0%

He calculated the chi-square value as

$$\chi^2 = \frac{(40.5-43.2)^2}{43.2} + \frac{(35.7-29.2)^2}{29.2} + \frac{(23.8-27.6)^2}{27.6} = 2.14$$

What is wrong with his chi-square calculation?

15.48 A professor showed 56 students four playing cards, numbered 1, 2, 3, and 4; shuffled them behind his back; and then selected one of the cards. While he looked at the selected card, the students tried to read his mind and wrote down their best guess as to which card was selected. Here are the results:

Card Number	Number of Students
1	10
2	16
3	27
4	3

Using the null hypothesis that each of these four numbers is equally likely to be chosen, he calculated the value of the chi-square statistic:

$$\chi^2 = \frac{\left(\frac{10}{56} - \frac{1}{4}\right)^2}{\frac{1}{4}} + \frac{\left(\frac{16}{56} - \frac{1}{4}\right)^2}{\frac{1}{4}} + \frac{\left(\frac{27}{56} - \frac{1}{4}\right)^2}{\frac{1}{4}} + \frac{\left(\frac{3}{56} - \frac{1}{4}\right)^2}{\frac{1}{4}}$$
$$= 0.3954$$

What is wrong with this calculation and why does it matter?

15.49 A researcher asked 76 college students to identify the food they grew up eating and their favorite food:

	Favorite Food				
Grew Up Eating	Asian	American	European	Latin American	Total
Asian	10	0	1	0	11
American	8	12	7	11	38
European	3	0	7	3	13
Latin American	3	2	2	7	14
Total	24	14	17	21	76

He used the binomial distribution to test these four null hypotheses:

H_0: probability[prefer Asian if grew up eating Asian] = 0.25
H_0: probability[prefer American if grew up eating American] = 0.25
H_0: probability[prefer European if grew up eating European] = 0.25
H_0: probability[prefer Latin American if grew up eating Latin American] = 0.25

What would have been a better statistical procedure? You do not need to do the procedure; just identify it.

15.50 Two professors looked at the birth months of 76 major league baseball players who committed suicide. They calculated the adjusted number of suicides in each birth month by dividing the number of suicides in that month by the total number of players with that birth month and multiplying by 1000. For example, there were a total of 638 players with January birth months; so, the adjusted January number is 1000(6/638) = 9.4, rounded off to 9. The expected value in each month is 126/12 = 10.5 and their calculated chi-square value was

$$\chi^2 = \frac{(9-10.5)^2}{10.5} + \frac{(13-10.5)^2}{10.5} + \ldots + \frac{(10-10.5)^2}{10.5} = 43.1$$

	Actual	Adjusted	Expected
January	6	9	10.5
February	7	13	10.5
March	5	8	10.5
April	5	10	10.5
May	5	9	10.5
June	6	11	10.5
July	2	3	10.5
August	19	29	10.5
September	5	8	10.5
October	7	11	10.5
November	3	5	10.5
December	6	10	10.5
Total	76	126	126

a. What is most serious error with their statistical procedure?
b. Do you think that their mistake increased or decreased the chi-square value?
c. How would you calculate the correct chi-square value?

Chapter 16

Nonparametric Tests

Chapter Outline

> *When I was a young man about to go out into the world, my father says to me a very valuable thing. He says to me like this: "One of these days in your travels, a guy is going to come up to you and show you a nice brand-new deck of cards on which the seal is not yet broken, and this guy is going to offer to bet you that he can make the jack of spades jump out of the deck and squirt cider in your ear. But, son, do not bet this man, for as sure as you stand there, you are going to wind up with an earful of cider."*
>
> —Sky Masterson
> in Damon Runyon's *Guys and Dolls*

Many statistical tests involve assumptions about the probability distribution for the data, for example, that the sampling distribution of the sample mean is normal or that the error term in the regression model is normally distributed. We can appeal to the central limit theorem and empirical histograms to support this convenient assumption, but there are situations where the samples are small or the data are clearly not normally distributed. Even if a normal distribution seems plausible, our enthusiasm may be restrained by prudent caution.

Some statistical tests are appropriate no matter what the distribution. They work for the normal distribution, but they also work for nonnormal distributions. For this reason, they are called *distribution-free* or *nonparametric* tests.

If the data are normally distributed, nonparametric tests are not as powerful as tests that use the known properties of the normal distribution. On the other hand, if the data do not come from a normal distribution, a nonparametric test has a clear advantage over a test that makes an erroneous assumption. The accuracy of tests that assume normality depends on the normality assumption. Nonparametric tests do not. For this reason, they are said to be "robust."

In this chapter, we will look at several nonparametric tests.

16.1 Sign Test for the Median

We have seen that the mean can be a misleading statistic for describing the center of a set of data. If 99 people earn $10,000 a year and one person earns $4,010,000, the $50,000 mean does not accurately describe the typical worker's pay. In such cases, the median is a popular alternative. With a bit of ingenuity, we can construct a hypothesis test for the median.

To illustrate the logic, suppose that the president of a large university claims that the median salary of professors is $100,000, a claim that cannot be directly verified since salaries are confidential. To test this claim, we can take a random salary of 20 professors and see how many earn more than $100,000 and how many earn less. If we were to find roughly half above and half below, these data would be consistent with the president's claim. On the other hand, if all 20 earn less than $100,000, this would cast considerable doubt on the president's claim. What about less extreme cases?

Our null hypothesis is that the median is $100,000 and we let the alternative hypothesis be two-sided:

$$H_0: \text{ median} = \$100{,}000$$
$$H_1: \text{ median} \neq \$100{,}000$$

Now suppose that we take a random sample of professors and discard any salaries that happen to be exactly equal to $100,000. If the null hypothesis is true and our sample is small relative to the population (not more than ten percent of the size of the population), then each observation is equally likely to be larger or smaller than $100,000.

Perhaps the sampled salaries (in $1,000s) turn out as follows:

Above $100,000		Below $100,000	
136	98	98	97
127	97	96	96
117	95	95	94
108	94	93	92
102	91	88	86

If each observation has a 0.5 probability of being larger than $100,000, we can use the binomial distribution (Chapter 12) with $p = 0.5$ to calculate the probability that the number of salaries above or below $100,000 would be as large (or larger) than the number actually observed. The binomial distribution tells us that with $p = 0.5$ and $n = 20$, the probability of 15 or more successes is 0.0207. Since our alternative hypothesis is two-sided, we double this probability to obtain the two-sided P value: $2(0.0207) = 0.0414$. Although these data do not reject the president's claim decisively, they certainly cast considerable doubt on it.

This clever procedure is called a *sign test* because it focuses on the signs of the deviations from the purported median. Those observations above the median are positive deviations, while those below are negative deviations. If the null hypothesis is true, the number of positive and negative deviations should be roughly equal. The binomial distribution tells us the probability that the number of positive or negative deviations would be as large (or larger) than the number we actually observe.

With a little ingenuity, a sign test can be applied to a variety of situations. For instance, a farmer may claim that he can predict whether rainfall in his area will be above or below average

each year. If "average" means median, then a sign test is clearly appropriate. Similarly a claim that a certain diet increases IQ scores would be confirmed statistically if the data reject the null hypothesis that the diet has no effect on IQ scores. Assuming, under the null hypothesis that IQ scores are equally likely to go up or down, a sign test can be used by having a random sample tested, placed on the diet, and then retested—and then comparing the number of scores that went up with the number that went down. Here is a third application. The null hypothesis that there is no gender discrimination in salaries could be tested by a matched-pair sample of men and women with similar qualifications and jobs. If the null hypothesis is true, then the higher paid person in any pair is equally likely to be male or female, and a sign test can be used.

16.2 Tests for Several CategoriesThe Wilcoxon Rank-Sum Test

Section 16.1 describes a one-sample nonparametric sign test for the median. The *Wilcoxon rank-sum test* is a nonparametric test for comparing two independent samples. (Another nonparametric test, the *Mann-Whitney U test*, is calculated somewhat differently but is equivalent to the Wilcoxon rank-sum test.)

To illustrate this test, consider that a grocery store has a limited amount of shelf space and thousands of products it might stock. The store manager must decide which products to carry and which to ignore. This is an important decision, because a store that carries the wrong products will lose sales and customers. It is also an important decision for product manufacturers, because a company can't sell products that customers can't find.

An additional decision for the store is where to display its products. It is no accident that supermarkets put batteries, candy, and magazines by the checkout counters, where they may remind customers that they need batteries or persuade them to make impulsive purchases. This visibility issue arises on every aisle of the store. Which items are put at the end of the aisle, where every passing shopper sees them, and which are put in the middle of the aisle, where they will be seen only by people who walk down that aisle? Which products are put at eye level, where they can be easily spotted, and which are put lower or higher, where it takes some effort to find them?

Two researchers made a formal study of the effect of shelf location on sales. In one experiment, they arranged for a certain breakfast cereal to be displayed at different shelf heights in 24 stores. In 12 stores, the selected brand was displayed at eye level; in the other 12 stores, the brand was either above or below eye level. Table 16.1 shows the rankings of these stores according to the sales of this cereal. (The two stores that tied for ranks 14 and 15 are both given the average of these two ranks, 14.5.) To avoid cluttering up this table, only the eye-level shelves are labeled; the unlabeled shelves are those not at eye level.

The two stores with the highest sales both had this cereal displayed at eye level; however, two other eye-level stores ranked 19th and 21st in sales. Overall, did the eye-level stores have a higher or lower ranking than the other stores? The *Wilcoxon rank-sum test statistic* is equal to the sum of the ranks of whichever sample we arbitrarily choose to denote as Sample 1. Here, the sum of the ranks for the eye-level stores turns out to be 130.5, while that for the other stores is 169.5:

$$\text{Eye level: sum of ranks} = 1 + 2 + \ldots + 19 + 21 = 130.5$$
$$\text{Other shelves: sum of ranks} = 3 + 5 + \ldots + 23 + 24 = 169.5$$

Table 16.1 Ranking of Cereal Sales in 24 Stores

Rank	Sales	Shelf
1	154	eye level
2	150	eye level
3	133	
4	130	eye level
5	126	
6	123	eye level
7	121	
8	112	eye level
9	111	eye level
10	109	
11	96	
12	93	
13	84	eye level
14.5	71	eye level
14.5	71	
16	67	eye level
17	62	eye level
18	58	
19	51	eye level
20	49	
21	38	eye level
22	37	
23	36	
24	27	

In general, a comparison of the average ranks tells us which sample was, on average, more highly ranked. Here,

$$\text{Eye level: average rank} = \frac{\text{sum of ranks}}{\text{sample size}} = \frac{130.5}{12} = 10.875$$

$$\text{Other shelves: average rank} = \frac{\text{sum of ranks}}{\text{sample size}} = \frac{169.5}{12} = 14.125$$

The eye-level stores tended to have lower rankings, signifying greater sales.

The statistical question is whether the observed difference in the sum of the ranks is sufficiently improbable to discredit the null hypothesis that shelf level doesn't matter. The answer requires knowledge of the the distribution of the sum of the ranks under the null hypothesis. For small samples, these probabilities can be found using statistical software or statistical reference books. If each sample has at least 10 observations and the null hypothesis is

true, the sum of the ranks for Sample 1 is approximately normally distributed with the following population mean and standard deviation.

$$\text{Sum of Sample 1 ranks} \sim \text{N}\left[\frac{n_1(n_1+n_2+1)}{2}, \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}\right] \tag{16.1}$$

where n_1 and n_2 are the number of observations in each sample.

Chapter 7 in the textbook showed how to convert a normally distributed test statistic into a standardized Z statistic:

$$Z = \frac{\text{observed statistic} - \text{null hypothesis parameter value}}{\text{standard deviation of statistic}}$$

Using Equation 16.1, the Z statistic for the Wilcoxon rank-sum test of the null hypothesis that two independent samples are drawn from identical populations is

$$Z = \frac{(\text{sum of Sample 1 ranks}) - \dfrac{n_1(n_1+n_2+1)}{2}}{\sqrt{\dfrac{n_1 n_2(n_1+n_2+1)}{12}}} \tag{16.2}$$

For the cereal-shelf study, we can let the eye-level data be Sample 1:

$$Z = \frac{130.5 - \dfrac{12(12+12+1)}{2}}{\sqrt{\dfrac{12(12)(12+12+1)}{12}}} = -1.126$$

Because the Z value is only –1.126, these data do not provide persuasive evidence against the null hypothesis that cereal sales do not depend on shelf heights: P[Z < –1.126] = 0.13, implying a two-sided P value of 2(0.13) = 0.26. (If the products on the shelves that were not at eye level are labeled Sample 1, the Z value is +1.126 and the two-sided P value is again 0.26.)

16.3 The Wilcoxon Signed-Rank Test

The Wilcoxon rank-sum Test explained in the preceding section is appropriate for comparing two independent samples. We've seen that sampling error in comparison tests can be reduced if some of the reasons for random variation can be controlled by choosing matched pairs—in essence, twins who receive identical treatments. If our data are matched pairs, we can use a nonparametric test based on a ranking of the absolute value of the differences between the matched pairs.

The 24 grocery stores analyzed in the previous section were, in fact, not two independent random samples, but rather 12 matched pairs of stores that were chosen for the similarity in their weekly sales figures. If the stores really are twins, then we can use the *Wilcoxon signed-rank test* in place of the Wilcoxon rank-sum test. Table 16.2 shows the necessary calculations. The first two columns show the sales in the 12 matched pairs of stores. The third column shows the difference in sales (eye-level sales minus other sales) and the fourth column gives the absolute values of these differences. The fifth column ranks the size of these absolute values, from 1 (smallest) to 12 (largest), and the sixth column affixes a positive or negative sign to these ranks, depending on whether the difference in sales was positive or negative. (Differences of zero are ignored completely, reducing the sample size accordingly.)

Table 16.2 Matched-Pair Analysis of Cereal Sales in Matched Stores

Eye-Level Shelf	Other Shelf	Difference in Sales	Absolute Value of Difference	Rank of Absolute Value	Signed Rank
111	71	+40	40	10	+10
150	121	+29	29	9	+9
130	133	–3	3	2	–2
154	126	+28	28	8	+8
67	93	–26	26	6	–6
112	49	+63	63	12	+12
84	109	–25	25	5	–5
123	96	+27	27	7	+7
71	27	+44	44	11	+11
62	58	+4	4	3	+3
38	36	+2	2	1	+1
51	37	+14	14	4	+4
Total					+52

The Wilcoxon signed-rank test is a nonparametric test based on the sum of the signed ranks of the absolute values of the matched-pair differences; for our example, this sum is 52, the sum of the values in the last column Table 16.2. For the null hypothesis that the two populations are identical, we can anticipate roughly half the signed ranks to be positive and half negative, with the magnitudes roughly equal too; thus Wilcoxon's sum of the signed ranks should be close to zero. A large value (either positive or negative) casts doubt on the null hypothesis. For small samples, the exact *P* value can be found using statistical software or statistical reference books. If the number of matched-pairs *n* is at least 10 and the null hypothesis is true, then Wilcoxon's sum of the signed ranks is approximately normally distributed with the following population mean and standard deviation:

$$\text{Sum of signed ranks} \sim \text{N}\left[0, \sqrt{\frac{n(n+1)(2n+1)}{6}}\right] \tag{16.3}$$

Again, we convert a normally distributed test statistic into a standardized *Z* statistic:

$$Z = \frac{\text{observed statistic} - \text{null hypothesis parameter value}}{\text{standard deviation of statistic}}$$

Using Equation 16.3, the *Z* statistic for the Wilcoxon signed-rank test of the null hypothesis that two matched-pair samples are drawn from identical populations is

$$Z = \frac{(\text{sum of signed ranks}) - 0}{\sqrt{\frac{n(n+1)(2n+1)}{6}}} \tag{16.4}$$

For the cereal-shelf data in Table 16.2,

$$Z = \frac{(\text{sum of signed ranks}) - 0}{\sqrt{\frac{n(n+1)(2n+1)}{6}}}$$

$$= \frac{52 - 0}{\sqrt{\frac{(12)(13)(25)}{6}}}$$

$$= 2.04$$

Now, P[$Z > 2.04$] = 0.0207, and the two-sided P value is 2(0.0207) = 0.0414, which is (barely) statistically significant at the 5 percent level.

This study is one of those borderline calls. If the stores were matched twins, as intended, then the data reject at the 5 percent level the null hypothesis the null hypothesis that shelf height is unimportant. If, at the other extreme, the stores are independent samples, the previous section showed the two-sided P value to be 0.26, which is not sufficient to reject the null hypothesis at the 5 percent level. The conclusion then hinges on whether we believe the stores to be matched pairs and how much we are persuaded by P values of 0.04 and 0.26. The safest decision, no doubt, is to gather additional data.

16.4 Rank Correlation Tests

In the regression model, the usual statistical tests assume that the error term is normally distributed. The reasonableness of this convenient assumption is based on the central limit theorem and a belief that the error term is the cumulative consequence of a great many separate influences.

A researcher who is reluctant to assume normality can instead work with rankings of the original data. Here, for example, are some midterm and final exam scores:

	Midterm		Final Exam	
Student	Score	Rank	Score	Rank
CC	75	3	84	1
JG	56	4	69	4
JL	25	5	49	5
SR	78	1	65	3
JV	76	2	83	2

Instead of estimating the correlation between the raw scores, we can look at the statistical correlation between the ranks:

Midterm Rank X	Final Rank Y
3	1
4	4
5	5
1	3
2	2

The correlation R between the ranks is called Spearman's rank correlation coefficient. As with any correlation coefficient, it is zero if there is no correlation and close to +1 or –1 if there

is a close association. The value of R can be determined by using the correlation formula (Equation 3.7 in Chapter 3 of the textbook) or by doing a least squares regression and taking the square root of R^2.

With ranked data, the average value of each variable is $(1 + 2 + ... + n)/n$. If there are no ties, the standard deviation is also fixed regardless of the sequence of ranks. Because of these constraints on the data, Spearman's rank correlation coefficient simplifies to

$$R = 1 - \frac{6\sum\left(X_i - Y_i\right)^2}{n\left(n^2 - 1\right)}$$

For the midterm and final exam scores

$$\sum\left(X_i - Y_i\right)^2 = (3-1)^2 + (4-4)^2 + (5-5)^2 + (1-3)^2 + (2-2)^2 = 8$$

so that

$$R = 1 - \frac{6(8)}{5\left(5^2 - 1\right)} = 0.60$$

and $R^2 = 0.60^2 = 0.36$. There is a positive association between rankings based on the midterm and final exam scores, but is hardly a perfect correlation. (For comparison, using the raw scores, the correlation is 0.82, with a R^2 of 0.675.)

For small sample like this, the exact probability distribution for R can be determined for the null hypothesis that the rankings are determined independently of each other. For $n = 5$, it turns out that R would have to exceed 0.90 to be statistically significant at the 5 percent level.

For larger samples, say at least 20, a statistical test can be based on the fact that the distribution of R is—you guessed it—approximately normal with an expected value of 0 under the null hypothesis of no correlation and a standard deviation of

$$\sigma_R = \sqrt{\frac{1}{n-1}}$$

Normality is such a plausible assumption for the regression model that researchers seldom question it. As a consequence, Spearman's rank correlation coefficient is seldom used unless the only available data are rankings. For example, instead of giving students numerical scores, a professor may rank them from best to worst. This is commonly done by graduate school admissions committees. When two people rank applicants, we can use Spearman's rank correlation coefficient to gauge the correlation between the rankings. If there isn't a strong positive correlation, then the people are using different criteria or the admissions process is pretty arbitrary!

***Sports Illustrated* Baseball Predictions**

Before the start of the major league baseball season, *Sports Illustrated* ranks the teams from best to worst. A researcher compared their 1986 rankings with the finish (as gauged by their winning percentage):

Team	SI Preseason	Actual winning %	Actual Rank
Mets	1	.667	1
Royals	2	.469	16.5
Blue Jays	3	.531	9.5
Yankees	4	.556	5
Reds	5	.531	9.5
Cardinals	6	.491	13
Tigers	7	.537	6.5
Dodgers	8	.451	19.5
Orioles	9	.451	19.5
Cubs	10	.438	23.5
A's	11	.469	16.5
Padres	12	.457	18
Twins	13	.438	23.5
Expos	14	.484	14
Braves	15	.447	21
Red Sox	16	.590	3
Mariners	17	.414	25
White Sox	18	.444	22
Phillies	19	.534	8
Angels	20	.568	4
Rangers	21	.537	6.5
Astros	22	.593	2
Brewers	23	.478	15
Indians	24	.519	11
Giants	25	.512	12
Pirates	26	.395	26

A scatter diagram is in Figure 16.1. The rank correlation coefficient is $R = 0.11$, confirming the visual impression conveyed by the scatter diagram that there was very little correlation between the *Sports Illustrated* predictions and actual results. For a formal statistical test, the standard deviation is

$$\sigma_R = \sqrt{\frac{1}{26-1}}$$
$$= 0.20$$

The Z value is

$$Z = \frac{0 - 0.11}{0.20}$$
$$= 0.55$$

and the two-sided P value is 0.58, which is not close to rejecting the null hypothesis that predictions and results are independent.

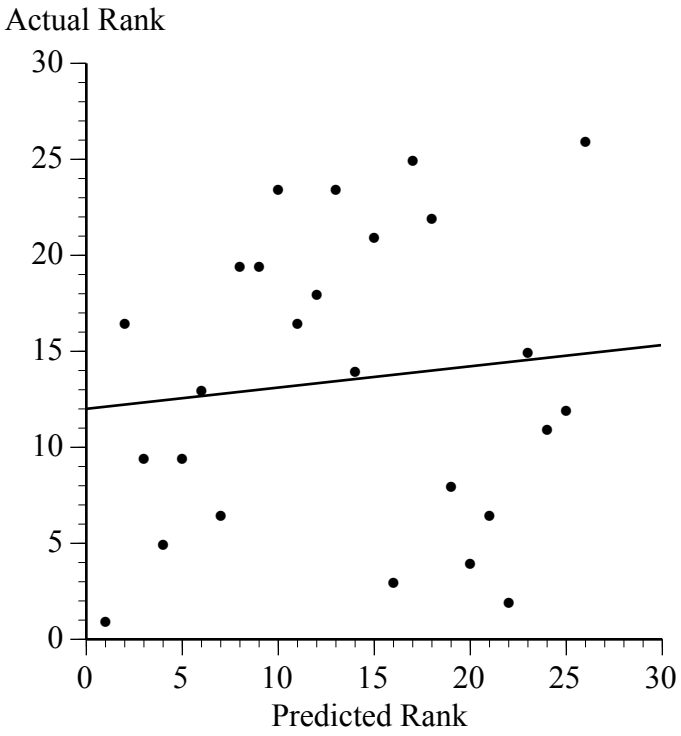


Figure 16.1 The Statistical Relationship Between Predicted Finish and Actual Finish

16.5 Runs Test

One of the foundations of probability and statistics is the Bernoulli-trial model, with its assumptions that each outcome is independent of the others. The usual analogy is a sequence of coin flips, in which the probability of heads or tails does not depend on what happened on other flips. This independence is also assumed in the central limit theorem, which underlies the normal distribution, and in the concept of a random sample, which underlies most statistical estimation and hypothesis tests.

There are many circumstances, though, in which we may suspect that the outcomes are not independent. Statisticians have devised tests for independence and I will show you one here—a *runs test*. Think of the coin-flip analogy again. I flipped a coin 20 times and here are the results:

T H H T H H T H T H H T T T T T H H T H

Heads happened to come up exactly 10 times, which is, of course, far from guaranteed. (The binomial distribution tells us that the probability of exactly 10 heads in 20 flips is 0.1762.) Ten heads in 20 flips are quite consistent with the the natural assumption that when a fair coin is fairly flipped, heads and tails are equally likely. The assumption of interest here is that the outcomes are independent, in that the chances of a head are not influenced by the outcomes of other flips.

One thing to examine would be what happened after each tail. Here, 6 tails were followed by

a head and 4 were followed by another tail—nothing suspicious there. We can also look at what happened after a head: 5 were followed by tails, 4 by heads and 1 by the end of the experiment. Again, no surprises. We could also look for more complicated patterns, such as the second flip after a tail or the third flip after a double heads. The possibilities are endless, which is not a pleasant prospect for researchers with better things to do with their time. In addition, we know that determined data grubbing will eventually discern some sort of pattern. So, if we look until we find a pattern, we will have demonstrated nothing but our stubborn determination.

What we would like is a single test that covers the most common kinds of departures from independence. Think about the question from the other side. What are the most common kinds of dependence? There may be *momentum* in that when a head appears, this increases the chances of another head. If there is momentum in coin flips, the data might look like this:

H H H H H H T T T T T T T T T H H H H H

We start with a head, which leads to heads on the next 5 throws. Then a tail appears, shifting momentum to the tails side. After 9 straight tails, momentum shifts back to heads for 5 flips.

At the other extreme—*reaction*—coin flips might obey an extreme version of the fallacious law of averages. Whenever a head comes up, a tail must follow to balance things out:

H T H T H T H T H T H T H T H T H T H T

Notice the difference between these two contrasting kinds of statistical dependence. With momentum, a run of heads is followed by a long run of tails, and then back to a run of heads: 3 runs in 20 flips. With reaction, on the other hand, each run of heads or tails only lasts 1 flip, and then the coin comes up on the opposite side: there are 20 runs. Thus, with momentum, the outcomes are characterized by a small number of long runs. With reaction, there are a large number of short runs.

A possible test for independence is to count the number of runs. If there are very few runs, the data exhibit momentum; if there are a large number of runs, reaction is shown.

With independence, the exact number of runs is a random variable that varies from one experiment to the next, but some clever statisticians figured out the probability distribution for the number of runs under the null hypothesis that the outcomes are independent. In a sequence of N observations, m are labeled positive (for example, heads) and n are labeled negative (for example, tails). If the observations are independent, then the number of runs X has an approximate normal distribution:

$$X \sim N\left[1+\frac{2mn}{N}, \sqrt{\frac{2mn(2mn-N)}{N^2(N-1)}}\right] \tag{16.5}$$

In the case of $N = 20$ independent coin flips, of which $m = 10$ are heads and $n = 10$ are tails, the expected value for the number of runs is $1 + 2(10)(10)/20 = 11$ and the standard deviation works out to be 2.18. In my 20 coin flips, there were 12 runs and the Z value is 0.459:

$$Z = \frac{X-\left(1+\frac{n}{2}\right)}{\frac{\sqrt{n-1}}{2}} = \frac{12-11}{2.18} = 0.459$$

The two-sided P value is 0.65, which is not persuasive evidence against the null hypothesis.

The runs test was not created for testing coin flips. There are many other, much more interesting situations in which there is a serious question of whether or not independence holds. In each case, we phrase the question so that it parallels the coin-flip model and then apply a runs test. For instance, it has been proposed that stock prices follow a *random walk* in that, at any moment, prices are equally likely to go up or down, regardless of previous price changes. If we think of a price increase as "heads" and a price decrease as "tails," the runs test can be applied.

A slightly more sophisticated model recognizes that stocks do not have expected returns of zero and that stock returns include price changes and dividends paid to shareholders. However, we can test for independence in the sense that whether a stock return is unusually high or low is unrelated to past returns. apply a runs test to annual stock returns, we can determine the median annual return during the period studied, so that there is an equal chance that a randomly selected return will be above or below the median. An above-median return is *heads*, a below-median return is *tails*. Then we can use a runs test to see if there is momentum or reaction in stock returns. This clever procedure can be applied to annual, daily, or hourly returns.

Runs in Stock Prices

Technical analysts claim they find profitable patterns in stock prices. By looking at where prices have been, they say they can predict where prices are going. Most finance professors dispute these claims. They argue that if any profitable pattern appears, it will be exploited out of existence. If stock prices can be counted on to go up at any point, eager buyers will jump the gun and buy in advance, and others would anticipate this jumping of the gun and buy even earlier. Still others would anticipate this anticipation, and so on. Chaos is the only sustainable pattern.

Part of the evidence against technical analysis is that they seem to make more money by selling their advice than by using it. There are so many people looking for patterns and so few who beat the market, that good luck seems to be the plausible explanation of any success they enjoy. Some other evidence against technical analysis is that technicians so often offer conflicting advice—some saying buy while others say sell. For instance, when a stock's price is rising rapidly, some technicians (those in the momentum camp) advise, "Buy. A body in motion tends to remain in motion." Yet others advise, "Sell. What goes up must come down." If there were persuasive patterns, people would surely agree what these patterns mean.

For some statistical evidence, here are the annual stock returns for the S&P 500, over the 20-year period 1961–1980:

Year	Percent Return	Year	Percent Return
1961	26.89	1971	14.31
1962	–8.73	1972	18.98
1963	22.80	1973	–14.66
1964	16.48	1974	–26.47
1965	12.45	1975	37.20
1966	–10.06	1976	23.84
1967	23.98	1977	–7.18
1968	11.06	1978	6.56
1969	–8.50	1979	18.44
1970	4.01	1980	32.42

Stock returns were positive about twice as often as they were negative. Stocks usually have a positive return. The two middle returns are 12.45 and 14.31 percent. Splitting the difference, the median return is 13.38 percent. Now we relabel the data depending on whether the annual return was above or below the median:

Year	Percent Return	Year	Percent Return
1961	+	1971	+
1962	–	1972	+
1963	+	1973	–
1964	+	1974	–
1965	–	1975	+
1966	–	1976	+
1967	+	1977	–
1968	–	1978	–
1969	–	1979	+
1970	–	1980	+

As a check, you should always count the number of positive and negative signs to see if they are, as intended, equal. Here, there are 10 positive signs and 10 negative signs.

The question of interest is whether the number of runs is unusually high or low. Equation 16.5 tells us that, with independence and $N = 20$, $m = 10$, and $n = 10$, the number of runs is approximately normally distributed with an expected value of 11 and a standard deviation of 2.18, just as in the case of 20 independent coin flips. As it turns out, there are exactly 11 runs in our annual stock return data, providing no evidence whatsoever against the null hypothesis.

On the other hand some studies have found evidence of momentum over short time periods, although this momentum has been too slight to be profitable after taxes and trading fees have been paid. (As the joke goes, the broker made money, the government made money, and two out of three isn't bad.) For some evidence of brief momentum, Table 16.3 shows 132 S&P 500 monthly returns over the period 1970 through 1980.

Table 16.3 Eleven Years of Monthly Stock Returns, percent

	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
1970	–7.43	5.86	0.30	–8.89	–5.47	–4.82	7.52	5.09	3.47	–0.97	5.36	5.84
1971	4.19	1.41	3.82	3.77	–3.67	0.21	–3.99	4.12	–0.56	–4.04	0.27	8.77
1972	1.94	2.99	0.72	0.57	2.19	–2.05	0.36	3.91	–0.36	1.07	5.05	1.31
1973	–1.59	–3.33	–0.02	–3.95	–1.39	–0.51	3.94	–3.18	4.15	0.03	–10.82	1.83
1974	–0.85	0.19	–2.17	–3.73	–2.72	–1.28	–7.59	–8.28	–11.70	16.57	–4.48	–1.77
1975	12.51	6.74	2.37	4.93	5.09	4.62	–6.59	–1.44	–3.28	6.37	3.13	–0.96
1976	11.99	–0.58	3.26	–0.99	–0.73	4.27	–0.68	0.14	2.47	–2.06	–0.09	5.40
1977	–4.89	–1.51	–1.19	0.14	–1.50	4.75	–1.51	–1.33	0.00	–4.15	3.70	0.48
1978	–5.96	–1.61	2.76	8.70	1.36	–1.52	5.60	3.40	–0.48	–8.91	2.60	1.72
1979	4.21	–2.84	5.75	0.36	–1.68	4.10	1.10	6.11	0.25	–6.56	5.14	1.92
1980	6.10	0.31	–9.87	4.29	5.62	2.96	6.76	1.31	2.81	1.87	10.95	–3.15

The median return is 0.34%, and Table 16.4 shows the months in which returns were above and below this median. Using Equation 16.5, with independence and $N = 132$, $m = 66$, and $n = 66$, the expected value of the number of runs is $1 + 2(66)(66)/132 = 67$ and the standard deviation works out to be 5.7. Table 16.4 shows that there were only 55 runs, so that the Z value is

$$Z = \frac{55 - 67}{5.7}$$
$$= -2.105$$

The two-sided P value is 0.035, which is borderline statistical evidence of momentum.

Whether a trading strategy based on momentum (and taking transaction costs and taxes into account) is more profitable than a buy-and-hold strategy is a separate question.

Table 16.4 Eleven Years of Monthly Stock Returns, above or below median

	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
1970	–	+	–	–	–	–	+	+	+	–	+	+
1971	+	+	+	+	–	–	–	+	–	–	–	+
1972	+	+	+	+	+	–	+	+	–	+	+	+
1973	–	–	–	–	–	–	+	–	+	–	–	+
1974	–	–	–	–	–	–	–	–	–	+	–	–
1975	+	+	+	+	+	+	–	–	–	+	+	–
1976	+	–	+	–	–	+	–	–	+	–	–	+
1977	–	–	–	–	–	+	–	–	–	–	+	+
1978	–	–	+	+	+	–	+	+	–	–	+	+
1979	+	–	+	+	–	+	+	+	–	–	+	+
1980	+	–	–	+	+	+	+	+	+	+	+	–

Exercises

16.1 Exercise 6.30 in the textbook shows the oral temperatures of 50 healthy adults. Use these data to calculate the two-sided P value for a test of the null hypothesis that the median temperature of healthy adults is 98.6.

16.2 Use the data in Exercise 7.19 in the textbook to determine the two-sided P value for a test of the null hypothesis that the population median is zero.

16.3 Use the data in Exercise 7.20 in the textbook to determine the two-sided P value for a test of the null hypothesis that the population median is zero.

16.4 At graduation, a college president announces that the median starting salary for graduating seniors who have taken jobs is $40,000. A dubious senior uses a random sample of 17 classmates to test this claim. These 17 students report salaries (in $1,000s) of 35, 28, 46, 23, 37, 38, 27, 37, 48, 43, 39, 54, 42, 38, 43, 37, and 38. Do these data support the president's claim?

16.5 (continuation) Is there any reason to think that the answers in this kind of survey might be biased one way or another? If there is a bias, will it increase or decrease the chances of

incorrectly rejecting the null hypothesis? Some people decline to answer salary surveys. Could these nonresponses bias the results or can they be safely ignored?

16.6 Twenty salaries were used in this chapter to test the null hypothesis that the median professor's salary is $100,000. Use these data to make a *t* test of the null hypothesis that the mean salary is $100,000. Is your conclusion the same as with the median? If there is a difference, how would you reconcile the competing results?

16.7 Below are birth weight data for a random sample of black and white babies born in a suburban hospital. Make a rank sum test of the null hypothesis that these weights came from the same distribution.

Black	White
7 lb, 3oz	7 lb, 10 oz
5 lb, 15 oz	6 lb, 14 oz
6 lb, 12 oz	5 lb, 3 oz
7 lb, 7 oz	8 lb, 12 oz
5 lb, 5 oz	7 lb, 2 oz
8 lb, 2 oz	7 lb, 8 oz
5 lb, 10 oz	7 lb, 10 oz
6 lb, 15 oz	6 lb, 2 oz
7 lb, 3 oz	7 lb, 13 oz
5 lb, 7 oz	7 lb, 5 oz
6 lb, 8 oz	6 lb, 5 oz
7 lb, 12 oz	8 lb, 9 oz
6 lb, 5 oz	6 lb, 12 oz
7 lb, 2 oz	8 lb, 6 oz
7 lb, 9 oz	7 lb, 14 oz

16.8 Two vice-presidents ranked ten prospective new products.

Product	Bob's Ranking	Ray's Ranking
1	4	2
2	2	1
3	9	8
4	7	7
5	8	9
6	3	4
7	10	10
8	5	5
9	1	3
10	6	6

Use Spearman's rank correlation coefficient to gauge the correspondence between these rankings. Interpret your results.

16.9 Rank the unemployment and voting data in Table 8.13 in the textbook and compute Spearman's rank correlation coefficient. Is the correlation between these rankings statistically significant at the 5 percent level?

16.10 In a cloud-seeding experiment, 26 of 52 clouds were randomly selected for seeding with silver nitrate, and the rainfall (in acre-feet) was recorded. Use the Wilcoxon rank-sum statistic to test the null hypothesis that these data are from identical distributions.

Seeded					Unseeded				
2745.6	430.0	242.5	115.3	7.7	1202.6	163.0	47.3	26.1	4.9
1697.8	334.1	200.7	92.4	4.1	830.1	147.8	41.1	24.4	1.0
1656.0	302.8	198.6	40.6		372.4	95.0	36.6	21.7	
978.0	274.7	129.6	32.7		345.5	87.0	29.0	17.3	
703.4	274.7	119.0	31.4		321.2	81.2	28.6	11.5	
489.1	255.0	118.3	17.5		244.3	68.5	26.3	4.9	

16.11 The National Transportation Safety Administration crashed twenty 1988–1991 medium-size cars into a wall at 35 miles per hour and measured the head injuries to dummies in the driver and front-passenger seats. Use the Wilcoxon signed-rank statistic to test the null hypothesis that these data are from identical distributions.

	Driver	Passenger		Driver	Passenger
Acura Legend LS	435	618	Lexus ES250	992	630
Audi 100	185	710	Mazda 929	273	858
Buick Park Ave	1467	794	Mercedes 190E	800	833
Chevrolet Camaro	585	583	Mercury Sable	712	410
Chevrolet Lumina	1200	0	Mitsubishi Starion	952	377
Chrysler Le Baron	298	2043	Nissan 300zx	765	0
Eagle Premier	877	868	Oldsmobile Delta 88	710	539
Ford Taurus	480	258	Peugeot 505S	1983	2192
Honda Accord Lx	562	539	Toyota Cressida	790	554
Infiniti M-30	466	443	Volvo 740 GLE	519	445

16.12 Table 7.1 in the textbook shows the matched-pair prices of inexpensive male and female haircuts in 20 stores. Calculate the sum of the signed ranks and determine the two-sided P value for a test of the null hypothesis that these data are from identical distributions.

16.13 Here are data on the interest rates on U. S. 3-month Treasury bills. Find the median interest rate and then identify monthly levels that are above (+) or below (–) this median. Count the number of runs and calculate the two-sided P value for a test of the null hypothesis that monthly Treasury-bill rates are independent.

	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
2006	4.24	4.43	4.51	4.60	4.72	4.79	4.95	4.96	4.81	4.92	4.94	4.85
2007	4.98	5.03	4.94	4.87	4.73	4.61	4.82	4.20	3.89	3.90	3.27	3.00
2008	2.75	2.12	1.26	1.29	1.73	1.86	1.63	1.72	1.13	0.67	0.19	0.03
2009	0.13	0.30	0.21	0.16	0.18	0.18	0.18	0.17	0.12	0.07	0.05	0.05
2010	0.06	0.11	0.15	0.16	0.16	0.12	0.16	0.16	0.15	0.13	0.14	0.14

16.14 (continuation) Now compute the monthly change in the Treasury bill rate. Find the median change and then identify monthly changes that are above (+) or below (–) this median. Count the number of runs and calculate the two-sided P value for a test of the null hypothesis that monthly changes in the Treasury-bill rate are independent.

16.15 A home inventor claims to have perfected a gizmo that increases miles per gallon (MPG). Tests are conducted involving two similar cars, one with the gizmo and one without, driven over the same route at the same time and at approximately the same speed. Ten tests were conducted, with a variety of car models, roads, weather conditions, and speeds:

Test	MPG With Gizmo	MPG Without
1	27	24
2	32	30
3	26	29
4	18	22
5	19	26
6	38	34
7	33	37
8	45	43
9	32	28
10	36	35

Assuming these are matched pairs,

a. Calculate the sum of the signed ranks and determine the two-sided P value for a test of the null hypothesis that the gizmo has no effect on gasoline mileage.
b. Calculate the t value and two-sided P value for a test of the null hypothesis that the expected value of the matched-pair difference in gasoline mileage is equal to 0.

16.16 A sample of ten people rated two colas on a scale of 1 to 10:

Person	Rating of First Cola	Rating of Second Cola	Difference
1	8	6	2
2	2	3	–1
3	6	7	–1
4	10	7	3
5	9	6	3
6	2	1	1
7	8	6	2
8	10	8	2
9	1	2	–1
10	6	4	2

Assuming these are matched pairs,

a. Calculate the sum of the signed ranks and determine the two-sided P value for a test of the null hypothesis that the ratings come from the same population.
b. Calculate the t value and two-sided P value for a test of the null hypothesis that the expected value of the matched-pair difference is equal to 0.

16.17 The deans at two business schools compared the starting salaries of their 2011 graduates. The first school had 120 graduates in 2011; the second had 225. A ranking of their salaries found that the average rank for the first school was 154.5. According to these salary ranks, which school's graduates did better in the job market? Make a rank sum test of the null hypothesis that the starting salaries are drawn from the same distribution.

16.18 Thirteen patients with advanced stomach cancer and 11 patients with advanced breast cancer were treated with ascorbate. Use the Wilcoxon rank-sum statistic to test the null hypothesis that the survival times (in days) are from identical populations.

Stomach			Breast		
124	412	103	1235	727	719
42	51	876	24	3808	
25	1112	146	1581	791	
45	46		1166	1804	
340	396		40	3460	

16.19 A Stanford University team examined changes in the triglyceride level in 30 blood samples that had been in frozen storage for eight months. Use the Wilcoxon rank-sum statistic to see if the observed differences are statistically significant at the 5% level.

Before	After	Before	After	Before	After	Before	After	Before	After
74	66	177	185	126	133	83	81	131	127
80	85	88	96	72	69	79	74	228	227
75	71	85	76	301	302	194	192	115	129
136	132	267	273	99	106	124	129	83	81
104	103	71	73	97	94	42	48	211	212
102	103	174	172	71	67	145	148	169	182

16.20 For each of the following sets of data, indicate whether you would expect to find a large or small number of runs.

a. whether cars passing a certain point are traveling faster than 55 miles per hour.
b. a weekly win or loss by a college football team.
c. which team scores in a college basketball game.
d. daily pollution readings in Los Angeles.
e. the monthly unemployment rate.

16.21 Do you think that the success of political parties exhibits momentum, reaction, or independence? Apply a runs test to the political parties of these 23 U. S. presidential election winners:

Year	Winner	Year	Winner
1920	Warrne Harding, Republican	1968	Richard Nixon, Republican
1924	Calvin Coolidge, Republican	1972	Richard Nixon, Republican
1928	Herbert Hoover, Republican	1976	Jimmy Carter, Democrat
1932	Franklin Roosevelt, Democrat	1980	Ronald Reagan, Republican
1936	Franklin Roosevelt, Democrat	1984	Ronald Reagan, Republican
1940	Franklin Roosevelt, Democrat	1988	George H. W. Bush, Republican
1944	Franklin Roosevelt, Democrat	1992	Bill Clinton, Democrat
1948	Harry Truman, Democrat	1996	Bill Clinton, Democrat
1952	Dwight Eisenhower, Republican	2000	George W. Bush, Republican
1956	Dwight Eisenhower, Republican	2004	George W. Bush. Republican
1960	John F. Kennedy, Democrat	2008	Barack Obama, Democrat
1964	Lyndon Johnson, Democrat		

16.22 Write down the results of 20 imaginary coin flips, Now use a runs test to see if these imaginary flips were statistically independent.

16.23 The deans of two small engineering schools compared a random sample of the starting salaries of their 1991 graduates who majored in engineering:

Eastern Institute of Technology	Western Institute of Technology
$45,000	$42,000
42,500	42,000
40,000	41,000
39,000	38,500
38,000	36,000
37,500	35,000
37,000	33,000
36,000	33,000
35,000	32,000
34,000	30,000
32,500	
31,000	
30,000	
30,000	

Make a rank sum test of the null hypothesis that the starting salaries are drawn from the same distribution.

16.24 Do the following annual unemployment data (for 1961 through 1990, reading left to right) indicate momentum of reaction? Calculate the two-sided P value for a runs test of the null hypothesis that annual unemployment rates are independent.

6.5	5.4	5.5	5.0	4.4	3.7	3.7	3.5	3.4	4.8
5.8	5.5	4.8	5.5	8.3	7.6	6.9	6.0	5.8	7.0
7.5	9.5	9.5	7.4	7.1	6.9	6.1	5.4	5.2	5.4

16.25 Construct 10 hypothetical observations to show how the level of the Treasury bill rate could have very few runs, while the change in the Treasury bill rate has a great many runs. (You do not need to do a formal statistical test; but you do need to show that, in each case, the number of runs is above or below the expected value if the data were independent.)

16.26 A count of the number of runs in the daily percentage changes in the Dow Jones industrial stock average during two great bull markets yielded the data shown below. In each case, compute the expected number of runs under the null hypothesis that the daily percentage changes are independent and use the difference between the actual and expected number of runs to test this null hypothesis.

	Number of Observations	Number of Runs
January 3, 1928 – September 3, 1929	495	233
January 2, 1986 – August 25, 1987	417	220

16.27 During the college football season, the Associated Press compiles weekly rankings of the top teams. Shown below are its preseason and postseason rankings of ten traditional powers during one randomly selected year (1973). Use these data to test the null hypothesis that there is no correlation between the Associated Press's preseason and postseason rankings.

	Preseason	Postseason
1.	Southern California	Notre Dame
2.	Ohio State	Ohio State
3.	Texas	Oklahoma
4.	Nebraska	Alabama
5.	Michigan	Penn State
6.	Alabama	Michigan
7.	Penn State	Nebraska
8.	Notre Dame	Southern California
9.	UCLA	UCLA
10.	Oklahoma	Texas

16.28 Twenty random digits generated by a computer are listed below. Letting the even numbers be + and the negative numbers –, make a runs test of their independence.

5	2	4	8	1	7	6	6	8	5
3	9	0	1	4	9	8	4	7	7

16.29 (continuation) Letting the numbers larger than 4 be + and the numbers less than 5 be minus –, make a runs test of their independence.

16.30 Redo Exercise 16.28, this time using a computer random number generator to generate twenty random digits.

16.31 Redo Exercise 16.29, obtaining your twenty random digits from a computer random number generator.

16.32 Redo Exercise 16.28, this time using a computer random number generator to generate 50 random digits. If the random number generator is functioning as intended, is a runs test at the 1 percent level more likely to reject independence when 20 digits are tested or 50?

16.33 A taxi manufacturer claims that the median number of breakdowns per year for its cars is two. A random sample of 84 taxis finds

Number of breakdowns	0	1	2	3	>3
Number of taxis	6	11	34	29	9

Use these data for a sign test of the null hypothesis that the median is two.

16.34 A Chief Executive Officer feels that sales are "stop and go," with intermittent lulls and hectic activity. Here are some daily sales data (in dollars), beginning on a randomly selected day and continuing for 20 days (in this order, 900, 950, ...):

900	950	840	1010	1090	1030	1210	1050	1110	970
850	930	980	910	820	940	1040	1180	1230	1060

a. Find the median and identify which days are above and which below the median.
b. How many runs are there? Does this suggest momentum or reaction?
c. Make a runs test of the null hypothesis that the outcomes are independent.

16.35 Below are some data on annual income (in $1000) and education (highest year completed). Calculate the rank correlation coefficient. Is its value statistically significant at the 5 percent level? What is the null hypothesis?

Income:	28	44	30	50	46	37	34	70	38	40	35	60	45	55
Education:	8	9	10	12	12	12	12	12	13	16	16	16	18	20

16.36 The scores that Soviet and American judges gave an American gymnast are shown below. Make a rank sum test of the hypothesis that these scores come from the same distribution.

Soviet	American
7.5	9.0
9.0	9.5
8.0	9.5
8.5	10.0
7.5	9.5
8.0	9.5

16.37 From 1950 through 2010, American League baseball teams won the Major League Baseball World Series championship 33 times and National League teams won 27 times. (No World Series was played in 1994 because of a strike) Here are the year-to-year results, reading from left to right:

A	A	A	A	N	N	A	N	A	N	N	A	A	N	N
N	A	N	A	N	A	N	A	A	A	N	N	A	A	N
N	N	N	A	A	A	N	A	N	A	N	A	A	A	N
A	N	A	A	A	N	A	N	A	A	N	A	N	A	N

Calculate the number of runs and the expected number of runs if these were 60 independent trials. Does the actual number of runs suggest momentum or reaction in World Series results? Is the difference between the actual and expected number of runs statistically significant at the 5 percent level?

16.38 (continuation) Do you think that daily or annual weather data are more likely to be reject the null hypothesis of independence? Explain your reasoning, with an example if possible.

16.39 A researcher examined the Friday prices of ten stocks over the 153-week period from August 8, 1986 to July 7, 1989. For each of these ten stocks, 152 weekly prices changes were computed, the median price change was determined, and each week was labeled + or – depending on whether its price change was above or below the median. The number of statistical runs were then determined, with the results shown below. What is the expected value of the number of runs for each stock if the weekly price changes are independent? Do any of these stocks show a statistically significant (at the 1 percent level) departure from independence?

Stock	Number of Runs
American South African	72
Burlington Northern	74
Dow Chemical	81
Georgia Pacific	80

Greyhound	68
Louisiana Land and Exploration	81
Pittston	80
Southern Company	73
Tandy	78
Unocal	75

16.40 Shown below are several 1989 preseason and post-season ratings of college football teams. (In the *New York Times* post-season rankings, Alabama and Southern Cal are tied and so are Brigham Young and Michigan State.) Compare the *Times* and Associated Press preseason rank ordering of the twenty teams ranked by the *Times*. (Ignore the five teams that are ranked by the Associated Press, but not by the *Times*, so that in each case the ranks range from 1 to 20.) Gauge the association between these twenty rank orderings by calculating Spearman's correlation coefficient. Is the value of R statistically significant at the 5 percent level? At the 1 percent level?

	Pre-season		Post-season	
Rank	New York Times	Associated Press	New York Times	Associated Press
1	Michigan	Michigan	Miami (Fla.)	Miami (Fla.)
2	Nebraska	Notre Dame	Notre Dame	Notre Dame
3	Miami (Fla.)	Nebraska	Florida State	Florida State
4	Notre Dame	Miami (Fla.)	Auburn	Colorado
5	Southern Cal	Southern Cal	Colorado	Tennessee
6	Louisiana State	Florida State	Houston	Auburn
7	Brigham Young	Louisiana State	Tennessee	Michigan
8	Florida State	Auburn	Michigan	Southern Cal
9	Auburn	U.C.L.A.	Clemson	Alabama
10	U.C.L.A.	Arkansas	Alabama	Illinois
11	Arizona	Penn State	Southern Cal	Nebraska
12	Ohio State	Clemson	Penn State	Clemson
13	Arkansas	Syracuse	Nebraska	Arkansas
14	Clemson	Colorado	Illinois	Houston
15	Alabama	Oklahoma	Arkansas	Penn State
16	Syracuse	Alabama	Texas Tech	Michigan State
17	Colorado	West Virginia	Washington	Pittsburgh
18	Illinois	Arizona	Virginia	Virginia
19	Penn State	Brigham Young	Brigham Young	Texas Tech
20	Oklahoma	Pittsburgh	Michigan State	Texas A&M
21		Houston		West Virginia
22		Illinois		Brigham Young
23		Iowa		Washington
24		North Carolina		Ohio State
25		Ohio State		Arizona

16.41 Repeat Exercise 16.40, this time comparing the post-season *New York Times* and Associated Press rank orderings of the top 20 teams ranked by the *Times*. Remember that the *Times* rankings have two ties.

16.42 Use the data in Exercise 16.40 to compare the preseason and post-season *New York Times* rank orderings of the top ten teams in the *Times* preseason rankings. Assume that U.C.L.A. had the lowest post season ranking among these ten teams and that Louisiana State had the next lowest ranking. (Ignore the other teams, so that in each case the ranks range from 1 to 10.) Calculate Spearman's rank correlation coefficient for these two sets of ten rankings. Is this value of R statistically significant at the 5 percent level? At the 1 percent level?

Now compare the preseason and post-season Associated Press rank orderings of the top ten teams in the Associated Press preseason rankings. Again assume that U.C.L.A. had the lowest post-season ranking among these ten teams and that Louisiana State had the next lowest ranking. Is the preseason/post-season value of R higher for the *New York Times* or the Associated Press?

16.43 (continuation) Why is the procedure used in Exercise 16.42, which focuses on the top ten teams in the preseason poll, an imperfect measure of the correlation between preseason and post-season polls? Give an example of preseason and post-season poll results for which the Spearman rank correlation coefficient is 1.0, yet most would consider the preseason poll a poor predictor of the post-season poll?

16.44 Each of the students in an investments class used newspaper and magazine articles to pick a stock to follow from September 9 to December 4, 1987. During this period, the stock market as a whole fell 28.4%. Twelve of these 38 students picked stocks that did better than this, while 26 picked stocks that did worse. Overall, the student stocks had an average return of –34.7% with a standard deviation of 13.9%.

Use the sample mean to test the null hypothesis that the student stocks were randomly selected from a distribution with a mean of –28.4%. Use the sample median to test the null hypothesis that the student stocks were randomly selected from a distribution with a median of –28.4%.

16.45 The rankings of ten large cities on per capita sales of two types of novels are shown below. Compute Spearman's rank correlation coefficient and interpret your result.

city	historical	occult
New York	1	2
Chicago	5	8
Los Angeles	2	1
Philadelphia	9.5	4
Houston	3	5.5
Detroit	7.5	10
Dallas	6	8
San Diego	9.5	5.5
Phoenix	7.5	3
Baltimore	4	8

16.46 Here are ten years of monthly percentage changes in the U. S. Consumer Price Index (CPI). (The value 0.23 is 23 hundredths of a percent.) The median is 0.215. Identify monthly changes that are above (+) or below (–) the median. Count the number of runs and calculate the two-sided *P* value for a test of the null hypothesis that monthly changes in the CPI are independent.

	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
2001	0.23	0.06	0.17	0.51	0.23	-0.17	0.00	0.39	-0.28	-0.06	-0.06	0.17
2002	0.17	0.28	0.45	0.11	0.06	0.22	0.28	0.17	0.22	0.17	0.17	0.44
2003	0.55	0.16	-0.38	-0.16	0.11	0.33	0.44	0.33	-0.11	0.05	0.27	0.43
2004	0.21	0.21	0.16	0.43	0.37	0.11	0.05	0.32	0.53	0.47	0.00	-0.05
2005	0.42	0.36	0.31	-0.05	0.05	0.62	0.62	1.38	0.15	-0.50	0.00	0.61
2006	0.05	0.15	0.50	0.30	0.25	0.55	0.44	-0.49	-0.44	0.05	0.54	0.14
2007	0.42	0.49	0.37	0.39	0.17	0.24	0.02	0.37	0.33	0.79	0.29	0.36
2008	0.24	0.37	0.31	0.55	0.90	0.87	-0.16	0.03	-0.88	-1.81	-0.79	0.27
2009	0.46	-0.14	0.11	0.13	0.68	0.12	0.34	0.20	0.23	0.24	0.09	0.14
2010	0.05	0.02	0.01	-0.14	-0.21	0.35	0.21	0.16	0.25	0.12	0.43	0.40

16.47 Exercise 8.8 in the textbook shows annual data on U.S. housing starts. Calculate the change in housing starts each year and calculate the two-sided *P* value for a runs test of the null hypothesis that annual changes in housing starts are independent.

16.48 Here are the data used in Figure 9.8 in the textbook for nine cigarette brands on each brand's percentage of total cigarette advertising and percentage of total cigarettes smoked by persons 12–18 years of age:

	Advertising (%)	Teenage Market (%)
Marlboro	59.5	12.7
Camel	8.7	4.9
Newport	11.1	4.7
Salem	3.8	5.4
Kool	4.3	4.8
Benson & Hedges	2.1	5.3
Winston	2.5	6.9
Merit	0.9	6.3
Virginia Slims	2.2	4.8

a. Calculate the simple correlation and the two-sided *P* value (this is equal to the two-sided *P* value for testing the null hypothesis that the slope in a linear regression is 0).
b. Calculate Spearman's rank correlation coefficient and the two-sided *P* value.
c. Identify and explain any substantive differences in your answers to questions a and b.

16.49 Exercise 8.20 in the textbook shows the prepregnancy weights of 25 mothers and the birth weights of their babies. Calculate Spearman's rank correlation coefficient. Is this value of *R* statistically significant at the 5 percent level? At the 1 percent level?

16.50 Exercise 8.25 in the textbook shows recruiting and performance ratings for 20 college football teams. Calculate Spearman's rank correlation coefficient and the two-sided *P* value.