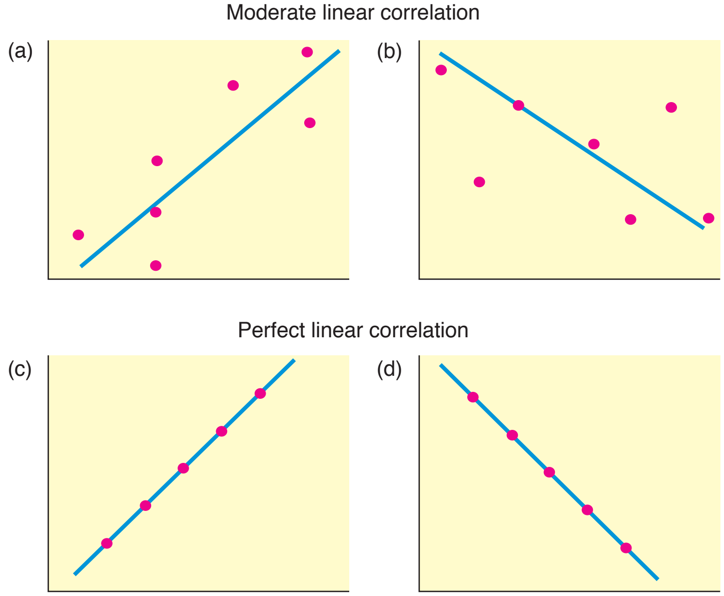
FIGURE 9-4

Scatter Diagrams with Moderate and Perfect Linear Correlation



Scatter diagrams for the same data look different from one another when they are graphed using different scales. Problem 19 at the end of this section explores how changing scales affect the look of a scatter diagram.

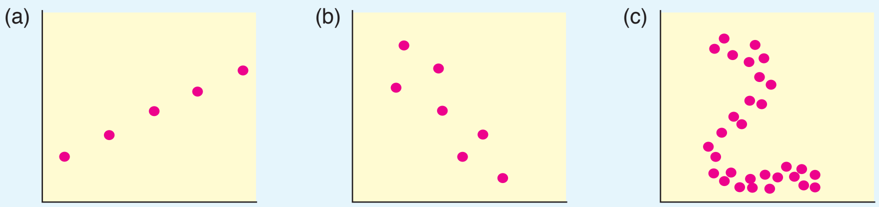
values of y. Figure 9-4 parts (a) and (c) show scatter diagrams in which the variables are positively correlated. On the other hand, if low values of x are associated with high values of y and high values of x are associated with low values of y, the variables are said to be *negatively correlated*. Figure 9-4 parts (b) and (d) show variables that are negatively correlated.

Negative correlation

GUIDED EXERCISE 2 Scatter diagram and linear correlation

Examine the scatter diagrams in Figure 9-5 and then answer the following questions.

FIGURE 9-5 Scatter Diagrams



(a) Which diagram has no linear correlation? → Figure 9-5(c) has no linear correlation. No straight-line fit should be attempted.

(b) Which has perfect linear correlation? → Figure 9-5(a) has perfect linear correlation and can be fitted exactly by a straight line.

(c) Which can be reasonably fitted by a straight line? → Figure 9-5(b) can be reasonably fitted by a straight line.

TECH NOTES

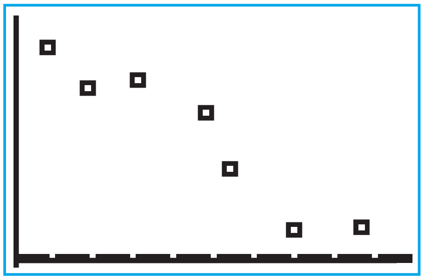
The TI-84Plus/TI-83Plus/TI-*n*spire calculators, Excel 2007, and Minitab all produce scatter plots. For each technology, enter the x values in one column and the corresponding y values in another column. The displays on the next page show the data from Guided Exercise 1 regarding safety training and hours lost because of accidents. Notice that the scatter plots do not necessarily show the origin.

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) Enter the data into two columns. Use **Stat Plot** and choose the first type. Use option **9: ZoomStat** under **Zoom**. To check the scale, look at the settings displayed under **Window**.

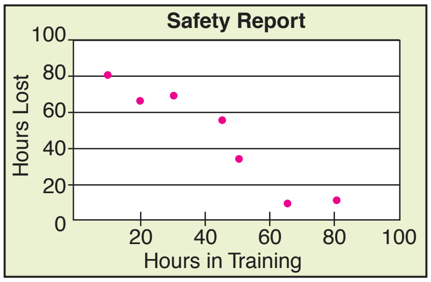
Excel 2007 Enter the data into two columns. On the home screen, click the **Insert** tab. In the Chart Group, select **Scatter** and choose the first type. In the next ribbon, the Chart Layout Group offers options for including titles and axes labels. Right clicking on data points provides other options such as data labels. Changing the size of the diagram box changes the scale on the axes.

Minitab Enter the data into two columns. Use the menu selections **Stat ➤ Regression ➤ Fitted Line Plot**. The best-fit line is automatically plotted on the scatter diagram.

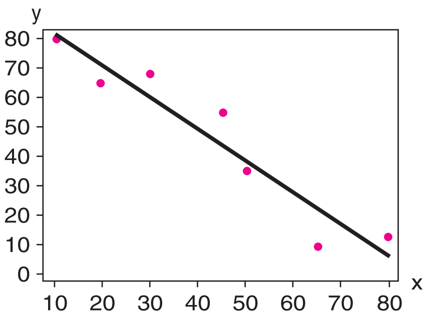
TI-84Plus/TI-83Plus/TI-*n*spire Display



Excel Display



Minitab Display



Sample Correlation Coefficient *r*

Looking at a scatter diagram to see whether a line best describes the relationship between the values of data pairs is useful. In fact, whenever you are looking for a relationship between two variables, making a scatter diagram is a good first step.

Sample correlation coefficient *r*

There is a mathematical measurement that describes the strength of the linear association between two variables. This measure is the *sample correlation coefficient r*. The full name for r is the *Pearson product-moment correlation coefficient*, named in honor of the English statistician Karl Pearson (1857–1936), who is credited with formulating r.

The **sample correlation coefficient** r is a numerical measurement that assesses the strength of a *linear* relationship between two variables x and y.

1. r is a unitless measurement between -1 and 1. In symbols, $-1 \le r \le 1$. If $r = 1$, there is perfect positive linear correlation. If $r = -1$, there is perfect negative linear correlation. If $r = 0$, there is no linear correlation. The closer r is to 1 or -1, the better a line describes the relationship between the two variables x and y.
2. Positive values of r imply that as x increases, y tends to increase. Negative values of r imply that as x increases, y tends to decrease.
3. The value of r is the same regardless of which variable is the explanatory variable and which is the response variable. In other words, the value of r is the same for the pairs (x, y) and the corresponding pairs (y, x).
4. The value of r does not change when either variable is converted to different units.

We'll develop the defining formula for r and then give a more convenient computation formula.

Development of Formula for *r*

If there is a *positive* linear relation between variables x and y, then high values of x are paired with high values of y, and low values of x are paired with low values of y. [See Figure 9-6(a).] In the case of *negative* linear correlation, high values of x are paired with low values of y, and low values of x are paired with high values of y. This relation is pictured in Figure 9-6(b). If there is *little or no linear correlation* between x and y, however, then we will find both high and low x values sometimes paired with high y values and sometimes paired with low y values. This relation is shown in Figure 9-6(c).

These observations lead us to the development of the formula for the sample correlation coefficient r. Taking *high* to mean "above the mean," we can express the relationships pictured in Figure 9-6 by considering the products

$$(x - \bar{x})(y - \bar{y})$$

If both x and y are high, both factors will be positive, and the product will be positive as well. The sign of this product will depend on the relative values of x and y compared with their respective means.

$$(x - \bar{x})(y - \bar{y}) \begin{cases} \text{is positive if } x \text{ and } y \text{ are both "high"} \\ \text{is positive if } x \text{ and } y \text{ are both "low"} \\ \\ \text{is negative if } x \text{ is "low," but } y \text{ is "high"} \\ \text{is negative if } x \text{ is "high," but } y \text{ is "low"} \end{cases}$$

In the case of positive linear correlation, most of the products $(x - \bar{x})(y - \bar{y})$ will be positive, and so will the sum over all the data pairs

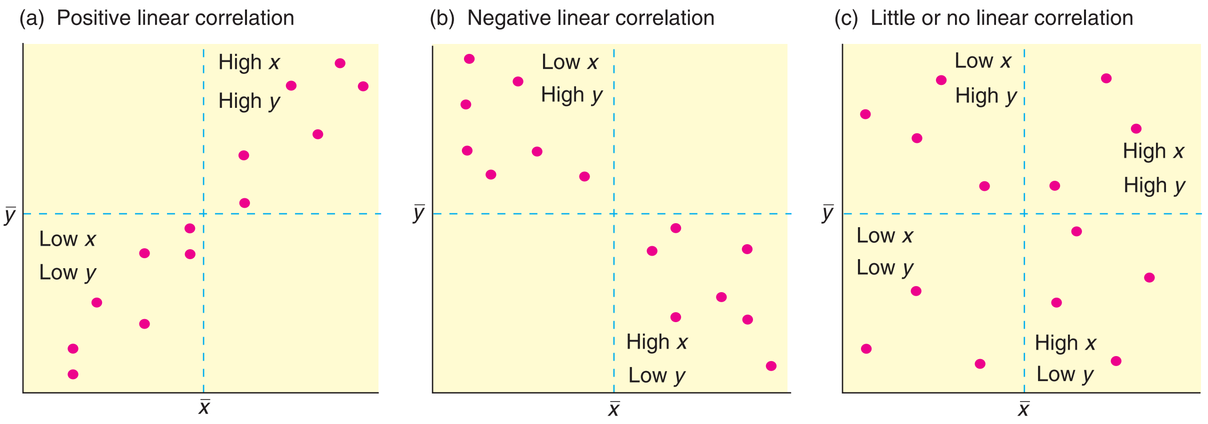
$$\Sigma(x - \bar{x})(y - \bar{y})$$

For negative linear correlation, the products will tend to be negative, so the sum also will be negative. On the other hand, in the case of little, if any, linear correlation, the sum will tend to be zero.

One trouble with the preceding sum is that it increases or decreases, depending on the units of x and y. Because we want r to be unitless, we standardize both x and y of a data pair by dividing each factor $(x - \bar{x})$ by the sample standard deviation s_x and each factor $(y - \bar{y})$ by s_y. Finally, we take an average of all the

FIGURE 9-6

Patterns for Linear Correlation



products. For technical reasons, we take the average by dividing by $n - 1$ instead of by n. This process leads us to the desired measurement, r.

$$r = \frac{1}{n-1} \Sigma \frac{(x - \bar{x})}{s_x} \cdot \frac{(y - \bar{y})}{s_y} \tag{1}$$

Computation Formula for *r*

The defining formula for r shows how the mean and standard deviation of each variable in the data pair enter into the formulation of r. However, the defining formula is technically difficult to work with because of all the subtractions and products. A computation formula for r uses the raw data values of x and y directly.

PROCEDURE

HOW TO COMPUTE THE SAMPLE CORRELATION COEFFICIENT *r*

Requirements

Obtain a random sample of n data pairs (x, y). The data pairs should have a *bivariate normal distribution*. This means that for a fixed value of x, the y values should have a normal distribution (or at least a mound-shaped and symmetric distribution), and for a fixed y, the x values should have their own (approximately) normal distribution.

Procedure

1. Using the data pairs, compute Σx, Σy, Σx^2, Σy^2, and Σxy.
2. With n = sample size, Σx, Σy, Σx^2, Σy^2, and Σxy, you are ready to compute the sample correlation coefficient r using the computation formula

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} \tag{2}$$

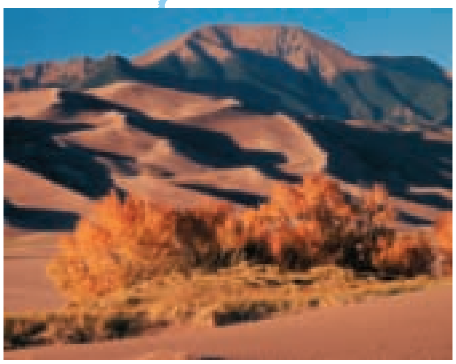
Be careful! The notation Σx^2 means first square x and then calculate the sum, whereas $(\Sigma x)^2$ means first sum the x values and then square the result.

When x and y values of the data pairs are exchanged, the sample correlation coefficient r remains the same. Problem 20 explores this result.

Interpretation It can be shown mathematically that r is always a number between $+1$ and -1 ($-1 \le r \le +1$). Table 9-2 gives a quick summary of some basic facts about r.

For most applications, you will use a calculator or computer software to compute r directly. However, to build some familiarity with the structure of the sample correlation coefficient, it is useful to do some calculations for yourself. Example 2 and Guided Exercise 3 show how to use the computation formula to compute r.

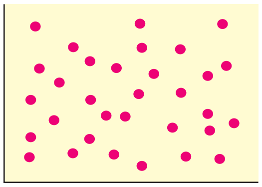
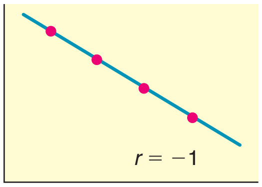
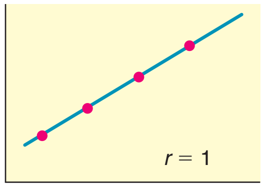
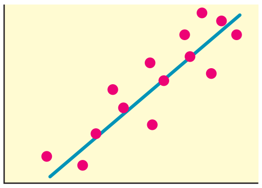
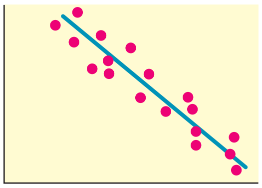
EXAMPLE 2 COMPUTING *r*



David Muench/Terra/CORBIS

Sand driven by wind creates large, beautiful dunes at the Great Sand Dunes National Monument, Colorado. Of course, the same natural forces also create large dunes in the Great Sahara and Arabia. Is there a linear correlation between wind velocity and sand drift rate? Let x be a random variable representing wind velocity (in 10 cm/sec) and let y be a random variable representing drift rate of sand (in 100 gm/cm/sec). A test site at the Great Sand Dunes National Monument gave the following information about x and y (Reference: *Hydrologic, Geologic, and Biologic Research at Great Sand Dunes National Monument*, Proceedings of the National Park Service Research Symposium).

TABLE 9-2 Some Facts about the Correlation Coefficient

If r Is	Then	The Scatter Diagram Might Look Something Like	
0	There is no linear relation among the points of the scatter diagram.		
1 or -1	There is a perfect linear relation between x and y values; all points lie on the least-squares line.		
Between 0 and 1 ($0 < r < 1$)	The x and y values have a *positive correlation*. By this, we mean that *large* x values are associated with *large* y values, and *small* x values are associated with *small* y values.		As we go from left to right, the least-squares line goes *up*.
Between -1 and 0 ($-1 < r < 0$)	The x and y values have a *negative correlation*. By this, we mean that *large* x values are associated with *small* y values, and *small* x values are associated with *large* y values.		As we go from left to right, the least-squares line goes *down*.

x	70	115	105	82	93	125	88
y	3	45	21	7	16	62	12

(a) Construct a scatter diagram. Do you expect r to be positive?

SOLUTION: Figure 9-7 displays the scatter diagram. From the scatter diagram, it appears that as x values increase, y values also tend to increase. Therefore, r should be positive.

FIGURE 9-7

Wind Velocity (10 cm/sec) and Drift Rate of Sand (100 gm/cm/sec)

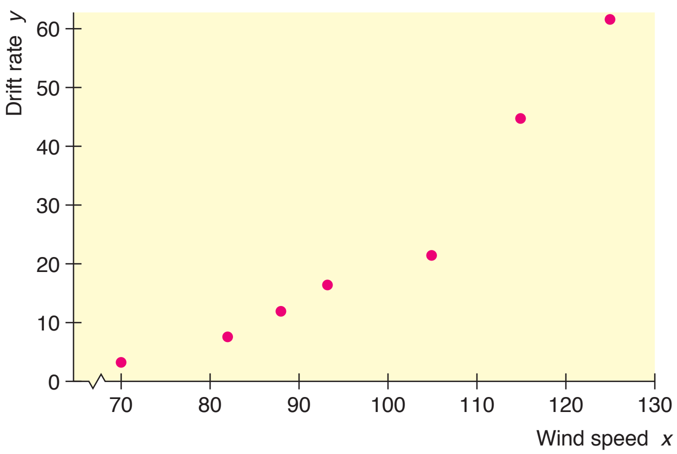


TABLE 9-3 Computation Table

x	y	x^2	y^2	xy
70	3	4900	9	210
115	45	13,225	2025	5175
105	21	11,025	441	2205
82	7	6724	49	574
93	16	8649	256	1488
125	62	15,625	3844	7750
88	12	7744	144	1056
$\Sigma x = 678$	$\Sigma y = 166$	$\Sigma x^2 = 67{,}892$	$\Sigma y^2 = 6768$	$\Sigma xy = 18{,}458$

(b) Compute r using the computation formula (formula 2).

SOLUTION: To find r, we need to compute Σx, Σx^2, Σy, Σy^2, and Σxy. It is convenient to organize the data in a table of five columns (Table 9-3) and then sum the entries in each column. Of course, many calculators give these sums directly. Using the computation formula for r, the sums from Table 9-3, and $n = 7$, we have

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\,\sqrt{n\Sigma y^2 - (\Sigma y)^2}} \qquad (2)$$

$$= \frac{7(18{,}458) - (678)(166)}{\sqrt{7(67{,}892) - (678)^2}\sqrt{7(6768) - (166)^2}} \approx \frac{16{,}658}{(124.74)(140.78)} \approx 0.949$$

Note: Using a calculator to compute r directly gives 0.949 to three places after the decimal.

(c) *Interpretation* What does the value of r tell you?

SOLUTION: Since r is very close to 1, we have an indication of a strong positive linear correlation between wind velocity and drift rate of sand. In other words, we expect that higher wind speeds tend to mean greater drift rates. Because r is so close to 1, the association between the variables appears to be linear.

Because it is quite a task to compute r for even seven data pairs, the use of columns as in Example 2 is extremely helpful. Your value for r should always be between -1 and 1, inclusive. Use a scatter diagram to get a rough idea of the value of r. If your computed value of r is outside the allowable range, or if it disagrees quite a bit with the scatter diagram, recheck your calculations. Be sure you distinguish between expressions such as (Σx^2) and $(\Sigma x)^2$. Negligible rounding errors may occur, depending on how you (or your calculator) round.

GUIDED EXERCISE 3 Computing r

In one of the Boston city parks, there has been a problem with muggings in the summer months. A police cadet took a random sample of 10 days (out of the 90-day summer) and compiled the following data. For each day, x represents the number of police officers on duty in the park and y represents the number of reported muggings on that day.

x	10	15	16	1	4	6	18	12	14	7
y	5	2	1	9	7	8	1	5	3	6

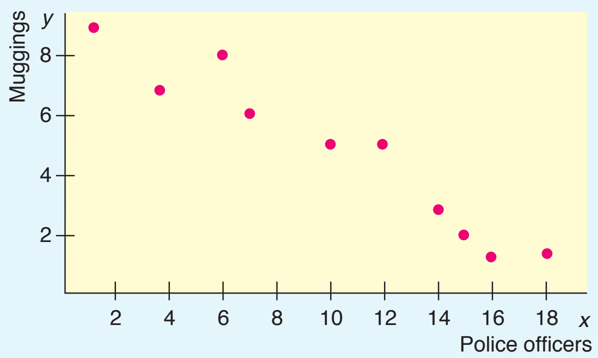
Continued

GUIDED EXERCISE 3 *continued*

(a) Construct a scatter diagram of x and y values.

Figure 9-8 shows the scatter diagram.

FIGURE 9-8 Scatter Diagram for Number of Police Officers versus Number of Muggings



(b) From the scatter diagram, do you think the computed value of r will be positive, negative, or zero? Explain.

r will be negative. The general trend is that large x values are associated with small y values, and vice versa. From left to right, the least-squares line goes down.

(c) Complete TABLE 9-4.

x	y	x^2	y^2	xy
10	5	100	25	50
15	2	225	4	30
16	1	256	1	16
1	9	1	81	9
4	7	16	49	28
6	8	____	____	____
18	1	____	____	____
12	5	____	____	____
14	3	____	____	____
7	6	49	36	42
$\Sigma x = 103$	$\Sigma y = 47$	$\Sigma x^2 =$ ___	$\Sigma y^2 =$ ___	$\Sigma xy =$ ___
$(\Sigma x)^2 =$ ___	$(\Sigma y)^2 =$ ___			

TABLE 9-5 **Completion of Table 9-4**

x	y	x^2	y^2	xy
6	8	36	64	48
18	1	324	1	18
12	5	144	25	60
14	3	196	9	42
		$\Sigma x^2 = 1347$	$\Sigma y^2 = 295$	$\Sigma xy = 343$

$(\Sigma x)^2 = 10{,}609$ $(\Sigma y)^2 = 2209$

(d) Compute r. Alternatively, find the value of r directly by using a calculator or computer software.

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{10(343) - (103)(47)}{\sqrt{10(1347) - (103)^2}\sqrt{10(295) - (47)^2}}$$

$$\approx \frac{-1411}{(53.49)(27.22)} \approx -0.969$$

(e) *Interpretation* What does the value of r tell you about the relationship between the number of police officers and the number of muggings in the park?

There is a strong negative linear relationship between the number of police officers and the number of muggings. It seems that the more officers there are in the park, the fewer the number of muggings.

LOOKING FORWARD

If the scatter diagram and the value of the sample correlation coefficient r indicate a linear relationship between the data pairs, how do we find a suitable linear equation for the data? This process, called linear regression, is presented in the next section, Section 9.2.

TECH NOTES

Most calculators that support two-variable statistics provide the value of the sample correlation coefficient r directly. Statistical software provides r, r^2, or both.

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) First use **CATALOG,** find **DiagnosticOn,** and press **Enter** twice. Then, when you use **STAT, CALC,** option **8:LinReg(a + bx),** the value of r will be given (data from Example 2). In the next section, we will discuss the line $y = a + bx$ and the meaning of r^2.

Excel 2007 Excel gives the value of the sample correlation coefficient r in several outputs. One way to find the value of r is to click the **Insert** function f_x. Then in the dialogue box, select **Statistical** for the category and **Correl** for the function.

Minitab Use the menu selection **Stat ➤ Basic Statistics ➤ Correlation.**

```
LinReg
 y=a+bx
 a=-79.97763496
 b=1.070565553
 r²=.8997719968
 r=.9485631222
```

LOOKING FORWARD

When the data are ranks (without ties) instead of measurements, the Pearson product-moment correlation coefficient can be reduced to a simpler equation called the Spearman rank correlation coefficient. This coefficient is used with nonparametric methods and is discussed in Section 11.3.

CRITICAL THINKING

Sample correlation compared to population correlation

Cautions about Correlation

The correlation coefficient can be thought of as a measure of how well a linear model fits the data points on a scatter diagram. The closer r is to $+1$ or -1, the better a line "fits" the data. Values of r close to 0 indicate a poor fit to any line.

Usually a scatter diagram does not contain *all* possible data points that could be gathered. Most scatter diagrams represent only a *random sample* of data pairs taken from a very large population of all possible pairs. Because r is computed on the basis of a random sample of (x, y) pairs, we expect the values of r to vary from one sample to the next (much as the sample mean $\bar{x}$ varies from sample to sample). This brings up the question of the *significance* of r. Or, put another way, what are the chances that our random sample of data pairs indicates a high correlation when, in fact, the population's x and y values are not so strongly correlated? Right now, let's just say that the significance of r is a separate issue that will be treated in Section 9.3, where we test the *population correlation coefficient* ρ (Greek letter *rho*, pronounced "row").

r = **sample** correlation coefficient computed from a random sample of (x, y) data pairs.

ρ = **population** correlation coefficient computed from all population data pairs (x, y).

Problem 21 demonstrates an informal process for determining whether or not r is significant. Problem 22 explores the effect of sample size on the significance of r. Problem 24 uses the value of ρ between two dependent variables to find the mean and standard deviation of a linear combination of the two variables.

There is a less formal way to address the significance of r using a table of "critical values" or "cut-off values" based on the r distribution and the number of data pairs. Problem 21 at the end of this section discusses this method.

Extrapolation

The value of the sample correlation coefficient r and the strength of the linear relationship between variables is computed based on the sample data. The situation may change for measurements larger than or smaller than the data values included in the sample. For instance, for infants, there may be a high positive correlation between age in months and weight. However, that correlation might not apply for people ages 20 to 30 years.

Causation

The correlation coefficient is a mathematical tool for measuring the strength of a linear relationship between two variables. As such, it makes no implication about cause or effect. The fact that two variables tend to increase or decrease together does not mean that a change in one is *causing* a change in the other. A strong correlation between x and y is sometimes due to other (either known or unknown) variables. Such variables are called *lurking variables*.

Lurking variables

In ordered pairs (x, y), x is called the **explanatory** variable and y is called the **response** variable. When r indicates a linear correlation between x and y, changes in values of y tend to respond to changes in values of x according to a linear model. A **lurking variable** is a variable that is neither an explanatory nor a response variable. Yet, a lurking variable may be responsible for changes in both x and y.

EXAMPLE 3 Causation and lurking variables

Over a period of years, the population of a certain town increased. It was observed that during this period the correlation between x, the number of people attending church, and y, the number of people in the city jail, was $r = 0.90$. Does going to church *cause* people to go to jail? Is there a *lurking variable* that might cause both variables x and y to increase?

SOLUTION: We hope church attendance does not cause people to go to jail! During this period, there was an increase in population. Therefore, it is not too surprising that both the number of people attending church and the number of people in jail increased. The high correlation between x and y is likely due to the lurking variable of population increase.

Correlation between averages

The correlation between two variables consisting of averages is usually higher than the correlation between two variables representing corresponding raw data. One reason is that the use of averages reduces the variation that exists between individual measurements (see Section 6.5 and the central limit theorem). A high correlation based on two variables consisting of averages does not necessarily imply a high correlation between two variables consisting of individual measurements. See Problem 23 at the end of this section.

Problem 23 at the end of this section explores the correlation of averages.

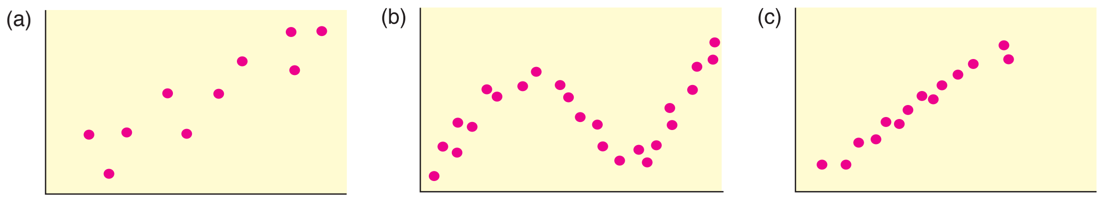
VIEWPOINT Low on Credit, High on Cost!!!

How do you measure automobile insurance risk? One way is to use a little statistics and customer credit ratings. Insurers say statistics show that drivers who have a history of bad credit are more likely to be in serious car accidents. According to a high-level executive at Allstate Insurance Company, financial instability is an extremely powerful predictor of future insurance losses. In short, there seems to be a strong correlation between bad credit ratings and auto insurance claims. Consequently, insurance companies want to charge higher premiums to customers with bad credit ratings. Consumer advocates object strongly because they say bad credit does not cause automobile accidents, and more than 20 states prohibit or restrict the use of credit ratings to determine auto insurance premiums. Insurance companies respond by saying that your best defense is to pay your bills on time!

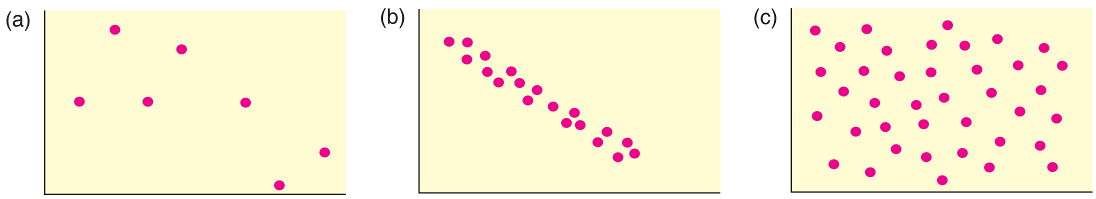
SECTION 9.1 PROBLEMS

Note: Answers may vary due to rounding.

1. *Statistical Literacy* When drawing a scatter diagram, along which axis is the explanatory variable placed? Along which axis is the response variable placed?
2. *Statistical Literacy* Suppose two variables are positively correlated. Does the response variable increase or decrease as the explanatory variable increases?
3. *Statistical Literacy* Suppose two variables are negatively correlated. Does the response variable increase or decrease as the explanatory variable increases?
4. *Statistical Literacy* Describe the relationship between two variables when the correlation coefficient r is
   (a) near -1.
   (b) near 0.
   (c) near 1.
5. *Critical Thinking: Linear Correlation* Look at the following diagrams. Does each diagram show high linear correlation, moderate or low linear correlation, or no linear correlation?



6. *Critical Thinking: Linear Correlation* Look at the following diagrams. Does each diagram show high linear correlation, moderate or low linear correlation, or no linear correlation?



7. *Critical Thinking: Lurking Variables* Over the past few years, there has been a strong positive correlation between the annual consumption of diet soda drinks and the number of traffic accidents.

(a) Do you think increasing consumption of diet soda drinks causes traffic accidents? Explain.
(b) What lurking variables might be causing the increase in one or both of the variables? Explain.

8. ***Critical Thinking: Lurking Variables*** Over the past decade, there has been a strong positive correlation between teacher salaries and prescription drug costs.
(a) Do you think paying teachers more causes prescription drugs to cost more? Explain.
(b) What lurking variables might be causing the increase in one or both of the variables? Explain.

9. ***Critical Thinking: Lurking Variables*** Over the past 50 years, there has been a strong negative correlation between average annual income and the record time to run 1 mile. In other words, average annual incomes have been rising while the record time to run 1 mile has been decreasing.
(a) Do you think increasing incomes cause decreasing times to run the mile? Explain.
(b) What lurking variables might be causing the increase in one or both of the variables? Explain.

10. ***Critical Thinking: Lurking Variables*** Over the past 30 years in the United States, there has been a strong negative correlation between the number of infant deaths at birth and the number of people over age 65.
(a) Is the fact that people are living longer causing a decrease in infant mortalities at birth?
(b) What lurking variables might be causing the increase in one or both of the variables? Explain.

11. ***Interpretation*** Trevor conducted a study and found that the correlation between the price of a gallon of gasoline and gasoline consumption has a linear correlation coefficient of -0.7. What does this result say about the relationship between price of gasoline and consumption? The study included gasoline prices ranging from \$2.70 to \$5.30 per gallon. Is it reliable to apply the results of this study to prices of gasoline higher than \$5.30 per gallon? Explain.

12. ***Interpretation*** Do people who spend more time on social networking sites spend more time using Twitter? Megan conducted a study and found that the correlation between the times spent on the two activities was 0.8. What does this result say about the relationship between times spent on the two activities? If someone spends more time than average on a social networking site, can you automatically conclude that he or she spends more time than average using Twitter? Explain.

13. ***Veterinary Science: Shetland Ponies*** How much should a healthy Shetland pony weigh? Let x be the age of the pony (in months), and let y be the average weight of the pony (in kilograms). The following information is based on data taken from *The Merck Veterinary Manual* (a reference used in most veterinary colleges).

x	3	6	12	18	24
y	60	95	140	170	185

(a) Make a scatter diagram and draw the line you think best fits the data.
(b) Would you say the correlation is low, moderate, or strong? positive or negative?
(c) Use a calculator to verify that $\Sigma x = 63$, $\Sigma x^2 = 1089$, $\Sigma y = 650$, $\Sigma y^2 = 95{,}350$, and $\Sigma xy = 9930$. Compute r. As x increases from 3 to 24 months, does the value of r imply that y should tend to increase or decrease? Explain.

14. *Health Insurance: Administrative Cost* The following data are based on information from *Domestic Affairs*. Let x be the average number of employees in a group health insurance plan, and let y be the average administrative cost as a percentage of claims.

x	3	7	15	35	75
y	40	35	30	25	18

(a) Make a scatter diagram and draw the line you think best fits the data.
(b) Would you say the correlation is low, moderate, or strong? positive or negative?
(c) Use a calculator to verify that $\Sigma x = 135$, $\Sigma x^2 = 7133$, $\Sigma y = 148$, $\Sigma y^2 = 4674$, and $\Sigma xy = 3040$. Compute r. As x increases from 3 to 75, does the value of r imply that y should tend to increase or decrease? Explain.

15. *Meteorology: Cyclones* Can a low barometer reading be used to predict maximum wind speed of an approaching tropical cyclone? Data for this problem are based on information taken from *Weatherwise* (Vol. 46, No. 1), a publication of the American Meteorological Society. For a random sample of tropical cyclones, let x be the lowest pressure (in millibars) as a cyclone approaches, and let y be the maximum wind speed (in miles per hour) of the cyclone.

x	1004	975	992	935	985	932
y	40	100	65	145	80	150

(a) Make a scatter diagram and draw the line you think best fits the data.
(b) Would you say the correlation is low, moderate, or strong? positive or negative?
(c) Use a calculator to verify that $\Sigma x = 5823$, $\Sigma x^2 = 5{,}655{,}779$, $\Sigma y = 580$, $\Sigma y^2 = 65{,}750$, and $\Sigma xy = 556{,}315$. Compute r. As x increases, does the value of r imply that y should tend to increase or decrease? Explain.

16. *Geology: Earthquakes* Is the magnitude of an earthquake related to the depth below the surface at which the quake occurs? Let x be the magnitude of an earthquake (on the Richter scale), and let y be the depth (in kilometers) of the quake below the surface at the epicenter. The following is based on information taken from the National Earthquake Information Service of the U.S. Geological Survey. Additional data may be found by visiting the Brase/Brase statistics site at **http://www.cengage.com/statistics/brase** and finding the link to Earthquakes.

x	2.9	4.2	3.3	4.5	2.6	3.2	3.4
y	5.0	10.0	11.2	10.0	7.9	3.9	5.5

(a) Make a scatter diagram and draw the line you think best fits the data.
(b) Would you say the correlation is low, moderate, or strong? positive or negative?
(c) Use a calculator to verify that $\Sigma x = 24.1$, $\Sigma x^2 = 85.75$, $\Sigma y = 53.5$, $\Sigma y^2 = 458.31$, and $\Sigma xy = 190.18$. Compute r. As x increases, does the value of r imply that y should tend to increase or decrease? Explain.

17. *Baseball: Batting Averages and Home Runs* In baseball, is there a linear correlation between batting average and home run percentage? Let x represent the batting average of a professional baseball player, and let y represent the player's home run percentage (number of home runs per 100 times at bat). A random sample of $n = 7$ professional baseball players gave the following information (Reference: *The Baseball Encyclopedia*, Macmillan Publishing Company).

x	0.243	0.259	0.286	0.263	0.268	0.339	0.299
y	1.4	3.6	5.5	3.8	3.5	7.3	5.0

(a) Make a scatter diagram and draw the line you think best fits the data.
(b) Would you say the correlation is low, moderate, or high? positive or negative?
(c) Use a calculator to verify that $\Sigma x = 1.957$, $\Sigma x^2 \approx 0.553$, $\Sigma y = 30.1$, $\Sigma y^2 = 150.15$, and $\Sigma xy \approx 8.753$. Compute r. As x increases, does the value of r imply that y should tend to increase or decrease? Explain.

18. *University Crime: FBI Report* Do larger universities tend to have more property crime? University crime statistics are affected by a variety of factors. The surrounding community, accessibility given to outside visitors, and many other factors influence crime rates. Let x be a variable that represents student enrollment (in thousands) on a university campus, and let y be a variable that represents the number of burglaries in a year on the university campus. A random sample of $n = 8$ universities in California gave the following information about enrollments and annual burglary incidents (Reference: *Crime in the United States*, Federal Bureau of Investigation).

x	12.5	30.0	24.5	14.3	7.5	27.7	16.2	20.1
y	26	73	39	23	15	30	15	25

(a) Make a scatter diagram and draw the line you think best fits the data.
(b) Would you say the correlation is low, moderate, or high? positive or negative?
(c) Using a calculator, verify that $\Sigma x = 152.8$, $\Sigma x^2 = 3350.98$, $\Sigma y = 246$, $\Sigma y^2 = 10{,}030$, and $\Sigma xy = 5488.4$. Compute r. As x increases, does the value of r imply that y should tend to increase or decrease? Explain.

19. *Expand Your Knowledge: Effect of Scale on Scatter Diagram* The initial visual impact of a scatter diagram depends on the scales used on the x and y axes. Consider the following data:

x	1	2	3	4	5	6
y	1	4	6	3	6	7

(a) Make a scatter diagram using the same scale on both the x and y axes (i.e., make sure the unit lengths on the two axes are equal).
(b) Make a scatter diagram using a scale on the y axis that is twice as long as that on the x axis.
(c) Make a scatter diagram using a scale on the y axis that is half as long as that on the x axis.
(d) On each of the three graphs, draw the straight line that you think best fits the data points. How do the slopes (or directions) of the three lines appear to change? *Note:* The actual slopes will be the same; they just appear different because of the choice of scale factors.

20. *Expand Your Knowledge: Effect on r of Exchanging x and y Values* Examine the computation formula for r, the sample correlation coefficient [formulas (1) and (2) of this section].
(a) In the formula for r, if we exchange the symbols x and y, do we get a different result or do we get the same (equivalent) result? Explain.
(b) If we have a set of x and y data values and we exchange corresponding x and y values to get a new data set, should the sample correlation coefficient be the same for both sets of data? Explain.

(c) Compute the sample correlation coefficient r for each of the following data sets and show that r is the same for both.

x	1	3	4
y	2	1	6

x	2	1	6
y	1	3	4

21. *Expand Your Knowledge: Using a Table to Test ρ* The correlation coefficient r is a *sample* statistic. What does it tell us about the value of the population correlation coefficient ρ (Greek letter rho)? We will build the formal structure of hypothesis tests of ρ in Section 9.3. However, there is a quick way to determine if the sample evidence based on r is strong enough to conclude that there is some population correlation between the variables. In other words, we can use the value of r to determine if $\rho \neq 0$. We do this by comparing the value $|r|$ to an entry in Table 9-6. The value of α in the table gives us the probability of concluding that $\rho \neq 0$ when, in fact, $\rho = 0$ and there is no population correlation. We have two choices for α: $\alpha = 0.05$ or $\alpha = 0.01$.

PROCEDURE

HOW TO USE TABLE 9-6 TO TEST ρ

1. First compute r from a random sample of n data pairs (x, y).
2. Find the table entry in the row headed by n and the column headed by your choice of α. Your choice of α is the risk you are willing to take of mistakenly concluding that $\rho \neq 0$ when, in fact, $\rho = 0$.
3. Compare $|r|$ to the table entry.
   (a) If $|r| \geq$ table entry, then there is sufficient evidence to conclude that $\rho \neq 0$, and we say that r is **significant**. In other words, we conclude that there is some population correlation between the two variables x and y.
   (b) If $|r| <$ table entry, then the evidence is insufficient to conclude that $\rho \neq 0$, and we say that r is **not significant**. We do not have enough evidence to conclude that there is any correlation between the two variables x and y.

TABLE 9-6 **Critical Values for Correlation Coefficient r**

n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$
3	1.00	1.00	13	0.53	0.68	23	0.41	0.53
4	0.95	0.99	14	0.53	0.66	24	0.40	0.52
5	0.88	0.96	15	0.51	0.64	25	0.40	0.51
6	0.81	0.92	16	0.50	0.61	26	0.39	0.50
7	0.75	0.87	17	0.48	0.61	27	0.38	0.49
8	0.71	0.83	18	0.47	0.59	28	0.37	0.48
9	0.67	0.80	19	0.46	0.58	29	0.37	0.47
10	0.63	0.76	20	0.44	0.56	30	0.36	0.46
11	0.60	0.73	21	0.43	0.55			
12	0.58	0.71	22	0.42	0.54			

(a) Look at Problem 13 regarding the variables x = age of a Shetland pony and y = weight of that pony. Is the value of $|r|$ large enough to conclude that weight and age of Shetland ponies are correlated? Use $\alpha = 0.05$.

(b) Look at Problem 15 regarding the variables x = lowest barometric pressure as a cyclone approaches and y = maximum wind speed of the cyclone. Is the value of $|r|$ large enough to conclude that lowest barometric pressure and wind speed of a cyclone are correlated? Use $\alpha = 0.01$.

22. ***Expand Your Knowledge: Sample Size and Significance of Correlation*** In this problem, we use Table 9-6 to explore the significance of r based on different sample sizes. See Problem 21.

(a) Is a sample correlation coefficient $r = 0.820$ significant at the $\alpha = 0.01$ level based on a sample size of $n = 7$ data pairs? What about $n = 9$ data pairs?

(b) Is a sample correlation coefficient $r = 0.40$ significant at the $\alpha = 0.05$ level based on a sample size of $n = 20$ data pairs? What about $n = 27$ data pairs?

(c) Is it true that in order to be significant, an r value must be larger than 0.90? larger than 0.70? larger than 0.50? What does sample size have to do with the significance of r? Explain.

23. ***Expand Your Knowledge: Correlation of Averages*** Fuming because you are stuck in traffic? Roadway congestion is a costly item, in both time wasted and fuel wasted. Let x represent the *average* annual hours per person spent in traffic delays and let y represent the *average* annual gallons of fuel wasted per person in traffic delays. A random sample of eight cities showed the following data (Reference: *Statistical Abstract of the United States*, 122nd Edition).

x (hr)	28	5	20	35	20	23	18	5
y (gal)	48	3	34	55	34	38	28	9

(a) Draw a scatter diagram for the data. Verify that $\Sigma x = 154$, $\Sigma x^2 = 3712$, $\Sigma y = 249$, $\Sigma y^2 = 9959$, and $\Sigma xy = 6067$. Compute r.

The data in part (a) represent *average* annual hours lost per person and *average* annual gallons of fuel wasted per person in traffic delays. Suppose that instead of using average data for different cities, you selected one person at random from each city and measured the annual number of hours lost x for that person and the annual gallons of fuel wasted y for the same person.

x (hr)	20	4	18	42	15	25	2	35
y (gal)	60	8	12	50	21	30	4	70

(b) Compute $\bar{x}$ and $\bar{y}$ for both sets of data pairs and compare the averages. Compute the sample standard deviations s_x and s_y for both sets of data pairs and compare the standard deviations. In which set are the standard deviations for x and y larger? Look at the defining formula for r, Equation 1. Why do smaller standard deviations s_x and s_y tend to increase the value of r?

(c) Make a scatter diagram for the second set of data pairs. Verify that $\Sigma x = 161$, $\Sigma x^2 = 4583$, $\Sigma y = 255$, $\Sigma y^2 = 12{,}565$, and $\Sigma xy = 7071$. Compute r.

(d) Compare r from part (a) with r from part (c). Do the data for averages have a higher correlation coefficient than the data for individual measurements? List some reasons why you think hours lost per individual and fuel wasted per individual might vary more than the same quantities averaged over all the people in a city.

24. ***Expand Your Knowledge: Dependent Variables*** In Section 5.1, we studied linear combinations of *independent* random variables. What happens if the variables are not independent? A lot of mathematics can be used to prove the following:

Let x and y be random variables with means μ_x and μ_y, variances σ_x^2 and σ_y^2, and population correlation coefficient ρ (the Greek letter rho). Let a and b be any constants and let $w = ax + by$. Then,

$$\mu_w = a\mu_x + b\mu_y$$
$$\sigma_w^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\sigma_x\sigma_y\rho$$

Covariance

In this formula, ρ is the population correlation coefficient, theoretically computed using the population of all (x, y) data pairs. The expression $\sigma_x\sigma_y\rho$ is called the *covariance* of x and y. If x and y are independent, then $\rho = 0$ and the formula for σ_w^2 reduces to the appropriate formula for independent variables (see Section 5.1). In most real-world applications, the population parameters are not known, so we use sample estimates with the understanding that our conclusions are also estimates.

Do you have to be rich to invest in bonds and real estate? No, mutual fund shares are available to you even if you aren't rich. Let x represent annual percentage return (after expenses) on the Vanguard Total Bond Index Fund, and let y represent annual percentage return on the Fidelity Real Estate Investment Fund. Over a long period of time, we have the following population estimates (based on *Morningstar Mutual Fund Report*).

$\mu_x \approx 7.32$ $\sigma_x \approx 6.59$ $\mu_y \approx 13.19$ $\sigma_y \approx 18.56$ $\rho \approx 0.424$

(a) Do you think the variables x and y are independent? Explain.
(b) Suppose you decide to put 60% of your investment in bonds and 40% in real estate. This means you will use a weighted average $w = 0.6x + 0.4y$. Estimate your expected percentage return μ_w and risk σ_w.
(c) Repeat part (b) if $w = 0.4x + 0.6y$.
(d) Compare your results in parts (b) and (c). Which investment has the higher expected return? Which has the greater risk as measured by σ_w?

SECTION 9.2

Linear Regression and the Coefficient of Determination

FOCUS POINTS

- State the least-squares criterion.
- Use sample data to find the equation of the least-squares line. Graph the least-squares line.
- Use the least-squares line to predict a value of the response variable y for a specified value of the explanatory variable x.
- Explain the difference between interpolation and extrapolation.
- Explain why extrapolation beyond the sample data range might give results that are misleading or meaningless.
- Use r^2 to determine *explained* and *unexplained* variation of the response variable y.

In Denali National Park, Alaska, the wolf population is dependent on a large, strong caribou population. In this wild setting, caribou are found in very large herds. The well-being of an entire caribou herd is not threatened by wolves. In fact, it is thought that wolves keep caribou herds strong by helping prevent overpopulation. Can the caribou population be used to predict the size of the wolf population?

Let x be a random variable that represents the fall caribou population (in hundreds) in Denali National Park, and let y be a random variable that represents the late-winter wolf population in the park. A random sample of recent years gave the following information (Reference: U.S. Department of the Interior, National Biological Service).

x	30	34	27	25	17	23	20
y	66	79	70	60	48	55	60

Looking at the scatter diagram in Figure 9-9, we can ask some questions.

1. Do the data indicate a linear relationship between x and y?
2. Can you find an equation for the best-fitting line relating x and y? Can you use this relationship to predict the size of the wolf population when you know the size of the caribou population?
3. What fractional part of the variability in y can be associated with the variability in x? What fractional part of the variability in y is not associated with a corresponding variability in x?

FIGURE 9-9

Caribou and Wolf Populations

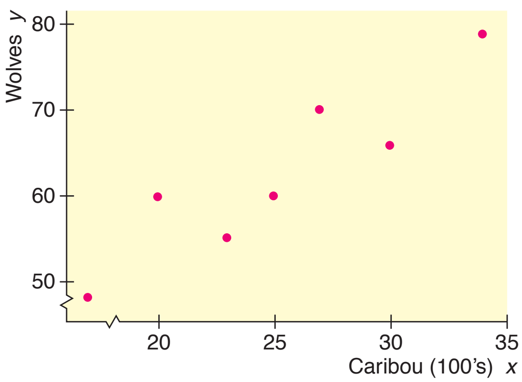
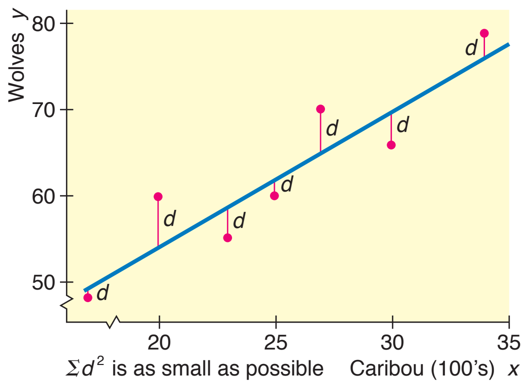


FIGURE 9-10

Least-Squares Criterion



The first step in answering these questions is to try to express the relationship as a mathematical equation. There are many possible equations, but the simplest and most widely used is the linear equation, or the equation of a straight line. Because we will be using this line to predict the y values from the x values, we call x the *explanatory variable* and y the *response variable*.

Explanatory variable
Response variable
Least-squares criterion

Our job is to find the linear equation that "best" represents the points of the scatter diagram. For our criterion of best-fitting line, we use the *least-squares criterion*, which states that the line we fit to the data points must be such that *the sum of the squares of the vertical distances from the points to the line be made as small as possible*. The least-squares criterion is illustrated in Figure 9-10.

Least-squares criterion

The sum of the squares of the vertical distances from the data points (x, y) to the line is made as small as possible.

In Figure 9-10, d represents the difference between the y coordinate of the data point and the corresponding y coordinate on the line. Thus, if the data point lies above the line, d is positive, but if the data point is below the line, d is negative. As a result, the sum of the d values can be small even if the points are widely spread in the scatter diagram. However, the squares d^2 cannot be negative. By minimizing the sum of the squares, we are, in effect, not allowing positive and negative d values to "cancel out" one another in the sum. It is in this way that we can meet the least-squares criterion of minimizing the sum of the squares of the vertical distances between the points and the line over *all* points in the scatter diagram.

Least-squares line

We use the notation $\hat{y} = a + bx$ for the *least-squares line*. A little algebra tells us that b is the slope and a is the intercept of the line. In this context, $\hat{y}$ (read "y hat") represents the value of the response variable y estimated using the least-squares line and a given value of the explanatory variable x.

Techniques of calculus can be applied to show that a and b may be computed using the following procedure.

PROCEDURE

HOW TO FIND THE EQUATION OF THE LEAST-SQUARES LINE $\hat{y} = a + bx$

Requirements for Statistical Inference

Obtain a random sample of n data pairs (x, y), where x is the *explanatory variable* and y is the *response variable*. The data pairs should have a *bivariate normal distribution*. This means that for a fixed value of x, the y values should have a normal distribution (or at least a mound-shaped and symmetric distribution), and for a fixed y, the x values should have their own (approximately) normal distribution.

Problem 21 demonstrates that for the same data set, the least-squares lines for predicting y or for predicting x are essentially different.

Procedure

1. Using the data pairs, compute Σx, Σy, Σx^2, Σy^2, and Σxy. Then compute the sample means $\bar{x}$ and $\bar{y}$.
2. With n = sample size, Σx, Σy, Σx^2, Σy^2, Σxy, $\bar{x}$, and $\bar{y}$, you are ready to compute the slope b and intercept a using the computation formulas

Slope b

Intercept a

$$\text{Slope:} \quad b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} \tag{3}$$

$$\text{Intercept:} \quad a = \bar{y} - b\bar{x} \tag{4}$$

Be careful! The notation Σx^2 means first square x and then calculate the sum, whereas (Σx^2) means first sum the x values and then square the result.

3. The equation of the least-squares line computed from your sample data is

$$\hat{y} = a + bx \tag{5}$$

For data following exponential growth or power law models, logarithmic transformations can be used to transform the data into linear models. Then linear regression can be used on the transformed data. Problems 22–25 show these methods.

COMMENT The computation formulas for the slope of the least-squares line, the sample correlation coefficient r, and the standard deviations s_x and s_y use many of the same sums. There is, in fact, a relationship between the sample correlation coefficient r and the slope of the least-squares line b. In instances where we know r, s_x, and s_y, we can use the following formula to compute b.

$$b = r\left(\frac{s_y}{s_x}\right) \tag{6}$$

COMMENT In other mathematics courses, the slope-intercept form of the equation of a line is usually given as $y = mx + b$, where m refers to the slope of the line and b to the y coordinate of the y intercept. In statistics, when there is only one explanatory variable, it is common practice to use the letter b to designate the slope of the least-squares line and the letter a to designate the y coordinate of the intercept. For example, these are the symbols used on the TI-84Plus/TI-83Plus/TI-*n*spire calculators as well as on many other calculators.

Using the formulas to find the values of a and b

For most applications, you can use a calculator or computer software to compute a and b directly. However, to build some familiarity with the structure of the computation formulas, it is useful to do some calculations yourself. Example 4 shows how to use the computation formulas to find the values of a and b and the equation of the least-squares line $\hat{y} = a + bx$.

Note: If you are using your calculator to find the values of a and b directly, you may omit the discussion regarding use of the formulas. Go to the margin header "Using the values of a and b to construct the equation of the least-squares line."

EXAMPLE 4 **LEAST-SQUARES LINE**

Let's find the least-squares equation relating the variables x = size of caribou population (in hundreds) and y = size of wolf population in Denali National Park. Use x as the explanatory variable and y as the response variable.


Joe McDonald/Encyclopedia/Corbis

(a) Use the computation formulas to find the slope b of the least-squares line and the y intercept a.

SOLUTION: Table 9-7 gives the data values x and y along with the values x^2, y^2, and xy. First compute the sample means.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{176}{7} \approx 25.14 \quad \text{and} \quad \bar{y} = \frac{\Sigma y}{n} = \frac{438}{7} \approx 62.57$$

Next compute the slope b.

$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{7(11{,}337) - (176)(438)}{7(4628) - (176)^2} = \frac{2271}{1420} \approx 1.60$$

Use the values of b, $\bar{x}$ and $\bar{y}$ to compute the y intercept a.

$$a = \bar{y} - b\bar{x} \approx 62.57 - 1.60(25.14) \approx 22.35$$

Note that calculators give the values $b \approx 1.599$ and $a \approx 22.36$. These values differ slightly from those you computed using the formulas because of rounding.

(b) Use the values of a and b (either computed or obtained from a calculator) to find the equation of the least-squares line.

SOLUTION:

$$\hat{y} = a + bx$$
$$\hat{y} \approx 22.35 + 1.60x \quad \text{since} \quad a \approx 22.35 \quad \text{and} \quad b \approx 1.60$$

TABLE 9-7 **Sums for Computing b, $\bar{x}$, and $\bar{y}$**

x	y	x^2	y^2	xy
30	66	900	4356	1980
34	79	1156	6241	2686
27	70	729	4900	1890
25	60	625	3600	1500
17	48	289	2304	816
23	55	529	3025	1265
20	60	400	3600	1200
$\Sigma x = 176$	$\Sigma y = 438$	$\Sigma x^2 = 4628$	$\Sigma y^2 = 28{,}026$	$\Sigma xy = 11{,}337$

Graphing the least-squares line

(c) Graph the equation of the least-squares line on a scatter diagram.

SOLUTION: To graph the least-squares line, we have several options available. The slope-intercept method of algebra is probably the quickest, but may not always be convenient if the intercept is not within the range of the sample data values. It is just as easy to select two x values in the range of the x data values and then use the least-squares line to compute two corresponding $\hat{y}$ values.

In fact, we already have the coordinates of one point on the least-squares line. By the formula for the intercept [Equation (4)], the point $(\bar{x}, \bar{y})$ is always on the least-squares line. For our example, $(\bar{x}, \bar{y}) = (25.14, 62.57)$.

The point $(\bar{x}, \bar{y})$ is always on the least-squares line.

Another x value within the data range is $x = 34$. Using the least-squares line to compute the corresponding $\hat{y}$ value gives

$$\hat{y} \approx 22.35 + 1.60(34) \approx 76.75$$

We place the two points (25.14, 62.57) and (34, 76.75) on the scatter diagram (using a different symbol than that used for the sample data points) and connect the points with a line segment (Figure 9-11).

Meaning of slope

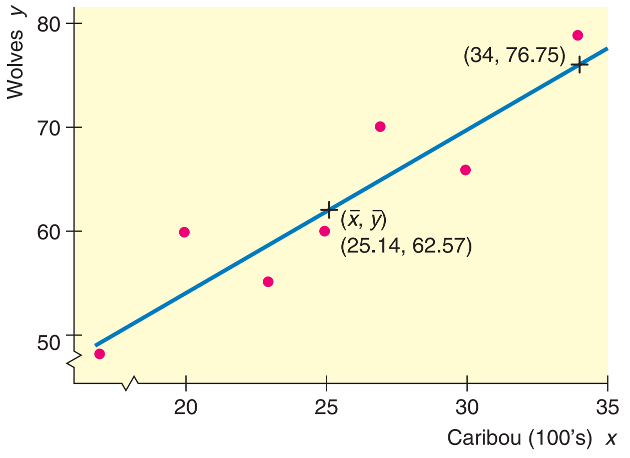
In the equation $\hat{y} = a + bx$, the slope b tells us how many units $\hat{y}$ changes for each unit change in x. In Example 4 regarding size of wolf and caribou populations,

$$\hat{y} \approx 22.35 + 1.60x$$

The slope 1.60 tells us that if the number of caribou (in hundreds) changes by 1 (hundred), then we expect the sustainable wolf population to change by 1.60. In other words, our model says that an increase of 100 caribou will increase the predicted wolf population by 1.60. If the caribou population decreases by 400, we predict the sustainable wolf population to decrease by 6.4.

FIGURE 9-11

Caribou and Wolf Populations



Marginal change

The slope of the least-squares line tells us how many units the response variable is expected to change for each unit change in the explanatory variable. The number of units change in the response variable for each unit change in the explanatory variable is called the **marginal change** of the response variable.

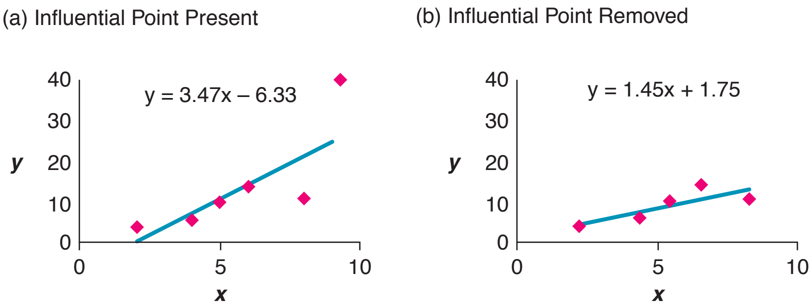
Some points in the data set have a strong influence on the equation of the least-squares line.

Influential points

A data pair is **influential** if removing it would substantially change the equation of the least-squares line or other calculations associated with linear regression. An influential point often has an x-value near the extreme high or low value of the data set.

Figure 9-12 shows two scatter diagrams produced in Excel. Figure 9-12(a) has an influential point. Figure 9-12(b) shows the scatter diagram with the influential point removed. Notice that the equations for the least-squares lines are quite different.

FIGURE 9-12



If a data set has an influential point, look at the influential point carefully to make sure it is not the result of a data collection error or a data-recording error. A valid influential point affects the equation of the least-squares line. The group project in Data Highlights at the end of this chapter further explores influential points.

CRITICAL THINKING

Using the Least-Squares Line for Prediction

Predicting y for a specified x

Making predictions is one of the main applications of linear regression. In other words, you use the equation of the least-squares line to predict the $\hat{y}$ value for a specified x value. The accuracy of the prediction depends on several components.

How well does the least-squares line fit the original data points?
The accuracy of the prediction depends on how well the least-squares line fits the original raw data points. Here are some tools to assess the fit of the line.

- Look at the scatter diagram, taking into account the scale of each axis.
- See if there are any influential points.
- Consider the value of the sample correlation coefficient r. The closer r is to 1 or -1, the better the least-squares line fits the data.
- Consider the value of the coefficient of determination r^2 (as discussed later in this section).
- Look at the residuals and a residual plot (see Problem 19 for a discussion of residual plots).

Residual

Problems 19 and 20 at the end of this section show how to make a residual plot.

The **residual** is the difference between the y value in a specified data pair (x, y) and the value $\hat{y} = a + bx$ predicted by the least-squares line for the same x.

$y - \hat{y}$ is the **residual**

If the residuals seem random about 0, the least-squares line provides a reasonable model for the data. Later in this section you will see the residual used in the development of the *coefficient of determination*, another important measurement associated with linear regression.

Does the prediction involve interpolation or extrapolation?

Interpolation
Extrapolation

Another issue that affects the validity of predictions is whether you are *interpolating* or *extrapolating*.

Predicting $\hat{y}$ values for x values that are **between** observed x values in the data set is called **interpolation.**

Predicting $\hat{y}$ values for x values that are **beyond** observed x values in the data set is called **extrapolation.** Extrapolation may produce unrealistic forecasts.

The least-squares line may not reflect the relationship between x and y for values of x outside the data range. For example, there is a fairly high correlation between height and age for boys ages 1 year to 10 years. In general, the older the boy, the taller the boy. A least-squares line based on such data would give good predictions of height for boys of ages between 1 and 10. However, it would be fairly meaningless to use the same linear regression line to predict the height of a 20-year-old or 50-year-old man.

The data are sample data.

Another consideration when working with predictions is the fact that the least-squares line is based on sample data. Each different sample will produce a slightly different equation for the least-squares line. Just as there are confidence intervals for parameters such as population means, there are confidence intervals for the prediction of y for a given x. We will examine confidence intervals for predictions in Section 9.3.

The least-squares line uses x as the explanatory variable and y as the response variables.

One more important fact about predictions: The least-squares line is developed with x as the explanatory variable and y as the response variable. This model can be used only to predict y values from specified x values. If you wish to begin with y values and predict corresponding x values, you must start all over and compute a new equation. Such an equation would be developed using a model with x as the response variable and y as the explanatory variable. See Problem 21 at the end of this section. Note that the equation for predicting x values *cannot* be derived from the least-squares line predicting y simply by solving the equation for x.

The least-squares line developed with x as the explanatory variable and y as the response variable can be used only to predict y values from specified x values.

The next example shows how to use the least-squares line for predictions.

EXAMPLE 5 PREDICTIONS

We continue with Example 4 regarding size of the wolf population as it relates to size of the caribou population. Suppose you want to predict the size of the wolf population when the size of the caribou population is 21 (hundred).

Joe McDonald/Corbis

(a) In the least-squares model developed in Example 4, which is the explanatory variable and which is the response variable? Can you use the equation to predict the size of the wolf population for a specified size of caribou population?

SOLUTION: The least-squares line $\hat{y} \approx 22.35 + 1.60x$ was developed using x = size of caribou population (in hundreds) as the explanatory variable and y = size of wolf population as the response variable. We can use the equation to predict the y value for a specified x value.

(b) The sample data pairs have x values ranging from 17 (hundred) to 34 (hundred) for the size of the caribou population. To predict the size of the wolf population when the size of the caribou population is 21 (hundred), will you be interpolating or extrapolating?

SOLUTION: Interpolating, since 21 (hundred) falls within the range of sample x values.

(c) Predict the size of the wolf population when the caribou population is 21 (hundred).

SOLUTION: Using the least-squares line from Example 4 and the value 21 in place of x gives

$$\hat{y} \approx 22.35 + 1.60x \approx 22.35 + 1.60(21) \approx 55.95$$

Rounding up to a whole number gives a prediction of 56 for the size of the wolf population.

GUIDED EXERCISE 4 Least-squares line

The Quick Sell car dealership has been using 1-minute spot ads on a local TV station. The ads always occur during the evening hours and advertise the different models and price ranges of cars on the lot that week. During a 10-week period, a Quick Sell dealer kept a weekly record of the number x of TV ads versus the number y of cars sold. The results are given in Table 9-8.

The manager decided that Quick Sell can afford only 12 ads per week. At that level of advertisement, how many cars can Quick Sell expect to sell each week? We'll answer this question in several steps.

TABLE 9-8

x	y
6	15
20	31
0	10
14	16
25	28
16	20
28	40
18	25
10	12
8	15

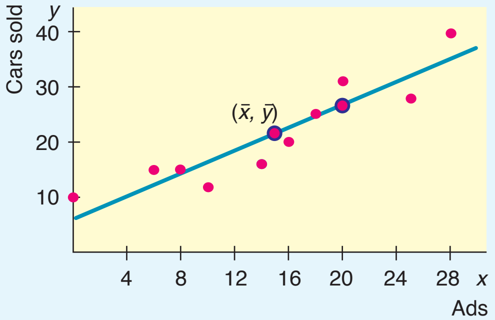
(a) Draw a scatter diagram for the data.

The scatter diagram is shown in Figure 9-13 on the next page. The plain red dots in Figure 9-13 are the points of the scatter diagram. Notice that the least-squares line is also shown with two extra points used to position that line.

Continued

GUIDED EXERCISE 4 *continued*

FIGURE 9-13 Scatter Diagram and Least-Squares Line for Table 9-8



(b) Look at Equations (3) to (5) pertaining to the least-squares line (page 522). Two of the quantities that we need to find b are (Σx) and (Σxy). List the others.

→ We also need n, (Σy), (Σx^2), and $(\Sigma x)^2$.

(c) Complete Table 9-9(a).

→ The missing table entries are shown in Table 9-9(b).

TABLE 9-9(a)

x	y	x^2	xy
6	15	36	90
20	31	400	620
0	10	0	0
14	16	196	224
25	28	625	700
16	20	256	320
28	40	___	___
18	25	___	___
10	12	___	___
8	15	64	120
$\Sigma x = 145$	$\Sigma y = 212$	$\Sigma x^2 =$ ___	$\Sigma xy =$ ___

TABLE 9-9(b)

x^2	xy
$(28)^2 = 784$	$28(40) = 1120$
$(18)^2 = 324$	$18(25) = 450$
$(10)^2 = 100$	$10(12) = 120$
$\Sigma x^2 = 2785$	$\Sigma xy = 3764$

(d) Compute the sample means $\bar{x}$ and $\bar{y}$

→
$$\bar{x} = \frac{\Sigma x}{n} = \frac{145}{10} = 14.5$$
$$\bar{y} = \frac{\Sigma y}{n} = \frac{212}{10} = 21.2$$

(e) Compute a and b for the equation $\hat{y} = a + bx$ of the least-squares line.

→
$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$
$$= \frac{10(3764) - (145)(212)}{10(2785) - (145)^2} = \frac{6900}{6825} \approx 1.01$$
$$a = \bar{y} - b\bar{x}$$
$$\approx 21.2 - 1.01(14.5) \approx 6.56$$

(f) What is the equation of the least-squares line $\hat{y} = a + bx$?

→ Using the values of a and b computed in part (e) or values of a and b obtained directly from a calculator,

$$\hat{y} \approx 6.56 + 1.01x$$

(g) Plot the least-squares line on your scatter diagram.

→ The least-squares line goes through the point $(\bar{x}, \bar{y}) = (14.5, 21.2)$. To get another point on the line,

Continued

GUIDED EXERCISE 4 *continued*

select a value for x and compute the corresponding $\hat{y}$ value using the equation $\hat{y} = 6.56 + 1.01x$. For $x = 20$, we get $\hat{y} = 6.56 + 1.01(20) = 26.8$, so the point (20, 26.8) is also on the line. The least-squares line is shown in Figure 9-13.

(h) Read the $\hat{y}$ value for $x = 12$ from your graph. Then use the equation of the least-squares line to calculate $\hat{y}$ when $x = 12$. How many cars can the manager expect to sell if 12 ads per week are aired on TV?

The graph gives $\hat{y} \approx 19$. From the equation, we get

$$\begin{aligned} \hat{y} &\approx 6.56 + 1.01x \\ &\approx 6.56 + 1.01(12) \quad \text{using 12 in place of } x \\ &\approx 18.68 \end{aligned}$$

To the nearest whole number, the manager can expect to sell 19 cars when 12 ads are aired on TV each week.

(i) *Interpretation* How reliable do you think the prediction is? Explain. (Guided Exercise 5 will show that $r \approx 0.919$.)

The prediction should be fairly reliable. The prediction involves interpolation, and the scatter diagram shows that the data points are clustered around the least-squares line. From the next Guided Exercise we have the information that r is high. Of course, other variables might affect the value of y for $x = 12$.

TECH NOTES

When we have more data pairs, it is convenient to use a technology tool such as the TI-84Plus/TI-83Plus/TI-*n*spire calculators, Excel 2007, or Minitab to find the equation of the least-squares line. The displays show results for the data of Guided Exercise 4 regarding car sales and ads.

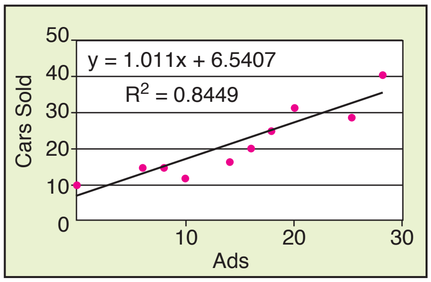
TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) Press **STAT**, choose **Calculate**, and use option **8:LinReg (a + bx)**. For a graph showing the scatter plot and the least-squares line, press the **STAT PLOT** key, turn on a plot, and highlight the first type. Then press the **Y =** key. To enter the equation of the least-squares line, press **VARS**, select **5:Statistics**, highlight **EQ**, and then highlight **1:RegEQ**. Press **ENTER**. Finally, press **ZOOM** and choose **9:ZoomStat**.

Excel 2007 There are several ways to find the equation of the least-squares line in Excel. One way is to make a scatter plot. On the home screen, click the **Insert** tab. In the Chart Group, select **Scatter** and choose the first type. In the next ribbon, the Chart Layout Group offers options for including titles and axes labels. Right click on any data point and select **Add Trendline**. In the dialogue box, select **linear** and check **Display Equation on Chart**. To display the value of the coefficient of determination, check **Display R-squared Value on Chart**.

TI-84Plus/TI-83/TI-*n*spire Display

```
LinReg
 y=a+bx
 a=6.5407
 b=1.0110
 r²=.8449
 r=.9192
```

Excel 2007 Display



Minitab There are a number of ways to generate the least-squares line. One way is to use the menu selection **Stat ➤ Regression ➤ Fitted Line Plot.** The least-squares equation is shown with the diagram.

Coefficient of Determination

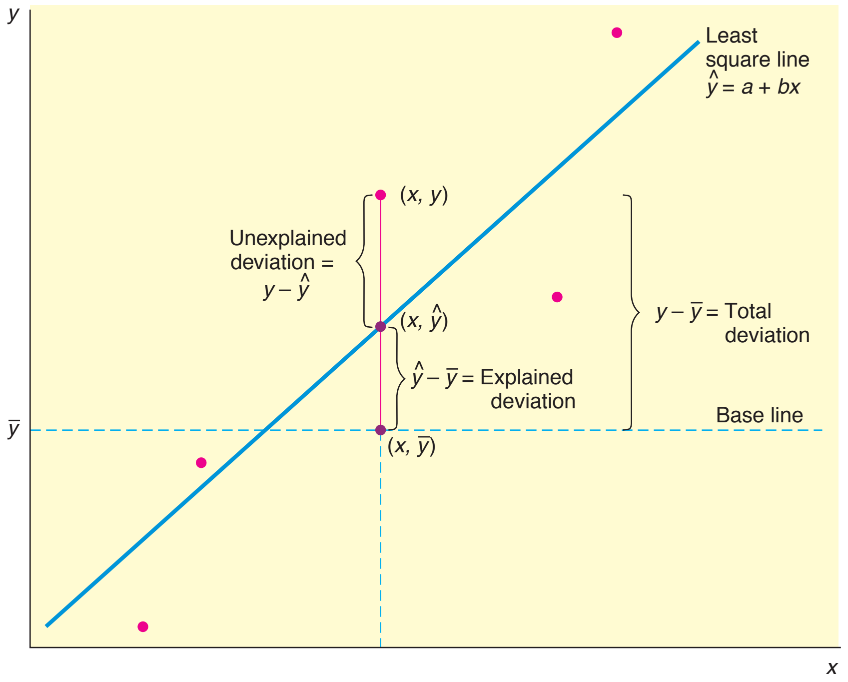
Coefficient of determination r^2

There is another way to answer the question "How good is the least-squares line as an instrument of regression?" The *coefficient of determination* r^2 is the square of the sample correlation coefficient r.

Suppose we have a scatter diagram and corresponding least-squares line, as shown in Figure 9-14.

FIGURE 9-14

Explained and Unexplained Deviations



Let us take the point of view that $\bar{y}$ is a kind of baseline for the y values. If we were given an x value, and if we were completely ignorant of regression and correlation but we wanted to predict a value of y corresponding to the given x, a reasonable guess for y would be the mean $\bar{y}$. However, since we do know how to construct the least-squares regression line, we can calculate $\hat{y} = a + bx$, the predicted value corresponding to x. In most cases, the predicted value $\hat{y}$ on the least-squares line will not be the same as the actual data value y. We will measure deviations (or differences) from the baseline $\bar{y}$. (See Figure 9-14.)

Total deviation $= y - \bar{y}$

Explained deviation $= \hat{y} - \bar{y}$

Unexplained deviation $= y - \hat{y}$ (also known as the *residual*)

The total deviation $y - \bar{y}$ is a measure of how far y is from the baseline $\bar{y}$. This can be broken into two parts. The explained deviation $\hat{y} - \bar{y}$ tells us how far the estimated y value "should" be from the baseline $\bar{y}$. (The "explanation" of this part of the deviation is the least-squares line, so to speak.) The unexplained deviation $y - \hat{y}$ tells us how far our data value y is "off." This amount is called *unexplained* because it is due to random chance and other factors that the least-squares line cannot account for.

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

$$\begin{pmatrix}\text{Total}\\ \text{deviation}\end{pmatrix} = \begin{pmatrix}\text{Explained}\\ \text{deviation}\end{pmatrix} + \begin{pmatrix}\text{Unexplained}\\ \text{deviation}\end{pmatrix}$$

At this point, we wish to include all the data pairs and we wish to deal only with nonnegative values (so that positive and negative deviations won't cancel out).

Therefore, we construct the following equation for the sum of squares. This equation can be derived using some lengthy algebraic manipulations, which we omit.

Explained variation

Unexplained variation

$$\Sigma(y - \bar{y})^2 = \Sigma(\hat{y} - \bar{y})^2 + \Sigma(y - \hat{y})^2$$

$$\begin{pmatrix}\text{Total}\\\text{variation}\end{pmatrix} = \begin{pmatrix}\text{Explained}\\\text{variation}\end{pmatrix} + \begin{pmatrix}\text{Unexplained}\\\text{variation}\end{pmatrix}$$

Note that the sum of *squares* is taken over all data points and is then referred to as *variation* (not deviation).

The preceding concepts are connected together in the following important statement (whose proof we omit):

If r is the sample correlation coefficient [see Equation (2)], then it can be shown that

$$r^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2} = \frac{\text{Explained variation}}{\text{Total variation}}$$

r^2 is called the *coefficient of determination*.

Let us summarize our discussion.

Coefficient of determination r^2

1. Compute the sample correlation coefficient r using the procedure of Section 9.1. Then simply compute r^2, the sample coefficient of determination.
2. *Interpretation* The value r^2 is the ratio of explained variation over total variation. That is, r^2 is the fractional amount of total variation in y that can be explained by using the linear model $\hat{y} = a + bx$.
3. *Interpretation* Furthermore, $1 - r^2$ is the fractional amount of total variation in y that is due to random chance or to the possibility of lurking variables that influence y.

In other words, the coefficient of determination r^2 is a measure of the proportion of variation in y that is explained by the regression line, using x as the explanatory variable. If $r = 0.90$, then $r^2 = 0.81$ is the coefficient of determination. We can say that about 81% of the (variation) behavior of the y variable can be explained by the corresponding (variation) behavior of the x variable if we use the equation of the least-squares line. The remaining 19% of the (variation) behavior of the y variable is due to random chance or to the possibility of lurking variables that influence y.

GUIDED EXERCISE 5 Coefficient of determination r^2

In Guided Exercise 4, we looked at the relationship between x = number of 1-minute spot ads on TV advertising different models of cars and y = number of cars sold each week by the sponsoring car dealership.

(a) Using the sums found in Guided Exercise 4, compute the sample correlation coefficient r. $n = 10$, $\Sigma x = 145$, $\Sigma y = 212$, $\Sigma x^2 = 2785$, and $\Sigma xy = 3764$. You also need $\Sigma y^2 = 5320$.

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$r = \frac{10(3764) - (145)(212)}{\sqrt{10(2785) - (145)^2}\sqrt{10(5320) - (212)^2}}$$

$$\approx \frac{6900}{(82.61)(90.86)}$$

$$\approx 0.919$$

Continued

GUIDED EXERCISE 5 *continued*

(b) Compute the coefficient of determination r^2. → $r^2 \approx 0.845$

(c) ***Interpretation*** What percentage of the variation in the number of car sales can be explained by the ads and the least-squares line? → 84.5%

(d) ***Interpretation*** What percentage of the variation in the number of car sales is not explained by the ads and the least-squares line? → 100% − 84.5%, or 15.5%

VIEWPOINT It's Freezing!

Can you use average temperatures in January to predict how bad the rest of the winter will be? Can you predict the number of days with freezing temperatures for the entire calendar year using conditions in January? How good would such a forecast be for predicting growing season or number of frost-free days? Methods of this section can help you answer such questions. For more information, visit the Brase/Brase statistics site at http://www.cengage.com/statistics/brase *and find the link to Temperatures.*

SECTION 9.2 PROBLEMS

1. ***Statistical Literacy*** In the least-squares line $\hat{y} = 5 - 2x$, what is the value of the slope? When x changes by 1 unit, by how much does $\hat{y}$ change?

2. ***Statistical Literacy*** In the least squares line $\hat{y} = 5 + 3x$, what is the marginal change in $\hat{y}$ for each unit change in x?

3. ***Critical Thinking*** When we use a least-squares line to predict y values for x values beyond the range of x values found in the data, are we extrapolating or interpolating? Are there any concerns about such predictions?

4. ***Critical Thinking*** If two variables have a negative linear correlation, is the slope of the least-squares line positive or negative?

5. ***Critical Thinking: Interpreting Computer Printouts*** We use the form $\hat{y} = a + bx$ for the least-squares line. In some computer printouts, the least-squares equation is not given directly. Instead, the value of the constant a is given, and the coefficient b of the explanatory or predictor variable is displayed. Sometimes a is referred to as the constant, and sometimes as the intercept. Data from *Climatology Report No. 77-3* of the Department of Atmospheric Science, Colorado State University, showed the following relationship between elevation (in thousands of feet) and average number of frost-free days per year in Colorado locations.

   A Minitab printout provides

```
Predictor      Coef      SE Coef      T        P
Constant      318.16      28.31     11.24    0.002
Elevation    -30.878      3.511     -8.79    0.003

s = 11.8603   R-Sq = 96.3%
```

   Notice that "Elevation" is listed under "Predictor." This means that elevation is the explanatory variable x. Its coefficient is the slope b. "Constant" refers to a in the equation $\hat{y} = a + bx$.

(a) Use the printout to write the least-squares equation.
(b) For each 1000-foot increase in elevation, how many fewer frost-free days are predicted?
(c) The printout gives the value of the coefficient of determination r^2. What is the value of r? Be sure to give the correct sign for r based on the sign of b.
(d) ***Interpretation*** What percentage of the variation in y can be *explained* by the corresponding variation in x and the least-squares line? What percentage is *unexplained*?

6. ***Critical Thinking: Interpreting Computer Printouts*** Refer to the description of a computer display for regression described in Problem 5. The following Minitab display gives information regarding the relationship between the body weight of a child (in kilograms) and the metabolic rate of the child (in 100 kcal/24 hr). The data is based on information from *The Merck Manual* (a commonly used reference in medical schools and nursing programs).

```
Predictor      Coef      SE Coef      T        P
Constant      0.8565     0.4148      2.06    0.084
Weight        0.40248    0.02978    13.52    0.000
s = 0.517508   R-Sq = 96.8%
```

(a) Write out the least-squares equation.
(b) For each 1-kilogram increase in weight, how much does the metabolic rate of a child increase?
(c) What is the value of the sample correlation coefficient r?
(d) ***Interpretation*** What percentage of the variation in y can be *explained* by the corresponding variation in x and the least-squares line? What percentage is *unexplained*?

For Problems 7–18, please do the following.
(a) Draw a scatter diagram displaying the data.
(b) Verify the given sums Σx, Σy, Σx^2, Σy^2, and Σxy and the value of the sample correlation coefficient r.
(c) Find $\bar{x}$, $\bar{y}$, a, and b. Then find the equation of the least-squares line $\hat{y} = a + bx$.
(d) Graph the least-squares line on your scatter diagram. Be sure to use the point $(\bar{x}, \bar{y})$ as one of the points on the line.
(e) ***Interpretation*** Find the value of the coefficient of determination r^2. What percentage of the variation in y can be *explained* by the corresponding variation in x and the least-squares line? What percentage is *unexplained*?
Answers may vary slightly due to rounding.

7. ***Economics: Entry-Level Jobs*** An economist is studying the job market in Denver-area neighborhoods. Let x represent the total number of jobs in a given neighborhood, and let y represent the number of entry-level jobs in the same neighborhood. A sample of six Denver neighborhoods gave the following information (units in hundreds of jobs).

x	16	33	50	28	50	25
y	2	3	6	5	9	3

Source: Neighborhood Facts, The Piton Foundation. To find out more, visit the Brase/Brase statistics site at http://www.cengage.com/statistics/brase *and find the link to the Piton Foundation.*

Complete parts (a) through (e), given $\Sigma x = 202$, $\Sigma y = 28$, $\Sigma x^2 = 7754$, $\Sigma y^2 = 164$, $\Sigma xy = 1096$, and $r \approx 0.860$.
(f) For a neighborhood with $x = 40$ jobs, how many are predicted to be entry-level jobs?

8. *Ranching: Cattle* You are the foreman of the Bar-S cattle ranch in Colorado. A neighboring ranch has calves for sale, and you are going to buy some to add to the Bar-S herd. How much should a healthy calf weigh? Let x be the age of the calf (in weeks), and let y be the weight of the calf (in kilograms). The following information is based on data taken from *The Merck Veterinary Manual* (a reference used by many ranchers).

x	1	3	10	16	26	26
y	42	50	75	100	150	200

Complete parts (a) through (e), given $\Sigma x = 92$, $\Sigma y = 617$, $\Sigma x^2 = 2338$, $\Sigma y^2 = 82{,}389$, $\Sigma xy = 13{,}642$, and $r \approx 0.998$.

(f) The calves you want to buy are 12 weeks old. What does the least-squares line predict for a healthy weight?

9. *Weight of Car: Miles per Gallon* Do heavier cars really use more gasoline? Suppose a car is chosen at random. Let x be the weight of the car (in hundreds of pounds), and let y be the miles per gallon (mpg). The following information is based on data taken from *Consumer Reports* (Vol. 62, No. 4).

x	27	44	32	47	23	40	34	52
y	30	19	24	13	29	17	21	14

Complete parts (a) through (e), given $\Sigma x = 299$, $\Sigma y = 167$, $\Sigma x^2 = 11{,}887$, $\Sigma y^2 = 3773$, $\Sigma xy = 5814$, and $r \approx -0.946$.

(f) Suppose a car weighs $x = 38$ (hundred pounds). What does the least-squares line forecast for y = miles per gallon?

10. *Basketball: Fouls* Data for this problem are based on information from *STATS Basketball Scoreboard*. It is thought that basketball teams that make too many fouls in a game tend to lose the game even if they otherwise play well. Let x be the number of fouls that were more than (i.e., over and above) the number of fouls made the opposing team made. Let y be the percentage of times the team with the larger number of fouls won the game.

x	0	2	5	6
y	50	45	33	26

Complete parts (a) through (e), given $\Sigma x = 13$, $\Sigma y = 154$, $\Sigma x^2 = 65$, $\Sigma y^2 = 6290$, $\Sigma xy = 411$, and $r \approx -0.988$.

(f) If a team had $x = 4$ fouls over and above the opposing team, what does the least-squares equation forecast for y?

11. *Auto Accidents: Age* Data for this problem are based on information taken from the *Wall Street Journal*. Let x be the age in years of a licensed automobile driver. Let y be the percentage of all fatal accidents (for a given age) due to speeding. For example, the first data pair indicates that 36% of all fatal accidents involving 17-year-olds are due to speeding.

x	17	27	37	47	57	67	77
y	36	25	20	12	10	7	5

Complete parts (a) through (e), given $\Sigma x = 329$, $\Sigma y = 115$, $\Sigma x^2 = 18{,}263$, $\Sigma y^2 = 2639$, $\Sigma xy = 4015$, and $r \approx -0.959$.

(f) Predict the percentage of all fatal accidents due to speeding for 25-year-olds.

12. ***Auto Accidents: Age*** Let x be the age of a licensed driver in years. Let y be the percentage of all fatal accidents (for a given age) due to failure to yield the right-of-way. For example, the first data pair states that 5% of all fatal accidents of 37-year-olds are due to failure to yield the right-of-way. The *Wall Street Journal* article referenced in Problem 11 reported the following data:

x	37	47	57	67	77	87
y	5	8	10	16	30	43

Complete parts (a) through (e), given $\Sigma x = 372$, $\Sigma y = 112$, $\Sigma x^2 = 24{,}814$, $\Sigma y^2 = 3194$, $\Sigma xy = 8254$, and $r \approx -0.943$.

(f) Predict the percentage of all fatal accidents due to failing to yield the right-of-way for 70-year-olds.

13. ***Income: Medical Care*** Let x be per capita income in thousands of dollars. Let y be the number of medical doctors per 10,000 residents. Six small cities in Oregon gave the following information about x and y (based on information from *Life in America's Small Cities* by G. S. Thomas, Prometheus Books).

x	8.6	9.3	10.1	8.0	8.3	8.7
y	9.6	18.5	20.9	10.2	11.4	13.1

Complete parts (a) through (e), given $\Sigma x = 53$, $\Sigma y = 83.7$, $\Sigma x^2 = 471.04$, $\Sigma y^2 = 1276.83$, $\Sigma xy = 755.89$, and $r \approx 0.934$.

(f) Suppose a small city in Oregon has a per capita income of 10 thousand dollars. What is the predicted number of M.D.s per 10,000 residents?

14. ***Violent Crimes: Prisons*** Does prison really deter violent crime? Let x represent percent change in the rate of violent crime and y represent percent change in the rate of imprisonment in the general U.S. population. For 7 recent years, the following data have been obtained (Source: *The Crime Drop in America*, edited by Blumstein and Wallman, Cambridge University Press).

x	6.1	5.7	3.9	5.2	6.2	6.5	11.1
y	−1.4	−4.1	−7.0	−4.0	3.6	−0.1	−4.4

Complete parts (a) through (e), given $\Sigma x = 44.7$, $\Sigma y = -17.4$, $\Sigma x^2 = 315.85$, $\Sigma y^2 = 116.1$, $\Sigma xy = -107.18$, and $r \approx 0.084$.

(f) ***Critical Thinking*** Considering the values of r and r^2, does it make sense to use the least-squares line for prediction? Explain.

15. ***Education: Violent Crime*** The following data are based on information from the book *Life in America's Small Cities* (by G. S. Thomas, Prometheus Books). Let x be the percentage of 16- to 19-year-olds not in school and not high school graduates. Let y be the reported violent crimes per 1000 residents. Six small cities in Arkansas (Blytheville, El Dorado, Hot Springs, Jonesboro, Rogers, and Russellville) reported the following information about x and y:

x	24.2	19.0	18.2	14.9	19.0	17.5
y	13.0	4.4	9.3	1.3	0.8	3.6

Complete parts (a) through (e), given $\Sigma x = 112.8$, $\Sigma y = 32.4$, $\Sigma x^2 = 2167.14$, $\Sigma y^2 = 290.14$, $\Sigma xy = 665.03$, and $r \approx 0.764$.

(f) If the percentage of 16- to 19-year-olds not in school and not graduates reaches 24% in a similar city, what is the predicted rate of violent crimes per 1000 residents?

16. *Research: Patents* The following data are based on information from the *Harvard Business Review* (Vol. 72, No. 1). Let x be the number of different research programs, and let y be the mean number of patents per program. As in any business, a company can spread itself too thin. For example, too many research programs might lead to a decline in overall research productivity. The following data are for a collection of pharmaceutical companies and their research programs:

x	10	12	14	16	18	20
y	1.8	1.7	1.5	1.4	1.0	0.7

Complete parts (a) through (e), given $\Sigma x = 90$, $\Sigma y = 8.1$, $\Sigma x^2 = 1420$, $\Sigma y^2 = 11.83$, $\Sigma xy = 113.8$, and $r \approx -0.973$.

(f) Suppose a pharmaceutical company has 15 different research programs. What does the least-squares equation forecast for y = mean number of patents per program?

17. *Archaeology: Artifacts* Data for this problem are based on information taken from *Prehistoric New Mexico: Background for Survey* (by D. E. Stuart and R. P. Gauthier, University of New Mexico Press). It is thought that prehistoric Indians did not take their best tools, pottery, and household items when they visited higher elevations for their summer camps. It is hypothesized that archaeological sites tend to lose their cultural identity and specific cultural affiliation as the elevation of the site increases. Let x be the elevation (in thousands of feet) of an archaeological site in the southwestern United States. Let y be the percentage of unidentified artifacts (no specific cultural affiliation) at a given elevation. The following data were obtained for a collection of archaeological sites in New Mexico:

x	5.25	5.75	6.25	6.75	7.25
y	19	13	33	37	62

Complete parts (a) through (e), given $\Sigma x = 31.25$, $\Sigma y = 164$, $\Sigma x^2 = 197.813$, $\Sigma y^2 = 6832$, $\Sigma xy = 1080$, and $r \approx 0.913$.

(f) At an archaeological site with elevation 6.5 (thousand feet), what does the least-squares equation forecast for y = percentage of culturally unidentified artifacts?

18. *Cricket Chirps: Temperature* Anyone who has been outdoors on a summer evening has probably heard crickets. Did you know that it is possible to use the cricket as a thermometer? Crickets tend to chirp more frequently as temperatures increase. This phenomenon was studied in detail by George W. Pierce, a physics professor at Harvard. In the following data, x is a random variable representing chirps per second and y is a random variable representing temperature (°F). These data are also available for download at the Online Study Center.

Dorling Kindersley/Getty Images

x	20.0	16.0	19.8	18.4	17.1	15.5	14.7	17.1
y	88.6	71.6	93.3	84.3	80.6	75.2	69.7	82.0

x	15.4	16.2	15.0	17.2	16.0	17.0	14.4
y	69.4	83.3	79.6	82.6	80.6	83.5	76.3

Source: Reprinted by permission of the publisher from *The Songs of Insects* by George W. Pierce, Cambridge, Mass.: Harvard University Press, Copyright © 1948 by the President and Fellows of Harvard College.

Complete parts (a) through (e), given $\Sigma x = 249.8$, $\Sigma y = 1200.6$, $\Sigma x^2 = 4200.56$, $\Sigma y^2 = 96.725.86$, $\Sigma xy = 20{,}127.47$, and $r \approx 0.835$.

(f) What is the predicted temperature when $x = 19$ chirps per second?

19. ***Expand Your Knowledge: Residual Plot*** The least-squares line usually does not go through all the sample data points (x, y). In fact, for a specified x value from a data pair (x, y), there is usually a difference between the predicted value and the y value paired with x. This difference is called the *residual*.

The **residual** is the difference between the y value in a specified data pair (x, y) and the value $\hat{y} = a + bx$ predicted by the least-squares line for the same x.

$y - \hat{y}$ is the **residual**.

Residual plot

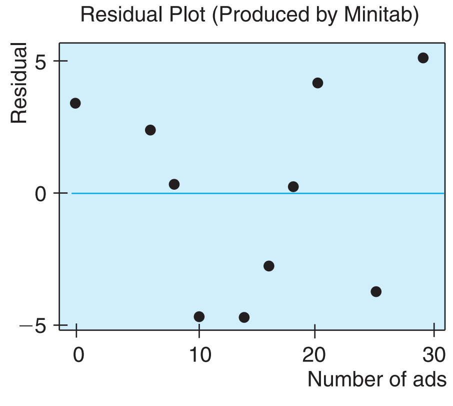
One way to assess how well a least-squares line serves as a model for the data is a **residual plot**. To make a residual plot, we put the x values in order on the horizontal axis and plot the corresponding residuals $y - \hat{y}$ in the vertical direction. Because the mean of the residuals is always zero for a least-squares model, we place a horizontal line at zero. The accompanying figure shows a residual plot for the data of Guided Exercise 4, in which the relationship between the number of ads run per week and the number of cars sold that week was explored. To make the residual plot, first compute all the residuals. Remember that x and y are the given data values, and $\hat{y}$ is computed from the least-squares line $\hat{y} \approx 6.56 + 1.01x$.

Residual

x	y	$\hat{y}$	$y - \hat{y}$
6	15	12.6	2.4
20	31	26.8	4.2
0	10	6.6	3.4
14	16	20.7	−4.7
25	28	31.8	−3.8

Residual

x	y	$\hat{y}$	$y - \hat{y}$
16	20	22.7	−2.7
28	40	34.8	5.2
18	25	24.7	0.3
10	12	16.7	−4.7
8	15	14.6	0.4



(a) If the least-squares line provides a reasonable model for the data, the pattern of points in the plot will seem random and unstructured about the horizontal line at 0. Is this the case for the residual plot?
(b) If a point on the residual plot seems far outside the pattern of other points, it might reflect an unusual data point (x, y), called an *outlier*. Such points may have quite an influence on the least-squares model. Do there appear to be any outliers in the data for the residual plot?

20. ***Residual Plot: Miles per Gallon*** Consider the data of Problem 9.
(a) Make a residual plot for the least-squares model.
(b) Use the residual plot to comment about the appropriateness of the least-squares model for these data. See Problem 19.

21. *Critical Thinking: Exchange x and y in Least-Squares Equation*

(a) Suppose you are given the following (x, y) data pairs:

x	1	3	4
y	2	1	6

Show that the least-squares equation for these data is $y = 1.071x + 0.143$ (rounded to three digits after the decimal).

(b) Now suppose you are given these (x, y) data pairs:

x	2	1	6
y	1	3	4

Show that the least-squares equation for these data is $y = 0.357x + 1.595$ (rounded to three digits after the decimal).

(c) In the data for parts (a) and (b), did we simply exchange the x and y values of each data pair?

(d) Solve $y = 0.143 + 1.071x$ for x. Do you get the least-squares equation of part (b) with the symbols x and y exchanged?

(e) In general, suppose we have the least-squares equation $y = a + bx$ for a set of data pairs (x, y). If we solve this equation for x, will we *necessarily* get the least-squares equation for the set of data pairs (y, x) (with x and y exchanged)? Explain using parts (a) through (d).

22. *Expand Your Knowledge: Logarithmic Transformations, Exponential Growth Model* There are several extensions of linear regression that apply to exponential growth and power law models. Problems 22–25 will outline some of these extensions. First of all, recall that a variable grows *linearly* over time if it *adds* a fixed increment during each equal time period. *Exponential* growth occurs when a variable is *multiplied* by a fixed number during each time period. This means that exponential growth increases by a fixed multiple or percentage of the previous amount. College algebra can be used to show that if a variable grows exponentially, then its logarithm grows linearly. The exponential growth model is $y = \alpha\beta^x$, where α and β are fixed constants to be estimated from data.

How do we know when we are dealing with exponential growth, and how can we estimate α and β? Please read on. Populations of living things such as bacteria, locusts, fish, panda bears, and so on, tend to grow (or decline) exponentially. However, these populations can be restricted by outside limitations such as food, space, pollution, disease, hunting, and so on. Suppose we have data pairs (x, y) for which there is reason to believe the scatter plot is not linear, but rather exponential, as described above. This means the increase in y values begins rather slowly but then seems to explode. *Note:* For exponential growth models, we assume all $y > 0$.

Consider the following data, where x = time in hours and y = number of bacteria in a laboratory culture at the end of x hours.

x	1	2	3	4	5
y	3	12	22	51	145

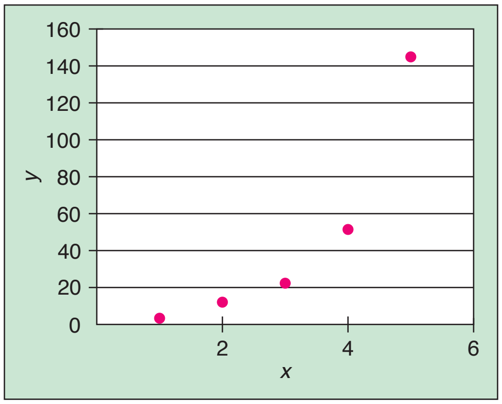
(a) Look at the Excel graph of the scatter diagram of the (x, y) data pairs. Do you think a straight line will be a good fit to these data? Do the y values seem almost to explode as time goes on?

(b) Now consider a transformation $y' = \log y$. We are using common logarithms of base 10 (however, natural logarithms of base e would work just as well).

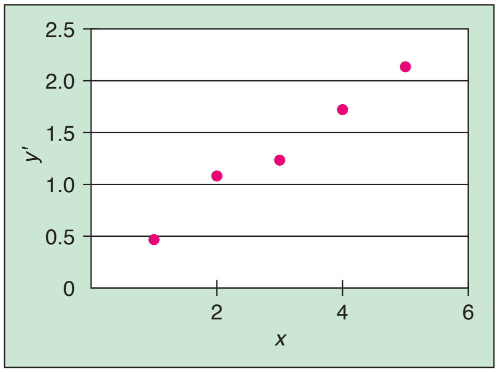
x	1	2	3	4	5
$y' = \log y$	0.477	1.079	1.342	1.748	2.161

Look at the Excel graph of the scatter diagram of the (x, y') data pairs and compare this diagram with the diagram in part (a). Which graph appears to better fit a straight line?

Excel Graphs



Part (a) Model with (x, y) Data Pairs



Part (b) Model with (x, y') Data Pairs

(c) Use a calculator with regression keys to verify the linear regression equation for the (x, y) data pairs, $\hat{y} = -50.3 + 32.3x$, with sample correlation coefficient $r = 0.882$.

(d) Use a calculator with regression keys to verify the linear regression equation for the (x, y') data pairs, $y' = 0.150 + 0.404x$, with sample correlation coefficient $r = 0.994$. The sample correlation coefficient $r = 0.882$ for the (x, y) pairs is not bad. But the sample correlation coefficient $r = 0.994$ for the (x, y') pairs is a lot better!

(e) The exponential growth model is $y = \alpha\beta^x$. Let us use the results of part (d) to estimate α and β for this strain of laboratory bacteria. The equation $y' = a + bx$ is the same as $\log y = a + bx$. If we raise both sides of this equation to the power 10 and use some college algebra, we get $y = 10^a(10^b)^x$. Thus, $\alpha \approx 10^a$ and $\beta \approx 10^b$. Use these results to approximate α and β and write the exponential growth equation for our strain of bacteria.

Note: The TI-84Plus/TI-83Plus/TI-*n*spire calculators fully support the exponential growth model. Place the original x data in list L1 and the corresponding y data in list L2. Then press **STAT,** followed by **CALC,** and scroll down to option **0: ExpReg.** The output gives values for α, β, and the sample correlation coefficient r.

23. ***Expand Your Knowledge: Logarithmic Transformations, Exponential Growth Model*** Let x = day of observation and y = number of locusts per square meter during a locust infestation in a region of North Africa.

x	2	3	5	8	10
y	2	3	12	125	630

(a) Draw a scatter diagram of the (x, y) data pairs. Do you think a straight line will be a good fit to these data? Do the y values almost seem to explode as time goes on?

(b) Now consider a transformation $y' = \log y$. We are using common logarithms of base 10. Draw a scatter diagram of the (x, y') data pairs and compare this diagram with the diagram of part (a). Which graph appears to better fit a straight line?

(c) Use a calculator with regression keys to find the linear regression equation for the data pairs (x, y'). What is the correlation coefficient?

(d) The exponential growth model is $y = \alpha\beta^x$. Estimate α and β and write the exponential growth equation. *Hint:* See Problem 22.

24. ***Expand Your Knowledge: Logarithmic Transformations, Power Law Model*** When we take measurements of the same general type, a power law of the form $y = \alpha x^{\beta}$ often gives an excellent fit to the data. A lot of research has been conducted as to why power laws work so well in business, economics, biology, ecology, medicine, engineering, social science, and so on. Let us just say that if you do not have a good straight-line fit to data pairs (x, y), and the scatter plot does not rise dramatically (as in exponential growth), then a power law is often a good choice. College algebra can be used to show that power law models become linear when we apply logarithmic transformations to both variables. To see how this is done, please read on. *Note:* For power law models, we assume all $x > 0$ and all $y > 0$.

Suppose we have data pairs (x, y) and we want to find constants α and β such that $y = \alpha x^{\beta}$ is a good fit to the data. First, make the logarithmic transformations $x' = \log x$ and $y' = \log y$. Next, use the (x', y') data pairs and a calculator with linear regression keys to obtain the least-squares equation $y' = a + bx'$. Note that the equation $y' = a + bx'$ is the same as $\log y = a + b(\log x)$. If we raise both sides of this equation to the power 10 and use some college algebra, we get $y = 10^{a}(x)^{b}$. In other words, for the power law model, we have $\alpha \approx 10^{a}$ and $\beta \approx b$.

In the electronic design of a cell phone circuit, the buildup of electric current (Amps) is an important function of time (microseconds). Let x = time in microseconds and let y = Amps built up in the circuit at time x.

x	2	4	6	8	10
y	1.81	2.90	3.20	3.68	4.11

(a) Make the logarithmic transformations $x' = \log x$ and $y' = \log y$. Then make a scatter plot of the (x', y') values. Does a linear equation seem to be a good fit to this plot?
(b) Use the (x', y') data points and a calculator with regression keys to find the least-squares equation $y' = a + bx'$. What is the sample correlation coefficient?
(c) Use the results of part (b) to find estimates for α and β in the power law $y = \alpha x^{\beta}$. Write the power law giving the relationship between time and Amp buildup.

Note: The TI-84Plus/TI-83Plus/TI-*n*spire calculators fully support the power law model. Place the original x data in list L1 and the corresponding y data in list L2. Then press **STAT,** followed by **CALC,** and scroll down to option **A: PwrReg.** The output gives values for α, β, and the sample correlation coefficient r.

25. ***Expand Your Knowledge: Logarithmic Transformations, Power Law Model*** Let x = boiler steam pressure in 100 lb/in.^{2} and let y = critical sheer strength of boiler plate steel joints in tons/in.^{2}. We have the following data for a series of factory boilers.

x	4	5	6	8	10
y	3.4	4.2	6.3	10.9	13.3

(a) Make the logarithmic transformations $x' = \log x$ and $y' = \log y$. Then make a scatter plot of the (x', y') values. Does a linear equation seem to be a good fit to this plot?
(b) Use the (x', y') data points and a calculator with regression keys to find the least-squares equation $y' = a + bx'$. What is the sample correlation coefficient?
(c) Use the results of part (b) to find estimates for α and β in the power law $y = \alpha x^{\beta}$. Write the power equation for the relationship between steam pressure and sheer strength of boiler plate steel. *Hint:* See Problem 24.

SECTION 9.3

Inferences for Correlation and Regression

FOCUS POINTS

- Test the correlation coefficient ρ.
- Use sample data to compute the standard error of estimate S_e.
- Find a confidence interval for the value of y predicted for a specified value of x.
- Test the slope β of the least-squares line.
- Find a confidence interval for the slope β of the least-squares line and interpret its meaning.

Learn more, earn more! We have probably all heard this platitude. The question is whether or not there is some truth in this statement. Do college graduates have an improved chance at a better income? Is there a trend in the general population to support the "learn more, earn more" statement?

Consider the following variables: x = percentage of the population 25 or older with at least four years of college and y = percentage *growth* in per capita income over the past seven years. A random sample of six communities in Ohio gave the information (based on *Life in America's Small Cities* by G. S. Thomas) shown in Table 9-10 on the next page.

If we use what we learned in Sections 9.1 and 9.2, we can compute the sample correlation coefficient r and the least-squares line $\hat{y} = a + bx$ using the data of Table 9-10. However, r is only a *sample* correlation coefficient, and $\hat{y} = a + bx$ is only a *"sample-based"* least-squares line. What if we used *all* possible data pairs (x, y) from *all* U.S. cities, not just six towns in Ohio? If we accomplished this seemingly impossible task, we would have the *population* of all (x, y) pairs.

Population correlation coefficient ρ

From this population of (x, y) pairs, we could (in theory) compute the *population correlation coefficient*, which we call ρ (Greek letter rho, pronounced like "row"). We could also compute the least-squares line for the entire population, which we denote as $y = \alpha + \beta x$ using more Greek letters, α (alpha) and β (beta).

Sample Statistic		Population Parameter
r	$\rightarrow$	ρ
a	$\rightarrow$	α
b	$\rightarrow$	β
$\hat{y} = a + bx$	$\rightarrow$	$y = \alpha + \beta x$

Requirements for inferences concerning linear regression

Requirements for statistical inference

To make inferences regarding correlation and linear regression, we need to be sure that

(a) The set (x, y) of ordered pairs is a *random sample* from the population of all possible such (x, y) pairs.

(b) For each fixed value of x, the y values have a normal distribution. All of the y distributions have the same variance, and, for a given x value, the distribution of y values has a mean that lies on the least-squares line. We also assume that for a fixed y, each x has its own normal distribution. In most cases the results are still accurate if the distributions are simply mound-shaped and symmetric and the y variances are approximately equal.

We assume these conditions are met for all inferences presented in this section.

Testing the Correlation Coefficient

The first topic we want to study is the statistical significance of the sample correlation coefficient r. To do this, we construct a statistical test of ρ, the population correlation coefficient. The test will be based on the following theorem.

THEOREM 9.1 Let r be the sample correlation coefficient computed using data pairs (x, y). We use the null hypothesis

H_0: x and y have no linear correlation, so $\rho = 0$

The alternate hypothesis may be

$H_1: \rho > 0$ or $H_1: \rho < 0$ or $H_1: \rho \neq 0$

The conversion of r to a Student's t distribution is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{with } d.f. = n - 2$$

where n is the number of sample data pairs (x, y) $(n \geq 3)$.

PROCEDURE

HOW TO TEST THE POPULATION CORRELATION COEFFICIENT ρ

1. Use the *null hypothesis* $H_0: \rho = 0$. In the context of the application, state the *alternate hypothesis* ($\rho > 0$ or $\rho < 0$ or $\rho \neq 0$) and set the *level of significance* α.
2. Obtain a random sample of $n \geq 3$ data pairs (x, y) and compute the *sample test statistic*

   $$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{with degrees of freedom } d.f. = n - 2$$

   where r is the sample correlation coefficient
   n is the sample size
3. Use a Student's t distribution and the type of test, one-tailed or two-tailed, to find (or estimate) the *P-value* corresponding to the test statistic.
4. *Conclude* the test. If P-value $\leq \alpha$, then reject H_0. If P-value $> \alpha$, then do not reject H_0.
5. *Interpret your conclusion* in the context of the application.

Problem 13 at the end of this section discusses how sample size might affect the significance of r.

EXAMPLE 6 TESTING ρ

Let's return to our data from Ohio regarding the percentage of the population with at least four years of college and the percentage of growth in per capita income (Table 9-10). We'll develop a test for the population correlation coefficient ρ.

SOLUTION: First, we compute the sample correlation coefficient r. Using a calculator, statistical software, or a "by-hand" calculation from Section 9.1, we find

$r \approx 0.887$

Now we test the population correlation coefficient ρ. Remember that x represents percentage college graduates and y represents percentage salary increases in

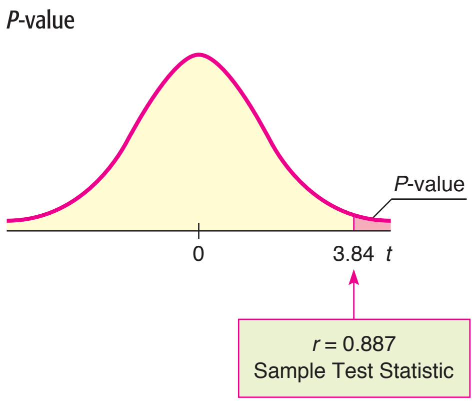
TABLE 9-10 Education and Income Growth Percentages

x	9.9	11.4	8.1	14.7	8.5	12.6
y	37.1	43.0	33.4	47.1	26.5	40.2

TABLE 9-11 Excerpt from Student's *t* Distribution

✓one-tail area	0.010	0.005
two-tail area	0.020	0.010
d.f. = 4	3.747	4.604
	↑ Sample t = 3.84	

FIGURE 9-15



the general population. We suspect the population correlation is positive, $\rho > 0$. Let's use a 1% level of significance:

H_0: $\rho = 0$ (no linear correlation)

H_1: $\rho > 0$ (positive linear correlation)

Convert the sample test statistic $r = 0.887$ to t using $n = 6$.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.887\sqrt{6-2}}{\sqrt{1-0.887^2}} \approx 3.84 \quad \text{with } d.f. = n - 2 = 6 - 2 = 4$$

The P-value for the sample test statistic $t = 3.84$ is shown in Figure 9-15. Since we have a right-tailed test, we use the one-tail area in the Student's t distribution (Table 6 of Appendix II).

From Table 9-11, we see that

$$0.005 < P\text{-value} < 0.010$$

Since the interval containing the P-value is less than the level of significance $\alpha = 0.01$, we reject H_0 and conclude that the population correlation coefficient between x and y is positive. Technology gives P-value ≈ 0.0092.

Caution: Although we have shown that x and y are positively correlated, we have not shown that an increase in education *causes* an increase in earnings.

GUIDED EXERCISE 6 *Testing ρ*

A medical research team is studying the effect of a new drug on red blood cells. Let x be a random variable representing milligrams of the drug given to a patient. Let y be a random variable representing red blood cells per cubic milliliter of whole blood. A random sample of $n = 7$ volunteer patients gave the following results.

x	9.2	10.1	9.0	12.5	8.8	9.1	9.5
y	5.0	4.8	4.5	5.7	5.1	4.6	4.2

Use a calculator to verify that $r \approx 0.689$. Then use a 1% level of significance to test the claim that $\rho \neq 0$.

Continued

GUIDED EXERCISE 6 *continued*

(a) State the null and alternate hypotheses. What is the level of significance α?

H_0: $\rho = 0$; H_1: $\rho \neq 0$; $\alpha = 0.01$

(b) Compute the sample test statistic.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \approx \frac{0.689\sqrt{7-2}}{\sqrt{1-0.689^2}} \approx \frac{1.5406}{0.7248} \approx 2.126$$

(c) Use the Student's t distribution, Table 6 of Appendix II, to estimate the P-value.

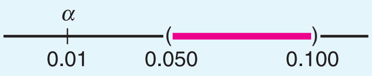
$d.f. = n - 2 = 7 - 2 = 5$; two-tailed test

✓ two-tail area	0.100	0.050
d.f. = 5	2.015	2.571
	↑ Sample t = 2.126	

$0.050 < P\text{-value} < 0.100$

(d) Do we reject or fail to reject H_0?

Since the interval containing the P-value lies to the right of $\alpha = 0.01$, we do not reject H_0. Technology gives P-value ≈ 0.0866.



(e) *Interpret* the conclusion in the context of the application.

At the 1% level of significance, the evidence is not strong enough to indicate any correlation between the amount of drug administered and the red blood cell count.

Standard Error of Estimate

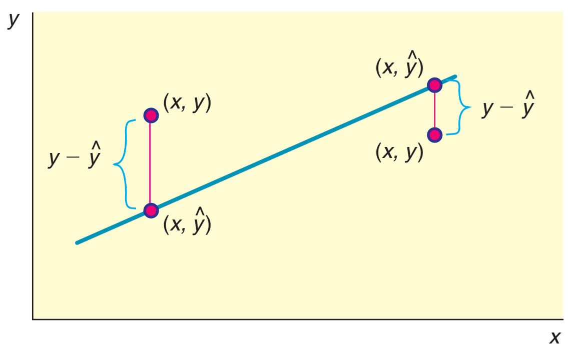
Sometimes a scatter diagram clearly indicates the existence of a linear relationship between x and y, but it can happen that the points are widely scattered about the least-squares line. We need a method (besides just looking) for measuring the spread of a set of points about the least-squares line. There are three common methods of measuring the spread. One method uses the *standard error of estimate*. The others use the *coefficient of correlation* and the *coefficient of determination*.

For the standard error of estimate, we use a measure of spread that is in some ways like the standard deviation of measurements of a single variable. Let

$$\hat{y} = a + bx$$

FIGURE 9-16

The Distance Between Points (x, y) and $(x, \hat{y})$



be the predicted value of y from the least-squares line. Then $y - \hat{y}$ is the difference between the y value of the *data point* (x, y) shown on the scatter diagram (Figure 9-16) and the $\hat{y}$ value of the point on the *least-squares line* with the same x value. The quantity $y - \hat{y}$ is known as the *residual*. To avoid the difficulty of having some positive and some negative values, we square the quantity $(y - \hat{y})$. Then we sum the squares and, for technical reasons, divide this sum by $n - 2$. Finally, we take the square root to obtain the *standard error of estimate*, denoted by S_e.

Residual

Standard error of estimate S_e

$$\textbf{Standard error of estimate} = S_e = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}} \qquad (7)$$

where $\hat{y} = a + bx$ and $n \geq 3$.

Note: To compute the standard error of estimate, we require that there be at least three points on the scatter diagram. If we had only two points, the line would be a perfect fit, since two points determine a line. In such a case, there would be no need to compute S_e.

The nearer the scatter points lie to the least-squares line, the smaller S_e will be. In fact, if $S_e = 0$, it follows that each $y - \hat{y}$ is also zero. This means that all the scatter points lie *on* the least-squares line if $S_e = 0$. The larger S_e becomes, the more scattered the points are.

The formula for the standard error of estimate is reminiscent of the formula for the standard deviation, which is also a measure of dispersion. However, the standard deviation involves differences of data values from a mean, whereas the standard error of estimate involves the differences between experimental and predicted y values for a given x (i.e., $y - \hat{y}$).

The actual computation of S_e using Equation (7) is quite long because the formula requires us to use the least-squares line equation to compute a predicted value $\hat{y}$ for *each* x value in the data pairs. There is a computational formula that we strongly recommend you use. However, as with all the computation formulas, be careful about rounding. This formula is sensitive to rounding, and you should carry as many digits as seem reasonable for your problem. Answers will vary, depending on the rounding used. We give the formula here and follow it with an example of its use.

PROCEDURE

How to find the standard error of estimate S_e

1. Obtain a random sample of $n \geq 3$ data pairs (x, y).
2. Use the procedures of Section 9.2 to find a and b from the sample least-squares line $\hat{y} = a + bx$.
3. The standard error of estimate is

$$S_e = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n - 2}} \qquad (8)$$

Computation formula for S_e

With a considerable amount of algebra, Equations (7) and (8) can be shown to be mathematically equivalent. Equation (7) shows the strong similarity between the standard error of estimate and the standard deviation. Equation (8) is a shortcut calculation formula because it involves few subtractions. The sums Σx, Σy, Σx^2, Σy^2, and Σxy are provided directly on most calculators that support two-variable statistics.

In the next example, we show you how to compute the standard error of estimate using the computation formula.

EXAMPLE 7

Least-squares line and S_e

June and Jim are partners in the chemistry lab. Their assignment is to determine how much copper sulfate ($CuSO_4$) will dissolve in water at 10, 20, 30, 40, 50, 60, and 70°C. Their lab results are shown in Table 9-12, where y is the weight in grams of copper sulfate that will dissolve in 100 grams of water at x°C.

Sketch a scatter diagram, find the equation of the least-squares line, and compute S_e.

Photo Researchers

SOLUTION: Figure 9-17 includes a scatter diagram for the data of Table 9-12. To find the equation of the least-squares line and the value of S_e, we set up a computational table (Table 9-13).

$$\bar{x} = \frac{\Sigma x}{n} = \frac{280}{7} = 40 \quad \text{and} \quad \bar{y} = \frac{\Sigma y}{n} = \frac{213}{7} \approx 30.429$$

$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{7(9940) - (280)(213)}{7(14{,}000) - (280)^2} = \frac{9940}{19{,}600} \approx 0.50714$$

$$a = \bar{y} - b\bar{x} \approx 30.429 - 0.507(40) \approx 10.149$$

The equation of the least-squares line is

$$\hat{y} = a + bx$$
$$\hat{y} \approx 10.14 + 0.51x$$

The graph of the least-squares line is shown in Figure 9-17. Notice that it passes through the point $(\bar{x}, \bar{y}) = (40, 30.4)$. Another point on the line can be found by using $x = 15$ in the equation of the line $\hat{y} = 10.14 + 0.51x$. When we use 15 in place of x, we obtain $\hat{y} = 10.14 + 0.51(15) = 17.8$. The point $(15, 17.8)$ is the other point we used to graph the least-squares line in Figure 9-17.

The standard error of estimate is computed using the computational formula

$$S_e = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n - 2}}$$

$$\approx \sqrt{\frac{7229 - 10.149(213) - 0.507(9940)}{7 - 2}} \approx \sqrt{\frac{27.683}{5}} \approx 2.35$$

Note: This formula is very sensitive to rounded values of a and b.

TABLE 9-12 **Lab Results (x = °C, y = amount of $CuSo_4$)**

x	y
10	17
20	21
30	25
40	28
50	33
60	40
70	49

FIGURE 9-17

Scatter Diagram and Least-Squares Line for Chemistry Experiment

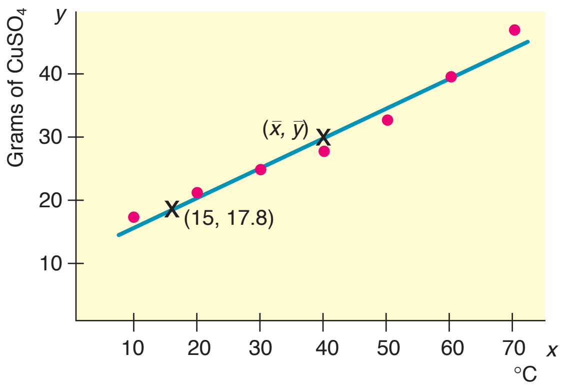


TABLE 9-13 **Computational Table**

x	y	x^2	y^2	xy
10	17	100	289	170
20	21	400	441	420
30	25	900	625	750
40	28	1600	784	1120
50	33	2500	1089	1650
60	40	3600	1600	2400
70	49	4900	2401	3430
$\Sigma x = 280$	$\Sigma y = 213$	$\Sigma x^2 = 14{,}000$	$\Sigma y^2 = 7229$	$\Sigma xy = 9940$

TECH NOTES Although many calculators that support two-variable statistics and linear regression do not provide the value of the standard error of estimate S_e directly, they do provide the sums required for the calculation of S_e. The TI-84Plus/TI-83Plus/TI-*n*spire, Excel 2007, and Minitab all provide the value of S_e.

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) The value for S_e is given as s under **STAT, TEST,** option **E: LinRegTTest.**

Excel 2007 Click the **Insert Function** f_x. In the dialogue box, use **Statistical** for the category, and select the function **STEYX.**

Minitab Use the menu choices **Stat ➤ Regression ➤ Regression**. The value for S_e is given as s in the display.

Confidence Intervals for *y*

Population slope β

The least-squares line gives us a predicted value $\hat{y}$ for a specified x value. However, we used sample data to get the equation of the line. The line derived from the population of all data pairs is likely to have a slightly different slope, which we designate by the symbol β for population slope, and a slightly different y intercept, which we designate by the symbol α for population intercept. In addition, there is some random error ε, so the true y value is

$$y = \alpha + \beta x + \varepsilon$$

Because of the random variable ε, for each x value there is a corresponding distribution of y values. The methods of linear regression were developed so that the distribution of y values for a given x is centered on the population regression line. Furthermore, the distributions of y values corresponding to each x value all have the same standard deviation, estimated by the standard error of estimate S_e.

Confidence interval for predicted y

Using all this background, the theory tells us that for a specific x, a c *confidence interval for* y is given by the next procedure.

PROCEDURE

How to find a confidence interval for a predicted y from the least-squares line

1. Obtain a random sample of $n \geq 3$ data pairs (x, y).
2. Use the procedure of Section 9.2 to find the least-squares line $\hat{y} = a + bx$. You also need to find $\bar{x}$ from the sample data and the standard error of estimate S_e using Equation (8) of this section.
3. The **c confidence interval for y** for a **specified value of x** is

$$\hat{y} - E < y < \hat{y} + E$$

where

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{x})^2}{n\Sigma x^2 - (\Sigma x)^2}}$$

$\hat{y} = a + bx$ is the predictd value of y from the least-squares line for a *specified* x value

c = confidence level $(0 < c < 1)$

n = number of data pairs $(n \geq 3)$

t_c = critical value from Student's t distribution for c confidence level using $d.f. = n - 2$

S_e = standard error of estimate

The formulas involved in the computation of a c confidence interval look complicated. However, they involve quantities we have already computed or values we can easily look up in tables. The next example illustrates this point.

EXAMPLE 8 **Confidence interval for prediction**

Using the data of Table 9-13 of Example 7, find a 95% confidence interval for the amount of copper sulfate that will dissolve in 100 grams of water at 45°C.

SOLUTION: First, we need to find $\hat{y}$ for $x = 45$°C. We use the equation of the least-squares line that we found in Example 7.

$$\hat{y} \approx 10.14 + 0.51x \quad \text{from Example 7}$$
$$\hat{y} \approx 10.14 + 0.51(45) \quad \text{use 45 in place of } x$$
$$\hat{y} \approx 33$$

A 95% confidence interval for y is then

$$\hat{y} - E < y < \hat{y} + E$$
$$33 - E < y < 33 + E$$

where $E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{x})^2}{n\Sigma x^2 - (\Sigma x)^2}}$.

From Example 7, we have $n = 7$, $\Sigma x = 280$, $\Sigma x^2 = 14{,}000$, $\bar{x} = 40$, and $S_e \approx 2.35$. Using $n - 2 = 7 - 2 = 5$ degrees of freedom, we find from Table 6 of Appendix II that $t_{0.95} = 2.571$.

$$E \approx (2.571)(2.35)\sqrt{1 + \frac{1}{7} + \frac{7(45 - 40)^2}{7(14{,}000) - (280)^2}}$$
$$\approx (2.571)(2.35)\sqrt{1.15179} \approx 6.5$$

A 95% confidence interval for y is

$$33 - 6.5 \le y \le 33 + 6.5$$
$$26.5 \le y \le 39.5$$

This means we are 95% sure that the interval between 26.5 grams and 39.5 grams is one that contains the predicted amount of copper sulfate that will dissolve in 100 grams of water at 45°C. The interval is fairly wide but would decrease with more sample data.

GUIDED EXERCISE 7 **Confidence interval for prediction**

Let's use the data of Example 7 to compute a 95% confidence interval for y = amount of copper sulfate that will dissolve at $x = 15$°C.

(a) From Example 7, we have

$\hat{y} \approx 10.14 + 0.51x$

Evaluate $\hat{y}$ for $x = 15$.

→ $\hat{y} \approx 10.14 + 0.51x$
$\approx 10.14 + 0.51(15)$
≈ 17.8

(b) The bound E on the error of estimate is

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{x})^2}{n\Sigma x^2 - (\Sigma x)^2}}$$

From Example 7, we know that $S_e \approx 2.35$, $\Sigma x = 280$, $\Sigma x^2 = 14{,}000$, $\bar{x} = 40$, and $n = 7$. Find $t_{0.95}$ and compute E.

→ $t_{0.95} = 2.571$ for $d.f. = n - 2 = 5$

$$E \approx (2.571)(2.35)\sqrt{1 + \frac{1}{7} + \frac{7(15 - 40)^2}{7(14{,}000) - (280)^2}}$$
$$\approx (2.571)(2.35)\sqrt{1.366071} \approx 7.1$$

Continued

GUIDED EXERCISE 7 *continued*

(c) Find a 95% confidence interval for y.

$\hat{y} - E \leq y \leq \hat{y} + E$

The confidence interval is

$$17.8 - 7.1 \leq y \leq 17.8 + 7.1$$
$$10.7 \leq y \leq 24.9$$

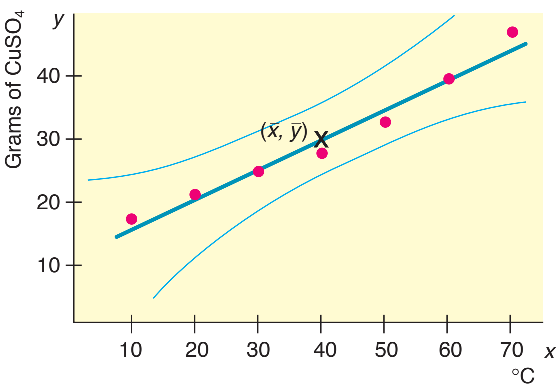
As we compare the results of Guided Exercise 7 and Example 8, we notice that the 95% confidence interval of y values for $x = 15°C$ is 7.1 units above and below the least-squares line, while the 95% confidence interval of y values for $x = 45°C$ is only 6.5 units above and below the least-squares line. This comparison reflects the general property that confidence intervals for y are narrower the nearer we are to the mean $\bar{x}$ of the x values. As we move near the extremes of the x distribution, the confidence intervals for y become wider. This is another reason that we should not try to use the least-squares line to predict y values for x values beyond the data extremes of the sample x distribution.

Confidence prediction band

If we were to compute a 95% confidence interval for all x values in the range of the sample x values, the *confidence interval band* would curve away from the least-squares line, as shown in Figure 9-18.

FIGURE 9-18

95% Confidence Band for Predicted Values $\hat{y}$



TECH NOTES

Minitab provides confidence intervals for predictions. Use the menu selection **Stat ➤ Regression ➤ Regression**. Under **Options**, enter the observed x value and set the confidence level. In the output, the confidence interval for predictions is designated by %PI.

Inferences about the Slope β

Recall that $\hat{y} = a + bx$ is the sample-based least-squares line and that $y = \alpha + \beta x$ is the population-based least-squares line computed (in theory) from the population of all (x, y) data pairs. In many real-world applications, the slope β is very important because β measures the rate at which y changes per unit change in x. Our next topic is to develop statistical tests and confidence intervals for β. Our work is based on the following theorem.

THEOREM 9.2 Let b be the slope of the sample least-squares line $\hat{y} = a + bx$ computed from a random sample of $n \geq 3$ data pairs (x, y). Let β be the slope of the population least-squares line $y = \alpha + \beta x$, which is in theory computed from the population of all (x, y) data pairs. Let S_e be the standard error of estimate computed from the sample. Then

$$t = \frac{b - \beta}{S_e \Big/ \sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}}$$

has a Student's t distribution with degrees of freedom $d.f. = n - 2$.

COMMENT The expression $S_e \Big/ \sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}$ is called the *standard error* for b.

Using this theorem, we can construct procedures for statistical tests and confidence intervals for β.

PROCEDURE

HOW TO TEST β AND FIND A CONFIDENCE INTERVAL FOR β

Requirements

Obtain a random sample of $n \geq 3$ data pairs (x, y). Use the procedure of Section 9.2 to find b, the slope of the sample least-squares line. Use Equation (8) of this section to find S_e, the standard error of estimate.

Procedure

Hypothesis test for β

For a statistical test of β

1. Use the *null hypothesis* H_0: $\beta = 0$. Use an *alternate hypothesis* H_1 appropriate to your application ($\beta > 0$ or $\beta < 0$ or $\beta \neq 0$). Set the level of significance α.
2. Use the null hypothesis H_0: $\beta = 0$ and the values of S_e, n, Σx, Σx^2, and b to compute the *sample test statistic*.

$$t = \frac{b}{S_e}\sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2} \qquad \text{with } d.f. = n - 2$$

3. Use a Student's t distribution and the type of test, one-tailed or two-tailed, to find (or estimate) the *P-value* corresponding to the test statistic.
4. *Conclude* the test. If P-value $\leq \alpha$, then reject H_0. If P-value $> \alpha$, then do not reject H_0.
5. *Interpret your conclusion* in the context of the application.

Confidence interval for β

To find a confidence interval for β

$$b - E < \beta < b + E$$

where
$$E = \frac{t_c S_e}{\sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}}$$

c = confidence level $(0 < c < 1)$

n = number of data pairs (x, y), $n \geq 3$

t_c = Student's t distribution critical value for confidence level c and $d.f. = n - 2$

S_e = standard error of estimate

EXAMPLE 9 TESTING β AND FINDING A CONFIDENCE INTERVAL FOR β

Plate tectonics and the spread of the ocean floor are very important topics in modern studies of earthquakes and earth science in general. A random sample of islands in the Indian Ocean gave the following information.

x = age of volcanic island in the Indian Ocean (units in 10^6 years)

y = distance of the island from the center of the midoceanic ridge (units in 100 kilometers)

x	120	83	60	50	35	30	20	17
y	30	16	15.5	14.5	22	18	12	0

Source: From King, Cuchaine A. M. *Physical Geography*. Oxford: Basil Blackwell, 1980, pp. 77–86 and 196–206. Reprinted by permission of the publisher.

Photo Researchers

(a) Starting from raw data values (x, y), the first step is simple but tedious. In short, you may verify (if you wish) that

$$\Sigma x = 415,\ \Sigma y = 128,\ \Sigma x^2 = 30{,}203,\ \Sigma y^2 = 2558.5,$$
$$\Sigma xy = 8133,\ \overline{x} = 51.875,\ \text{and}\ \overline{y} = 16$$

(b) The next step is to compute b, a, and S_e. Using a calculator, statistical software, or the formulas, we get

$$b \approx 0.1721 \quad \text{and} \quad a \approx 7.072$$

Since $n = 8$, we get

$$S_e = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}}$$
$$\approx \sqrt{\frac{2558.5 - 7.072(128) - 0.1721(8133)}{8-2}} \approx 6.50$$

(c) Use an $\alpha = 5\%$ level of significance to test the claim that β is positive.

SOLUTION: $\alpha = 0.05$; H_0: $\beta = 0$; H_1: $\beta > 0$. The sample test statistic is

$$t = \frac{b}{S_e}\sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2} \approx \frac{0.1721}{6.50}\sqrt{30{,}203 - \frac{(415)^2}{8}} \approx 2.466$$

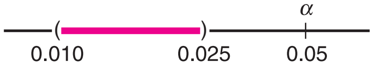
with $d.f. = n - 2 = 8 - 2 = 6$.

We use the Student's t distribution (Table 6 of Appendix II) to find an interval containing the P-value. The test is a one-tailed test. Technology gives P-value ≈ 0.0244.

TABLE 9-14 Excerpt from Table 6, Appendix II

✓ one-tail area	0.025	0.010
$d.f.$ = 6	2.447	3.143
	↑ Sample t = 7.563	

$$0.010 < P\text{-value} < 0.025$$



Since the interval containing the P-value is less than $\alpha = 0.05$, we reject H_0 and conclude that, at the 5% level of significance, the slope is positive.

(d) Find a 75% confidence interval for β.

SOLUTION: For $c = 0.75$ and $d.f. = n - 2 = 8 - 2 = 6$, the critical value $t_c = 1.273$. The margin of error E for the confidence interval is

$$E = \frac{t_c S_e}{\sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}} \approx \frac{1.273(6.50)}{\sqrt{30{,}203 - \frac{(415)^2}{8}}} \approx 0.0888$$

Using $b \approx 0.17$, a 75% confidence interval for β is

$$b - E < \beta < b + E$$
$$0.17 - 0.09 < \beta < 0.17 + 0.09$$
$$0.08 < \beta < 0.26$$

(e) ***Interpretation*** What does the confidence interval mean?
Recall the units involved (x in 10^6 years and y in 100 kilometers). It appears that, in this part of the world, we can be 75% confident that we have an interval showing that the ocean floor is moving at a rate of between 8 mm and 26 mm per year.

GUIDED EXERCISE 8 Inference for β

How fast do puppies grow? That depends on the puppy. How about male wolf pups in the Helsinki Zoo (Finland)? Let x = age in weeks and y = weight in kilograms for a random sample of male wolf pups. The following data are based on the article "Studies of the Wolf in Finland *Canis lupus L*" (*Ann. Zool. Fenn.*, Vol. 2, pp. 215–259) by E. Pulliainen, University of Helsinki.

x	8	10	14	20	28	40	45
y	7	13	17	23	30	34	35

$\Sigma x = 165$, $\Sigma y = 159$, $\Sigma y^2 = 5169$, $\Sigma y^2 = 4317$, $\Sigma xy = 4659$

(a) Verify the following values.
$\bar{x} \approx 23.571$, $\bar{y} \approx 22.714$,
$b \approx 0.7120$, $a \approx 5.932$,
$S_e \approx 3.368$

Use the formulas for $\bar{x}$, $\bar{y}$, b, a and S_e or find the results directly using your calculator or computer software.

(b) Use a 1% level of significance to test the claim that $\beta \neq 0$, and ***interpret*** the results in the context of this application.

$\alpha = 0.01$; H_0: $\beta = 0$; H_1: $\beta \neq 0$
Convert $b \approx 0.7120$ to a t value.

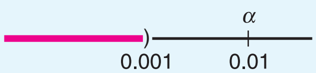
$$t = \frac{b}{S_e}\sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}$$

$$\approx \frac{0.7120}{3.368}\sqrt{5169 - \frac{(165)^2}{7}} \approx 7.563$$

From Table 6, Appendix II, for a two-tailed test with $d.f. = n - 2 = 7 - 2 = 5$,

✓two-tail area	0.001
$d.f.$ = 5	6.869
	↑ Sample t = 7.563

Noting that areas decrease as t values increase, we have $0.001 > P$-value. Technology gives P-value ≈ 0.0006.



Since the P-value is less than $\alpha = 0.01$, we reject H_0 and conclude that the population slope β is not zero.

Continued

GUIDED EXERCISE 8 *continued*

(c) Compute an 80% confidence interval for β and ***interpret*** the results in the context of this application.

d.f. = 5. For an 80% confidence interval, the critical value $t_c = 1.476$. The confidence interval is

$$b - E < \beta < b + E$$

where $b = 0.712$ and

$$E = \frac{t_c S_c}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}} \approx \frac{1.476(3.368)}{\sqrt{5169 - \frac{(165)^2}{7}}}$$

$$\approx 0.139$$

The interval is from 0.57 kg to 0.85 kg. We can be 80% confident that the interval computed is one that contains β. For each week's change in age, the weight change is between 0.57 kg and 0.85 kg.

Computation Hints for Sample Test Statistic t Used in Testing ρ and Testing β

Problem 14 discusses the fact that for the same data, the values of the sample test statistics for r and b are equal.

In Section 9.2 we saw that the values for the sample correlation coefficient r and slope b of the least-squares line are related by the formula

$$b = r\left(\frac{s_y}{s_x}\right)$$

Using this relationship and some algebra, it can be shown that

For the same sample, the sample correlation coefficient r and the slope b of the least-squares line have the same sample test statistic t, with $d.f. = n - 2$, where n is the number of data pairs.

Consequently, when doing calculations or using results from technology, we can use the following strategies.

(a) ***Calculations "by hand":*** Find the sample test statistic t corresponding to r. The sample test statistic t corresponding to b is the same.

(b) ***Using computer results:*** Most computer-based statistical packages provide the sample test statistic t corresponding to b. The sample test statistic t corresponding to r is the same.

TECH NOTES

The sample test statistic t corresponding to the sample correlation coefficient r is the same as the t value corresponding to b, the slope of the least-squares line (see Problem 14 at the end of this section). Consequently, the two tests H_0: $\rho = 0$ and H_0: $\beta = 0$ (with similar corresponding alternate hypotheses) have the same conclusions. The TI-84Plus/TI-83Plus/TI-*n*spire calculators use this fact explicitly. Minitab and Excel 2007 show t and the two-tailed P-value for the slope b of the least-squares line. Excel also shows confidence intervals for β. The displays show data from Guided Exercise 8 regarding the age and weight of wolf pups.

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) Under STAT, select TEST and use option **E:LinRegTTest.**

```
LinRegTTest
 y=a+bx
 β≠0 and ρ≠0
↑b=.7120
 s=3.3676
 r²=.9196
 r=.9590
```

```
LinRegTTest
 y=a+bx
 β≠0 and ρ≠0
 t=7.5632
 p=6.4075E-4
 df=5.0000
↓a=5.9317
```

Note that the value of S_e is given as s.

Excel 2007 On the home screen, click the **Data** tab. Select **Data Analysis** in the Analysis group. In the dialogue box, select **Regression**. Widen columns of the output as necessary to see all the results.

Regression Statistics	
Multiple R	0.958966516
R Square	0.919616778
Adjusted R Square	0.903540133
Standard Error	3.367628886
Observations	7

Value of S_e (Standard Error)

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.931681179	2.558126184	2.318760198	0.068158803	-0.644180779	12.50754314
X Variable 1	0.711989283	0.094138596	7.563202697	0.000640746	0.469998714	0.953979853

b (Coefficients, X Variable 1) — for two-tailed test (P-value)

Minitab Use the menu selection **Stat ➤ Regression ➤ Regression**. The value of S_e is S; P is the P-value of a two-tailed test. For a one-tailed test, divide the P-value by 2.

```
Regression Analysis
The regression equation is
y = 5.93 + 0.712 x
Predictor       Coef      StDev       T       P
Constant       5.932      2.558    2.32   0.068
x            0.71199    0.09414    7.56   0.001
S = 3.368     R-Sq = 92.0%     R-Sq(adj) = 90.4%
```

VIEWPOINT Hawaiian Island Hopping!

Suppose you want to go camping in Hawaii. Yes! Hawaii has both state and federal parks where you can enjoy camping on the beach or in the mountains. However, you will probably need to rent a car to get to the different campgrounds. How much will the car rental cost? That depends on the islands you visit. For car rental data and regression statistics you can compute regarding costs on different Hawaiian Islands, visit the Brase/Brase statistics site at http://www.cengage.com/statistics/brase *and find the link to Hawaiian Islands.*

SECTION 9.3 PROBLEMS

1. *Statistical Literacy* What is the symbol used for the population correlation coefficient?

2. *Statistical Literacy* What is the symbol used for the slope of the population least-squares line?

3. *Statistical Literacy* For a fixed confidence level, how does the length of the confidence interval for predicted values of y change as the corresponding x values become further away from $\bar{x}$?

4. *Statistical Literacy* How does the t value for the sample correlation coefficient r compare to the t value for the corresponding slope b of the sample least-squares line?

Using Computer Printouts Problems 5 and 6 use the following information. Prehistoric pottery vessels are usually found as sherds (broken pieces) and are carefully reconstructed if enough sherds can be found. Information taken from *Mimbres Mogollon Archaeology* by A. I. Woosley and A. J. McIntyre (University of New Mexico Press) provides data relating x = body diameter in centimeters and y = height in centimeters of prehistoric vessels reconstructed from sherds found at a prehistoric site. The following Minitab printout provides an analysis of the data.

```
Predictor      Coef     SE Coef       T        P
Constant     −0.223      2.429     −0.09    0.929
Diameter     0.7848     0.1471      5.33    0.001
S  =  4.07980    R-Sq  =  80.3%
```

5. *Critical Thinking: Using Information from a Computer Display to Test for Significance* Refer to the Minitab printout regarding prehistoric pottery.
   (a) Minitab calls the explanatory variable the predictor variable. Which is the predictor variable, the diameter of the pot or the height?
   (b) For the least-squares line $\hat{y} = a + bx$, what is the value of the constant a? What is the value of the slope b? (*Note*: The slope is the coefficient of the predictor variable.) Write the equation of the least-squares line.
   (c) The P-value for a two-tailed test corresponding to each coefficient is listed under "P." The t value corresponding to the coefficient is listed under "T." What is the P-value of the slope? What are the hypotheses for a two-tailed test of $\beta = 0$? Based on the P-value in the printout, do we reject or fail to reject the null hypothesis for $\alpha = 0.01$?
   (d) Recall that the t value and resulting P-value of the slope b equal the t value and resulting P-value of the corresponding sample correlation coefficient r. To find the value of the sample correlation coefficient r, take the square root of the R-Sq value shown in the display. What is the value of r? Consider a two-tailed test for ρ. Based on the P-value shown in the Minitab display, is the correlation coefficient significant at the 1% level of significance?

6. *Critical Thinking: Using Information from a Computer Display to Find a Confidence Interval* Refer to the Minitab printout regarding prehistoric pottery.
   (a) The standard error S_e of the linear regression model is given in the printout as "S." What is the value of S_e?
   (b) The standard error of the coefficient of the predictor variable is found under "SE Coef." Recall that the standard error for b is $S_e\Big/\sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}$. From the Minitab display, what is the value of the standard error for the slope b?

(c) The formula for the margin of error E for a $c\%$ confidence interval for the slope β can be written as $E = t_c(\text{SE Coef})$. The Minitab display is based on $n = 9$ data pairs. Find the critical value t_c for a 95% confidence interval in Table 6 of Appendix II. Then find a 95% confidence interval for the population slope β.

In Problems 7–12, parts (a) and (b) relate to testing ρ. Part (c) requests the value of S_e. Parts (d) and (e) relate to confidence intervals for prediction. Parts (f) and (g) relate to testing β and finding confidence intervals for β.

Answers may vary due to rounding.

7. ***Basketball: Free Throws and Field Goals*** Let x be a random variable that represents the percentage of successful free throws a professional basketball player makes in a season. Let y be a random variable that represents the percentage of successful field goals a professional basketball player makes in a season. A random sample of $n = 6$ professional basketball players gave the following information (Reference: *The Official NBA Basketball Encyclopedia*, Villard Books).

x	67	65	75	86	73	73
y	44	42	48	51	44	51

(a) Verify that $\Sigma x = 439$, $\Sigma y = 280$, $\Sigma x^2 = 32{,}393$, $\Sigma y^2 = 13{,}142$, $\Sigma xy = 20{,}599$, and $r \approx 0.784$.
(b) Use a 5% level of significance to test the claim that $\rho > 0$.
(c) Verify that $S_e \approx 2.6964$, $a \approx 16.542$, $b \approx 0.4117$, and $\bar{x} \approx 73.167$.
(d) Find the predicted percentage $\hat{y}$ of successful field goals for a player with $x = 70\%$ successful free throws.
(e) Find a 90% confidence interval for y when $x = 70$.
(f) Use a 5% level of significance to test the claim that $\beta > 0$.
(g) Find a 90% confidence interval for β and ***interpret*** its meaning.

8. ***Baseball: Batting Average and Strikeouts*** Let x be a random variable that represents the batting average of a professional baseball player. Let y be a random variable that represents the percentage of strikeouts of a professional baseball player. A random sample of $n = 6$ professional baseball players gave the following information (Reference: *The Baseball Encyclopedia*, Macmillan).

x	0.328	0.290	0.340	0.248	0.367	0.269
y	3.2	7.6	4.0	8.6	3.1	11.1

(a) Verify that $\Sigma x = 1.842$, $\Sigma y = 37.6$, $\Sigma x^2 = 0.575838$, $\Sigma y^2 = 290.78$, $\Sigma xy = 10.87$, and $r \approx -0.891$.
(b) Use a 5% level of significance to test the claim that $\rho \neq 0$.
(c) Verify that $S_e \approx 1.6838$, $a \approx 26.247$, and $b \approx -65.081$.
(d) Find the predicted percentage of strikeouts for a player with an $x = 0.300$ batting average.
(e) Find an 80% confidence interval for y when $x = 0.300$.
(f) Use a 5% level of significance to test the claim that $\beta \neq 0$.
(g) Find a 90% confidence interval for β and ***interpret*** its meaning.

9. ***Scuba Diving: Depth*** What is the optimal amount of time for a scuba diver to be on the bottom of the ocean? That depends on the depth of the dive. The U.S. Navy has done a lot of research on this topic. The Navy defines the "optimal time" to be the time at each depth for the best balance between length of work period and decompression time after surfacing. Let x = depth of dive in meters, and let y = optimal time in hours. A random sample of divers gave the following data (based on information taken from *Medical Physiology* by A. C. Guyton, M.D.).

x	14.1	24.3	30.2	38.3	51.3	20.5	22.7
y	2.58	2.08	1.58	1.03	0.75	2.38	2.20

(a) Verify that $\Sigma x = 201.4$, $\Sigma y = 12.6$, $\Sigma x^2 = 6734.46$, $\Sigma y^2 = 25.607$, $\Sigma xy = 311.292$, and $r \approx -0.976$.
(b) Use a 1% level of significance to test the claim that $\rho < 0$.
(c) Verify that $S_e \approx 0.1660$, $a \approx 3.366$, and $b \approx -0.0544$.
(d) Find the predicted optimal time in hours for a dive depth of $x = 18$ meters.
(e) Find an 80% confidence interval for y when $x = 18$ meters.
(f) Use a 1% level of significance to test the claim that $\beta < 0$.
(g) Find a 90% confidence interval for β and ***interpret*** its meaning.

10. ***Physiology: Oxygen*** Aviation and high-altitude physiology is a specialty in the study of medicine. Let x = partial pressure of oxygen in the alveoli (air cells in the lungs) when breathing naturally available air. Let y = partial pressure when breathing pure oxygen. The (x, y) data pairs correspond to elevations from 10,000 feet to 30,000 feet in 5000-foot intervals for a random sample of volunteers. Although the medical data were collected using airplanes, they apply equally well to Mt. Everest climbers (summit 29,028 feet).

x	6.7	5.1	4.2	3.3	2.1 (units: mm Hg/10)
y	43.6	32.9	26.2	6.2	13.9 (units: mm Hg/10)

(Based on information taken from *Medical Physiology* by A. C. Guyton, M.D.)
(a) Verify that $\Sigma x = 21.4$, $\Sigma y = 132.8$, $\Sigma x^2 = 103.84$, $\Sigma y^2 = 4125.46$, $\Sigma xy = 652.6$, and $r \approx 0.984$.
(b) Use a 1% level of significance to test the claim that $\rho > 0$.
(c) Verify that $S_e \approx 2.5319$, $a \approx -2.869$, and $b \approx 6.876$.
(d) Find the predicted pressure when breathing pure oxygen if the pressure from breathing available air is $x = 4.0$.
(e) Find a 90% confidence interval for y when $x = 4.0$.
(f) Use a 1% level of significance to test the claim that $\beta > 0$.
(g) Find a 95% confidence interval for β and ***interpret*** its meaning.

11. ***New Car: Negotiating Price*** Suppose you are interested in buying a new Toyota Corolla. You are standing on the sales lot looking at a model with different options. The list price is on the vehicle. As a salesperson approaches, you wonder what the dealer invoice price is for this model with its options. The following data are based on information taken from *Consumer Guide* (Vol. 677). Let x be the list price (in thousands of dollars) for a random selection of Toyota Corollas of different models and options. Let y be the dealer invoice (in thousands of dollars) for the given vehicle.

x	12.6	13.0	12.8	13.6	13.4	14.2
y	11.6	12.0	11.5	12.2	12.0	12.8

(a) Verify that $\Sigma x = 79.6$, $\Sigma y = 72.1$, $\Sigma x^2 = 1057.76$, $\Sigma y^2 = 867.49$, $\Sigma xy = 957.84$, and $r \approx 0.956$.
(b) Use a 1% level of significance to test the claim that $\rho > 0$.
(c) Verify that $S_e \approx 0.1527$, $a \approx 1.965$, and $b \approx 0.758$.
(d) Find the predicted dealer invoice when the list price is $x = 14$ (thousand dollars).
(e) Find an 85% confidence interval for y when $x = 14$ (thousand dollars).
(f) Use a 1% level of significance to test the claim that $\beta > 0$.
(g) Find a 95% confidence interval for β and ***interpret*** its meaning.

12. *New Car: Negotiating Price* Suppose you are interested in buying a new Lincoln Navigator or Town Car. You are standing on the sales lot looking at a model with different options. The list price is on the vehicle. As a salesperson approaches, you wonder what the dealer invoice price is for this model with its options. The following data are based on information taken from *Consumer Guide* (Vol. 677). Let x be the list price (in thousands of dollars) for a random selection of these cars of different models and options. Let y be the dealer invoice (in thousands of dollars) for the given vehicle.

x	32.1	33.5	36.1	44.0	47.8
y	29.8	31.1	32.0	42.1	42.2

(a) Verify that $\Sigma x = 193.5$, $\Sigma y = 177.2$, $\Sigma x^2 = 7676.71$, $\Sigma y^2 = 6432.5$, $\Sigma xy = 7023.19$, and $r \approx 0.977$.
(b) Use a 1% level of significance to test the claim that $\rho > 0$.
(c) Verify that $S_e \approx 1.5223$, $a \approx 1.4084$, and $b \approx 0.8794$.
(d) Find the predicted dealer invoice when the list price is $x = 40$ (thousand dollars).
(e) Find a 95% confidence interval for y when $x = 40$ (thousand dollars).
(f) Use a 1% level of significance to test the claim that $\beta > 0$.
(g) Find a 90% confidence interval for β and ***interpret*** its meaning.

13. *Expand Your Knowledge: Sample Size and Significance of r*
(a) Suppose $n = 6$ and the sample correlation coefficient is $r = 0.90$. Is r significant at the 1% level of significance (based on a two-tailed test)?
(b) Suppose $n = 10$ and the sample correlation coefficient is $r = 0.90$. Is r significant at the 1% level of significance (based on a two-tailed test)?
(c) Explain why the test results of parts (a) and (b) are different even though the sample correlation coefficient $r = 0.90$ is the same in both parts. Does it appear that sample size plays an important role in determining the significance of a correlation coefficient? Explain.

14. *Expand Your Knowledge: Student's t Value for Sample r and for Sample b* It is not obvious from the formulas, but the values of the sample test statistic t for the correlation coefficient and for the slope of the least-squares line are equal for the same data set. This fact is based on the relation

$$b = r\frac{s_y}{s_x}$$

where s_y and s_x are the sample standard deviations of the x and y values, respectively.
(a) Many computer software packages give the t value and corresponding P-value for b. If β is significant, is ρ significant?
(b) When doing statistical tests "by hand," it is easier to compute the sample test statistic t for the sample correlation coefficient r than it is to compute the sample test statistic t for the slope b of the sample least-squares line. Compare the results of parts (b) and (f) for Problems 7–12 of this problem set. Is the sample test statistic t for r the same as the corresponding test statistic for b? If you conclude that ρ is positive, can you conclude that β is positive at the same level of significance? If you conclude that ρ is not significant, is β also not significant at the same level of significance?

SECTION 9.4

Multiple Regression

FOCUS POINTS

- Learn about the advantages of multiple regression.
- Learn the basic ingredients that go into a multiple regression model.
- Discuss standard error for computed coefficients and the coefficient of multiple determination.
- Test coefficients in the model for statistical significance.
- Compute confidence intervals for predictions.

Advantages of Multiple Regression

There are many examples in statistics in which one variable can be predicted very accurately in terms of another *single* variable. However, predictions usually improve if we consider additional relevant information. For example, the sugar content y of golden delicious apples taken from an apple orchard in Colorado could be predicted from x_1 = number of days in growing season. If we also included information regarding x_2 = soil quality rating and x_3 = amount of available water, then we would expect our prediction of y = sugar content to be more accurate.

Likewise, the annual net income y of a new franchise auto parts store could be predicted using only x_1 = population size of sales district. However, we would probably get a better prediction of y values if we included the explanatory variables x_2 = size of store inventory, x_3 = dollar amount spent on advertising in local newspapers, and x_4 = number of competing stores in the sales district.

Multiple regression

For most statistical applications, we gain a definite advantage in the reliability of our predictions if we include more *relevant* data and corresponding (relevant) random variables in the computation of our predictions. In this section, we will give you an idea of how this can be done by methods of *multiple regression*. You should be aware that an in-depth study of multiple regression requires the use of advanced mathematics. However, if you are willing to let the computer be a "friend who gives you useful information," then you will learn a great deal about multiple regression in this section. We will let the computer do most of the calculating work while we interpret the results.

Basic Terminology and Notation

In statistics, the most commonly used mathematical formulas for expressing linear relationships among more than two variables are *equations* of the form

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k \tag{9}$$

Here, y is the variable that we want to predict or forecast. We will employ the usual terminology and call y the *response variable*. The k variables $x_1, x_2, \ldots, x_k$ are specified variables on which the predictions are going to be based. Once again, we will employ the popular terminology and call $x_1, x_2, \ldots, x_k$ the *explanatory variables*. This terminology is easy to remember if you just think of the explanatory variables $x_1, x_2, \ldots, x_k$ as "explaining" the response y.

In Equation (9), $b_0, b_1, b_2, \ldots, b_k$ are numerical constants (called *coefficients*) that must be mathematically determined from given data. The numerical values of these coefficients are obtained from the *least-squares criterion*, which we will discuss after the following exercise.

GUIDED EXERCISE 9 Components of multiple regression equation

An industrial psychologist working for a hospital-supply company is studying the following variables for a random sample of company employees:

x_1 = number of years the employee has been with the company

x_2 = job-training level (0 = lowest level and 5 = highest level)

x_3 = interpersonal skills (0 = lowest level and 10 = highest level)

y = job-performance rating from supervisor (1 = lowest rating, 20 = highest rating)

The psychologist wants to predict y using x_1, x_2, and x_3 together in a least-squares equation.

(a) Identify the response variable and the explanatory variables.

→ The response variable is what we want to predict. This is y, job performance. The explanatory variables are years of experience x_1, training level x_2, and interpersonal skills x_3. In a sense, these variables "explain" the response variable.

(b) After collecting data, the psychologist used a computer with appropriate software to obtain the least-squares linear equation

$$y = 1 + 0.2x_1 + 2.3x_2 + 0.7x_3$$

Identify the constant term and each of the coefficients with its corresponding variable.

→ The constant term is 1.

Explanatory Variable	Coefficient
x_1	0.2
x_2	2.3
x_3	0.7

(c) Use the equation to predict the job-performance rating of an employee with 3 years of experience, a training level of 4, and an interpersonal skill rating of 2.

→ Substituting $x_1 = 3$, $x_2 = 4$, and $x_3 = 2$ into the least-squares equation and multiplying by the respective coefficients, we obtain the predicted job performance rating of

$$y = 1 + 0.2(3) + 2.3(4) + 0.7(2) = 12.2$$

Of course, the *predicted* value for job performance might differ from the actual rating given by the supervisor.

Theory for the least-squares criterion (optional)

This material is a little sophisticated, so you may wish to skip ahead to the discussion of regression models and computers and omit the following explanation of basic theory.

In multiple regression, the least-squares criterion states that the following sum (over all data points),

$$\Sigma[y_i - (b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{ki})]^2 \quad (10)$$

must be made as small as possible. In this formula,

y_i = ith data value for y

x_{1i} = ith data value for x_1

x_{2i} = ith data value for x_2

$\vdots$

x_{ki} = ith data value for x_k

Recall that Equation (9) gives the predicted y value; therefore,

$$y_i - (b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{ki}) \quad (11)$$

represents the *difference* between the *observed* y value (that is, y_i) and the *predicted* y value based on the data values $x_{1i}, x_{2i}, \ldots, x_{ki}$. When we square this difference, total the result over all data points, and choose the values of $b_0, b_1, b_2, \ldots, b_k$ to minimize the sum [i.e., minimize Equation (10)], then we are satisfying the least-squares criterion.

Residual

COMMENT The algebraic expression in Equation (11) is very important. In fact, it has a special name in the theory of regression. It is called a *residual*. The residual is simply the difference between the actual data value and the predicted value of the response variable based on given data values for the explanatory variables. Advanced topics in the theory of regression study residuals in great detail. Such a detailed treatment is beyond the scope of this text. However, from the discussion presented so far, we see that the method of least squares chooses the values of the coefficients b_i to make the sum of the squares of the residuals as small as possible.

After a good deal of mathematics has been done (involving a considerable amount of calculus), the least-squares criterion can be reduced to solving a system of linear equations. These are usually called *normal equations* (not to be confused with the normal distribution).

In the simplest case, where there are only *two* explanatory variables x_1 and x_2 and we want to fit the equation

$$y = b_0 + b_1x_1 + b_2x_2$$

to given data, there are three normal equations that must be solved for b_0, b_1, and b_2. These normal equations are

$$\left.\begin{aligned} \Sigma y_i &= nb_0 + b_1(\Sigma x_{1i}) + b_2(\Sigma x_{2i}) \\ \Sigma x_{1i}y_i &= b_0(\Sigma x_{1i}) + b_1(\Sigma x_{1i}{}^2) + b_2(\Sigma x_{1i}x_{2i}) \\ \Sigma x_{2i}y_i &= b_0(\Sigma x_{2i}) + b_1(\Sigma x_{1i}x_{2i}) + b_2(\Sigma x_{2i}{}^2) \end{aligned}\right\} \qquad (12)$$

In the system of Equations (12), n represents the number of data points and x_{1i}, x_{2i}, and y_i all represent given data values.

Therefore, the only unknowns are the coefficients b_0, b_1, and b_2; we can use the system of Equations (12) to solve for these unknowns. This is the procedure that lets us obtain the least-squares regression equation in Equation (9) when we have only *two* explanatory variables.

As you can see, this is all rather complicated, and the more explanatory variables $x_1, x_2, \ldots, x_k$ we have, the more involved the calculations become. In the general case, if you have k explanatory variables, there will be $k + 1$ normal equations that must be solved for the coefficients $b_0, b_1, b_2, \ldots, b_k$.

Regression Models and Computers

As you can see from the preceding optional discussion, the work required to find an equation satisfying the least-squares criterion is tremendous and can be very complex. Today, such work is conveniently left to computers. In this text, we use two computer software packages that specialize in statistical applications.

Minitab is a widely used statistical software package. It fully supports multiple regression. Excel 2007 has a multiple regression component that performs much of the multiple regression analysis. We will use Minitab in our example. Many other software packages, including SPSS, support multiple regression and have outputs similar to those of Minitab.

Ingredients of the regression model

In this section, we will often refer to a *regression model*. What do we mean by this? We mean a mathematical package that consists of the following ingredients:

1. The model will have a collection of random variables, *one* of which has been identified as the response variable, with *any or all* of the remaining variables being identified as explanatory variables.
2. Associated with a given application will be a collection of numerical data values for each of the variables of part 1.
3. Using the numerical data values, the least-squares criterion, and the declared response and explanatory variables, a *least-squares equation* (also called a *regression equation*) will be constructed. In Section 9.2, we were able to construct the least-squares equation using only a hand calculator. However, in multiple regression, we will use a computer to construct the least-squares equation.
4. The model usually includes additional information about the variables used, the coefficients and regression equation, and a measure of "goodness of fit" of the regression equation to the data values. In modern practice, this information usually comes to you in the form of computer displays.
5. Finally, the regression model enables you to supply given values of the explanatory variables for the purpose of predicting or forecasting the corresponding value of the response variable. You also should be able to construct a $c\%$ confidence interval for your least-squares prediction. In multiple regression, this will be done by the computer at your request.

Problem 7 at the end of this section discusses *curvilinear regression* (also known as *polynomial regression*).

The next example demonstrates computer applications of a typical multiple regression problem. In the context of the example, we will introduce some of the basic techniques of multiple regression.

Example Utilizing Minitab

EXAMPLE 10 MULTIPLE REGRESSION

Antelope are beautiful and graceful animals that live on the high plains of the western United States. Thunder Basin National Grasslands in Wyoming is home to hundreds of antelope. The Bureau of Land Management (BLM) has been studying the Thunder Basin antelope population for the past 8 years. The variables used are

x_1 = spring fawn count (in hundreds of fawns)

x_2 = size of adult antelope population (in hundreds)

x_3 = annual precipitation (in inches)

x_4 = winter severity index (1 = mild and 5 = extremely severe) (This is an index based on temperature and wind chill factors.)

The data obtained in the study over the 8-year period are shown in Table 9-15.

D. Robert & Lorri Franz/CORBIS

TABLE 9-15 Data for Thunder Basin Antelope Study

Year	x_1	x_2	x_3	x_4
1	2.9	9.2	13.2	2
2	2.4	8.7	11.5	3
3	2.0	7.2	10.8	4
4	2.3	8.5	12.3	2
5	3.2	9.6	12.6	3
6	1.9	6.8	10.6	5
7	3.4	9.7	14.1	1
8	2.1	7.9	11.2	3

Summary Statistics for Each Variable

It is a good idea to first look at the summary statistics for each variable. Figure 9-19 shows the Minitab display of the summary statistics.

Menu selection: **Stat ➤ Basic Statistic ➤ Display Descriptive Statistics**

FIGURE 9-19

Minitab Display of Summary Statistics for Each Variable

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
x1	8	2.525	2.350	2.525	0.570	0.202
x2	8	8.450	8.600	8.450	1.076	0.380
x3	8	12.037	11.900	12.037	1.229	0.435
x4	8	2.875	3.000	2.875	1.246	0.441

Variable	Minimum	Maximum	Q1	Q3
x1	1.900	3.400	2.025	3.125
x2	6.800	9.700	7.375	9.500
x3	10.600	14.100	10.900	13.050
x4	1.000	5.000	2.000	3.750

This type of information can be very useful because it tells you basic information about the variables you are studying. Sample means and sample standard deviations with a Student's t distribution are essential ingredients for estimating or testing population means (Chapters 7 and 8).

For example, if μ_2 represents the *population mean* of x_2 (adult antelope population), then by using the methods of Section 7.2 we can quickly estimate a 90% confidence interval for μ_2:

$$7.729 < \mu_2 < 9.171$$

Since our units are in hundreds, this means we can be 90% sure that the *population mean* μ_2 of adult antelope in the Thunder Basin National Grasslands is between 773 and 917.

Correlation Between Variables

It is also useful to examine how the variables relate to each other. Figure 9-20 shows the sample correlation coefficients r between each of the two variables. A natural question arises: Which of the variables are closely related to each other, and which are not as closely related? Recall (from Section 9.1) that if the correlation coefficient is near 1 or -1, then the corresponding variables have a lot in common. If the correlation coefficient is near zero, the variables have much less influence on each other.

Menu selection: **Stat ➤ Basic Statistics ➤ Correlation**

FIGURE 9-20

Minitab Display of Correlation Coefficients Between Variables

Correlations (Pearson)

	x1	x2	x3
x2	0.939		
x3	0.924	0.903	
x4	−0.739	−0.836	−0.901

Look at Figure 9-20. Which of the variables has the greatest influence on x_1? The sample correlation coefficient between x_1 and x_2 is $r = 0.939$, with a corresponding coefficient of determination of $r^2 \approx 0.88$. This means that if we consider only x_1 and x_2 (and none of the other variables), then about 88% of the variation in x_1 can be explained by the corresponding variation in x_2 (by itself). Similarly, if we consider only x_1 and x_3, we see the sample correlation coefficient $r = 0.924$, with a corresponding coefficient of determination of $r^2 \approx 0.85$. About

85% of the variation in x_1 can be explained by the corresponding variation in x_3. The variable x_4 has much less influence on x_1 because the sample correlation coefficient between these two variables is $r = -0.739$, with corresponding coefficient of determination $r^2 \approx 0.55$, or only 55%.

These relationships are very reasonable in the context of our problem. It is common sense that the number of spring fawns x_1 is strongly related to x_2, the size of the adult antelope population. Furthermore, the spring fawn count x_1 is very much influenced by available food for the fawn (and its mother). Thunder Basin National Grasslands is a semiarid region, and available food (grass) is almost completely determined by annual precipitation x_3. Antelope are naturally strong and hardy animals. Therefore, the temperature and wind chill index x_4 will have much less effect on the adult does and corresponding number of spring fawns provided there is plenty of available food.

Least-Squares Equation

Figure 9-21 shows a display that gives an expression for the actual least-squares equation and a lot of information about the equation. To get this display or a similar display, the user needs to declare which variable is the response variable and which are the explanatory variables. For Figure 9-21, we designated x_1 as the response variable. This means that x_1 is the variable we choose to predict. We also designated variables x_2, x_3, and x_4 as explanatory variables. This means that x_2, x_3, and x_4 will be used *together* to predict x_1. There is a lot of flexibility here. We could have designated any *one* of the variables x_1, x_2, x_3, x_4 as the response variable and *any or all* of the remaining variables as explanatory variables. So there are several possible regression models the computer can construct for you, depending on the type of information you want. In this example, we want to predict x_1 (spring fawn count) by using x_2 (adult population), x_3 (annual precipitation), and x_4 (winter index) *together*.

Menu selection: **Stat ➤ Regression ➤ Regression.** In the dialogue box, select x_1 as the response and x_2, x_3, x_4 as the predictors.

FIGURE 9-21

Minitab Display of Regression Analysis

```
Regression Analysis
The regression equation is
x1 = -5.92 + 0.338 x2 + 0.402 x3 + 0.263 x4
Predictor        Coef        StDev         T         P
Constant       -5.922        1.256     -4.72     0.009
x2            0.33822      0.09947      3.40     0.027
x3             0.4015       0.1099      3.65     0.022
x4            0.26295      0.08514      3.09     0.037
S = 0.1209    R-Sq = 97.4%     R-Sq(adj) = 95.5%
```

The least-squares regression equation is given near the top of the display. Then more information is given about the constant and coefficients. The parts of the equation are

$$x_1 = -5.92 + 0.338x_2 + 0.402x_3 + 0.263x_4 \qquad (13)$$

response variable   constant   coefficient of associated explanatory variable

COMMENT In the case of a simple regression model in which we have only one explanatory variable, the coefficient of that variable is the *slope* of the least-squares line. This slope (or coefficient) represents the change in the response variable per unit change in the explanatory variable. In a multiple regression model such as Equation (13), the coefficients also can be thought of as a slope, *provided* we hold the other variables as arbitrary and fixed constants. For example, the coefficient of x_2 in Equation (13) is $b_2 = 0.338$. This means that if x_3 (precipitation) and x_4 (winter index) are taken into account but held constant, then $b_2 = 0.338$ represents the change in x_1 (spring fawn count) per unit change in x_2 (adult antelope count). Since our units are in hundreds, this indicates that if x_3 and x_4 are taken into account as arbitrary but fixed values, then an increase of 100 adult antelope would give an expected increase of 33.8, or 34, spring fawns.

A natural question arises: How good a fit is the least-squares regression Equation (13) for our given data?

Coefficient of multiple determination

One way to answer this question is to examine the *coefficient of multiple determination*. The coefficient of multiple determination is a direct generalization of the concept of coefficient of determination (between *two* variables) as discussed in Section 9.2, and it has essentially the same meaning. The coefficient of multiple determination is given in the display of Figure 9-21 as a percent. We see R-sq = 97.4%. This means that about 97.4% of the variation in the response variable x_1 can be explained from the least-squares Equation (13) and the corresponding *joint* variation of the variables x_2, x_3, and x_4 taken together. The remaining 100% − 97.4% = 2.6% of the variation in x_1 is due to random chance or possibly the presence of other variables not included in this regression equation. (We will discuss the *standard error* associated with each coefficient later in this section.)

Predictions

Let's use the current regression model to predict the response variable x_1. Recall that in Section 9.2 we first made predictions from the least-squares line and then constructed a confidence interval for our predictions. Although the exact details are beyond the scope of this text, this process can be generalized to multiple regression. The calculations are very tedious, but that's why we use a computer!

Suppose we ask the following question: In a year when $x_2 = 8.2$ (hundreds of adult antelope), $x_3 = 11.7$ (inches of precipitation), and $x_4 = 3$ (winter index), what do we predict for x_1 (spring fawn count)? Furthermore, let's suppose we want an 85% confidence interval for our prediction.

To answer this question, we look at Figure 9-22, which shows the Minitab prediction result for x_1 from the specified values of x_2, x_3, and x_4.

Menu selection: **Stat ➤ Regression ➤ Regression.** In the dialogue box, select Options. List the new observations for x_2, x_3, and x_4 in order, separated by spaces. Specify the confidence level. Be sure that Fit Intercept is checked.

FIGURE 9-22

Minitab Display Showing the Predicted Value of x_1

```
Predicted Values
Fit        StDev Fit         85.0% CI              85.0% PI
2.3378        0.0472     (2.2539, 2.4217)     (2.1069, 2.5687)
```

The value for Fit is 2.3378. This is the predicted value for x_1. The 85% confidence interval for the prediction is designated as 85% PI. We see that the interval for x_1 (rounded to two digits after the decimal) is $2.11 \le x_1 \le 2.57$. This means we are 85% confident that the number of spring fawns will be in the range of 211 to 257.

Please note that this is *not* a confidence interval for the population mean of x_1. Rather, we have constructed a confidence interval for the *actual value* of x_1 under the conditions $x_2 = 8.2$, $x_3 = 11.7$, and $x_4 = 3$.

COMMENT Extrapolation much beyond the data range for any of the variables in a multiple regression model can produce results that might be meaningless and unrealistic. Many computer software packages warn about computing a confidence interval for a prediction when some of the values of the explanatory variables are beyond the data range in either direction.

Testing a Coefficient for Significance

In applications of multiple regression, it is possible to have many different variables. Occasionally, you might suspect that one of the explanatory variables x_i is not very useful as a tool for predicting the response variable. It simply may not influence the response variable much at all. To decide whether or not this is the case, we construct a test for the significance of the coefficient of x_i in the least-squares equation.

Recall that the general least-squares equation is

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k \tag{14}$$

where y = response variable
x_i = explanatory variable for $i = 1, 2, \ldots, k$
b_i = numerical coefficient for $i = 0, 1, 2, \ldots, k$

Equation (14) was constructed from given data. Usually, the data are only a small subset of all possible data that could have been collected.

Let us suppose (in theory) that we used *all possible data* that could ever be obtained for our regression problem and that we constructed the regression equation using the entire population of all possible data. Then we would get the *theoretical* regression equation

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k \tag{15}$$

where y and x_i are as in Equation (14), but β_i is the *theoretical* coefficient of x_i.

Now look back at the regression analysis in Figure 9-21. Beside the constant and each coefficient is a number in the StDev column. This is the *standard error* corresponding to that coefficient. The standard error can be thought of as similar to a standard deviation that corresponds to the coefficient. The calculation of the number is beyond the scope of this text, but it is available on computer printouts, and we will use it to construct our test.

Let us call S_i the standard error for coefficient x_i (S_0 is the standard error for the constant). Under very basic and general assumptions, it can be proved that

$$t = \frac{b_i - \beta_i}{S_i} \tag{16}$$

has a Student's t distribution with degrees of freedom $d.f. = n - k - 1$, where n = number of data points and k = number of explanatory variables in the least-squares equation.

Now let us return to the question: Is x_i useful as an explanatory variable in the least-squares equation?

The answer is that it is *not* useful if $\beta_i = 0$. In that case, the (theoretical) coefficient of x_i would be zero and x_i would contribute nothing to the least-squares equation. However, if $\beta_i \neq 0$, then the explanatory variable x_i does contribute information in the least-squares equation.

Consider the following hypotheses,

$$H_0: \beta_i = 0 \qquad \text{and} \qquad H_1: \beta_i \neq 0$$

If we accept H_0, we conclude that $\beta_i = 0$ and x_i probably should be dropped as an explanatory variable in the least-squares equation. If we accept H_1, we conclude that $\beta_i \neq 0$ and x_i should be included as an explanatory variable in the least-squares equation.

EXAMPLE 11 TESTING A COEFFICIENT

We'll use the data and printouts of Example 10 and test the significance of x_3 as an explanatory variable using level of significance $\alpha = 0.05$.

H_0: $\beta_3 = 0$ and H_1: $\beta_3 \neq 0$

To find the t value corresponding to b_3, we use Equation (16) and the null hypothesis H_0: $\beta_3 = 0$. This gives us the equation

$$t = \frac{b_3}{S_3} \tag{17}$$

In the regression analysis shown in Figure 9-21, we see a t value for the constant and each coefficient. This t value is exactly the value of $t = b_i/S_i$. This is the t value corresponding to the sample test statistic. For the coefficient of x_3, we see

t value ≈ 3.65

Notice in Figure 9-21 that we are also given the P-value based on a two-tailed test of the sample test statistic for each coefficient. This is the value in the column headed "P." For the sample test statistic $t \approx 3.65$, the corresponding P-value is 0.022. Since the P-value is less than the level of significance $\alpha = 0.05$, we reject H_0. In other words, at the 5% level of significance, we can say that the population correlation coefficient β_3 of x_3 is not 0.

We conclude at the 5% level of significance that x_3 (annual precipitation) should be included as an explanatory variable in the least-squares equation. Notice that Figure 9-21 also gives the P-value for each ratio, so we can conclude the test using P-values directly. Using the P-values, we see that x_2 and x_3 are also significant at the 5% level.

Confidence Intervals for Coefficients

Equation (16) also gives us the basis for finding *confidence intervals* for β_i. A c% confidence interval for β_i will be

$$b_i - tS_i < \beta_i < b_i + tS_i$$

where $d.f. = n - k - 1$, t is selected according to the specified confidence level, b_i is the numerical value of the coefficient from Figure 9-21, S_i is the numerical value of the standard error from Figure 9-21, n is the number of data points, and k is the number of explanatory variables in the least-squares equation.

EXAMPLE 12 CONFIDENCE INTERVAL FOR A COEFFICIENT

Suppose we want to compute a 90% confidence interval for β_2, the coefficient of x_2. From Figure 9-21, we have (rounding to three digits after the decimal)

$b_2 = 0.338$, $S_2 = 0.099$, and $d.f. = 4$

From the t table (Table 6, Appendix II), we find $t = 2.132$, so,

$$b_2 - tS_2 < \beta_2 < b_2 + tS_2$$
$$0.338 - 2.132(0.099) < \beta_2 < 0.338 + 2.132(0.099)$$
$$0.127 < \beta_2 < 0.549$$

Excel 2007 Displays

Although Excel gives information very similar to that supplied by Minitab, the least-squares equation is not explicitly displayed. However, the intercept (constant) and coefficients of the variables are shown with the corresponding standard errors and t values with P-values (two-tailed test). Excel shows the confidence interval for each coefficient. However, there is no built-in function to provide predicted values or confidence intervals for predicted values. Note that as in the Minitab regression analysis, we will not make use of the ANOVA information in the Excel display.

On the home screen, click the **Data** tab. Select **Data Analysis** from the Analysis group. In the dialogue box, select **Regression**. Note that when you enter data into the worksheet, all the explanatory variables must be together in a block. Figure 9-23 shows the Excel display for Examples 10 and 11.

FIGURE 9-23

Excel Display of Regression Analysis

Regression Statistics						
Multiple R	0.987060478					
R Square	0.974288388					
Adjusted R Square	0.955004679					
Standard Error	0.120927579					
Observations	8					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	2.216506083	0.738835361	50.5239104	0.001228863	
Residual	4	0.058493917	0.014623479			
Total	7	2.275				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-5.922011616	1.255623292	-4.716391972	0.009196085	-9.40818798	-2.435835251
x2	0.338217487	0.099470083	3.400193085	0.027272474	0.062043691	0.614391283
x3	0.401503945	0.109900277	3.653347874	0.021707246	0.096371226	0.706636664
x4	0.262946128	0.085136028	3.088541172	0.036626194	0.02657013	0.499322125

VIEWPOINT Synoptic Climatology

Synoptic *means "giving a summary from the same basic point of view." In this case, the point of view is Niwot Ridge, high above the timberline in the Rocky Mountains. Vegetation, water, temperature, and wind all affect the delicate balance of this alpine environment. How do these elements of nature interact to sustain life in such a harsh land? One answer can be found by collecting data at the location and using multiple regression to study the interaction of variables. For more information, visit the Brase/Brase statistics site at* **http://www.cengage.com/statistics/brase** *and find the link to Niwot Ridge Climate Study.*

SECTION 9.4 PROBLEMS

1. *Statistical Literacy* Given the linear regression equation

$$x_1 = 1.6 + 3.5x_2 - 7.9x_3 + 2.0x_4$$

(a) Which variable is the response variable? Which variables are the explanatory variables?

(b) Which number is the constant term? List the coefficients with their corresponding explanatory variables.

(c) If $x_2 = 2$, $x_3 = 1$, and $x_4 = 5$, what is the predicted value for x_1?

(d) Explain how each coefficient can be thought of as a "slope" under certain conditions. Suppose x_3 and x_4 were held at fixed but arbitrary values and x_2 was increased by 1 unit. What would be the corresponding change in x_1? Suppose x_2 increased by 2 units. What would be the expected change in x_1? Suppose x_2 decreased by 4 units. What would be the expected change in x_1?

(e) Suppose that $n = 12$ data points were used to construct the given regression equation and that the standard error for the coefficient of x_2 is 0.419. Construct a 90% confidence interval for the coefficient of x_2.

(f) Using the information of part (e) and level of significance 5%, test the claim that the coefficient of x_2 is different from zero. Explain how the conclusion of this test would affect the regression equation.

2. *Statistical Literacy* Given the linear regression equation

$$x_3 = -16.5 + 4.0x_1 + 9.2x_4 - 1.1x_7$$

(a) Which variable is the response variable? Which variables are the explanatory variables?

(b) Which number is the constant term? List the coefficients with their corresponding explanatory variables.

(c) If $x_1 = 10$, $x_4 = -1$, and $x_7 = 2$, what is the predicted value for x_3?

(d) Explain how each coefficient can be thought of as a "slope." Suppose x_1 and x_7 were held as fixed but arbitrary values. If x_4 increased by 1 unit, what would we expect the corresponding change in x_3 to be? If x_4 increased by 3 units, what would be the corresponding expected change in x_3? If x_4 decreased by 2 units, what would we expect for the corresponding change in x_3?

(e) Suppose that $n = 15$ data points were used to construct the given regression equation and that the standard error for the coefficient of x_4 is 0.921. Construct a 90% confidence interval for the coefficient of x_4.

(f) Using the information of part (e) and level of significance 1%, test the claim that the coefficient of x_4 is different from zero. Explain how the conclusion has a bearing on the regression equation.

For Problems 3–6, use appropriate multiple regression software of your choice and enter the data. Note that the data are also available for download at the Online Study Center in formats for Excel, Minitab portable files, SPSS files, and ASCII files.

3. *Medical: Blood Pressure* The systolic blood pressure of individuals is thought to be related to both age and weight. For a random sample of 11 men, the following data were obtained:

Systolic Blood Pressure x_1	Age (years) x_2	Weight (pounds) x_3	Systolic Blood Pressure x_1	Age (years) x_2	Weight (pounds) x_3
132	52	173	137	54	188
143	59	184	149	61	188
153	67	194	159	65	207
162	73	211	128	46	167
154	64	196	166	72	217
168	74	220			

(a) Generate summary statistics, including the mean and standard deviation of each variable. Compute the coefficient of variation (see Section 3.2) for each variable. Relative to its mean, which variable has the greatest spread of data values? Which variable has the smallest spread of data values relative to its mean?

(b) For each pair of variables, generate the sample correlation coefficient r. Compute the corresponding coefficient of determination r^2. Which variable (other than x_1) has the greatest influence (by itself) on x_1? Would you say that both variables x_2 and x_3 show a strong influence on x_1? Explain your answer. What percent of the variation in x_1 can be explained by the corresponding variation in x_2? Answer the same question for x_3.

(c) Perform a regression analysis with x_1 as the response variable. Use x_2 and x_3 as explanatory variables. Look at the coefficient of multiple determination. What percentage of the variation in x_1 can be explained by the corresponding variations in x_2 and x_3 *taken together*?

(d) Look at the coefficients of the regression equation. Write out the regression equation. Explain how each coefficient can be thought of as a slope. If age were held fixed, but a person put on 10 pounds, what would you expect for the corresponding change in systolic blood pressure? If a person kept the same weight but got 10 years older, what would you expect for the corresponding change in systolic blood pressure?

(e) Test each coefficient to determine if it is zero or not zero. Use level of significance 5%. Why would the outcome of each test help us determine whether or not a given variable should be used in the regression model?

(f) Find a 90% confidence interval for each coefficient.

(g) Suppose Michael is 68 years old and weighs 192 pounds. Predict his systolic blood pressure, and find a 90% confidence range for your prediction (if your software produces prediction intervals).

4. ***Education: Exam Scores*** Professor Gill has taught general psychology for many years. During the semester, she gives three multiple-choice exams, each worth 100 points. At the end of the course, Dr. Gill gives a comprehensive final worth 200 points. Let x_1, x_2, and x_3 represent a student's scores on exams 1, 2, and 3, respectively. Let x_4 represent the student's score on the final exam. Last semester Dr. Gill had 25 students in her class. The student exam scores are shown below.

x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
73	80	75	152	79	70	88	164	81	90	93	183
93	88	93	185	69	70	73	141	88	92	86	177
89	91	90	180	70	65	74	141	78	83	77	159
96	98	100	196	93	95	91	184	82	86	90	177
73	66	70	142	79	80	73	152	86	82	89	175
53	46	55	101	70	73	78	148	78	83	85	175
69	74	77	149	93	89	96	192	76	83	71	149
47	56	60	115	78	75	68	147	96	93	95	192
87	79	90	175								

Since Professor Gill has not changed the course much from last semester to the present semester, the preceding data should be useful for constructing a regression model that describes this semester as well.

(a) Generate summary statistics, including the mean and standard deviation of each variable. Compute the coefficient of variation (see Section 3.2) for each variable. Relative to its mean, would you say that each exam had about the same spread of scores? Most professors do not wish to give an exam that is extremely easy or extremely hard. Would you say that all of the exams were about the same level of difficulty? (Consider both means and spread of test scores.)

(b) For each pair of variables, generate the sample correlation coefficient r. Compute the corresponding coefficient of determination r^2. Of the three exams 1, 2, and 3, which do you think had the most influence on the final exam 4? Although one exam had more influence on the final exam, did the other two exams still have a lot of influence on the final? Explain each answer.

(c) Perform a regression analysis with x_4 as the response variable. Use x_1, x_2, and x_3 as explanatory variables. Look at the coefficient of multiple determination. What percentage of the variation in x_4 can be explained by the corresponding variations in x_1, x_2, and x_3 taken together?

(d) Write out the regression equation. Explain how each coefficient can be thought of as a slope. If a student were to study "extra hard" for exam 3 and increase his or her score on that exam by 10 points, what corresponding change would you expect on the final exam? (Assume that exams 1 and 2 remain "fixed" in their scores.)

(e) Test each coefficient in the regression equation to determine if it is zero or not zero. Use level of significance 5%. Why would the outcome of each hypothesis test help us decide whether or not a given variable should be used in the regression equation?

(f) Find a 90% confidence interval for each coefficient.

(g) This semester Susan has scores of 68, 72, and 75 on exams 1, 2, and 3, respectively. Make a prediction for Susan's score on the final exam and find a 90% confidence interval for your prediction (if your software supports prediction intervals).

5. *Entertainment: Movies* A motion picture industry analyst is studying movies based on epic novels. The following data were obtained for 10 Hollywood movies made in the past five years. Each movie was based on an epic novel. For these data, x_1 = first-year box office receipts of the movie, x_2 = total production costs of the movie, x_3 = total promotional costs of the movie, and x_4 = total book sales prior to movie release. All units are in millions of dollars.

x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
85.1	8.5	5.1	4.7	30.3	3.5	1.2	3.5
106.3	12.9	5.8	8.8	79.4	9.2	3.7	9.7
50.2	5.2	2.1	15.1	91.0	9.0	7.6	5.9
130.6	10.7	8.4	12.2	135.4	15.1	7.7	20.8
54.8	3.1	2.9	10.6	89.3	10.2	4.5	7.9

(a) Generate summary statistics, including the mean and standard deviation of each variable. Compute the coefficient of variation (see Section 3.2) for each variable. Relative to its mean, which variable has the largest spread of data values? Why would a variable with a large coefficient of variation be expected to change a lot relative to its average value? Although x_1 has the largest standard deviation, it has the smallest coefficient of variation. How does the mean of x_1 help explain this?

(b) For each pair of variables, generate the sample correlation coefficient r. Compute the corresponding coefficient of determination r^2. Which of the three variables x_2, x_3, and x_4 has the *least* influence on box office receipts? What percent of the variation in box office receipts can be attributed to the corresponding variation in production costs?

(c) Perform a regression analysis with x_1 as the response variable. Use x_2, x_3, and x_4 as explanatory variables. Look at the coefficient of multiple determination. What percentage of the variation in x_1 can be explained by the corresponding variations in x_2, x_3, and x_4 taken together?

(d) Write out the regression equation. Explain how each coefficient can be thought of as a slope. If x_2 (production costs) and x_4 (book sales) were held fixed but x_3 (promotional costs) was increased by $1 million, what would you expect for the corresponding change in x_1 (box office receipts)?

(e) Test each coefficient in the regression equation to determine if it is zero or not zero. Use level of significance 5%. Explain why book sales x_4 probably are not contributing much information in the regression model to forecast box office receipts x_1.

(f) Find a 90% confidence interval for each coefficient.

(g) Suppose a new movie (based on an epic novel) has just been released. Production costs were $x_2 = 11.4$ million; promotion costs were $x_3 = 4.7$ million; book sales were $x_4 = 8.1$ million. Make a prediction for $x_1 =$ first-year box office receipts and find an 85% confidence interval for your prediction (if your software supports prediction intervals).

(h) Construct a new regression model with x_3 as the response variable and x_1, x_2, and x_4 as explanatory variables. Suppose Hollywood is planning a new epic movie with projected box office sales $x_1 = 100$ million and production costs $x_2 = 12$ million. The book on which the movie is based had sales of $x_4 = 9.2$ million. Forecast the dollar amount (in millions) that should be budgeted for promotion costs x_3 and find an 80% confidence interval for your prediction.

6. *Franchise Business: Market Analysis* All Greens is a franchise store that sells house plants and lawn and garden supplies. Although All Greens is a franchise, each store is owned and managed by private individuals. Some friends have asked you to go into business with them to open a new All Greens store in the suburbs of San Diego. The national franchise headquarters sent you the following information at your request. These data are about 27 All Greens stores in California. Each of the 27 stores has been doing very well, and you would like to use the information to help set up your own new store. The variables for which we have data are

x_1 = annual net sales, in thousands of dollars
x_2 = number of square feet of floor display in store, in thousands of square feet
x_3 = value of store inventory, in thousands of dollars
x_4 = amount spent on local advertising, in thousands of dollars
x_5 = size of sales district, in thousands of families
x_6 = number of competing or similar stores in sales district

A sales district was defined to be the region within a 5-mile radius of an All Greens store.

x_1	x_2	x_3	x_4	x_5	x_6	x_1	x_2	x_3	x_4	x_5	x_6
231	3	294	8.2	8.2	11	65	1.2	168	4.7	3.3	11
156	2.2	232	6.9	4.1	12	98	1.6	151	4.6	2.7	10
10	0.5	149	3	4.3	15	398	4.3	342	5.5	16.0	4
519	5.5	600	12	16.1	1	161	2.6	196	7.2	6.3	13
437	4.4	567	10.6	14.1	5	397	3.8	453	10.4	13.9	7
487	4.8	571	11.8	12.7	4	497	5.3	518	11.5	16.3	1
299	3.1	512	8.1	10.1	10	528	5.6	615	12.3	16.0	0
195	2.5	347	7.7	8.4	12	99	0.8	278	2.8	6.5	14
20	1.2	212	3.3	2.1	15	0.5	1.1	142	3.1	1.6	12
68	0.6	102	4.9	4.7	8	347	3.6	461	9.6	11.3	6
570	5.4	788	17.4	12.3	1	341	3.5	382	9.8	11.5	5
428	4.2	577	10.5	14.0	7	507	5.1	590	12.0	15.7	0
464	4.7	535	11.3	15.0	3	400	8.6	517	7.0	12.0	8
15	0.6	163	2.5	2.5	14						

(a) Generate summary statistics, including the mean and standard deviation of each variable. Compute the coefficient of variation (see Section 3.2) for each variable. Relative to its mean, which variable has the largest spread of data values? Which variable has the least spread of data values relative to its mean?

(b) For each pair of variables, generate the sample correlation coefficient r. For all pairs involving x_1, compute the corresponding coefficient of determination r^2. Which variable has the greatest influence on annual net sales? Which variable has the least influence on annual net sales?

(c) Perform a regression analysis with x_1 as the response variable. Use x_2, x_3, x_4, x_5, and x_6 as explanatory variables. Look at the coefficient of multiple determination. What percentage of the variation in x_1 can be explained by the corresponding variations in x_2, x_3, x_4, x_5, and x_6 taken together?

(d) Write out the regression equation. If two new competing stores moved into the sales district but the other explanatory variables did not change, what would you expect for the corresponding change in annual net sales? Explain your answer. If you increased the local advertising by a thousand dollars but the other explanatory variables did not change, what would you expect for the corresponding change in annual net sales? Explain.

(e) Test each coefficient to determine if it is or is not zero. Use level of significance 5%.

(f) Suppose you and your business associates rent a store, get a bank loan to start up your business, and do a little research on the size of your sales district and the number of competing stores in the district. If $x_2 = 2.8$, $x_3 = 250$, $x_4 = 3.1$, $x_5 = 7.3$, and $x_6 = 2$, use a computer to forecast $x_1 =$ annual net sales and find an 80% confidence interval for your forecast (if your software produces prediction intervals).

(g) Construct a new regression model with x_4 as the response variable and x_1, x_2, x_3, x_5, and x_6 as explanatory variables. Suppose an All Greens store in Sonoma, California, wants to estimate a range of advertising costs appropriate to its store. If it spends too little on advertising, it will not reach enough customers. However, it does not want to overspend on advertising for this type and size of store. At this store, $x_1 = 163$, $x_2 = 2.4$, $x_3 = 188$, $x_5 = 6.6$, and $x_6 = 10$. Use these data to predict x_4 (advertising costs) and find an 80% confidence interval for your prediction. At the 80% confidence level, what range of advertising costs do you think is appropriate for this store?

7. ***Expand Your Knowledge: Curvilinear Polynomial Regression*** In this section we studied multiple linear regression. Our basic linear model has been

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

Since all the variables $x_1, x_2, \ldots, x_k$ are of first degree, this is an example of linear regression. However, the same basic methods of linear regression can be used for *curvilinear regression* (also known as *polynomial regression*). The interested reader can find a great deal of information on this topic in the book *Applied Numerical Methods* by Carnahan, Luther, and Wilkes from page 573 on.

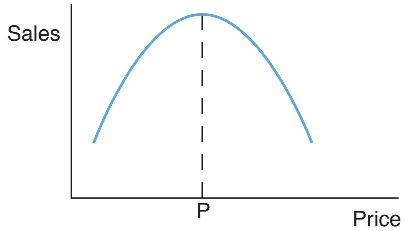
Assume we have at least $k + 2$ data pairs (x, y) and we want to approximate y using a polynomial of degree k. To do this, we make the following identification.

$$x_1 = x; x_2 = x^2; x_3 = x^3; \ldots ; x_k = x^k$$

Then we use our known methods of multiple regression to obtain coefficients $b_0, b_1, b_2, b_3, \ldots, b_k$ and the equation $y = b_0 + b_1x_1 + b_2x^2 + b_3x^3 + \cdots + b_kx^k$. This is called the *least-squares curvilinear regression model*.

Marketing studies show that price increases often have a point of diminishing returns. For a popular product, the price can often increase with sales.

However, when the price becomes too high, sales start to drop off. In the following graph, P = point of diminishing returns.



To estimate the point of diminishing returns, we use a quadratic polynomial, $y = b_0 + b_1x + b_2x^2$. A very popular women's knit T-shirt was tested for sales appeal and price in six large department stores. In each city, the T-shirts were advertised extensively in the local media, so price and sales initially went up. However, as price increased, sales eventually dropped off. Let x = price per T-shirt in dollars and y = number of T-shirts sold in a day at that price. We have the following data.

City	A	B	C	D	E	F
x	12.97	13.88	15.95	18.50	19.99	22.50
y	23	31	33	29	25	17

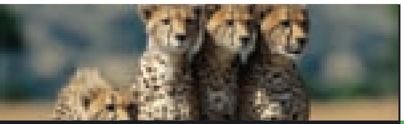
To construct our quadratic polynomial, we use multilinear regression with the following table of data values.

$x_1 = x$	12.97	13.88	15.95	18.50	19.99	22.50
$x_2 = x^2$	168.22	192.65	254.40	342.25	399.60	506.25
y	23	31	33	29	25	17

Computer software gives us coefficients for the model $y = b_0 + b_1x_1 + b_2x_2 = b_0 + b_1x + b_2x^2$, which becomes $y = -93.80 + 15.10x - 0.45x^2$. The coefficient of determination is $r^2 = 0.88$ (not too bad!). The curvilinear regression equation $y = -93.80 + 15.10x - 0.45x^2$ is a quadratic curve that opens downward. A little extra mathematics shows that the top point on the curve (point of diminishing returns) occurs when the cost per shirt of $x = \$16.78$ with $y = 32.87$ shirts sold per day. This suggests the knit t-shirts should be priced at \$16.78 and that about 33 of them will sell per day in a large department store.

Use the Internet, school library, popular magazines, or any other source to collect (x, y) data pairs regarding variables of interest to you. Construct a curvilinear regression model from your data and interpret the results.

TECH NOTES In the Using Technology section at the end of this chapter, you will find a "mini case study" of seven important variables from the economy of the United States for the years 1976 to 1987. Readers interested in applications of multiple regression and the U.S. economy are referred to this material.



Chapter Review

SUMMARY

This chapter discusses linear regression models and inferences related to these models.

- A scatter diagram of data pairs (x, y) gives a graphical display of the relationship (if any) between x and y data. We are looking for a linear relationship.
- For data pairs (x, y), x is called the *explanatory variable* and is plotted along the horizontal axis. The *response variable* y is plotted along the vertical axis.
- The Pearson product-moment *correlation coefficient* r gives a numerical measurement assessing the strength of a linear relationship between x and y. It is based on a random sample of (x, y) data pairs.
- The value of r ranges from -1 to 1, with 1 indicating perfect positive linear correlation, -1 indicating perfect negative linear correlation, and 0 indicating no linear correlation.
- If the scatter diagram and sample correlation coefficient r indicate a linear relationship between x and y values of the data pairs, we use the least-squares criteria to develop the equation of the least-squares line

  $$\hat{y} = a + bx$$

  where $\hat{y}$ is the value of y predicted by the least-squares line for a given x value, a is the y intercept, and b is the slope.
- Methods of testing the population correlation coefficient ρ show whether or not the sample statistic r is significant. We test the null hypothesis H_0: $\rho = 0$ against a suitable alternate hypothesis ($\rho > 0$, $\rho < 0$, or $\rho \neq 0$).
- Methods of testing the population slope β show whether or not the sample slope b is significant. We test the null hypothesis H_0: $\beta = 0$ against a suitable alternate hypothesis ($\beta > 0$, $\beta < 0$, or $\beta \neq 0$).
- Confidence intervals for β give us a range of values for β based on the sample statistic b and specified confidence level c.
- Confidence intervals for the predicted value of y give us a range of values for y for a specific x value. The interval is based on the sample prediction $\hat{y}$ and confidence level c.
- The *coefficient of determination* r^2 is a value that measures the proportion of variation in y explained by the least-squares line, the linear regression model, and the variation in the explanatory variable x.
- The difference $y - \hat{y}$ between the y value in the data pair (x, y) and the corresponding predicted value $\hat{y}$ for the same x is called the *residual*.
- The *standard error of estimate* S_e is a measure of data spread about the least-squares line. It is based on the residuals.
- Techniques of multiple regression (with computer assistance) help us analyze a linear relation involving several variables.

IMPORTANT WORDS & SYMBOLS

Section 9.1

Paired data values 502
Scatter diagram 502
Explanatory variable 502
Response variable 502
No linear correlation 503
Perfect linear correlation 504
Positive correlation 504
Negative correlation 505
Sample correlation coefficient r 506
Extrapolation 513
Causation 513
Lurking variable 513

Section 9.2

Least-squares criterion 521
Least-squares line $\hat{y} = a + bx$ 522
Slope b 522
Intercept a 522
Meaning of slope 524
Marginal change 525
Influential point 525
Residual 525
Interpolation 526
Extrapolation 526
Coefficient of determination r^2 530

VIEWPOINT Living Arrangements

Male, female, married, single, living alone, living with friends or relatives—all these categories are of interest to the U.S. Census Bureau. In addition to these categories, there are others, such as age, income, and health needs. How strongly correlated are these variables? Can we use one or more of these variables to predict the others? How good is such a prediction? Methods of this chapter can help you answer such questions. For more information regarding such data, visit the Brase/Brase statistics site at **http://www.cengage.com/statistics/brase** *and find the link to Census Bureau.*

CHAPTER REVIEW PROBLEMS

1. *Statistical Literacy* Suppose the scatter diagram of a random sample of data pairs (x, y) shows no linear relationship between x and y. Do you expect the value of the sample correlation coefficient r to be close to 1, -1, or 0?

2. *Statistical Literacy* What does it mean to say that the sample correlation coefficient r is significant?

3. *Statistical Literacy* When using the least-squares line for prediction, are results usually more reliable for extrapolation or interpolation?

4. *Statistical Literacy* Suppose that for $x = 3$, the predicted value is $\hat{y} = 6$. The data pair (3, 8) is part of the sample data. What is the value of the residual for $x = 3$?

In Problems 5–10, parts (a)–(e) involve scatter diagrams, least-squares lines, correlation coefficients with coefficients of determination, tests of ρ, and predictions. Parts (f)–(i) involve standard error of estimate, confidence intervals for predictions, tests of β, and confidence intervals for β.

When solving problems involving the standard error of estimate, testing of the correlation coefficient, or testing of β or confidence intervals for β, make the assumption that x and y are normally distributed random variables. Answers may vary slightly due to rounding.

5. *Desert Ecology: Wildlife* Bighorn sheep are beautiful wild animals found throughout the western United States. Data for this problem are based on information taken from *The Desert Bighorn*, edited by Monson and Sumner (University of Arizona Press). Let x be the age of a bighorn sheep (in years), and let y be the mortality rate (percent that die) for this age group. For example, $x = 1$, $y = 14$ means that 14% of the bighorn sheep between 1 and 2 years old die. A random sample of Arizona bighorn sheep gave the following information:

x	1	2	3	4	5
y	14	18.9	14.4	19.6	20.0

$\Sigma x = 15$; $\Sigma y = 86.9$; $\Sigma x^2 = 55$; $\Sigma y^2 = 1544.73$; $\Sigma xy = 273.4$

(a) Draw a scatter diagram.
(b) Find the equation of the least-squares line.
(c) Find r. Find the coefficient of determination r^2. Explain what these measures mean in the context of the problem.

(d) Test the claim that the population correlation coefficient is positive at the 1% level of significance.
(e) Given the lack of significance of r, is it practical to find estimates of y for a given x value based on the least-squares line model? Explain.

6. ***Sociology: Job Changes*** A sociologist is interested in the relation between x = number of job changes and y = annual salary (in thousands of dollars) for people living in the Nashville area. A random sample of 10 people employed in Nashville provided the following information:

x (Number of job changes)	4	7	5	6	1	5	9	10	10	3
y (Salary in \$1000)	33	37	34	32	32	38	43	37	40	33

$\Sigma x = 60$; $\Sigma y = 359$; $\Sigma x^2 = 442$; $\Sigma y^2 = 13{,}013$; $\Sigma xy = 2231$
(a) Draw a scatter diagram for the data.
(b) Find $\bar{x}$, $\bar{y}$, b, and the equation of the least-squares line. Plot the line on the scatter diagram of part (a).
(c) Find the sample correlation coefficient r and the coefficient of determination. What percentage of variation in y is explained by the least-squares model?
(d) Test the claim that the population correlation coefficient ρ is positive at the 5% level of significance.
(e) If someone had $x = 2$ job changes, what does the least-squares line predict for y, the annual salary?
(f) Verify that $S_e \approx 2.56$.
(g) Find a 90% confidence interval for the annual salary of an individual with $x = 2$ job changes.
(h) Test the claim that the slope β of the population least-squares line is positive at the 5% level of significance.
(i) Find a 90% confidence interval for β and interpret its meaning.

7. ***Medical: Fat Babies*** Modern medical practice tells us not to encourage babies to become too fat. Is there a positive correlation between the weight x of a 1-year-old baby and the weight y of the mature adult (30 years old)? A random sample of medical files produced the following information for 14 females:

x (lb)	21	25	23	24	20	15	25	21	17	24	26	22	18	19
y (lb)	125	125	120	125	130	120	145	130	130	130	130	140	110	115

$\Sigma x = 300$; $\Sigma y = 1775$; $\Sigma x^2 = 6572$; $\Sigma y^2 = 226{,}125$; $\Sigma xy = 38{,}220$
(a) Draw a scatter diagram for the data.
(b) Find $\bar{x}$, $\bar{y}$, b, and the equation of the least-squares line. Plot the line on the scatter diagram of part (a).
(c) Find the sample correlation coefficient r and the coefficient of determination. What percentage of the variation in y is explained by the least-squares model?
(d) Test the claim that the population correlation coefficient ρ is positive at the 1% level of significance.
(e) If a female baby weighs 20 pounds at 1 year, what do you predict she will weigh at 30 years of age?
(f) Verify that $S_e \approx 8.38$.
(g) Find a 95% confidence interval for weight at age 30 of a female who weighed 20 pounds at 1 year of age.
(h) Test the claim that the slope β of the population least-squares line is positive at the 1% level of significance.
(i) Find an 80% confidence interval for β and interpret its meaning.

8. *Sales: Insurance* Dorothy Kelly sells life insurance for the Prudence Insurance Company. She sells insurance by making visits to her clients' homes. Dorothy believes that the number of sales should depend, to some degree, on the number of visits made. For the past several years, she has kept careful records of the number of visits (x) she makes each week and the number of people (y) who buy insurance that week. For a random sample of 15 such weeks, the x and y values follow:

x	11	19	16	13	28	5	20	14	22	7	15	29	8	25	16
y	3	11	8	5	8	2	5	6	8	3	5	10	6	10	7

$\Sigma x = 248$; $\Sigma y = 97$; $\Sigma x^2 = 4856$; $\Sigma y^2 = 731$; $\Sigma xy = 1825$

(a) Draw a scatter diagram for the data.
(b) Find $\bar{x}$, $\bar{y}$, b, and the equation of the least-squares line. Plot the line on the scatter diagram of part (a).
(c) Find the sample correlation coefficient r and the coefficient of determination. What percentage of the variation in y is explained by the least-squares model?
(d) Test the claim that the population correlation coefficient ρ is positive at the 1% level of significance.
(e) In a week during which Dorothy makes 18 visits, how many people do you predict will buy insurance from her?
(f) Verify that $S_e \approx 1.731$.
(g) Find a 95% confidence interval for the number of sales Dorothy would make in a week during which she made 18 visits.
(h) Test the claim that the slope β of the population least-squares line is positive at the 1% level of significance.
(i) Find an 80% confidence interval for β and interpret its meaning.

9. *Marketing: Coupons* Each box of Healthy Crunch breakfast cereal contains a coupon entitling you to a free package of garden seeds. At the Healthy Crunch home office, they use the weight of incoming mail to determine how many of their employees are to be assigned to collecting coupons and mailing out seed packages on a given day. (Healthy Crunch has a policy of answering all its mail on the day it is received.)

Let x = weight of incoming mail and y = number of employees required to process the mail in one working day. A random sample of 8 days gave the following data:

x (lb)	11	20	16	6	12	18	23	25
y (Number of employees)	6	10	9	5	8	14	13	16

$\Sigma x = 131$; $\Sigma y = 81$; $\Sigma x^2 = 2435$; $\Sigma y^2 = 927$; $\Sigma xy = 1487$

(a) Draw a scatter diagram for the data.
(b) Find $\bar{x}$, $\bar{y}$, b, and the equation of the least-squares line. Plot the line on the scatter diagram of part (a).
(c) Find the sample correlation coefficient r and the coefficient of determination. What percentage of the variation in y is explained by the least-squares model?
(d) Test the claim that the population correlation coefficient ρ is positive at the 1% level of significance.
(e) If Healthy Crunch receives 15 pounds of mail, how many employees should be assigned mail duty that day?
(f) Verify that $S_e \approx 1.726$.
(g) Find a 95% confidence interval for the number of employees required to process mail for 15 pounds of mail.

(h) Test the claim that the slope β of the population least-squares line is positive at the 1% level of significance.
(i) Find an 80% confidence interval for β and interpret its meaning.

10. ***Focus Problem: Changing Population and Crime Rate*** Let x be a random variable representing percentage change in neighborhood population in the past few years, and let y be a random variable representing crime rate (crimes per 1000 population). A random sample of six Denver neighborhoods gave the following information (Source: *Neighborhood Facts*, The Piton Foundation).

x	29	2	11	17	7	6
y	173	35	132	127	69	53

$\Sigma x = 72$; $\Sigma y = 589$; $\Sigma x^2 = 1340$; $\Sigma y^2 = 72{,}277$; $\Sigma xy = 9499$
(a) Draw a scatter diagram for the data.
(b) Find $\bar{x}$, $\bar{y}$, b, and the equation of the least-squares line. Plot the line on the scatter diagram of part (a).
(c) Find the sample correlation coefficient r and the coefficient of determination. What percentage of the variation in y is explained by the least-squares model?
(d) Test the claim that the population correlation coefficient ρ is not zero at the 1% level of significance.
(e) For a neighborhood with $x = 12\%$ change in population in the past few years, predict the change in the crime rate (per 1000 residents).
(f) Verify that $S_e \approx 22.5908$.
(g) Find an 80% confidence interval for the change in crime rate when the percentage change in population is $x = 12\%$.
(h) Test the claim that the slope β of the population least-squares line is not zero at the 1% level of significance.
(i) Find an 80% confidence interval for β and interpret its meaning.

DATA HIGHLIGHTS: GROUP PROJECTS

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

Scatter diagrams! Are they really useful? Scatter diagrams give a first impression of a data relationship and help us assess whether a linear relation provides a reasonable model for the data. In addition, we can spot *influential points*. A data point with an extreme x value can heavily influence the position of the least-squares line. In this project, we look at data sets with an influential point.

x	1	4	5	9	10	15
y	3	7	6	10	12	4

(a) Compute r and b, the slope of the least-squares line. Find the equation of the least-squares line, and sketch the line on the scatter diagram.

(b) Notice the point boxed in blue in Figure 9-24. Does it seem to lie away from the linear pattern determined by the other points? The coordinates of that point are (15, 4). Is it an influential point? Remove that point from the model and recompute r, b, and the equation of the least-squares line. Sketch this least-squares line on the diagram. How does the removal of the influential point affect the values of r and b and the position of the least-squares line?

(c) Consider the scatter diagram of Figure 9-25. Is there an influential point? If you remove the influential point, will the slope of the new least-squares line be larger or smaller than the slope of the line from the original data? Will the correlation coefficient be larger or smaller?

FIGURE 9-24

Scatter Diagram

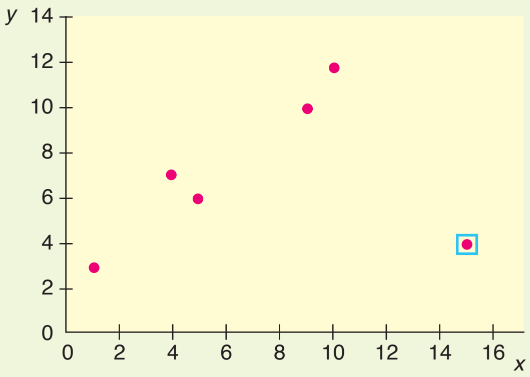
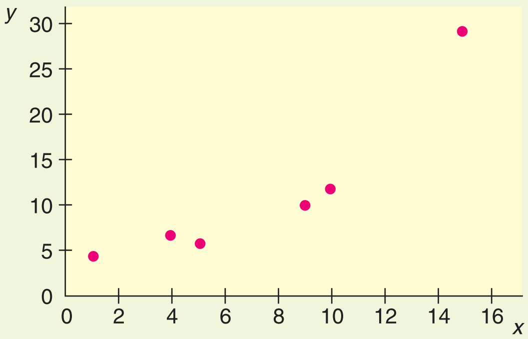


FIGURE 9-25

Scatter Diagram



LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

1. What do we mean when we say that two variables have a strong positive (or negative) linear correlation? What would a scatter diagram for these variables look like? Is it possible that two variables could be strongly related somehow but have a low *linear* correlation? Explain and draw a scatter diagram to demonstrate your point.

2. What do we mean by the least-squares criterion? Give a very general description of how the least-squares criterion is involved in the construction of the least-squares line. Why do we say the least-squares line is the "best-fitting" line for the data set?

3. In this chapter, we discussed three measures for "goodness of fit" of the least-squares line for given data. These measures were standard error of estimate, correlation coefficient, and coefficient of determination. Discuss the ways in which these measurements are different and the ways in which they are similar to each other. Be sure to include a discussion of explained variation, unexplained variation, and total variation in your answer. Draw a sketch and include appropriate formulas.

4. Look at the formula for confidence bounds for least-squares predictions. Which of the following conditions do you think will result in a *shorter* confidence interval for a prediction?
   (a) Larger or smaller values for the standard error of estimate
   (b) Larger or smaller number of data pairs
   (c) A value of x near $\bar{x}$ or a value of x far away from $\bar{x}$
   Why would a shorter confidence interval for a prediction be more desirable than a longer interval?

5. If you did not cover Section 9.4, Multiple Regression, omit this problem.

   For many applications in statistics, more data lead to more accurate results. In multiple regression, we have more variables (and data) than we have in most simple regression problems. Why will this usually lead to more accurate predictions? Will additional variables *always* lead to more accurate predictions? Explain your answer. Discuss the coefficient of multiple determination and its meaning in the context of multiple regression. How do we know if an explanatory variable has a statistically significant influence on the response variable? What do we mean by a regression model?

6. Use the Internet or go to the library and find a magazine or journal article in your field of major interest to which the content of this chapter could be applied. List the variables used, method of data collection, and general type of information and conclusions drawn.

Using Technology

Simple Linear Regression (One Explanatory Variable)

Application 1

The data in this section are taken from this source:

Based on King, Cuchlaine A. M. *Physical Geography*. Oxford: Basil Blackwell, 1980, pp. 77–86, 196–206.

Throughout the world, natural ocean beaches are beautiful sights to see. If you have visited natural beaches, you may have noticed that when the gradient or dropoff is steep, the grains of sand tend to be larger. In fact, a man-made beach with the "wrong"size granules of sand tends to be washed away and eventually replaced when the proper size grain is selected by the action of the ocean and the gradient of the bottom. Since man-made beaches are expensive, grain size is an important consideration.

In the data that follow, x = median diameter (in millimeters) of granules of sand, and y = gradient of beach slope in degrees on natural ocean beaches.

x	y
0.17	0.63
0.19	0.70
0.22	0.82
0.235	0.88
0.235	1.15
0.30	1.50
0.35	4.40
0.42	7.30
0.85	11.30

1. Find the sample mean and standard deviation for x and y.
2. Make a scatter plot. Would you expect a moderately high correlation and a good fit for the least-squares line?
3. Find the equation of the least-squares line, and graph the line on the scatter plot.
4. Find the sample correlation coefficient r and the coefficient of determination r^2. Is r significant at the 1% level of significance (two-tailed test)?
5. Test that $\beta > 0$ at the 1% level of significance. Find the standard error of estimate S_e and form an 80% confidence interval for β. As the diameter of granules of sand changes by 0.10 mm, by how much does the gradient of beach slope change?
6. Suppose you have a truckload of sifted sand in which the median size of granules is 0.38 mm. If you want to put this sand on a beach and you don't want the sand to wash away, then what does the least-squares line predict for the angle of the beach? *Note:* Heavy storms that produce abnormal waves may also wash out the sand. However, in the long run, the size of sand granules that remain on the beach or that are brought back to the beach by long-term wave action are determined to a large extent by the angle at which the beach drops off. What range of angles should the beach have if we want to be 90% confident that we are matching the size of our sand granules (0.38 mm) to the proper angle of the beach?
7. Suppose we now have a truckload of sifted sand in which the median size of the granules is 0.45 mm. Repeat Problem 6.

Technology Hints (Simple Regression)

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad)

Be sure to set **DiagnosticOn** (under **Catalog**).

(a) Scatter diagram: Use **STAT PLOT,** select the first type, use **ZOOM** option **9:ZoomStat.**

(b) Least-squares line and r: Use **STAT, CALC,** option **8:LinReg**(a + bx).

(c) Graph least-squares line and predict: Press Y=. Then, under **VARS,** select **5:Statistics,** then select **EQ,** and finally select item **1:RegEQ.** Press enter. This sequence of steps will automatically set $Y_1 =$ your regression equation. Press **GRAPH.** To find a predicted value, when the graph is showing press the **CALC** key and select item **1:Value**. Enter the x value, and the corresponding y value will appear.

(d) Testing ρ and β, value for S_e: Use **STAT, TEST,** option **E:LinRegTTest**. The value of S_e is in the display as s.

(e) Confidence intervals for β or predictions: Use formulas from Section 9.3.

Excel 2007

(a) Scatter plot, least-squares line, r^2: On home screen, click **Insert** tab. In the Charts group, select **Scatter** and choose the first type. Once plot is displayed, *right* click on any data point. Select **trend line.** Under options, check display line and display r^2.

(b) Prediction: Use **insert function** f_x ➤ **Statistical** ➤ **Forecast**.

(c) Coefficient r: Use f_x ➤ **Statistical** ➤ **Correl.**

(d) Testing β and confidence intervals for b: Use menu selection **Tools** ➤ **Data Analysis** ➤ **Regression.**

(e) Confidence interval for prediction: Use formulas from Section 9.3.

Minitab

(a) Scatter plot, least-squares line, r^2, S_e: Use menu selection **Stat** ➤ **Regression** ➤ **Fitted line plot.** The value of S_e is displayed as the value of s.

(b) Coefficient r: Use menu selection **Stat** ➤ **Basic Statistics** ➤ **Correlation.**

(c) Testing β, predictions, confidence interval for predictions: Use menu selection **Stat** ➤ **Regression** ➤ **Regression.**

(d) Confidence interval for β: Use formulas from Section 9.3.

SPSS

SPSS offers several options for finding the correlation coefficient r and the equation of the least-squares line. First enter the data in the data editor and label the variables appropriately in the variable view window. Use the menu choices **Analyze** ➤ **Regression** ➤ **Linear** and select dependent and independent variables. The output includes the correlation coefficient, the standard error of estimate, the constant, and the coefficient of the dependent variable with corresponding t values and P-values for two-tailed tests. The display shows the results for the data in this chapter's Focus Problem regarding crime rate and percentage change in population.

SPSS Display

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.927[a]	.859	.823	22.59076

a. Predictors: (Constant), % change in population

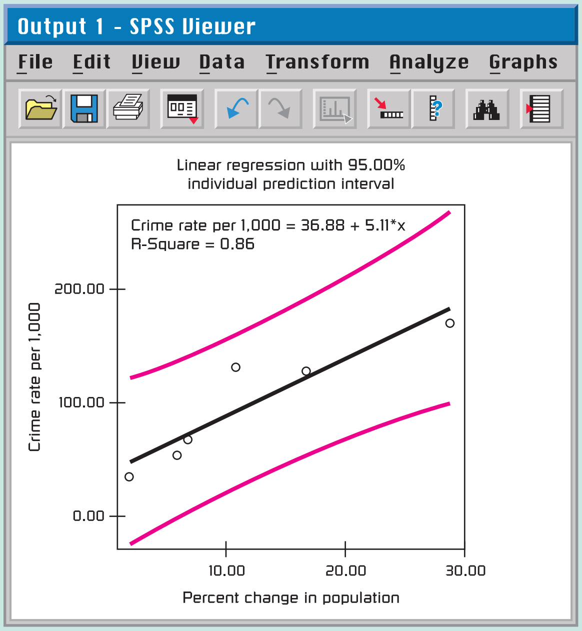
Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	36.881	15.474		2.383	.076
	% change in population	5.107	1.035	.927	4.932	.008

a. Dependent Variable: Crime rate per 1,000

With the menu choices **Graph ➤ Legacy Dialogues ➤ Interactive ➤ Scatterplot,** SPSS produces a scatter diagram with the least-squares line, least-squares equation, coefficient of determination r^2, and optional prediction bands. In the dialogue box, move the dependent variable to the box along the vertical axis and the independent variable to the box along the horizontal axis. Click the "fit" tab, highlight Regression as method, and check the box to include the constant in the equation. For optional prediction band, check individual, enter the confidence level, and check total. The following display shows a scatter diagram for the data in this chapter's Focus Problem regarding crime rate and percentage change in population.

SPSS Display for Focus Problem



Multiple Regression

Application 2

Data values in the following study are taken from *Statistical Abstract of the United States*, U.S. Department of Commerce, 103rd and 109th Editions (see Table 9-16). All data values represent annual averages as determined by the U.S. Department of Commerce.

1. Construct a regression model with

   Response variable: x_3 (foreign investments)

   Explanatory variables: x_5 (GNP), x_6 (U.S. dollar), and x_7 (consumer credit)

What is the coefficient of multiple determination?

(a) Use a 1% level of significance and test each coefficient for significance (two-tailed test).

(b) Examine the coefficients of the regression equation. Then explain why you think the following statement is true or false: "If the purchasing power of the U.S. dollar did not change and the GNP did not change, then an increase in consumer credit would likely be accompanied by a reduction in foreign investments."

(c) Suppose $x_5 = 3500$, $x_6 = 0.975$, and $x_7 = 450$. Predict the level of foreign investment. Find a 90% confidence interval for your prediction.

TABLE 9-16 Economic Data, 1976–1987 (on the data disk)

Year	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1976	10.9	7.61	31	974.9	1718	1.757	234.4
1977	12.0	7.42	35	894.6	1918	1.649	263.8
1978	12.5	8.41	42	820.2	2164	1.532	308.3
1979	17.7	9.44	54	844.4	2418	1.380	347.5
1980	28.1	11.46	83	891.4	2732	1.215	349.4
1981	35.6	13.91	109	932.9	3053	1.098	366.6
1982	31.8	13.00	125	884.4	3166	1.035	381.1
1983	29.0	11.11	137	1190.3	3406	1.000	430.4
1984	28.6	12.44	165	1178.5	3772	0.961	511.8
1985	26.8	10.62	185	1328.2	4015	0.928	592.4
1986	14.6	7.68	209	1792.8	4240	0.913	646.1
1987	17.9	8.38	244	2276.0	4527	0.880	685.5

We will use the following notation:

x_1 = price of a barrel of crude oil, in dollars per barrel

x_2 = percent interest on 10-year U.S. Treasury notes

x_3 = total foreign investments in U.S., in billions of dollars

x_4 = Dow Jones Industrial Average (DJIA)

x_5 = Gross National Product, GNP, in billions of dollars

x_6 = purchasing power of U.S. dollar with base 1983 corresponding to \$1.000

x_7 = consumer credit (i.e., consumer debt), in billions of dollars

2. Construct a new regression model with

   Response variable: x_4 (DJIA)

   Explanatory variables: x_3 (foreign investments), x_5 (GNP), and x_7 (consumer credit)

What is the coefficient of multiple determination?

   (a) Use a 5% level of significance and test each coefficient for significance (two-tailed test).
   (b) Examine the coefficients of the regression equation; then explain why you think the following statement is true or false: "If the GNP and consumer credit didn't change but foreign investments increased, the DJIA would likely show a strong increase."
   (c) Suppose $x_3 = 210$, $x_5 = 4260$, and $x_7 = 650$. Predict the DJIA and find an 85% confidence interval for your prediction.

3. Construct a new regression model with

   Response variable: x_7 (consumer credit)

   Explanatory variables: x_3 (foreign investments), x_5 (GNP), and x_6 (U.S. dollar)

What is the coefficient of multiple determination?

   (a) Use a 1% level of significance and test each coefficient for significance (two-tailed test).
   (b) Examine the coefficients of the regression equation; then explain why you think each of the following statements is true or false: "If both GNP and purchasing power of the U.S. dollar didn't change, then an increase in foreign investments would likely be accompanied by a reduction in consumer credit." "If both foreign investments and purchasing power of the U.S. dollar remained fixed, then an increase in GNP would likely be accompanied by an increase in consumer credit."
   (c) Suppose $x_3 = 88$, $x_5 = 2750$, and $x_6 = 1.250$. Predict consumer credit and find an 80% confidence interval for your prediction.

Technology Hints (Multiple Regression)

TI-84Plus/TI-83Plus/TI-*n*spire

Does not support multiple regression.

Excel 2007

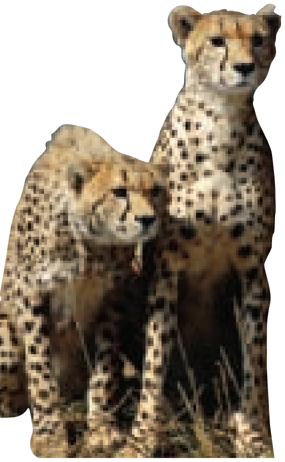
On the home screen, click the **Data** tab. Select **Data Analysis** from the Analysis group. In the dialogue box, select **Regression**. On the spreadsheet, the columns containing the explanatory variables need to be adjacent.

Minitab

Use the menu selection **Stat** ➤ **Regression** ➤ **Regression.**

SPSS

Use the menu selection **Analyze** ➤ **Regression** ➤ **Linear** and select dependent and independent variables.



Cumulative Review Problems

CHAPTERS 7–9

In Problems 1–6, please use the following steps (i) through (v) for all hypothesis tests.

(i) What is the level of significance? State the null and alternate hypotheses.

(ii) *Check Requirements* What sampling distribution will you use? What assumptions are you making? What is the value of the sample test statistic?

(iii) Find (or estimate) the *P*-value. Sketch the sampling distribution and show the area corresponding to the *P*-value.

(iv) Based on your answers in parts (i) to (iii), will you reject or fail to reject the null hypothesis? Are the data statistically significant at level α?

(v) *Interpret* your conclusion in the context of the application.

Note: For degrees of freedom *d.f.* not in the Student's t table, use the closest *d.f.* that is *smaller*. In some situations, this choice of *d.f.* may increase the *P*-value a small amount and thereby produce a slightly more "conservative" answer.

1. *Testing and Estimating μ, σ Known* Let x be a random variable that represents micrograms of lead per liter of water (ug/l). An industrial plant discharges water into a creek. The Environmental Protection Agency has studied the discharged water and found x to have a normal distribution, with $\sigma = 0.7$ ug/l (Reference: *EPA Wetlands Case Studies*).
   (a) The industrial plant says that the population mean value of x is $\mu = 2.0$ ug/l. However, a random sample of $n = 10$ water samples showed that $\bar{x} = 2.56$ ug/l. Does this indicate that the lead concentration population mean is higher than the industrial plant claims? Use $\alpha = 1\%$.
   (b) Find a 95% confidence interval for μ using the sample data and the EPA value for σ.
   (c) How large a sample should be taken to be 95% confident that the sample mean $\bar{x}$ is within a margin of error $E = 0.2$ ug/l of the population mean?

2. *Testing and Estimating μ, σ Unknown* Carboxyhemoglobin is formed when hemoglobin is exposed to carbon monoxide. Heavy smokers tend to have a high percentage of carboxyhemoglobin in their blood (Reference: *Laboratory and Diagnostic Tests* by F. Fishbach). Let x be a random variable representing percentage of carboxyhemoglobin in the blood. For a person who is a regular heavy smoker, x has a distribution that is approximately normal. A random sample of $n = 12$ blood tests given to a heavy smoker gave the following results (percent carboxyhemoglobin in the blood).

9.1	9.5	10.2	9.8	11.3	12.2
11.6	10.3	8.9	9.7	13.4	9.9

   (a) Use a calculator to verify that $\bar{x} \approx 10.49$ and $s \approx 1.36$.
   (b) A long-term population mean $\mu = 10\%$ is considered a health risk. However, a long-term population mean above 10% is considered a clinical alert that the person may be asymptomatic. Do the data indicate that the population mean percentage is higher than 10% for this patient? Use $\alpha = 0.05$.
   (c) Use the given data to find a 99% confidence interval for μ for this patient.

3. *Testing and Estimating a Proportion p* Although older Americans are most afraid of crime, it is young people who are more likely to be the actual victims of crime. It seems that older people are more cautious about the people with whom they associate. A national survey showed that 10% of all people ages 16–19 have been victims of crime (Reference: *Bureau*

of Justice Statistics). At Jefferson High School, a random sample of $n = 68$ students (ages 16–19) showed that $r = 10$ had been victims of a crime.

(a) Do these data indicate that the population proportion of students in this school (ages 16–19) who have been victims of a crime is different (either way) from the national rate for this age group? Use $\alpha = 0.05$. Do you think the conditions $np > 5$ and $nq > 5$ are satisfied in this setting? Why is this important?

(b) Find a 90% confidence interval for the proportion of students in this school (ages 16–19) who have been victims of a crime.

(c) How large a sample size should be used to be 95% sure that the sample proportion $\hat{p}$ is within a margin of error $E = 0.05$ of the population proportion of all students in this school (ages 16–19) who have been victims of a crime? *Hint:* Use sample data $\hat{p}$ as a preliminary estimate for p.

4. ***Testing Paired Differences*** Phosphorous is a chemical that is found in many household cleaning products. Unfortunately, phosphorous also finds its way into surface water, where it can harm fish, plants, and other wildlife. Two methods of phosphorous reduction are being studied. At a random sample of 7 locations, both methods were used and the total phosphorous reduction (mg/l) was recorded (Reference: *Environmental Protection Agency Case Study 832-R-93-005)*.

Site	1	2	3	4	5	6	7
Method I:	0.013	0.030	0.015	0.055	0.007	0.002	0.010
Method II:	0.014	0.058	0.017	0.039	0.017	0.001	0.013

Do these data indicate a difference (either way) in the average reduction of phosphorous between the two methods? Use $\alpha = 0.05$.

5. ***Testing and Estimating $\mu_1 - \mu_2$, σ_1 and σ_2 Unknown*** In the airline business, "on-time" flight arrival is important for connecting flights and general customer satisfaction. Is there a difference between summer and winter average on-time flight arrivals? Let x_1 be a random variable that represents percentage of on-time arrivals at major airports in the summer. Let x_2 be a random variable that represents percentage of on-time arrivals at major airports in the winter. A random sample of $n_1 = 16$ major airports showed that $\bar{x}_1 = 74.8\%$, with $s_1 = 5.2\%$. A random sample of $n_2 = 18$ major airports showed that $\bar{x}_2 = 70.1\%$, with $s_2 = 8.6\%$ (Reference: *Statistical Abstract of the United States*).

(a) Does this information indicate a difference (either way) in the population mean percentage of on-time arrivals for summer compared to winter? Use $\alpha = 0.05$.

(b) Find an 95% confidence interval for $\mu_1 - \mu_2$.

(c) What assumptions about the original populations have you made for the methods used?

6. ***Testing and Estimating a Difference of Proportions $p_1 - p_2$*** How often do you go out dancing? This question was asked by a professional survey group on behalf of the National Arts Survey. A random sample of $n_1 = 95$ single men showed that $r_1 = 23$ went out

Corbis

dancing occasionally. Another random sample of $n_2 = 92$ single women showed that $r_2 = 19$ went out dancing occasionally.

(a) Do these data indicate that the proportion of single men who go out dancing occasionally is higher than the proportion of single women who do so? Use a 5% level of significance. List the assumptions you made in solving this problem. Do you think these assumptions are realistic?

(b) Compute a 90% confidence interval for the population difference of proportions $p_1 - p_2$ of single men and single women who occasionally go out dancing.

7. *Essay and Project* In Chapters 7 and 8 you studied, estimation and hypothesis testing.

(a) Write a brief essay in which you discuss using information from samples to infer information about populations. Be sure to include methods of estimation and hypothesis testing in your discussion. What two sampling distributions are used in estimation and hypothesis testing of population means, proportions, paired differences, differences of means, and differences of proportions? What are the criteria for determining the appropriate sampling distribution? What is the level of significance of a test? What is the P-value? How is the P-value related to the alternate hypothesis? How is the null hypothesis related to the sample test statistic? Explain.

(b) Suppose you want to study the length of time devoted to commercial breaks for two different types of television programs. Identify the types of programs you want to study (e.g., sitcoms, sports events, movies, news, children's programs, etc.). Write a brief outline for your study. Consider whether you will use paired data (such as same time slot on two different channels) or independent samples. Discuss how to obtain random samples. How large should the sample be for a specified margin of error? Describe the protocol you will follow to measure the times of the commercial breaks. Determine whether you are going to compare the average time devoted to commercials or the proportion of time devoted to commercials. What assumptions will you make regarding population distributions? What graphics might be appropriate? What methods of estimation will you use? What methods of testing will you use?

8. *Critical Thinking* Explain hypothesis testing to a friend, using the following scenario as a model. Describe the hypotheses, the sample statistic, the P-value, the meanings of type I and type II errors, and the level of significance. Discuss the significance of the results. Formulas are not required.

   A team of research doctors designed a new knee surgery technique utilizing much smaller incisions than the standard method. They believe recovery times are shorter when the new method is used. Under the old method, the average recovery time for full use of the knee is 4.5 months. A random sample of 38 surgeries using the new method showed the average recovery time to be 3.6 months, with sample standard deviation of 1.7 months. The P-value for the test is 0.0011. The research team states that the results are statistically significant at the 1% level of significance.

9. *Linear Regression: Blood Glucose* Let x be a random variable that represents blood glucose level after a 12-hour fast. Let y be a random variable representing blood glucose level 1 hour after drinking sugar water (after the 12-hour fast). Units are in mg/10 ml. A random sample of eight adults gave the following information (Reference: *American Journal of Clinical Nutrition*, Vol. 19, pp. 345–351).

Jacob Halaska/Photolibrary/Getty Images

$\Sigma x = 63.8$; $\Sigma x^2 = 521.56$; $\Sigma y = 90.7$; $\Sigma y^2 = 1070.87$; $\Sigma xy = 739.65$

x	6.2	8.4	7.0	7.5	8.1	6.9	10.0	9.7
y	9.8	10.7	10.3	11.9	14.2	7.0	14.6	12.2

(a) Draw a scatter diagram for the data.
(b) Find the equation of the least-squares line and graph it on the scatter diagram.
(c) Find the sample correlation coefficient r and the sample coefficient of determination r^2. Explain the meaning of r^2 in the context of the application.
(d) If $x = 9.0$, use the least-squares line to predict y. Find an 80% confidence interval for your prediction.
(e) Use level of significance 1% and test the claim that the population correlation coefficient ρ is not zero. Interpret the results.
(f) Find an 85% confidence interval for the slope β of the population-based least-squares line. Explain its meaning in the context of the application.

10

John Bryson/Time Life Pictures/Getty Images

Norman Rockwell

"So what!"

—ANONYMOUS

We have all heard the exclamation, "So what!" Philologists (people who study cultural linguistics) tell us that this expression is a shortened version of "So what is the difference!" They also tell us that there are similar popular or slang expressions about differences in all languages and cultures. Norman Rockwell (1894–1978) painted everyday people and situations. In this cover, "Girl with Black Eye," for the *Saturday Evening Post* (May 23, 1953), a young lady is about to have a conference with her school principal. So what! It is human nature to challenge the claim that something is better, worse, or just simply different. In this chapter, we will focus on this very human theme by studying a variety of topics regarding questions of whether or not differences exist between two population variances or among several population means.

For online student resources, visit the Brase/Brase, *Understandable Statistics,* 10th edition web site at **http://www.cengage.com/statistics/brase**.

CHI-SQUARE AND *F* DISTRIBUTIONS

How do you decide if random variables are dependent or independent? (SECTION 10.1)

How do you decide if different populations share the same proportions of specified characteristics? (SECTION 10.1)

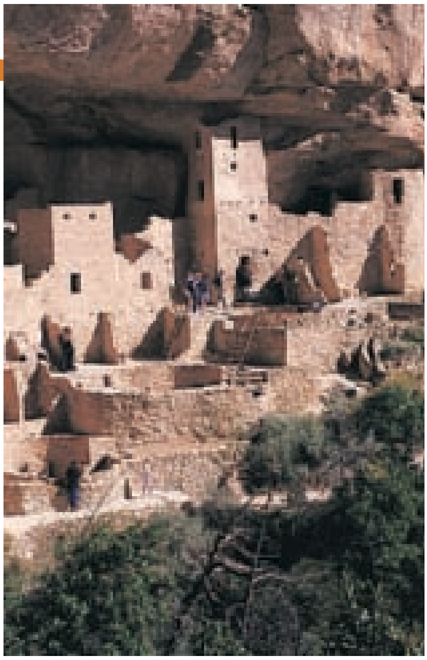
How do you decide if two distributions are not only dependent, but actually the same distribution? (SECTION 10.2)

How do you compute confidence intervals and tests for σ? (SECTION 10.3)

How do you test two variances σ_1^2 and σ_2^2 ? (SECTION 10.4)

What is one-way ANOVA? Where is it used? (SECTION 10.5)

What about two-way ANOVA? Where is it used? (SECTION 10.6)



The Image Works

Mesa Verde National Park

FOCUS PROBLEM

Archaeology in Bandelier National Monument

Archaeologists at Washington State University did an extensive summer excavation at Burnt Mesa Pueblo in Bandelier National Monument. Their work is published in the book *Bandelier Archaeological Excavation Project: Summer 1990 Excavations at Burnt Mesa Pueblo and Casa del Rito,* edited by T. A. Kohler.

One question the archaeologists asked was: Is raw material used by prehistoric Indians for stone tool manufacture independent of the archaeological excavation site? Two different excavation sites at Burnt Mesa Pueblo gave the information in the table below.

Use a chi-square test with 5% level of significance to test the claim that raw material used for construction of stone tools and excavation site are independent. (See Problem 17 of Section 10.1.)

Stone Tool Construction Material, Burnt Mesa Pueblo

Material	Site A	Site B	Row Total
Basalt	731	584	1315
Obsidian	102	93	195
Pedernal chert	510	525	1035
Other	85	94	179
Column Total	**1428**	**1296**	**2724**

Robert Brenner/PhotoEdit

Archaeological excavation site

PART I: INFERENCES USING THE CHI-SQUARE DISTRIBUTION

Overview of the Chi-Square Distribution

So far, we have used several probability distributions for hypothesis testing and confidence intervals, with the most frequently used being the normal distribution and the Student's t distribution. In this chapter, we will use two other probability distributions, namely, the chi-square distribution (where *chi* is pronounced like the first two letters in the word *kite*) and the F distribution. In Part I, we will see applications of the chi-square distribution, whereas in Part II, we will see some important applications of the F distribution.

Chi is a Greek letter denoted by the symbol χ, so chi-square is denoted by the symbol χ^2. Because the distribution is of chi-*square* values, the χ^2 values begin at 0 and then are all positive. The graph of the χ^2 distribution is not symmetrical, and like the Student's t distribution, it depends on the number of degrees of freedom. Figure 10-1 shows χ^2 distributions for several degrees of freedom (*d.f.*).

As the degrees of freedom increase, the graph of the chi-square distribution becomes more bell-like and begins to look more and more symmetric.

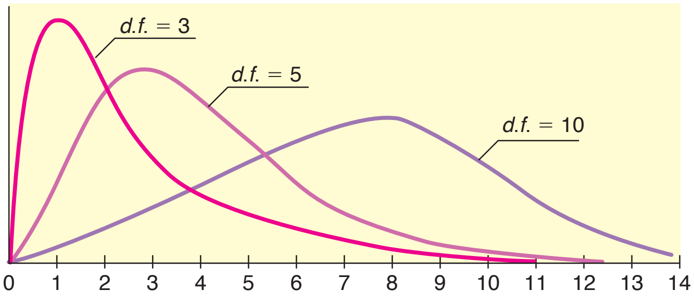
The **mode (high point)** of a chi-square distribution with n degrees of freedom occurs over $n - 2$ (for $n \geq 3$).

Table 7 of Appendix II shows critical values of chi-square distributions for which a designated area falls to the *right* of the critical value. Table 10-1 gives an excerpt from Table 7. Notice that the row headers are degrees of freedom, and the column headers are areas in the *right* tail of the distribution. For instance, according to the table, for a χ^2 distribution with 3 degrees of freedom, the area occurring to the *right* of $\chi^2 = 0.072$ is 0.995. For a χ^2 distribution with 4 degrees of freedom, the area falling to the *right* of $\chi^2 = 13.28$ is 0.010.

In the next three sections, we will see how to apply the chi-square distribution to different applications.

FIGURE 10-1

The χ^2 Distribution



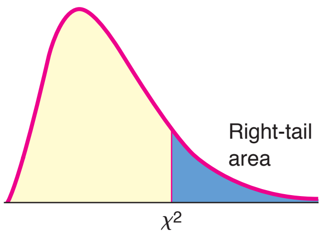


TABLE 10-1 **Excerpt from Table 7 (Appendix II): The χ^2 Distribution**

	Area of the Right Tail					
d.f.	0.995	0.990	0.975	. . .	0.010	0.005
⋮	⋮	⋮	⋮		⋮	⋮
3	0.072	0.115	0.216		11.34	12.84
4	0.207	0.297	0.484		13.28	14.86

SECTION 10.1

Chi-Square: Tests of Independence and of Homogeneity

FOCUS POINTS

- Set up a test to investigate independence of random variables.
- Use contingency tables to compute the sample χ^2 statistic.
- Find or estimate the *P*-value of the sample χ^2 statistic and complete the test.
- Conduct a test of homogeneity of populations.

Innovative Machines Incorporated has developed two new letter arrangements for computer keyboards. The company wishes to see if there is any relationship between the arrangement of letters on the keyboard and the number of hours it takes a new typing student to learn to type at 20 words per minute. Or, from another point of view, is the time it takes a student to learn to type *independent* of the arrangement of the letters on a keyboard?

To answer questions of this type, we test the hypotheses

Hypotheses

H_0: Keyboard arrangement and learning times *are independent*.
H_1: Keyboard arrangement and learning times *are not independent*.

Test of independence

Chi-square distribution

In problems of this sort, we are testing the *independence* of two factors. The probability distribution we use to make the decision is the *chi-square distribution*. Recall from the overview of the chi-square distribution that *chi* is pronounced like the first two letters of the word *kite* and is a Greek letter denoted by the symbol χ. Thus, chi-square is denoted by χ^2.

Innovative Machines' first task is to gather data. Suppose the company took a random sample of 300 beginning typing students and randomly assigned them to learn to type on one of three keyboards. The learning times for this sample are shown in Table 10-2. These learning times are the observed frequencies O.

Observed frequency O

Contingency table

Table 10-2 is called a *contingency table*. The *shaded boxes* that contain observed frequencies are called *cells*. The row and column totals are not considered to be cells. This contingency table is of size 3 × 3 (read "three-by-three") because there are three rows of cells and three columns. When giving the size of a contingency table, we always list the number of *rows first*.

To determine the **size** of a contingency table, count the number of rows containing data and the number of columns containing data. The size is

Number of rows × Number of columns

where the symbol "×" is read "by." The number of rows is always given first.

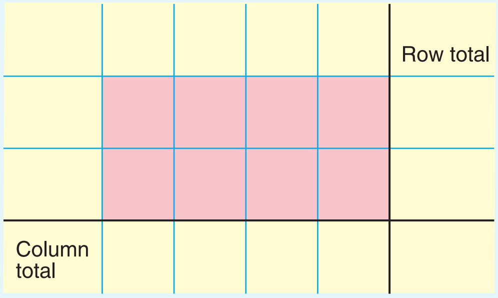
TABLE 10-2 **Keyboard versus Time to Learn to Type at 20 wpm**

Keyboard	21–40 h	41–60 h	61–80 h	Row Total
A	#1 25	#2 30	#3 25	80
B	#4 30	#5 71	#6 19	120
Standard	#7 35	#8 49	#9 16	100
Column Total	90	150	60	300 Sample size

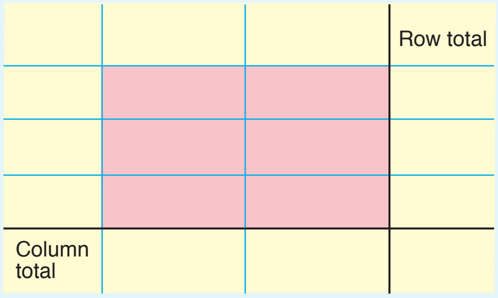
GUIDED EXERCISE 1 Size of contingency table

Give the sizes of the contingency tables in Figures 10-2(a) and (b). Also, count the number of cells in each table. (Remember, each pink shaded box is a cell.)

(a) FIGURE 10-2(a) Contingency Table



(b) FIGURE 10-2(b) Contingency Table



(a) There are two rows and four columns, so this is a 2×4 table. There are eight cells.

(b) Here we have three rows and two columns, so this is a 3×2 table with six cells.

Expected frequency E

We are testing the null hypothesis that the keyboard arrangement and the time it takes a student to learn to type are *independent*. We use this hypothesis to determine the *expected frequency* of each cell.

For instance, to compute the expected frequency of cell 1 in Table 10-2, we observe that cell 1 consists of all the students in the sample who learned to type on keyboard A and who mastered the skill at the 20-words-per-minute level in 21 to 40 hours. By the assumption (null hypothesis) that the two events are independent, we use the multiplication law to obtain the probability that a student is in cell 1.

$$\begin{aligned} P(\text{cell } 1) &= P(\text{keyboard A } \textit{and} \text{ skill in } 21-40 \text{ h}) \\ &= P(\text{keyboard A}) \cdot P(\text{skill in } 21-40 \text{ h}) \end{aligned}$$

Because there are 300 students in the sample and 80 used keyboard A,

$$P(\text{keyboard A}) = \frac{80}{300}$$

Also, 90 of the 300 students learned to type in 21–40 hours, so

$$P(\text{skill in } 21-40 \text{ h}) = \frac{90}{300}$$

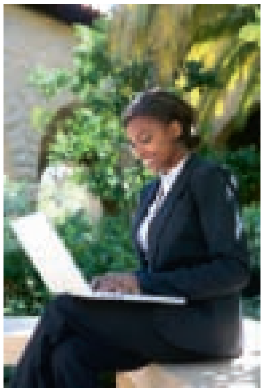
Using these two probabilities and the assumption of independence,

$$P(\text{keyboard A } \textit{and} \text{ skill in } 21-40 \text{ h}) = \frac{80}{300} \cdot \frac{90}{300}$$

Finally, because there are 300 students in the sample, we have the *expected frequency E* for cell 1.

$$\begin{aligned} E &= P(\text{student in cell } 1) \cdot (\text{no. of students in sample}) \\ &= \frac{80}{300} \cdot \frac{90}{300} \cdot 300 = \frac{80 \cdot 90}{300} = 24 \end{aligned}$$

We can repeat this process for each cell. However, the last step yields an easier formula for the expected frequency *E*.



Stephen Coburn, 2010/Used under license from Shutterstock.com

Row total
Column total

Formula for expected frequency E

$$E = \frac{\text{(Row total)(Column total)}}{\text{Sample size}}$$

Note: If the expected value is not a whole number, do *not* round it to the nearest whole number.

Let's use this formula in Example 1 to find the expected frequency for cell 2.

EXAMPLE 1 **Expected frequency**

Find the expected frequency for cell 2 of contingency Table 10-2.

SOLUTION: Cell 2 is in row 1 and column 2. The *row total* is 80, and the *column total* is 150. The size of the sample is still 300.

$$E = \frac{\text{(Row total)(Column total)}}{\text{Sample size}}$$

$$= \frac{(80)(150)}{300} = 40$$

GUIDED EXERCISE 2 **Expected frequency**

Table 10-3 contains the *observed frequencies* O and *expected frequencies* E for the contingency table giving keyboard arrangement and number of hours it takes a student to learn to type at 20 words per minute. Fill in the missing expected frequencies.

TABLE 10-3 **Complete Contingency Table of Keyboard Arrangement and Time to Learn to Type**

Keyboard	21–40 h	41–60 h	61–80 h	Row Total
A	#1 $O = 25$ $E = 24$	#2 $O = 30$ $E = 40$	#3 $O = 25$ $E = $ ___	80
B	#4 $O = 30$ $E = 36$	#5 $O = 71$ $E = $ ___	#6 $O = 19$ $E = $ ___	120
Standard	#7 $O = 35$ $E = $ ___	#8 $O = 49$ $E = 50$	#9 $O = 16$ $E = 20$	100
Column Total	90	150	60	300 Sample Size

For cell 3, we have

$$E = \frac{(80)(60)}{300} = 16$$

For cell 5, we have

$$E = \frac{(120)(150)}{300} = 60$$

For cell 6, we have

$$E = \frac{(120)(60)}{300} = 24$$

For cell 7, we have

$$E = \frac{(100)(90)}{300} = 30$$

Computing the sample test statistic χ^2

Now we are ready to compute the sample statistic χ^2 for the typing students. The χ^2 value is a measure of the sum of the differences between *observed frequency* O and *expected frequency* E in each cell. These differences are listed in Table 10-4.

As you can see, if we sum the differences between the observed frequencies and the expected frequencies of the cells, we get the value zero. This total certainly does not reflect the fact that there were differences between the observed and expected frequencies. To obtain a measure whose sum does reflect the magnitude of the differences, we square the differences and work with the quantities $(O - E)^2$. But instead of using the terms $(O - E)^2$, we use the values $(O - E)^2/E$.

TABLE 10-4 Differences Between Observed and Expected Frequencies

Cell	Observed O	Expected E	Difference $(O - E)$
1	25	24	1
2	30	40	−10
3	25	16	9
4	30	36	−6
5	71	60	11
6	19	24	−5
7	35	30	5
8	49	50	−1
9	16	20	−4
			$\Sigma(O - E) - 0$

We use this expression because a small difference between the observed and expected frequencies is not nearly as important when the expected frequency is large as it is when the expected frequency is small. For instance, for both cells 1 and 8, the squared difference $(O - E)^2$ is 1. However, this difference is more meaningful in cell 1, where the expected frequency is 24, than it is in cell 8, where the expected frequency is 50. When we divide the quantity $(O - E)^2$ by E, we take the size of the difference with respect to the size of the expected value. We use the sum of these values to form the sample statistic χ^2:

Sample test statistic χ^2

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

where the sum is over all cells in the contingency table.

COMMENT If you look up the word *irony* in a dictionary, you will find one of its meanings to be "the difference between actual (or observed) results and expected results." Because irony is so prevalent in much of our human experience, it is not surprising that statisticians have incorporated a related chi-square distribution into their work.

GUIDED EXERCISE 3 Sample χ^2

(a) Complete Table 10-5.

TABLE 10-5 Data of Table 10-4

Cell	O	E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
1	25	24	1	1	0.04
2	30	40	−10	100	2.50
3	25	16	9	81	5.06
4	30	36	−6	36	1.00
5	71	60	11	121	2.02
6	19	24	−5	25	1.04
7	35	30	5	25	0.83
8	49	50	____	____	____
9	16	20	____	____	____

$$\Sigma \frac{(O - E)^2}{E} = \_\_\_\_$$

The last two rows of Table 10-5 are

Cell	O	E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
8	49	50	−1	1	0.02
9	16	20	−4	16	0.80

$\Sigma \frac{(O - E)^2}{E}$ = total of last column = 13.31

(b) Compute the statistic χ^2 for this sample.

Since $\chi^2 = \Sigma \frac{(O - E)^2}{E}$, then $\chi^2 = 13.31$.

Notice that when the observed frequency and the expected frequency are very close, the quantity $(O - E)^2$ is close to zero, and so the statistic χ^2 is near zero. As the difference increases, the statistic χ^2 also increases. To determine how large the sample statistic can be before we must reject the null hypothesis of independence, we find the P-value of the statistic in the chi-square distribution, Table 7 of Appendix II, and compare it to the specified level of significance α. The P-value depends on the number of degrees of freedom. To test independence, the degrees of freedom $d.f.$ are determined by the following formula.

Degrees of freedom for test of independence

Degrees of freedom for test of independence

$$\text{Degrees of freedom} = (\text{Number of rows} - 1) \cdot (\text{Number of columns} - 1)$$

or $$d.f. = (R - 1)(C - 1)$$

where R = number of cell rows

C = number of cell columns

GUIDED EXERCISE 4 Degrees of freedom

Determine the number of degrees of freedom in the example of keyboard arrangements (see Table 10-2). Recall that the contingency table had three rows and three columns.

$$d.f. = (R - 1)(C - 1)$$
$$= (3 - 1)(3 - 1) = (2)(2) = 4$$

Finding the *P*-value for tests of independence

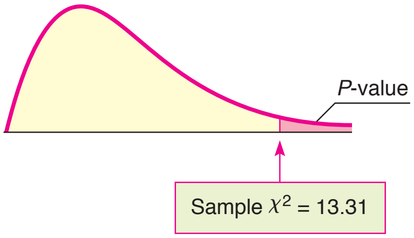
To test the hypothesis that the letter arrangement on a keyboard and the time it takes to learn to type at 20 words per minute are independent at the $\alpha = 0.05$ level of significance, we estimate the P-value shown in Figure 10-3 below for the sample test statistic $\chi^2 = 13.31$ (calculated in Guided Exercise 3). We then compare the P-value to the specified level of significance α.

For tests of independence, we always use a *right-tailed* test on the chi-square distribution. This is because we are testing to see if the χ^2 measure of the difference between the observed and expected frequencies is too large to be due to chance alone.

In Guided Exercise 4, we found that the degrees of freedom for the example of keyboard arrangements is 4. From Table 7 of Appendix II, in the row headed by $d.f. = 4$, we see that the sample $\chi^2 = 13.31$ falls between the entries 13.28 and 14.86.

FIGURE 10-3

P-value



Right-tail Area	0.010	0.005
$d.f. = 4$	13.28	14.86
	Sample $\chi^2 = 13.31$	

The corresponding P-value falls between 0.005 and 0.010. From technology, we get P-value ≈ 0.0098.

Since the P-value is less than the level of significance $\alpha = 0.05$, we reject the null hypothesis of independence and conclude that keyboard arrangement and learning time are *not* independent.

Tests of independence for two statistical variables involve a number of steps. A summary of the procedure follows.

PROCEDURE

HOW TO TEST FOR INDEPENDENCE OF TWO STATISTICAL VARIABLES

Setup

Construct a contingency table in which the rows represent one statistical variable and the columns represent the other. Obtain a random sample of observations, which are assigned to the cells described by the rows and columns. These assignments are called the **observed values *O*** from the sample.

Procedure

1. Set the level of significance α and use the hypotheses

   H_0: The variables are independent.

   H_1: The variables are not independent.

2. For each cell, compute the **expected frequency *E*** (do not round but give as a decimal number).

$$E = \frac{(\text{Row total})(\text{Column total})}{\text{Sample size}}$$

Requirement

You need a sample size large enough so that, for each cell, $E \geq 5$.

Now each cell has two numbers, the observed frequency O from the sample and the expected frequency E.

Next, compute the sample *chi-square test statistic*

$$\chi^2 = \Sigma \frac{(O - E)^2}{E} \quad \text{with degrees of freedom } d.f. = (R - 1)(C - 1)$$

where the sum is over all cells in the contingency table and

R = number of rows in contingency table

C = number of columns in contingency table

3. Use the chi-square distribution (Table 7 of Appendix II) and a *right-tailed test* to find (or estimate) the *P-value* corresponding to the test statistic.

4. *Conclude* the test. If P-value $\leq \alpha$, then reject H_0. If P-value $> \alpha$, then do not reject H_0.

5. *Interpret your conclusion* in the context of the application.

GUIDED EXERCISE 5 Testing independence

Super Vending Machines Company is to install soda pop machines in elementary schools and high schools. The market analysts wish to know if flavor preference and school level are independent. A random sample of 200 students was taken. Their school level and soda pop preferences are given in Table 10-6. Is independence indicated at the $\alpha = 0.01$ level of significance?

STEP 1: State the null and alternate hypotheses.

H_0: School level and soda pop preference are independent.

H_1: School level and soda pop preference are not independent.

STEP 2:

(a) Complete the contingency Table 10-6 by filling in the required expected frequencies.

Table 10-6 School Level and Soda Pop Preference

Soda Pop	High School	Elementary School	Row Total
Kula Kola	O = 33 #1 E = 36	O = 57 #2 E = 54	90
Mountain Mist	O = 30 #3 E = 20	O = 20 #4 E = 30	50
Jungle Grape	O = 5 #5 E = _____	O = 35 #6 E = _____	40
Diet Pop	O = 12 #7 E = _____	O = 8 #8 E = _____	20
Column Total	80	120	200 Sample Size

The expected frequency

for cell 5 is $\dfrac{(40)(80)}{200} = 16$

for cell 6 is $\dfrac{(40)(120)}{200} = 24$

for cell 7 is $\dfrac{(20)(80)}{200} = 8$

for cell 8 is $\dfrac{(20)(120)}{200} = 12$

Note: In this example, the expected frequencies are all whole numbers. If the expected frequency has a decimal part, such as 8.45, do *not* round the value to the nearest whole number; rather, give the expected frequency as the decimal number.

(b) Fill in Table 10-7 and use the table to find the sample statistic χ^2.

Table 10-7 Computational Table for χ^2

Cell	O	E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
1	33	36	−3	9	0.25
2	57	54	3	9	0.17
3	30	20	10	100	5.00
4	20	30	−10	100	3.33
5	5	16	−11	121	7.56
6	35	24	11	_____	_____
7	12	8	_____	_____	_____
8	8	12	_____	_____	_____

The last three rows of Table 10-7 should read as follows:

Cell	O	E	O − E	$(O - E)^2$	$(O - E)^2/E$
6	35	24	11	121	5.04
7	12	8	4	16	2.00
8	8	12	−4	16	1.33

χ^2 = total of last column

$$= \Sigma \frac{(O - E)^2}{E} = 24.68$$

(c) What is the size of the contingency table? Use the number of rows and the number of columns to determine the degrees of freedom.

The contingency table is of size 4 × 2. Since there are four rows and two columns,

$d.f. = (4 - 1)(2 - 1) = 3$

Continued

GUIDED EXERCISE 5 *continued*

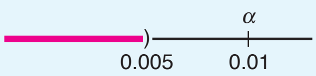
STEP 3: Use Table 7 of Appendix II to estimate the P-value of the sample statistic $\chi^2 = 24.68$ with $d.f. = 3$.

Right-tail Area	0.005
$d.f. = 3$	12.84 ↑ Sample $\chi^2 = 24.68$

As the χ^2 values increase, the area to the right decreases, so

P-value < 0.005

STEP 4: Conclude the test by comparing the P-value of the sample statistic to the level of significance $\alpha = 0.01$.



Since the P-value is less than α, we reject the null hypothesis of independence. Technology gives P-value ≈ 0.00002.

STEP 5: *Interpret* the test result in the context of the application.

At the 1% level of significance, we conclude that school level and soda pop preference are dependent.

TECH NOTES

The TI-84Plus/TI-83Plus/TI-*n*spire calculators, Excel 2007, and Minitab all support chi-square tests of independence. In each case, the observed data are entered in the format of the contingency table.

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) Enter the observed data into a matrix. Set the dimension of matrix [**B**] to match that of the matrix of observed values. Expected values will be placed in matrix [**B**]. Press **STAT, TESTS,** and select option **C:χ^2-Test.** The output gives the sample χ^2 with the P-value.

Excel 2007 Enter the table of observed values. Use the formulas of this section to compute the expected values. Enter the corresponding table of expected values. Finally, click **insert function** (f_x). Select **Statistical** for the category and **CHITEST** for the function. Excel returns the P-value of the sample χ^2 value.

Minitab Enter the contingency table of observed values. Use the menu selection **Stat ➤ Tables ➤ Chi-Square Test.** The output shows the contingency table with expected values and the sample χ^2 with P-value.

Tests of Homogeneity

We've seen how to use contingency tables and the chi-square distribution to test for independence of two random variables. The same process enables us to determine whether several populations share the same proportions of distinct categories. Such a test is called a *test of homogeneity*.

According to the dictionary, among the definitions of the word *homogeneous* are "of the same structure" and "composed of similar parts." In statistical jargon, this translates as a test of homogeneity to see if two or more populations share specified characteristics in the same proportions.

Test of homogeneity

A **test of homogeneity** tests the claim that *different populations* share the *same proportions* of specified characteristics.

The computational processes for conducting tests of independence and tests of homogeneity are the same. However, there are two main differences in the initial setup of the two types of tests, namely, the sampling method and the hypotheses.

Tests of independence compared to tests of homogeneity

1. **Sampling method**
   For tests of independence, we use one random sample and observe how the sample members are distributed among distinct categories.
   For tests of homogeneity, we take random samples from each different population and see how members of each population are distributed over distinct categories.

2. **Hypotheses**
   For tests of independence,

   H_0: The variables are independent.
   H_1: The variables are not independent.

   For tests of homogeneity,

   H_0: Each population shares respective characteristics in the same proportion.
   H_1: Some populations have different proportions of respective characteristics.

EXAMPLE 2 TEST OF HOMOGENEITY

Pets—who can resist a cute kitten or puppy? Tim is doing a research project involving pet preferences among students at his college. He took random samples of 300 female and 250 male students. Each sample member responded to the survey question "If you could own only one pet, what kind would you choose?" The possible responses were: "dog," "cat," "other pet," "no pet." The results of the study follow.

Image Source/Getty Images

Pet Preference

Gender	Dog	Cat	Other Pet	No Pet
Female	120	132	18	30
Male	135	70	20	25

Does the same proportion of males as females prefer each type of pet? Use a 1% level of significance.

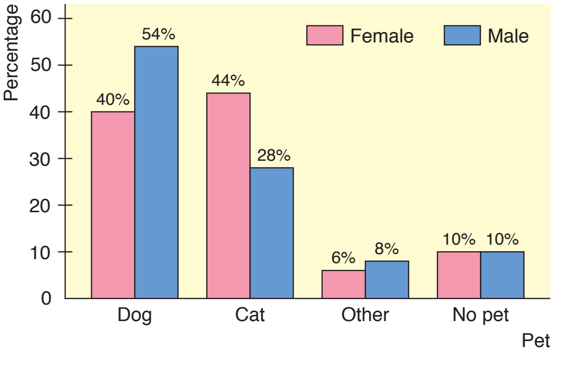
We'll answer this question in several steps.

(a) First make a cluster bar graph showing the percentages of females and the percentages of males favoring each category of pet. From the graph, does it appear that the proportions are the same for males and females?

SOLUTION: The cluster graph shown in Figure 10-4 was created using Minitab. Looking at the graph, it appears that there are differences in the proportions of females and males preferring each type of pet. However, let's conduct a statistical test to verify our visual impression.

FIGURE 10-4

Pet Preference by Gender



(b) Is it appropriate to use a test of homogeneity?

SOLUTION: Yes, since there are separate random samples for each designated population, male and female. We also are interested in whether each population shares the same proportion of members favoring each category of pet.

(c) State the hypotheses and conclude the test by using the Minitab printout.

H_0: The proportions of females and males naming each pet preference are the same.

H_1: The proportions of females and males naming each pet preference are not the same.

```
Chi-Square Test: Dog, Cat, Other, No Pet
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts
              Dog         Cat       Other      No Pet      Total
1             120         132          18          30        300
           139.09      110.18       20.73       30.00
            2.620       4.320       0.359       0.000

2             135          70          20          25        250
           115.91       91.82       17.27       25.00
            3.144       5.185       0.431       0.000

Total         255         202          38          55        550
Chi-Sq = 16.059, DF = 3, P-Value = 0.001
```

Since the P-value is less than α, we reject H_0 at the 1% level of significance.

(d) *Interpret* the results.

SOLUTION: It appears from the sample data that male and female students at Tim's college have different preferences when it comes to selecting a pet.

PROCEDURE

HOW TO TEST FOR HOMOGENEITY OF POPULATIONS

Setup

Obtain random samples from each of the populations. For each population, determine the number of members that share a distinct specified characteristic. Make a contingency table with the different populations as the rows (or columns) and the characteristics as the columns (or rows). The values recorded in the cells of the table are the **observed values *O*** taken from the samples.

Procedure

1. Set the level of significance and use the hypotheses

   H_0: The proportion of each population sharing specified characteristics is the same for all populations.

   H_1: The proportion of each population sharing specified characteristics is not the same for all populations.

2. Follow steps 2–5 of the procedure used to test for independence.

It is important to observe that when we reject the null hypothesis in a test of homogeneity, we don't know which proportions differ among the populations. We know only that the populations differ in some of the proportions sharing a characteristic.

Multinomial Experiments (Optional Reading)

Here are some observations that may be considered "brain teasers." In Chapters 6, 7, and 8, you studied normal approximations to binomial experiments. This concept resulted in some important statistical applications. Is it possible to extend this idea and obtain even more applications? Well, read on!

Consider a *binomial experiment* with n trials. The probability of success on each trial is p, and the probability of failure is $q = 1 - p$. If r is the number of successes out of n trials, then, from Chapter 5, you know that

$$P(r) = \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

The binomial setting has just two outcomes: success or failure. What if you want to consider more than just two outcomes on each trial (for instance, the outcomes shown in a contingency table)? Well, you need a new statistical tool.

Multinomial experiments

Consider a *multinomial experiment*. This means that

1. The trials are independent and repeated under identical conditions.
2. The outcome on each trial falls into exactly one of $k \geq 2$ categories or cells.
3. The probability that the outcome of a single trial will fall into the ith category or cell is p_i (where $i = 1, 2, \ldots, k$) and remains the same for each trial. Furthermore, $p_1 + p_2 + \cdots + p_k = 1$.
4. Let r_i be a random variable that represents the number of trials in which the outcome falls into category or cell i. If you have n trials, then $r_1 + r_2 + \cdots + r_k = n$. The multinomial probability distribution is then

$$P(r_1, r_2, \ldots r_k) = \frac{n!}{r_1! r_2! \ldots r_k!} p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k}$$

How are the multinomial distribution and the binomial distribution related? For the special case $k = 2$, we use the notation $r_1 = r$, $r_2 = n - r$, $p_1 = p$, and $p_2 = q$. In this special case, the multinomial distribution becomes the binomial distribution.

There are two important tests regarding the cell probabilities of a multinomial distribution.

I. **Test of Independence** (Section 10.1)
In this test, the null hypothesis of independence claims that each cell probability p_i will equal the product of its respective row and column probabilities. The alternate hypothesis claims that this is not so.

II. **Goodness-of-Fit Test** (Section 10.2)
In this test, the null hypothesis claims that each category or cell probability p_i will equal a prespecified value. The alternate hypothesis claims that this is not so.

So why don't we use the multinomial probability distribution in Sections 10.1 and 10.2? The reason is that the exact calculation of probabilities associated with type I errors using the multinomial distribution is very tedious and cumbersome. Fortunately, the British statistician Karl Pearson discovered that the chi-square distribution can be used for this purpose, provided the expected value of each cell or category is at least 5.

It is Pearson's chi-square methods that are presented in Sections 10.1 and 10.2. In a sense, you have seen a similar application to statistical tests in Section 8.3, where you used the normal approximation to the binomial when np, the expected number of successes, and nq, the expected number of failures, were both at least 5.

VIEWPOINT Loyalty! Going, Going, Gone!

Was there a time in the past when people worked for the same company all their lives, regularly purchased the same brand names, always voted for candidates from the same political party, and loyally cheered for the same sports team? One way to look at this question is to consider tests of statistical independence. Is customer loyalty independent of company profits? Can a company maintain its productivity independent of loyal workers? Can politicians do whatever they please independent of the voters back home? Americans may be ready to act on a pent-up desire to restore a sense of loyalty in their lives. For more information, see American Demographics*, Vol. 19, No. 9.*

SECTION 10.1 PROBLEMS

1. *Statistical Literacy* In general, are chi-square distributions symmetric or skewed? If skewed, are they skewed right or left?

2. *Statistical Literacy* For chi-square distributions, as the number of degrees of freedom increases, does any skewness increase or decrease? Do chi-square distributions become more symmetric (and normal) as the number of degrees of freedom becomes larger and larger?

3. *Statistical Literacy* For chi-square tests of independence and of homogeneity, do we use a right-tailed, left-tailed, or two-tailed test?

4. *Critical Thinking* In general, how do the hypotheses for chi-square tests of independence differ from those for chi-square tests of homogeneity? Explain.

5. *Critical Thinking* Zane is interested in the proportion of people who recycle each of three distinct products: paper, plastic, electronics. He wants to test the hypothesis that the proportion of people recycling each type of product differs by age group: 12–18 years old, 19–30 years old, 31–40 years old, over 40 years old. Describe the sampling method appropriate for a test of homogeneity regarding recycled products and age.

6. *Critical Thinking* Charlotte is doing a study on fraud and identity theft based both on source (checks, credit cards, debit cards, online banking/finance sites, other) and on gender of the victim. Describe the sampling method appropriate for a test of independence regarding source of fraud and gender.

7. *Interpretation: Test of Homogeneity* Consider Zane's study regarding products recycled and age group (see Problem 5). Suppose he found a sample $\chi^2 = 16.83$.
   (a) How many degrees of freedom are used? Recall that there were 4 age groups and 3 products specified. Approximate the *P*-value and conclude the test at the 1% level of significance. Does it appear that the proportion of people who recycle each of the specified products differ by age group? Explain.
   (b) From this study, can Zane identify how the different age groups differ regarding the proportion of those recycling the specified product? Explain.

8. *Interpretation: Test of Independence* Consider Charlotte's study of source of fraud/identity theft and gender (see Problem 6). She computed sample $\chi^2 = 10.2$.
   (a) How many degrees of freedom are used? Recall that there were 5 sources of fraud/identity theft and, of course, 2 genders. Approximate the *P*-value and conclude the test at the 5% level of significance. Would it seem that gender and source of fraud/identity theft are independent?
   (b) From this study, can Charlotte identify which source of fraud/identity theft is dependent with respect to gender? Explain.

For Problems 9–19, please provide the following information.
(a) What is the level of significance? State the null and alternate hypotheses.
(b) Find the value of the chi-square statistic for the sample. Are all the expected frequencies greater than 5? What sampling distribution will you use? What are the degrees of freedom?

(c) Find or estimate the P-value of the sample test statistic.

(d) Based on your answers in parts (a) to (c), will you reject or fail to reject the null hypothesis of independence?

(e) ***Interpret*** your conclusion in the context of the application.

Use the expected values E to the hundredths place.

9. ***Psychology: Myers–Briggs*** The following table shows the Myers–Briggs personality preferences for a random sample of 406 people in the listed professions (*Atlas of Type Tables* by Macdaid, McCaulley, and Kainz). E refers to extroverted and I refers to introverted.

Occupation	Personality Preference Type E	I	Row Total
Clergy (all denominations)	62	45	107
M.D.	68	94	162
Lawyer	56	81	137
Column Total	186	220	406

Use the chi-square test to determine if the listed occupations and personality preferences are independent at the 0.05 level of significance.

10. ***Psychology: Myers–Briggs*** The following table shows the Myers–Briggs personality preferences for a random sample of 519 people in the listed professions (*Atlas of Type Tables* by Macdaid, McCaulley, and Kainz). T refers to thinking and F refers to feeling.

Occupation	Personality Preference Type T	F	Row Total
Clergy (all denominations)	57	91	148
M.D.	77	82	159
Lawyer	118	94	212
Column Total	252	267	519

Use the chi-square test to determine if the listed occupations and personality preferences are independent at the 0.01 level of significance.

11. ***Archaeology: Pottery*** The following table shows site type and type of pottery for a random sample of 628 sherds at a location in Sand Canyon Archaeological Project, Colorado (*The Sand Canyon Archaeological Project,* edited by Lipe).

Site Type	Pottery Type: Mesa Verde Black-on-White	McElmo Black-on-White	Mancos Black-on-White	Row Total Black-on-White
Mesa Top	75	61	53	189
Cliff-Talus	81	70	62	213
Canyon Bench	92	68	66	226
Column Total	248	199	181	628

Use a chi-square test to determine if site type and pottery type are independent at the 0.01 level of significance.

12. ***Archaeology: Pottery*** The following table shows ceremonial ranking and type of pottery sherd for a random sample of 434 sherds at a location in the Sand Canyon Archaeological Project, Colorado (*The Architecture of Social Integration in Prehistoric Pueblos,* edited by Lipe and Hegmon).

Use a chi-square test to determine if ceremonial ranking and pottery type are independent at the 0.05 level of significance.

Ceremonial Ranking	Cooking Jar Sherds	Decorated Jar Sherds (Noncooking)	Row Total
A	86	49	135
B	92	53	145
C	79	75	154
Column Total	257	177	434

13. *Ecology: Buffalo* The following table shows age distribution and location of a random sample of 166 buffalo in Yellowstone National Park (based on information from *The Bison of Yellowstone National Park,* National Park Service Scientific Monograph Series).

Age	Lamar District	Nez Perce District	Firehole District	Row Total
Calf	13	13	15	41
Yearling	10	11	12	33
Adult	34	28	30	92
Column Total	57	52	57	166

Use a chi-square test to determine if age distribution and location are independent at the 0.05 level of significance.

14. *Psychology: Myers–Briggs* The following table shows the Myers–Briggs personality preference and area of study for a random sample of 519 college students (*Applications of the Myers–Briggs Type Indicator in Higher Education,* edited by Provost and Anchors). In the table, IN refers to introvert, intuitive; EN refers to extrovert, intuitive; IS refers to introvert, sensing; and ES refers to extrovert, sensing.

Myers–Briggs Preference	Arts & Science	Business	Allied Health	Row Total
IN	64	15	17	96
EN	82	42	30	154
IS	68	35	12	115
ES	75	42	37	154
Column Total	289	134	96	519

Use a chi-square test to determine if Myers–Briggs preference type is independent of area of study at the 0.05 level of significance.

15. *Sociology: Movie Preference* Mr. Acosta, a sociologist, is doing a study to see if there is a relationship between the age of a young adult (18 to 35 years old) and the type of movie preferred. A random sample of 93 young adults revealed the following data. Test whether age and type of movie preferred are independent at the 0.05 level.

	Person's Age			
Movie	18–23 yr	24–29 yr	30–35 yr	Row Total
Drama	8	15	11	34
Science fiction	12	10	8	30
Comedy	9	8	12	29
Column Total	29	33	31	93

16. *Sociology: Ethnic Groups* After a large fund drive to help the Boston City Library, the following information was obtained from a random sample of contributors to the library fund. Using a 1% level of significance, test the claim that the amount contributed to the library fund is independent of ethnic group.

Ethnic Group	Number of People Making Contribution: \$1–50	\$51–100	\$101–150	\$151–200	Over \$200	Row Total
A	83	62	53	35	18	251
B	94	77	48	25	20	264
C	78	65	51	40	32	266
D	105	89	63	54	29	340
Column Total	360	293	215	154	99	1121

17. ***Focus Problem: Archaeology*** The Focus Problem at the beginning of the chapter refers to excavations at Burnt Mesa Pueblo in Bandelier National Monument. One question the archaeologists asked was: Is raw material used by prehistoric Indians for stone tool manufacture independent of the archaeological excavation site? Two different excavation sites at Burnt Mesa Pueblo gave the information in the following table. Use a chi-square test with 5% level of significance to test the claim that raw material used for construction of stone tools and excavation site are independent.

Material	Stone Tool Construction Material, Burnt Mesa Pueblo: Site A	Site B	Row Total
Basalt	731	584	1315
Obsidian	102	93	195
Pedernal chert	510	525	1035
Other	85	94	179
Column Total	1428	1296	2724

18. ***Political Affiliation: Spending*** Two random samples were drawn from members of the U.S. Congress. One sample was taken from members who are Democrats and the other from members who are Republicans. For each sample, the number of dollars spent on federal projects in each congressperson's home district was recorded.
(i) Make a cluster bar graph showing the percentages of Congress members from each party who spent each designated amount in their respective home districts.
(ii) Use a 1% level of significance to test whether congressional members of each political party spent designated amounts in the same proportions.

Party	Dollars Spent on Federal Projects in Home Districts: Less than 5 Billion	5 to 10 Billion	More than 10 Billion	Row Total
Democratic	8	15	22	45
Republican	12	19	16	47
Column Total	20	34	38	92

19. ***Sociology: Methods of Communication*** Random samples of people ages 15–24 and 25–34 were asked about their preferred method of (remote) communication with friends. The respondents were asked to select one of the methods from the following list: cell phone, instant message, e-mail, other.

(i) Make a cluster bar graph showing the percentages in each age group who selected each method.

(ii) Test whether the two populations share the same proportions of preferences for each type of communication method. Use $\alpha = 0.05$.

Age	Preferred Communication Method: Cell Phone	Instant Message	E-mail	Other	Row Total
15–24	48	40	5	7	100
25–34	41	30	15	14	100
Column Total	89	70	20	21	200

SECTION 10.2

Chi-Square: Goodness of Fit

FOCUS POINTS

- Set up a test to investigate how well a sample distribution fits a given distribution.
- Use observed and expected frequencies to compute the sample χ^2 statistic.
- Find or estimate the P-value and complete the test.

Last year, the labor union bargaining agents listed five categories and asked each employee to mark the *one* most important to her or him. The categories and corresponding percentages of favorable responses are shown in Table 10-8. The bargaining agents need to determine if the *current* distribution of responses "fits" last year's distribution or if it is different.

In questions of this type, we are asking whether a population follows a specified distribution. In other words, we are testing the hypotheses

Hypotheses

H_0: The population fits the given distribution.

H_1: The population has a different distribution.

Computing sample χ^2

We use the chi-square distribution to test "goodness-of-fit" hypotheses. Just as with tests of independence, we compute the sample statistic:

$$\chi^2 = \Sigma \frac{(O - E)^2}{E} \text{ with degrees of freedom } = k - 1$$

where $E =$ expected frequency

$O =$ observed frequency

$\frac{(O - E)^2}{E}$ is summed for each category in the distribution

$k =$ number of categories in the distribution

Next we use the chi-square distribution table (Table 7, Appendix II) to estimate the P-value of the sample χ^2 statistic. Finally, we compare the P-value to the level of significance α and conclude the test.

TABLE 10-8 Bargaining Categories (last year)

Category	Percentage of Favorable Responses
Vacation time	4%
Salary	65%
Safety regulations	13%
Health and retirement benefits	12%
Overtime policy and pay	6%

Goodness-of-fit test

In the case of a *goodness-of-fit test,* we use the null hypothesis to compute the expected values for the categories. Let's look at the bargaining category problem to see how this is done.

In the bargaining category problem, the two hypotheses are

H_0: The present distribution of responses is the same as last year's.
H_1: The present distribution of responses is different.

The null hypothesis tells us that the *expected frequencies* of the present response distribution should follow the percentages indicated in last year's survey. To test this hypothesis, a random sample of 500 employees was taken. If the null hypothesis is true, then there should be 4%, or 20 responses, out of the 500 rating vacation time as the most important bargaining issue. Table 10-9 gives the other expected values and all the information necessary to compute the sample statistic χ^2. We see that the sample statistic is

$$\chi^2 = \Sigma \frac{(O - E)^2}{E} = 14.15$$

Steve Skjold/PhotoEdit

Larger values of the sample statistic χ^2 indicate greater differences between the proposed distribution and the distribution followed by the sample. The larger the χ^2 statistic, the stronger the evidence to reject the null hypothesis that the population distribution fits the given distribution. Consequently, goodness-of-fit tests are always *right-tailed* tests.

Type of test

For *goodness-of-fit tests,* we use a *right-tailed* test on the chi-square distribution. This is because we are testing to see if the χ^2 measure of the difference between the observed and expected frequencies is too large to be due to chance alone.

To test the hypothesis that the present distribution of responses to bargaining categories is the same as last year's, we use the chi-square distribution (Table 7

TABLE 10-9 Observed and Expected Frequencies for Bargaining Categories

Category	O	E	$(O - E)^2$	$(O - E)^2/E$
Vacation time	30	4% of 500 = 20	100	5.00
Salary	290	65% of 500 = 325	1225	3.77
Safety	70	13% of 500 = 65	25	0.38
Health and retirement	70	12% of 500 = 60	100	1.67
Overtime	40	6% of 500 = 30	100	3.33
	$\Sigma O = 500$	$\Sigma E = 500$		$\Sigma \frac{(O - E)^2}{E} = 14.15$

of Appendix II) to estimate the *P*-value of the sample statistic $\chi^2 = 14.15$. To estimate the *P*-value, we need to know the number of *degrees of freedom*. In the case of a goodness-of-fit test, the degrees of freedom are found by the following formula.

Degrees of freedom

> Degrees of freedom for goodness-of-fit test
>
> $d.f. = k - 1$
>
> where k = number of categories

Notice that when we compute the expected values *E*, we must use the null hypothesis to compute all but the last one. To compute the last one, we can subtract the previous expected values from the sample size. For instance, for the bargaining issues, we could have found the number of responses for overtime policy by adding the other expected values and subtracting that sum from the sample size 500. We would again get an expected value of 30 responses. The degrees of freedom, then, is the number of *E* values that *must* be computed by using the null hypothesis.

For the bargaining issues, we have

$$d.f. = 5 - 1 = 4$$

where $k = 5$ is the number of categories.

P-value

We now have the tools necessary to use Table 7 of Appendix II to estimate the *P*-value of $\chi^2 = 14.15$. Figure 10-5 shows the *P*-value. In Table 7, we use the row headed by $d.f. = 4$. We see that $\chi^2 = 14.15$ falls between the entries 13.28 and 14.86. Therefore, the *P*-value falls between the corresponding right-tail areas 0.005 and 0.010. Technology gives the *P*-value ≈ 0.0068.

To test the hypothesis that the distribution of responses to bargaining issues is the same as last year's at the 1% level of significance, we compare the *P*-value of the statistic to $\alpha = 0.01$.

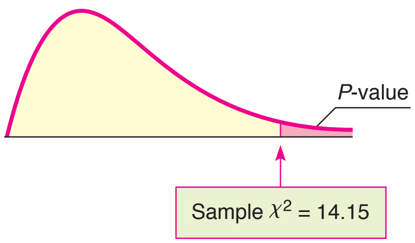
We see that the *P*-value is less than α, so we reject the null hypothesis that the distribution of responses to bargaining issues is the same as last year's.

Interpretation At the 1% level of significance, we can say that the evidence supports the conclusion that this year's responses to the issues are different from last year's.

Goodness-of-fit tests involve several steps that can be summarized as follows.

FIGURE 10-5

P-value



Right-tail area	0.010	0.005
$d.f. = 4$	13.28	14.86
	↑ Sample $\chi^2 = 14.15$	

PROCEDURE

How to test for goodness of fit

Setup

First, each member of a population needs to be classified into exactly one of several different categories. Next, you need a specific (theoretical) distribution that assigns a fixed probability (or percentage) that a member of the population will fall into one of the categories. You then need a random sample size n from the population. Let O represent the *observed number* of data from the sample that fall into each category. Let E represent the *expected number* of data from the sample that, in theory, would fall into each category.

$O =$ observed frequency count of a category using sample data

$E =$ expected frequency of a category

$=$ (sample size n)(probability assigned to category)

Requirement

The sample size n should be large enough that $E \geq 5$ in each category.

Procedure

1. Set the *level of significance* α and use the *hypotheses*

   H_0: The population fits the specified distribution of categories.

   H_1: The population has a different distribution.

2. For each category, compute $(O - E)^2/E$, and then compute the *sample test statistic*

   $$\chi^2 = \Sigma \frac{(O-E)^2}{E} \quad \text{with } d.f. = k - 1$$

   where the sum is taken over all categories and $k =$ number of categories.

3. Use the chi-square distribution (Table 7 of Appendix II) and a *right-tailed test* to find (or estimate) the *P-value* corresponding to the sample test statistic.
4. *Conclude* the test. If P-value $\leq \alpha$, then reject H_0. If P-value $> \alpha$, then do not reject H_0.
5. *Interpret your conclusion* in the context of the application.

One important application of goodness-of-fit tests is to genetics theory. Such an application is shown in Guided Exercise 6.

GUIDED EXERCISE 6 **Goodness-of-fit test**

According to genetics theory, red-green colorblindness in humans is a recessive sex-linked characteristic. In this case, the gene is carried on the X chromosome only. We will denote an X chromosome with the colorblindness gene by X_c and one without the gene by X_n. Women have two X chromosomes, and they will be red-green colorblind only if both chromosomes have the gene, designated X_cX_c. A woman can have normal vision but still carry the colorblind gene if only one of the chromosomes has the gene, designated X_cX_n. A man carries an X and a Y chromosome; if the X chromosome carries the colorblind gene (X_cY), the man is colorblind.

According to genetics theory, if a man with normal vision (X_nY) and a woman carrier (X_cX_n) have a child, the probabilities that the child will have red-green colorblindness, will have normal vision and not carry the gene, or will have normal vision and carry the gene are given by the *equally likely* events in Table 10-10.

Continued

GUIDED EXERCISE 6 *continued*

$P(\text{child has normal vision and is not a carrier}) = P(X_nY) + P(X_nX_n) = \frac{1}{2}$

$P(\text{child has normal vision and is a carrier}) = P(X_cX_n) = \frac{1}{4}$

$P(\text{child is red-green colorblind}) = P(X_cY) = \frac{1}{4}$

Table 10-10 Red-Green Colorblindness

	Father	
Mother	X_n	Y
X_c	X_cX_n	X_cY
X_n	X_nX_n	X_nY

To test this genetics theory, Genetics Labs took a random sample of 200 children whose mothers were carriers of the colorblind gene and whose fathers had normal vision. The results are shown in Table 10-11. We wish to test the hypothesis that the population follows the distribution predicted by the genetics theory (see Table 10-10). Use a 1% level of significance.

(a) State the null and alternate hypotheses. What is α?

H_0: The population fits the distribution predicted by genetics theory.

H_1: The population does not fit the distribution predicted by genetics theory.

$\alpha = 0.01$

(b) Fill in the rest of Table 10-11 and use the table to compute the sample statistic χ^2.

TABLE 10-11 Colorblindness Sample

Event	O	E	$(O - E)^2$	$(O - E)^2/E$
Red-green colorblind	35	50	225	4.50
Normal vision, noncarrier	105	_____	_____	_____
Normal vision, carrier	60	_____	_____	_____

TABLE 10-12 Completion of Table 10-11

Event	O	E	$(O - E)^2$	$(O - E)^2/E$
Red-green colorblind	35	50	225	4.50
Normal vision, noncarrier	105	100	25	0.25
Normal vision, carrier	60	50	100	2.00

The sample statistic is $\chi^2 = \Sigma \frac{(O - E)^2}{E} = 6.75$

(c) There are $k = 3$ categories listed in Table 10-11. Use this information to compute the degrees of freedom.

$d.f. = k - 1$
$= 3 - 1 = 2$

(d) Find the P-value for $\chi^2 = 6.75$.

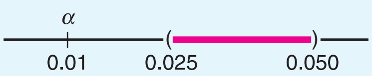
Using Table 7 of Appendix II and the fact that goodness-of-fit tests are right-tailed tests, we see that

Right-tail area	0.050	0.025
$d.f. = 2$	5.99	7.38
	↑ Sample $\chi^2 = 6.75$	

$0.025 < P\text{-value} < 0.050$
Technology gives P-value ≈ 0.0342.

(e) Conclude the test for $\alpha = 0.01$.

For $\alpha = 0.01$, we have



Since P-value $> \alpha$, do not reject H_0.

(f) *Interpret* the conclusion in the context of the application.

At the 1% level of significance, there is insufficient evidence to conclude that the population follows a distribution different from that predicted by genetics theory.

VIEWPOINT Run! Run! Run!

What description would you use for marathon runners? How about age distribution? Body weight? Length of stride? Heart rate? Blood pressure? What countries do these runners come from? What are their best running times? Make your own estimated distribution for these variables, and then consider a goodness-of-fit test for your distribution compared with available data. For more information on marathon runners, visit the Brase/Brase statistics site at **http://www.cengage.com/statistics/brase** *and find links to the Honolulu marathon site and to the* Runners World *site.*

SECTION 10.2 PROBLEMS

1. *Statistical Literacy* For a chi-square goodness-of-fit test, how are the degrees of freedom computed?

2. *Statistical Literacy* How are expected frequencies computed for goodness-of-fit tests?

3. *Statistical Literacy* Explain why goodness-of-fit tests are always right-tailed tests.

4. *Critical Thinking* When the sample evidence is sufficient to justify rejecting the null hypothesis in a goodness-of-fit test, can we tell exactly how the distribution of observed values over the specified categories differs from the expected distribution? Explain.

For Problems 5–16, please provide the following information.

(a) What is the level of significance? State the null and alternate hypotheses.
(b) Find the value of the chi-square statistic for the sample. Are all the expected frequencies greater than 5? What sampling distribution will you use? What are the degrees of freedom?
(c) Find or estimate the P-value of the sample test statistic.
(d) Based on your answers in parts (a) to (c), will you reject or fail to reject the null hypothesis that the population fits the specified distribution of categories?
(e) *Interpret* your conclusion in the context of the application.

5. *Census: Age* The age distribution of the Canadian population and the age distribution of a random sample of 455 residents in the Indian community of Red Lake Village (Northwest Territories) are shown below (based on *U.S. Bureau of the Census, International Data Base*).

Age (years)	Percent of Canadian Population	Observed Number in Red Lake Village
Under 5	7.2%	47
5 to 14	13.6%	75
15 to 64	67.1%	288
65 and older	12.1%	45

Use a 5% level of significance to test the claim that the age distribution of the general Canadian population fits the age distribution of the residents of Red Lake Village.

6. *Census: Type of Household* The type of household for the U.S. population and for a random sample of 411 households from the community of Dove Creek, Montana, are shown (based on *Statistical Abstract of the United States*).

Type of Household	Percent of U.S. Households	Observed Number of Households in Dove Creek
Married with children	26%	102
Married, no children	29%	112
Single parent	9%	33
One person	25%	96
Other (e.g., roommates, siblings)	11%	68

Use a 5% level of significance to test the claim that the distribution of U.S. households fits the Dove Creek distribution.

7. ***Archaeology: Stone Tools*** The types of raw materials used to construct stone tools found at the archaeological site Casa del Rito are shown below (*Bandelier Archaeological Excavation Project*, edited by Kohler and Root). A random sample of 1486 stone tools was obtained from a current excavation site.

Raw Material	Regional Percent of Stone Tools	Observed Number of Tools at Current Excavation Site
Basalt	61.3%	906
Obsidian	10.6%	162
Welded tuff	11.4%	168
Pedernal chert	13.1%	197
Other	3.6%	53

Use a 1% level of significance to test the claim that the regional distribution of raw materials fits the distribution at the current excavation site.

8. ***Ecology: Deer*** The types of browse favored by deer are shown in the following table (*The Mule Deer of Mesa Verde National Park*, edited by Mierau and Schmidt). Using binoculars, volunteers observed the feeding habits of a random sample of 320 deer.

Type of Browse	Plant Composition in Study Area	Observed Number of Deer Feeding on This Plant
Sage brush	32%	102
Rabbit brush	38.7%	125
Salt brush	12%	43
Service berry	9.3%	27
Other	8%	23

Use a 5% level of significance to test the claim that the natural distribution of browse fits the deer feeding pattern.

9. ***Meteorology: Normal Distribution*** The following problem is based on information from the *National Oceanic and Atmospheric Administration (NOAA) Environmental Data Service*. Let x be a random variable that represents the average daily temperature (in degrees Fahrenheit) in July in the town of Kit Carson, Colorado. The x distribution has a mean μ of approximately 75°F and standard deviation σ of approximately 8°F. A 20-year study (620 July days) gave the entries in the rightmost column of the following table.

I	II	III	IV
Region under Normal Curve	x°F	Expected % from Normal Curve	Observed Number of Days in 20 Years
$\mu - 3\sigma \le x < \mu - 2\sigma$	$51 \le x < 59$	2.35%	16
$\mu - 2\sigma \le x < \mu - \sigma$	$59 \le x < 67$	13.5%	78
$\mu - \sigma \le x < \mu$	$67 \le x < 75$	34%	212
$\mu \le x < \mu + \sigma$	$75 \le x < 83$	34%	221
$\mu + \sigma \le x < \mu + 2\sigma$	$83 \le x < 91$	13.5%	81
$\mu + 2\sigma \le x < \mu + 3\sigma$	$91 \le x < 99$	2.35%	12

(i) Remember that $\mu = 75$ and $\sigma = 8$. Examine Figure 6-5 in Chapter 6. Write a brief explanation for Columns I, II, and III in the context of this problem.
(ii) Use a 1% level of significance to test the claim that the average daily July temperature follows a normal distribution with $\mu = 75$ and $\sigma = 8$.

10. *Meteorology: Normal Distribution* Let x be a random variable that represents the average daily temperature (in degrees Fahrenheit) in January for the town of Hana, Maui. The x variable has a mean μ of approximately 68°F and standard deviation σ of approximately 4°F (see reference in Problem 9). A 20-year study (620 January days) gave the entries in the rightmost column of the following table.

I	II	III	IV
Region Under Normal Curve	x°F	Expected % from Normal Curve	Observed Number of Days in 20 Years
$\mu - 3\sigma \le x < \mu - 2\sigma$	$56 \le x < 60$	2.35%	14
$\mu - 2\sigma \le x < \mu - \sigma$	$60 \le x < 64$	13.5%	86
$\mu - \sigma \le x < \mu$	$64 \le x < 68$	34%	207
$\mu \le x < \mu + \sigma$	$68 \le x < 76$	34%	215
$\mu + \sigma \le x < \mu + 2\sigma$	$72 \le x < 76$	13.5%	83
$\mu + 2\sigma \le x < \mu + 3\sigma$	$76 \le x < 80$	2.35%	15

(i) Remember that $\mu = 68$ and $\sigma = 4$. Examine Figure 6-5 in Chapter 6. Write a brief explanation for Columns I, II, and III in the context of this problem.
(ii) Use a 1% level of significance to test the claim that the average daily January temperature follows a normal distribution with $\mu = 68$ and $\sigma = 4$.

11. *Ecology: Fish* The Fish and Game Department stocked Lake Lulu with fish in the following proportions: 30% catfish, 15% bass, 40% bluegill, and 15% pike. Five years later it sampled the lake to see if the distribution of fish had changed. It found that the 500 fish in the sample were distributed as follows.

Catfish	Bass	Bluegill	Pike
120	85	220	75

In the 5-year interval, did the distribution of fish change at the 0.05 level?

12. *Library: Book Circulation* The director of library services at Fairmont College did a survey of types of books (by subject) in the circulation library. Then she used library records to take a random sample of 888 books checked out last

term and classified the books in the sample by subject. The results are shown below.

Subject Area	Percent of Books in Circulation Library on This Subject	Number of Books in Sample on This Subject
Business	32%	268
Humanities	25%	214
Natural science	20%	215
Social science	15%	115
All other subjects	8%	76

Using a 5% level of significance, test the claim that the subject distribution of books in the library fits the distribution of books checked out by students.

13. ***Census: California*** The accuracy of a census report on a city in southern California was questioned by some government officials. A random sample of 1215 people living in the city was used to check the report, and the results are shown here:

Ethnic Origin	Census Percent	Sample Result
Black	10%	127
Asian	3%	40
Anglo	38%	480
Latino/Latina	41%	502
Native American	6%	56
All others	2%	10

Using a 1% level of significance, test the claim that the census distribution and the sample distribution agree.

14. ***Marketing: Compact Discs*** Snoop Incorporated is a firm that does market surveys. The Rollum Sound Company hired Snoop to study the age distribution of people who buy compact discs. To check the Snoop report, Rollum used a random sample of 519 customers and obtained the following data:

Customer Age (years)	Percent of Customers from Snoop Report	Number of Customers from Sample
Younger than 14	12%	88
14–18	29%	135
19–23	11%	52
24–28	10%	40
29–33	14%	76
Older than 33	24%	128

Using a 1% level of significance, test the claim that the distribution of customer ages in the Snoop report agrees with that of the sample report.

15. ***Accounting Records: Benford's Law*** Benford's law states that the first nonzero digits of numbers drawn at random from a large complex data file have the following probability distribution (Reference: American Statistical Association, *Chance*, Vol. 12, No. 3, pp. 27–31; see also the Focus Problem of Chapter 9).

First nonzero digit	1	2	3	4	5	6	7	8	9
Probability	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Suppose that $n = 275$ numerical entries were drawn at random from a large accounting file of a major corporation. The first nonzero digits were recorded for the sample.

First nonzero digit	1	2	3	4	5	6	7	8	9
Sample frequency	83	49	32	22	25	18	13	17	16

Use a 1% level of significance to test the claim that the distribution of first nonzero digits in this accounting file follows Benford's law.

16. ***Fair Dice: Uniform Distribution*** A gambler complained about the dice. They seemed to be loaded! The dice were taken off the table and tested one at a time. One die was rolled 300 times and the following frequencies were recorded.

Outcome	1	2	3	4	5	6
Observed frequency O	62	45	63	32	47	51

Do these data indicate that the die is unbalanced? Use a 1% level of significance. *Hint:* If the die is balanced, all outcomes should have the same expected frequency.

17. ***Highway Accidents: Poisson Distribution*** A civil engineer has been studying the frequency of vehicle accidents on a certain stretch of interstate highway. Long-term history indicates that there has been an average of 1.72 accidents per day on this section of the interstate. Let r be a random variable that represents number of accidents per day. Let O represent the number of observed accidents per day based on local highway patrol reports. A random sample of 90 days gave the following information.

r	0	1	2	3	4 or more
O	22	21	15	17	15

(a) The civil engineer wants to use a Poisson distribution to represent the probability of r, the number of accidents per day. The Poisson distribution is

$$P(r) = \frac{e^{-\lambda}\lambda^r}{r!}$$

where $\lambda = 1.72$ is the average number of accidents per day. Compute $P(r)$ for $r = 0, 1, 2, 3$, and 4 or more.

(b) Compute the expected number of accidents $E = 90P(r)$ for $r = 0, 1, 2, 3$, and 4 or more.

(c) Compute the sample statistic $\chi^2 = \Sigma \frac{(O - E)^2}{E}$ and the degrees of freedom.

(d) Test the statement that the Poisson distribution fits the sample data. Use a 1% level of significance.

18. ***Bacteria Colonies: Poisson Distribution*** A pathologist has been studying the frequency of bacterial colonies within the field of a microscope using samples of throat cultures from healthy adults. Long-term history indicates that there is an average of 2.80 bacteria colonies per field. Let r be a random variable that represents the number of bacteria colonies per field. Let O represent the number of observed bacteria colonies per field for throat cultures from healthy adults. A random sample of 100 healthy adults gave the following information.

r	0	1	2	3	4	5 or more
O	12	15	29	18	19	7

(a) The pathologist wants to use a Poisson distribution to represent the probability of r, the number of bacteria colonies per field. The Poisson distribution is

$$P(r) = \frac{e^{-\lambda}\lambda^r}{r!}$$

where $\lambda = 2.80$ is the average number of bacteria colonies per field. Compute $P(r)$ for $r = 0, 1, 2, 3, 4$, and 5 or more.

(b) Compute the expected number of colonies $E = 100P(r)$ for $r = 0, 1, 2, 3, 4$, and 5 or more.

(c) Compute the sample statistic $\chi^2 = \Sigma\frac{(O - E)^2}{E}$ and the degrees of freedom.

(d) Test the statement that the Poisson distribution fits the sample data. Use a 5% level of significance.

SECTION 10.3

Testing and Estimating a Single Variance or Standard Deviation

FOCUS POINTS

- Set up a test for a single variance σ^2.
- Compute the sample χ^2 statistic.
- Use the χ^2 distribution to estimate a P-value and conclude the test.
- Compute confidence intervals for σ^2 or σ.

Testing σ^2

Many problems arise that require us to make decisions about variability. In this section, we will study two kinds of problems: (1) we will test hypotheses about the variance (or standard deviation) of a population, and (2) we will find confidence intervals for the variance (or standard deviation) of a population. It is customary to talk about variance instead of standard deviation because our techniques employ the sample variance rather than the standard deviation. Of course, the standard deviation is just the square root of the variance, so any discussion about variance is easily converted to a similar discussion about standard deviation.

Test of variance

Let us consider a specific example in which we might wish to test a hypothesis about the variance. Almost everyone has had to wait in line. In a grocery store, bank, post office, or registration center, there are usually several checkout or service areas. Frequently, each service area has its own independent line. However, many businesses and government offices are adopting a "single-line" procedure.

In a single-line procedure, there is only one waiting line for everyone. As any service area becomes available, the next person in line gets served. The old, independent-lines procedure has a line at each service center. An incoming customer simply picks the shortest line and hopes it will move quickly. In either procedure, the number of clerks and the rate at which they work is the same, so the average waiting time is the *same*. What is the advantage of the single-line procedure? The difference is in the *attitudes* of people who wait in the lines. A lengthy waiting line will be more acceptable, even though the average waiting time is the same, if the variability of waiting times is smaller. When the variability is small, the inconvenience of waiting (although it might not be reduced) does become more predictable. This means impatience is reduced and people are happier.

To test the hypothesis that variability is less in a single-line process, we use the chi-square distribution. The next theorem tells us how to use the sample and population variance to compute values of χ^2.

THEOREM 10.1 If we have a *normal* population with variance σ^2 and a random sample of n measurements is taken from this population with sample variance s^2, then

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

has a chi-square distribution with degrees of freedom $d.f. = n - 1$.

Recall that the chi-square distribution is *not* symmetrical and that there are different chi-square distributions for different degrees of freedom. Table 7 of Appendix II gives chi-square values for which the area α is to the *right* of the given chi-square value.

EXAMPLE 3 χ^2 DISTRIBUTION

(a) Find the χ^2 value such that the area to the right of χ^2 is 0.05 when $d.f. = 10$.

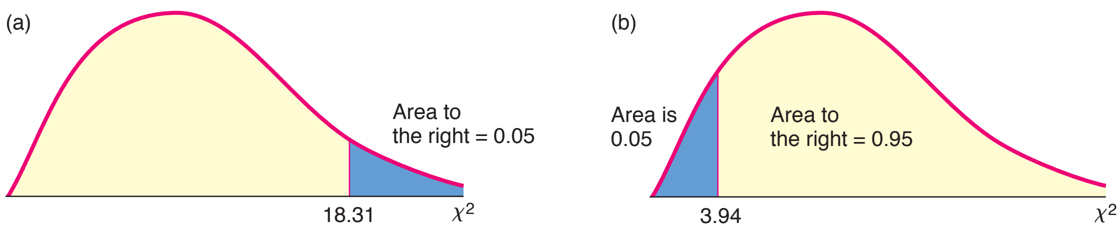
SOLUTION: Since the area to the right of χ^2 is to be 0.05, we look in the right-tail area = 0.050 column and the row with $d.f. = 10$. $\chi^2 = 18.31$ (see Figure 10-6a).

(b) Find the χ^2 value such that the area to the *left* of χ^2 is 0.05 when $d.f. = 10$.

SOLUTION: When the area to the left of χ^2 is 0.05, the corresponding area to the *right* is $1 - 0.05 = 0.95$, so we look in the right-tail area = 0.950 column and the row with $d.f. = 10$. We find $\chi^2 = 3.94$ (see Figure 10-6b).

FIGURE 10-6

χ^2 Distribution with $d.f. = 10$.

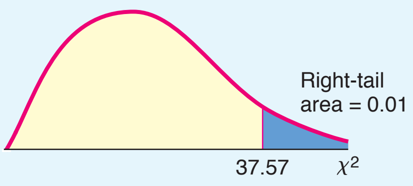


GUIDED EXERCISE 7 χ^2 distribution

(a) Find the area to the *right* of $\chi^2 = 37.57$ when $d.f. = 20$.

Use Table 7 of Appendix II. In the row headed by $d.f. = 20$, find the column with $\chi^2 = 37.57$. The column header 0.010 is the area to the right (see Figure 10-7).

FIGURE 10-7 χ^2 Distribution with $d.f. = 20$



Continued

GUIDED EXERCISE 7 *continued*

(b) Find the area to the *left* of $\chi^2 = 8.26$ when $d.f. = 20$.

(i) First use Table 7 of Appendix II to find the area to the right of $\chi^2 = 8.26$.

(ii) To get the area to the *left* of $\chi^2 = 8.26$, subtract the area to the right from 1.

Find the column with $\chi^2 = 8.26$ in the row headed by $d.f. = 20$. The column header 0.990 gives the area to the *right* of $\chi^2 = 8.26$.

Next, subtract the area to the right from 1.
Area to *left* = 1 − Area to *right* = 1 − 0.990 = 0.010 (see Figure 10-8).

FIGURE 10-8 χ^2 Distribution with $d.f. = 20$

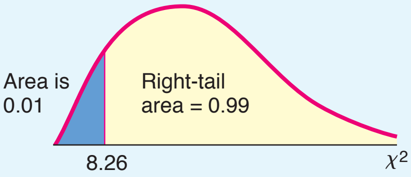


Table 10-13 summarizes the techniques for using the chi-square distribution (Table 7 of Appendix II) to find P-values for a right-tailed test, a left-tailed test, and a two-tailed test. Example 4 demonstrates the technique of finding P-values for a left-tailed test. Example 5 demonstrates the technique for a two-tailed test, and Guided Exercise 8 uses a right-tailed test.

Now let's use Theorem 10.1 and our knowledge of the chi-square distribution to determine if a single-line procedure has less variance of waiting times than independent lines.

TABLE 10-13 ***P*-values for Chi-Square Distribution Table (Table 7, Appendix II)**

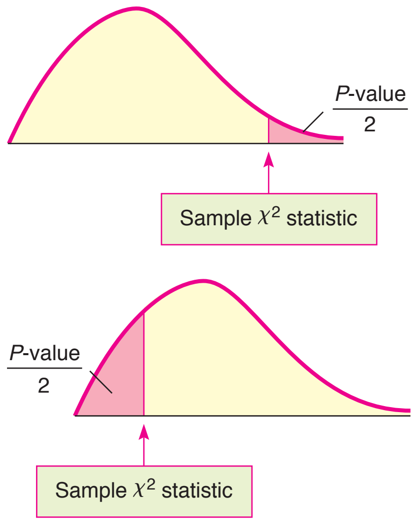
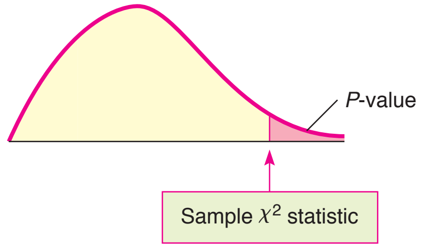
(a) Two-tailed test Remember that the P-value is the probability of getting a test statistic as extreme as, or more extreme than, the test statistic computed from the sample. For a two-tailed test, we need to account for corresponding equal areas in both the upper and lower tails. This means that in each tail, we have an area of P-value/2. The total P-value is then $P\text{-value} = 2\left(\frac{P\text{-value}}{2}\right)$	 Be sure to choose the area in the appropriate tail (left or right) so that $\frac{P\text{-value}}{2} \leq 0.5$.

TABLE 10-13 ***P*-values for Chi-Square Distribution Table (Table 7, Appendix II)** *(continued)*

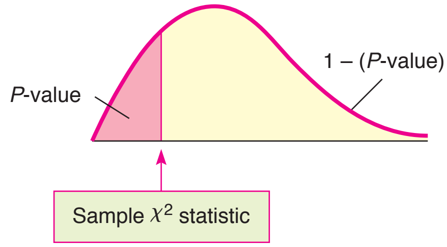
(b) Right-tailed test

Since the chi-square table gives right-tail probabilities, you can use the table directly to find or estimate the *P*-value.



(c) Left-tailed test

Since the chi-square table gives right-tail probabilities, you first find or estimate the quantity $1 - (P\text{-value})$ right tail. Then subtract from 1 to get the *P*-value of the left-tail.



EXAMPLE 4 TESTING THE VARIANCE (LEFT-TAILED TEST)

For years, a large discount store has used independent lines to check out customers. Historically, the standard deviation of waiting times is 7 minutes. The manager tried a new, single-line procedure. A random sample of 25 customers using the single-line procedure was monitored, and it was found that the standard deviation for waiting times was only $s = 5$ minutes. Use $\alpha = 0.05$ to test the claim that the variance in waiting times is reduced for the single-line method. Assume the waiting times are normally distributed.

Establish H_0 and H_1

SOLUTION: As a null hypothesis, we assume that the variance of waiting times is the same as that of the former independent-lines procedure. The alternate hypothesis is that the variance for the single-line procedure is less than that for the independent-lines procedure. If we let σ be the standard deviation of waiting times for the single-line procedure, then σ^2 is the variance, and we have

Sample χ^2 value and degrees of freedom

$$H_0\colon \sigma^2 = 49 \qquad H_1\colon \sigma^2 < 49 \qquad (\text{use } 7^2 = 49)$$

We use the chi-square distribution to test the hypotheses. Assuming that the waiting times are normally distributed, we compute our observed value of χ^2 by using Theorem 10.1, with $n = 25$.

$$s = 5 \quad \text{so} \quad s^2 = 25 \quad (\text{observed from sample})$$

$$\sigma = 7 \quad \text{so} \quad \sigma^2 = 49 \quad (\textit{from } H_0\colon \sigma^2 = 49)$$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)25}{49} \approx 12.24$$

$$d.f. = n - 1 = 25 - 1 = 24$$

Mark Richards/PhotoEdit

Checkout lines

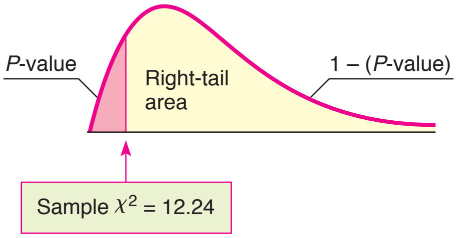
Next we estimate the *P*-value for $\chi^2 = 12.24$. Since we have a left-tailed test, the *P*-value is the area of the chi-square distribution that lies to the *left* of $\chi^2 = 12.24$, as shown in Figure 10-9 on the next page.

P-value

To estimate the *P*-value on the left, we consider the fact that the area of the right tail is between 0.975 and 0.990. To find an estimate for the area of the left tail, we *subtract* each right-tail endpoint from 1. The *P*-value (area of the left tail) is in the interval

FIGURE 10-9

P-value

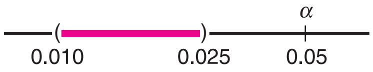


Right-tail Area	0.990	0.975
d.f. = 24	10.86	12.40
	Sample $\chi^2 = 12.24$	

$$1 - 0.990 < P\text{-value of left tail} < 1 - 0.975$$
$$0.010 < P\text{-value} < 0.025$$

Test conclusion

To conclude the test, we compare the *P*-value to the level of significance $\alpha = 0.05$.



Since the *P*-value is less than α, we reject H_0.

Interpretation

Interpretation At the 5% level of significance, we conclude that the variance of waiting times for a single line is less than the variance of waiting times for multiple lines.

The steps used in Example 4 for testing the variance σ^2 are summarized as follows.

PROCEDURE

HOW TO TEST σ^2

Requirements

You first need to know that a random variable x has a normal distribution. In testing σ^2, the normal assumption must be strictly observed (whereas in testing means, we can say "normal" or "approximately normal"). Next you need a random sample (size $n \geq 2$) of values from the x distribution for which you compute the sample variance s^2.

Procedure

1. In the context of the problem, state the *null hypothesis* H_0 and the *alternate hypothesis* H_1, and set the *level of significance* α.
2. Use the value of σ^2 given in the null hypothesis H_0, the sample variance s^2, and the sample size n to compute the *sample test statistic*

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \quad \text{with degrees of freedom } d.f. = n - 1$$

3. Use a chi-square distribution and the type of test to find or estimate the *P-value*. Use the procedures shown in Table 10-13 and Table 7 of Appendix II.
4. *Conclude* the test. If *P*-value $\leq \alpha$, then reject H_0. If *P*-value $> \alpha$, then do not reject H_0.
5. *Interpret your conclusion* in the context of the application.

EXAMPLE 5

TESTING THE VARIANCE (TWO-TAILED TEST)

Let x be a random variable that represents weight loss (in pounds) after following a certain diet for 6 months. After extensive study, it is found that x has a normal distribution with $\sigma = 5.7$ pounds. A new modification of the diet has been implemented. A random sample of $n = 21$ people use the modified diet for 6 months. For these people, the sample standard deviation of weight loss is $s = 4.1$ pounds. Does this result indicate that the variance of weight loss for the modified diet is different (either way) from the variance of weight loss for the original diet? Use $\alpha = 0.01$. Assume weight loss for each diet follows a normal distribution.

(a) What is the level of significance? State the null and alternate hypotheses.

SOLUTION: We are using $\alpha = 0.01$. The standard deviation of weight loss for the original diet is $\sigma = 5.7$ pounds, so the variance is $\sigma^2 = 32.49$. The null hypothesis is that the weight loss variance for the modified diet is the same as that for the original diet. The alternate hypothesis is that the variance is different.

$$H_0: \sigma^2 = 32.49 \quad H_1: \sigma^2 \neq 32.49$$

(b) Compute the sample test statistic χ^2 and the degrees of freedom.

SOLUTION: Using sample size $n = 21$, sample standard deviation $s = 4.1$ pounds, and $\sigma^2 = 32.49$ from the null hypothesis, we have

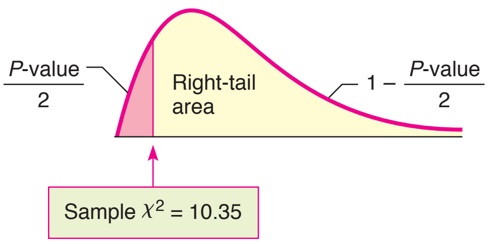
$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(21-1)4.1^2}{32.49} \approx 10.35$$

with degrees of freedom $d.f. = n - 1 = 21 - 1 = 20$.

(c) Use the chi-square distribution (Table 7 of Appendix II) to estimate the *P*-value.

FIGURE 10-10

***P*-value**



Right-tail Area	0.975	0.950
$d.f. = 20$	8.59	10.85
	Sample $\chi^2 = 10.35$	

SOLUTION: For a *two-tailed* test, the area beyond the sample χ^2 represents *half* the total *P*-value or (*P*-value)/2. Figure 10-10 shows this region, which is to the left of $\chi^2 \approx 10.35$. However, Table 7 of Appendix II gives the areas in the *right tail*. We use Table 7 to find the area in the right tail and then subtract from 1 to find the corresponding area in the left tail.

From the table, we see that the right-tail area falls in the interval between 0.950 and 0.975. Subtracting each endpoint of the interval from 1 gives us an interval containing (*P*-value)/2. Multiplying by 2 gives an interval for the *P*-value.

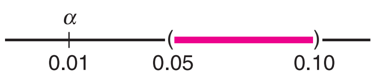
$$1 - 0.975 < \frac{P\text{-value}}{2} < 1 - 0.950 \quad \text{Subtract right-tail-area endpoints from 1.}$$

$$0.025 < \frac{P\text{-value}}{2} < 0.050$$

$$0.05 < P\text{-value} < 0.10 \quad \text{Multiply each part by 2.}$$

(d) *Conclude* the test.

SOLUTION: The *P*-value is greater than $\alpha = 0.01$, so we do not reject H_0.



(e) ***Interpret*** the conclusion in the context of the application.

SOLUTION: At the 1% level of significance, there is insufficient evidence to conclude that the variance of weight loss using the modified diet is different from the variance of weight loss using the original diet.

GUIDED EXERCISE 8 Testing the variance (right-tailed test)

Certain industrial machines require overhaul when wear on their parts introduces too much variability to pass inspection. A government official is visiting a dentist's office to inspect the operation of an x-ray machine. If the machine emits too little radiation, clear photographs cannot be obtained. However, too much radiation can be harmful to the patient. Government regulations specify an average emission of 60 millirads with standard deviation σ of 12 millirads, and the machine has been set for these readings. After examining the machine, the inspector is satisfied that the average emission is still 60 millirads. However, there is wear on certain mechanical parts. To test variability, the inspector takes a random sample of 30 x-ray emissions and finds the sample standard deviation to be $s = 15$ millirads. Does this support the claim that the variance is too high (i.e., the machine should be overhauled)? Use a 1% level of significance. Assume the emissions follow a normal distribution.

Let σ be the (population) standard deviation of emissions (in millirads) of the machine in its present condition.

(a) What is α? State H_0 and H_1.

$\alpha = 0.01$. Government regulations specify that $\sigma = 12$. This means that the variance $\sigma^2 = 144$. We are to test the claim that the variance is higher than government specifications allow.

H_0: $\sigma^2 = 144$ and H_1: $\sigma^2 > 144$

(b) Compute the sample test statistic χ^2 and corresponding degrees of freedom.

Using $n = 30$, $s = 15$, and $\sigma^2 = 144$ from H_0,

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(30-1)15^2}{144} \approx 45.3$$

Degrees of freedom $d.f. = n - 1 = 30 - 1 = 29$

(c) Estimate the P-value for the sample $\chi^2 = 45.3$ with $d.f. = 29$.

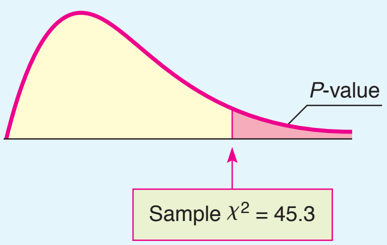
Since this is a *right-tailed* test, we look up P-values directly in the chi-square table (Table 7 of Appendix II).

Right-tail Area	0.050	0.025
$d.f. = 29$	42.56	45.72
	↑ Sample $\chi^2 = 45.3$	

$0.025 < P\text{-value} < 0.050$

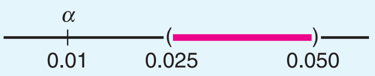
FIGURE 10-11

P-value



(d) *Conclude* the test.

The P-value for $\chi^2 = 45.3$ is greater than $\alpha = 0.01$.



Fail to reject H_0.

(e) *Interpret* the conclusion in the context of the application.

At the 1% level of significance, there is insufficient evidence to conclude that the variance of the radiation emitted by the machine is greater than that specified by government regulations. The evidence does not indicate that an adjustment is necessary at this time.

Confidence Interval for σ^2

Confidence Interval

Sometimes it is important to have a *confidence interval* for the variance or standard deviation. Let us look at another example.

Mr. Wilson is a truck farmer in California who makes his living on a large single-vegetable crop of green beans. Because modern machinery is being used, the entire crop must be harvested at the same time. Therefore, it is important to plant a variety of green beans that mature all at once. This means that Mr. Wilson wants a small standard deviation between maturing times of individual plants. A seed company is trying to develop a new variety of green bean with a small standard deviation of maturing times. To test the new variety, Mr. Wilson planted 30 of the new seeds and carefully observed the number of days required for each plant to arrive at its peak of maturity. The maturing times for these plants had a sample standard deviation of $s = 3.4$ days. How can we find a 95% confidence interval for the population standard deviation of maturing times for this variety of green bean? The answer to this question is based on the following procedure.

PROCEDURE

How to find a confidence interval for σ^2 and σ

Requirements

Let x be a random variable with a normal distribution and unknown population standard deviation σ. Take a random sample of size n from the x distribution and compute the sample standard deviation s.

Procedure

A **confidence interval for the population variance σ^2** is

$$\frac{(n-1)s^2}{\chi^2_U} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L} \qquad (1)$$

and a **confidence interval for the population standard deviation σ** is

$$\sqrt{\frac{(n-1)s^2}{\chi^2_U}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_L}}$$

where

$c =$ confidence level $(0 < c < 1)$

$n =$ sample size $(n \geq 2)$

χ^2_U

$\chi^2_U =$ chi-square value from Table 7 of Appendix II using $d.f. = n - 1$ and right-tail area $= (1 - c)/2$

χ^2_L

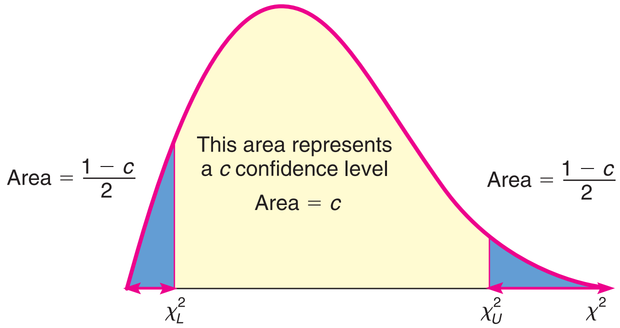
$\chi^2_L =$ chi-square value from Table 7 of Appendix II using $d.f. = n - 1$ and right-tail area $= (1 + c)/2$

From Figure 10-12, on the next page, we see that a c confidence level on a chi-square distribution with equal probability in each tail does not center the middle of the corresponding interval under the peak of the curve. This is to be expected because a chi-square curve is skewed to the right.

COMMENT Note that the method of computing confidence intervals for variances is different from the method of computing confidence intervals for means or proportions as studied in Chapter 7. Confidence intervals for σ^2 do not involve a maximal error of estimate E. Rather, the endpoints of the confidence interval are computed directly using the sample statistic s^2, the sample size, and the critical values.

FIGURE 10-12

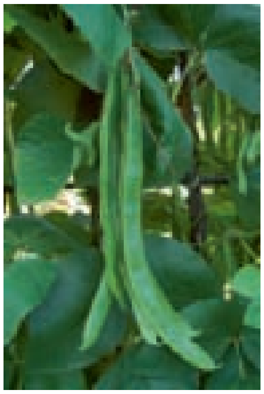
Area Representing a c Confidence Level on a Chi-Square Distribution with $d.f. = n - 1$



Now let us finish our example regarding the variance of maturing times for green beans.

EXAMPLE 6 CONFIDENCE INTERVALS FOR σ^2 AND σ

A random sample of $n = 30$ green bean plants has a sample standard deviation of $s = 3.4$ days for maturity. Find a 95% confidence interval for the population variance σ^2. Assume the distribution of maturity times is normal.



Iain Frazer, 2009/Used under license from Shutterstock.com

SOLUTION: To find the confidence interval, we use the following values:

$c = 0.95$ confidence level

$n = 30$ sample size

$d.f. = n - 1 = 30 - 1 = 29$ degrees of freedom

$s = 3.4$ sample standard deviation

To find the value of χ^2_U, we use Table 7 of Appendix II with $d.f. = 29$ and right-tail area $= (1 - c)/2 = (1 - 0.95)/2 = 0.025$. From Table 7, we get

$$\chi^2_U = 45.72$$

To find the value of χ^2_L, we use Table 7 of Appendix II with $d.f. = 29$ and right-tail area $= (1 + c)/2 = (1 + 0.95)/2 = 0.975$. From Table 7, we get

$$\chi^2_L = 16.05$$

Formula (1) tells us that our desired 95% confidence interval for σ^2 is

$$\frac{(n-1)s^2}{\chi^2_U} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}$$

$$\frac{(30-1)(3.4)^2}{45.72} < \sigma^2 < \frac{(30-1)(3.4)^2}{16.05}$$

$$7.33 < \sigma^2 < 20.89$$

To find a 95% confidence interval for σ, we simply take square roots; therefore, a 95% confidence interval for σ is

$$\sqrt{7.33} < \sigma < \sqrt{20.89}$$

$$2.71 < \sigma < 4.57$$

GUIDED EXERCISE 9 Confidence intervals for σ^2 and σ

A few miles off the Kona coast of the island of Hawaii, a research vessel lies anchored. This ship makes electrical energy from the solar temperature differential of (warm) surface water versus (cool) deep water. The basic idea is that the warm water is flushed over coils to vaporize a special fluid. The vapor is under pressure and drives electrical turbines. Then some electricity is used to pump up cold water to cool the vapor back to a liquid, and the process is repeated. Even though some electricity is used to pump up the cold water, there is plenty left to supply a moderate-sized Hawaiian town.

Continued

GUIDED EXERCISE 9 *continued*

The subtropic sun always warms up surface water to a reliable temperature, but ocean currents can change the temperature of the deep, cooler water. If the deep-water temperature is too variable, the power plant cannot operate efficiently or possibly cannot operate at all. To estimate the variability of deep ocean water temperatures, a random sample of 25 near-bottom readings gave a sample standard deviation of 7.3°C. Find a 99% confidence interval for the variance σ^2 and standard deviation σ of the deep-water temperatures. Assume deep-water temperatures are normally distributed.

(a) Determine the following values: $c =$ _____; $n =$ _____; $d.f =$ _____; $s =$ _____.

→ $c = 0.99$; $n = 25$; $d.f. = 24$; $s = 7.3$

(b) What is the value of χ^2_U? of χ^2_L?

→ We use Table 7 of Appendix II with $d.f. = 24$.

For χ^2_U, right-tail area $= (1 - 0.99)/2 = 0.005$

$$\chi^2_U = 45.56$$

For χ^2_L, right-tail area $= (1 + 0.99)/2 = 0.995$

$$\chi^2_L = 9.89$$

(c) Find a 99% confidence interval for σ^2.

→
$$\frac{(n-1)s^2}{\chi^2_U} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}$$
$$\frac{(24)(7.3)^2}{45.56} < \sigma^2 < \frac{24(7.3)^2}{9.89}$$
$$28.07 < \sigma^2 < 129.32$$

(d) Find a 99% confidence interval for σ.

→
$$\sqrt{28.07} < \sqrt{\sigma^2} < \sqrt{129.32}$$
$$5.30 < \sigma < 11.37$$

VIEWPOINT Adoption—A Good Choice!

Cuckoos are birds that are known to lay their eggs in the nests of other (host) birds. The host birds then hatch the eggs and adopt the cuckoo chicks as their own. Birds such as the meadow pipit, tree pipit, hedge sparrow, robin, and wren have all played host to cuckoo eggs and adopted their chicks. L. H. C. Tippett (1902–1985) was a pioneer in the field of statistical quality control who collected data on cuckoo eggs found in the nests of other birds. For more information and data from Tippett's study, visit the Brase/Brase statistics site at **http://www.cengage.com/statistics/brase** *and find a link to DASL, the Carnegie Mellon University Data and Story Library. Find Biology under Data Subjects, and then select the Cuckoo Egg Length Data file.*

SECTION 10.3 PROBLEMS

1. *Statistical Literacy* Does the x distribution need to be normal in order to use the chi-square distribution to test the variance? Is it acceptable to use the chi-square distribution to test the variance if the x distribution is simply mound-shaped and more or less symmetric?

2. *Critical Thinking* The x distribution must be normal in order to use a chi-square distribution to test the variance. What are some methods you can use to assess whether the x distribution is normal? *Hint:* See Chapter 6 and goodness-of-fit tests.

For Problems 3–11, please provide the following information.

(a) What is the level of significance? State the null and alternate hypotheses.
(b) Find the value of the chi-square statistic for the sample. What are the degrees of freedom? What assumptions are you making about the original distribution?
(c) Find or estimate the P-value of the sample test statistic.

(d) Based on your answers in parts (a) to (c), will you reject or fail to reject the null hypothesis of independence?

(e) *Interpret* your conclusion in the context of the application.

(f) Find the requested confidence interval for the population variance or population standard deviation. Interpret the results in the context of the application.

In each of the following problems, assume a normal population distribution.

3. *Archaeology: Chaco Canyon* The following problem is based on information from *Archaeological Surveys of Chaco Canyon, New Mexico,* by A. Hayes, D. Brugge, and W. Judge, University of New Mexico Press. A *transect* is an archaeological study area that is 1/5 mile wide and 1 mile long. A *site* in a transect is the location of a significant archaeological find. Let x represent the number of sites per transect. In a section of Chaco Canyon, a large number of transects showed that x has a population variance $\sigma^2 = 42.3$. In a different section of Chaco Canyon, a random sample of 23 transects gave a sample variance $s^2 = 46.1$ for the number of sites per transect. Use a 5% level of significance to test the claim that the variance in the new section is greater than 42.3. Find a 95% confidence interval for the population variance.

4. *Sociology: Marriage* The following problem is based on information from an article by N. Keyfitz in the *American Journal of Sociology* (Vol. 53, pp. 470–480). Let x = age in years of a rural Quebec woman at the time of her first marriage. In the year 1941, the population variance of x was approximately $\sigma^2 = 5.1$. Suppose a recent study of age at first marriage for a random sample of 41 women in rural Quebec gave a sample variance $s^2 = 3.3$. Use a 5% level of significance to test the claim that the current variance is less than 5.1. Find a 90% confidence interval for the population variance.

5. *Mountain Climbing: Accidents* The following problem is based on information taken from *Accidents in North American Mountaineering* (jointly published by The American Alpine Club and The Alpine Club of Canada). Let x represent the number of mountain climbers killed each year. The long-term variance of x is approximately $\sigma^2 = 136.2$. Suppose that for the past 8 years, the variance has been $s^2 = 115.1$. Use a 1% level of significance to test the claim that the recent variance for number of mountain climber deaths is less than 136.2. Find a 90% confidence interval for the population variance.

6. *Professors: Salaries* The following problem is based on information taken from *Academe, Bulletin of the American Association of University Professors*. Let x represent the average annual salary of college and university professors (in thousands of dollars) in the United States. For all colleges and universities in the United States, the population variance of x is approximately $\sigma^2 = 47.1$. However, a random sample of 15 colleges and universities in Kansas showed that x has a sample variance $s^2 = 83.2$. Use a 5% level of significance to test the claim that the variance for colleges and universities in Kansas is greater than 47.1. Find a 95% confidence interval for the population variance.

7. *Medical: Clinical Test* A new kind of typhoid shot is being developed by a medical research team. The old typhoid shot was known to protect the population for a mean time of 36 months, with a standard deviation of 3 months. To test the time variability of the new shot, a random sample of 23 people were given the new shot. Regular blood tests showed that the sample standard deviation of protection times was 1.9 months. Using a 0.05 level of significance, test the claim that the new typhoid shot has a smaller variance of protection times. Find a 90% confidence interval for the population standard deviation.

8. ***Veterinary Science: Tranquilizer*** Jim Mead is a veterinarian who visits a Vermont farm to examine prize bulls. In order to examine a bull, Jim first gives the animal a tranquilizer shot. The effect of the shot is supposed to last an average of 65 minutes, and it usually does. However, Jim sometimes gets chased out of the pasture by a bull that recovers too soon, and other times he becomes worried about prize bulls that take too long to recover. By reading journals, Jim has found that the tranquilizer should have a mean duration time of 65 minutes, with a standard deviation of 15 minutes. A random sample of 10 of Jim's bulls had a mean tranquilized duration time of close to 65 minutes but a standard deviation of 24 minutes. At the 1% level of significance, is Jim justified in the claim that the variance is larger than that stated in his journal? Find a 95% confidence interval for the population standard deviation.

9. ***Engineering: Jet Engines*** The fan blades on commercial jet engines must be replaced when wear on these parts indicates too much variability to pass inspection. If a single fan blade broke during operation, it could severely endanger a flight. A large engine contains thousands of fan blades, and safety regulations require that variability measurements on the population of all blades not exceed $\sigma^2 = 0.18$ mm^2. An engine inspector took a random sample of 61 fan blades from an engine. She measured each blade and found a sample variance of 0.27 mm^2. Using a 0.01 level of significance, is the inspector justified in claiming that all the engine fan blades must be replaced? Find a 90% confidence interval for the population standard deviation.

10. ***Law: Bar Exam*** A factor in determining the usefulness of an examination as a measure of demonstrated ability is the amount of spread that occurs in the grades. If the spread or variation of examination scores is very small, it usually means that the examination was either too hard or too easy. However, if the variance of scores is moderately large, then there is a definite difference in scores between "better," "average," and "poorer" students. A group of attorneys in a Midwest state has been given the task of making up this year's bar examination for the state. The examination has 500 total possible points, and from the history of past examinations, it is known that a standard deviation of around 60 points is desirable. Of course, too large or too small a standard deviation is not good. The attorneys want to test their examination to see how good it is. A preliminary version of the examination (with slight modifications to protect the integrity of the real examination) is given to a random sample of 24 newly graduated law students. Their scores give a sample standard deviation of 72 points.
    (i) Using a 0.01 level of significance, test the claim that the population standard deviation for the new examination is 60 against the claim that the population standard deviation is different from 60.
    (ii) Find a 99% confidence interval for the population variance.
    (iii) Find a 99% confidence interval for the population standard deviation.

11. ***Engineering: Solar Batteries*** A set of solar batteries is used in a research satellite. The satellite can run on only one battery, but it runs best if more than one battery is used. The variance σ^2 of lifetimes of these batteries affects the useful lifetime of the satellite before it goes dead. If the variance is too small, all the batteries will tend to die at once. Why? If the variance is too large, the batteries are simply not dependable. Why? Engineers have determined that a variance of $\sigma^2 = 23$ months (squared) is most desirable for these batteries. A random sample of 22 batteries gave a sample variance of 14.3 months (squared).
    (i) Using a 0.05 level of significance, test the claim that $\sigma^2 = 23$ against the claim that σ^2 is different from 23.
    (ii) Find a 90% confidence interval for the population variance σ^2.
    (iii) Find a 90% confidence interval for the population standard deviation σ.

PART II: INFERENCES USING THE *F* DISTRIBUTION

Overview of the *F* Distribution

The F probability distribution was first developed by the English statistician Sir Ronald Fisher (1890–1962). Fisher had a long and distinguished career in statistics, including research work at the agricultural station at Rothamsted (see Problems 5 and 6 at the end of Section 10.4). During his time there he developed the subjects of experimental design and ANOVA (see Sections 10.5 and 10.6).

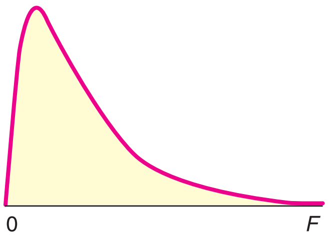
The F distribution is a ratio of two independent chi-square random variables, each with its own degrees of freedom,

$d.f._N$ = degrees of freedom in the numerator
$d.f._D$ = degrees of freedom in the denominator

The F distribution depends on these *two* degrees of freedom, $d.f._N$ and $d.f._D$. Figure 10-13 shows a typical F distribution.

FIGURE 10-13

Typical F Distribution ($d.f._N = 4$, $d.f._D = 7$)



Properties of the *F* distribution

- The F distribution is not symmetrical. It is skewed to the right.
- Values of F are always greater than or equal to zero.
- A specific F distribution (see Table 8 of Appendix II) is determined from *two* degrees of freedom. These are called *degrees of freedom for the numerator* $d.f._N$ and *degrees of freedom for the denominator* $d.f._D$.
- Area under the entire F distribution is one.

The degrees of freedom used in the F distribution depend on the particular application. Table 8 of Appendix II shows areas in the right-tail of different F distributions according to the degrees of freedom in both the numerator, $d.f._N$, and the denominator, $d.f._D$.

Table 10-14 shows an excerpt from Table 8. Notice that $d.f._D$ are row headers. For each $d.f._D$, right-tail areas from 0.100 down to 0.001 are provided in the next column. Then, under column headers for $d.f._N$ values of F are given corresponding to $d.f._D$, the right-tail area, and $d.f._N$. For example, for $d.f._D = 2$, Right-tail area $= 0.010$, and $d.f._N = 3$, the corresponding value of F is 99.17

In this text we present three applications of the F distribution: testing two variances (Section 10.4), one-way ANOVA (Section 10.5), and two-way ANOVA (Section 10.6). Sections 10.4 and 10.5 are self-contained and can be studied independently. Section 10.5 should be studied before Section 10.6.

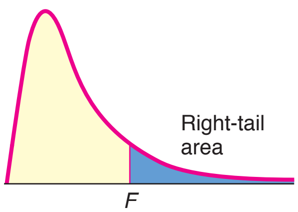


TABLE 10-14 **Excerpt from Table 8 (Appendix II): The *F* Distribution**

		Degrees of Freedom for Numerator $d.f._N$			
$d.f._D$	Right-tail Area	1	2	3	4 ...
⋮	⋮	⋮	⋮	⋮	⋮
	0.100	8.53	9.00	9.16	9.24
	0.050	18.51	19.00	19.16	19.25
✓ 2	0.025	38.51	39.00	39.17	39.25
	0.010	98.50	99.00	99.17	99.25
	0.001	998.50	999.00	999.17	9999.25

SECTION 10.4

Testing Two Variances

FOCUS POINTS

- Set up a test for two variances σ_1^2 and σ_2^2.
- Use sample variances to compute the sample F statistic.
- Use the F distribution to estimate a P-value and conclude the test.

In this section, we present a method for testing two variances (or, equivalently, two standard deviations). We use *independent random samples* from two populations to test the claim that the population variances are equal. The concept of variation among data is very important, and there are many possible applications in science, industry, business administration, social science, and so on.

In Section 10.3, we tested a *single* variance. The main mathematical tool we used was the chi-square probability distribution. In this section, the main tool is the F probability distribution.

Let us begin by stating what we need to assume for a test of two population variances.

Basic requirements

- The two populations are independent of each other. Recall from Section 8.5 that two sampling distributions are *independent* if there is no relation whatsoever between specific values of the two distributions.
- The two populations each have a *normal* probability distribution. This is important because the test we will use is sensitive to changes away from normality.

Setup for test

Now that we know the basic requirements, let's consider the setup.

How to Set Up the Test

STEP 1: Get Two Independent Random Samples, One from Each Population

We use the following notation:

Population I (larger s^2)	Population II (smaller s^2)
n_1 = sample size	n_2 = sample size
s_1^2 = sample variance	s_2^2 = sample variance
σ_1^2 = population variance	σ_2^2 = population variance

To simplify later discussion, we make the notational choice that

$$s_1^2 \geq s_2^2$$

This means that we *define* population I as the population with the *larger* (or equal, as the case may be) sample variance. This is only a notational convention and does not affect the general nature of the test.

STEP 2: Set Up the Hypotheses

Hypothesis tests about two variances

The null hypothesis will be that we have equal population variances.

$$H_0: \sigma_1^2 = \sigma_2^2$$

Reflecting on our notation setup, it makes sense to use an alternate hypothesis, either

$$H_1: \sigma_1^2 \neq \sigma_2^2 \qquad \text{or} \qquad H_1: \sigma_1^2 > \sigma_2^2$$

Notice that the test makes claims about variances. However, we can also use it for corresponding claims about standard deviations.

Hypotheses about Variances	Equivalent Hypotheses about Standard Deviations
$H_0: \sigma_1^2 = \sigma_2^2$	$H_0: \sigma_1 = \sigma_2$
$H_1: \sigma_1^2 \neq \sigma_2^2$	$H_1: \sigma_1 \neq \sigma_2$
$H_1: \sigma_1^2 > \sigma_2^2$	$H_1: \sigma_1 > \sigma_2$

STEP 3: Compute the Sample Test Statistic

Sample test statistic *F*

$$F = \frac{s_1^2}{s_2^2}$$

For two normally distributed populations with equal variances ($H_0: \sigma_1^2 = \sigma_2^2$), the sampling distribution we will use is the *F distribution* (see Table 8 of Appendix II).

The *F* distribution depends on *two* degrees of freedom.

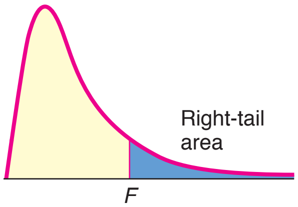
For **tests of two variances,** it can be shown that

$$d.f._N = n_1 - 1 \qquad \text{and} \qquad d.f._D = n_2 - 1$$

STEP 4: Find (or Estimate) the *P*-value of the Sample Test Statistic

Use the *F* distribution (Table 8 of Appendix II) to find the *P*-value of the sample test statistic. You need to know the degrees of freedom for the numerator, $d.f._N = n_1 - 1$, and the degrees of freedom for the denominator, $d.f._D = n_2 - 1$. Find the block of entries with your $d.f._D$ as row header and your $d.f._N$ as column header. Within that block of values, find the position of the sample test statistic *F*. Then find the corresponding right-tail area. For instance, using Table 10-15 (Excerpt from Table 8), we see that for $d.f._D = 2$ and $d.f._N = 3$, sample $F = 55.2$ lies between 39.17 and 99.17, with corresponding right-tail areas of 0.025 and 0.010. The interval containing the *P*-value for $F = 55.2$ is $0.010 < P\text{-value} < 0.025$.

TABLE 10-15 Excerpt from Table 8 (Appendix II): The *F* Distribution



$d.f._D$	Right-tail Area	Degrees of Freedom for Numerator $d.f._N$: 1	2	3	4 . . .
⋮	⋮	⋮	⋮	⋮	⋮
	0.100	8.53	9.00	9.16	9.24
	0.050	18.51	19.00	19.16	19.25
✓ 2	0.025	38.51	39.00	39.17	39.25
	0.010	98.50	99.00	99.17	99.25
	0.001	998.50	999.00	999.17	9999.25

TABLE 10-16 ***P*-values for Testing Two Variances (Table 8, Appendix II)**

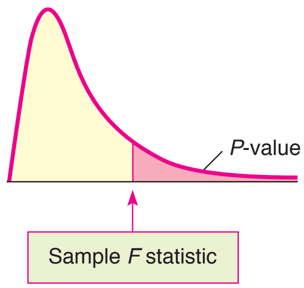
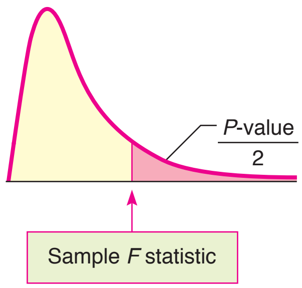
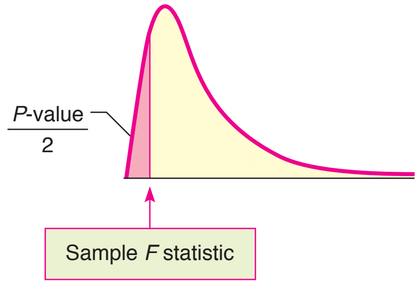
(a) Right-tailed test Since the *F*-distribution table gives right-tail probabilities, you can use the table directly to find or estimate the *P*-value.	
(b) Two-tailed test Remember that the *P*-value is the probability of getting a test statistic as extreme as or more extreme than the test statistic computed from the sample. For a two-tailed test, we need to account for corresponding equal areas in *both* the upper and lower tails. This means that in each tail, we have an area of (*P*-value)/2. The total *P*-value is then $P\text{-value} = 2\left(\frac{P\text{-value}}{2}\right)$	Be sure to choose the area in the appropriate tail (left or right) so that $\frac{P\text{-value}}{2} \le 0.5$.

Table 10-16 gives a summary for computing the *P*-value for both right-tailed and two-tailed tests for two variances.

Now that we have steps 1 to 4 as an outline, let's look at a specific example.

EXAMPLE 7 Testing two variances

Prehistoric Native Americans smoked pipes for ceremonial purposes. Most pipes were either carved-stone pipes or ceramic pipes made from clay. Clay pipes were easier to make, whereas stone pipes required careful drilling using hollow-core-bone drills and special stone reamers. An anthropologist claims that because clay pipes were easier to make, they show a greater variance in their construction. We want to test this claim using a 5% level of significance. Data for this example are taken from the Wind Mountain Archaeological Region (Source: *Mimbres Mogollon Archaeology* by A. I. Woosley and A. J. McIntyre, University of New Mexico Press). Assume the diameters of each type of pipe are normally distributed.

Ceramic Pipe Bowl Diameters (cm)

1.7	5.1	1.4	0.7	2.5	4.0
3.8	2.0	3.1	5.0	1.5	

Stone Pipe Bowl Diameters (cm)

1.6	2.1	3.1	1.4	2.2	2.1
2.6	3.2	3.4			

SOLUTION:

(a) *Check requirements* Assume that the pipe bowl diameters follow normal distributions and that the given data make up independent random samples of pipe measurements taken from archaeological excavations at Wind Mountain.

Use a calculator to verify the following:

Population I: Ceramic Pipes	Population II: Stone Pipes
$n_1 = 11$	$n_2 = 9$
$s_1^2 \approx 2.266$	$s_2^2 \approx 0.504$
σ_1^2 = population variance	σ_2^2 = population variance

Richard A. Cooke/CORBIS

Note: Because the sample variance for ceramic pipes (2.266) is larger than the sample variance for stone pipes (0.504), we designate population I as ceramic pipes.

(b) Set up the null and alternate hypotheses.

H_0: $\sigma_1^2 = \sigma_2^2$ (or the equivalent, $\sigma_1 = \sigma_2$)
H_1: $\sigma_1^2 > \sigma_2^2$ (or the equivalent, $\sigma_1 > \sigma_2$)

The null hypothesis states that there is no difference. The alternate hypothesis supports the anthropologist's claim that clay pipes have a larger variance.

(c) The sample test statistic is

$$F = \frac{s_1^2}{s_2^2} \approx \frac{2.266}{0.504} \approx 4.496$$

Now, if $\sigma_1^2 = \sigma_2^2$, then s_1^2 and s_2^2 also should be close in value. If this were the case, $F = s_1^2/s_2^2 \approx 1$. However, if $\sigma_1^2 > \sigma_2^2$, then we see that the sample statistic $F = s_1^2/s_2^2$ should be larger than 1.

(d) Find an interval containing the P-value for $F = 4.496$.
This is a right-tailed test (see Figure 10-14) with degrees of freedom

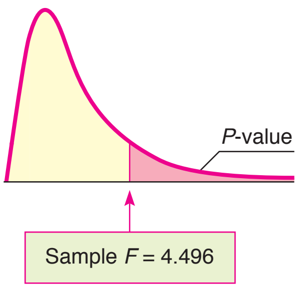
$$d.f._N = n_1 - 1 = 11 - 1 = 10 \quad \text{and} \quad d.f._D = n_2 - 1 = 9 - 1 = 8$$

The interval containing the P-value is

$$0.010 < P\text{-value} < 0.025$$

FIGURE 10-14

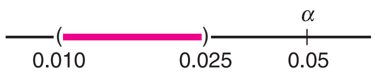
P-value



	Right-tail Area	$d.f._N$ 10
	0.100	2.54
$d.f._D$	0.050	3.35
✓ 8	0.025	4.30
	0.010	5.81
	0.001	11.54

(e) Conclude the test and *interpret* the results.

Since the P-value is less than $\alpha = 0.05$, we reject H_0.



Technology gives P-value ≈ 0.0218. At the 5% level of significance, the evidence is sufficient to conclude that the variance for the ceramic pipes is larger.

We summarize the steps involved in testing two variances with the following procedure.

PROCEDURE

How to Test Two Variances σ_1^2 and σ_2^2

Requirements

Assume that x_1 and x_2 are random variables that have *independent normal distributions* with unknown variances σ_1^2 and σ_2^2. Next, you need independent random samples of x_1 values and x_2 values, from which you compute sample variances s_1^2 and s_2^2. Use samples of sizes n_1 and n_2, respectively, with both samples of size at least 2. Without loss of generality, we may assume the notational setup is such that $s_1^2 \geq s_2^2$.

Procedure

1. Set the *level of significance* α. Use the *null hypothesis* H_0: $\sigma_1^2 = \sigma_2^2$. In the context of the problem, choose the *alternate hypothesis* to be H_1: $\sigma_1^2 > \sigma_2^2$ or H_1: $\sigma_1^2 \neq \sigma_2^2$.
2. Compute the *sample test statistic*

$$F = \frac{s_1^2}{s_2^2}$$

where $d.f._N = n_1 - 1$ (degrees of freedom numerator)

$d.f._D = n_2 - 1$ (degrees of freedom denominator)

3. Use the F distribution and the type of test to find or estimate the *P-value*. Use Table 8 of Appendix II and the procedure shown in Table 10-16.
4. *Conclude* the test. If P-value $\leq \alpha$, then reject H_0. If P-value $> \alpha$, then do not reject H_0.
5. *Interpret your conclusion* in the context of the application.

GUIDED EXERCISE 10 **Testing two variances**

A large variance in blood chemistry components can result in health problems as the body attempts to return to equilibrium. J. B. O'Sullivan and C. M. Mahan conducted a study reported in the *American Journal of Clinical Nutrition* (Vol. 19, pp. 345–351) that concerned the glucose (blood sugar) levels of pregnant and nonpregnant women at Boston City Hospital. For both groups, a fasting (12-hour fast) blood glucose test was done. The following data are in units of milligrams of glucose per 100 milliliters of blood. Assume blood glucose levels for each group are normally distributed.

Continued

GUIDED EXERCISE 10 *continued*

Glucose Test: Nonpregnant Women

73	61	104	75	85	65	62	98	92	106

Glucose Test: Pregnant Women

72	84	90	95	66	70	79	85

Medical researchers question if the variance of the glucose test results for nonpregnant women is *different* (either way) from the variance for pregnant women. Let's conduct a test using a 5% level of significance.

(a) ***Check requirements*** What assumptions must be made about the two populations and the samples?

→ The population measurements must follow independent normal distributions. The samples must be random samples from each population.

(b) Use a calculator to compute the sample variance for each data group. Then complete the following:

Population I	Population II
n_1 = _____	n^2 = _____
s_1^2 = _____	s_2^2 = _____

→ Recall that we choose our notation so that population I has the *larger* sample variance.

Population I	Population II
$n_1 = 10$	$n_2 = 8$
$s_1^2 \approx 298.25$	$s_2^2 \approx 103.84$

(c) What is α? State the null and alternate hypotheses.

→ $\alpha = 0.05$; H_0: $\sigma_1^2 = \sigma_2^2$; H_1: $\sigma_1^2 \neq \sigma_2^2$

(d) Compute the sample test statistic F, $d.f._N$, and $d.f._D$.

→ $$F = \frac{s_1^2}{s_2^2} \approx \frac{298.25}{103.84} \approx 2.87$$

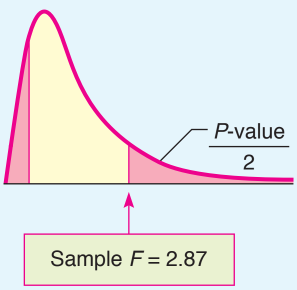
$d.f._N = n_1 - 1 = 10 - 1 = 9$

$d.f._D = n_2 - 1 = 8 - 1 = 7$

(e) Estimate the P-value.

→ Because this is a two-tailed test, we look up the area to the right of $F = 2.87$ and double it.

FIGURE 10-15 *P*-value



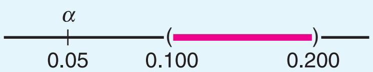
	Right-tail Area	$d.f._N$ 9
	0.100	2.72
$d.f._D$	0.050	3.68
✓ 7	0.025	4.82
	0.010	6.72
	0.001	14.33

$$0.050 < \frac{P\text{-value}}{2} < 0.100$$

$0.100 < P\text{-value} < 0.200$

(f) Conclude the test.

→ Since the P-value is greater than $\alpha = 0.05$, we do not reject H_0.



Technology gives P-value ≈ 0.1780.

(g) ***Interpret*** the results.

→ At the 5% level of significance, the evidence is insufficient to reject the claim of equal variances.

TECH NOTES The TI-84Plus/TI-83Plus/TI-*n*spire calculators support tests of two variances. Use **STAT**, **TESTS**, and option **D:2-SampFTest**. For results consistent with the notational convention that the larger variance goes in the numerator of the sample *F* statistic, put the data with the larger variance in **List1.**

Minitab Release 15 of Minitab supports tests of two variances. Use menu choices **Stat ➤ *Basic Statistics* ➤ 2 Variances.**

Variety is said to be the spice of life. However, in statistics, when we want to compare two populations, we will often need the assumption that the population variances are the same. As long as the two populations follow normal distributions, we can use the methods of this section and random samples from the populations to determine if the assumption of equal variances is reasonable at a given level of significance.

VIEWPOINT Mercury in Bass?

Largemouth bass were studied in 53 different lakes to examine factors that influence the level of mercury contamination. In many cases, the contamination was fairly low, except for older (trophy) fish, in which the mercury levels were much higher. Using information you have learned in this section, you can test variances of mercury contamination for different lakes and/or different regions. For more information, see the article by Lange, Royals, and Connor, "Transactions of the American Fisheries Society." To find this article, visit the Brase/Brase statistics site at **http://www.cengage.com/statistics/brase** *and find a link to DASL, the Carnegie Mellon University Data and Story Library. From the DASL site, link to Nature under Data Subjects, and select Mercury in the Bass Data file.*

SECTION 10.4 PROBLEMS

1. *Statistical Literacy* When using the *F* distribution to test variances from two populations, should the random variables from each population be independent or dependent?
2. *Statistical Literacy* When using the *F* distribution to test two variances, is it essential that each of the two populations be normally distributed? Would it be all right if the populations had distributions that were mound-shaped and more or less symmetrical?
3. *Statistical Literacy* In general, is the *F* distribution symmetrical? Can values of *F* be negative?
4. *Statistical Literacy* To use the *F* distribution, what degrees of freedom need to be calculated?

For Problems 5–12, please provide the following information.
(a) What is the level of significance? State the null and alternate hypotheses.
(b) Find the value of the sample *F* statistic. What are the degrees of freedom? What assumptions are you making about the original distribution?
(c) Find or estimate the *P*-value of the sample test statistic.
(d) Based on your answers in parts (a) to (c), will you reject or fail to reject the null hypothesis?
(e) *Interpret* your conclusion in the context of the application.

Assume that the data values in each problem come from independent populations and that each population follows a normal distribution.

5. ***Agriculture: Wheat*** Rothamsted Experimental Station (England) has studied wheat production since 1852. Each year, many small plots of equal size but different soil/fertilizer conditions are planted with wheat. At the end of the growing season, the yield (in pounds) of the wheat on the plot is measured. The following data are based on information taken from an article by G. A. Wiebe in the *Journal of Agricultural Research* (Vol. 50, pp. 331–357). For a random sample of years, one plot gave the following annual wheat production (in pounds):

4.15	4.21	4.27	3.55	3.50	3.79	4.09	4.42
3.89	3.87	4.12	3.09	4.86	2.90	5.01	3.39

Use a calculator to verify that, for this plot, the sample variance is $s^2 \approx 0.332$.

Another random sample of years for a second plot gave the following annual wheat production (in pounds):

4.03	3.77	3.49	3.76	3.61	3.72	4.13	4.01
3.59	4.29	3.78	3.19	3.84	3.91	3.66	4.35

Use a calculator to verify that the sample variance for this plot is $s^2 \approx 0.089$.

Test the claim that the population variance of annual wheat production for the first plot is larger than that for the second plot. Use a 1% level of significance.

6. ***Agriculture: Wheat*** Two plots at Rothamsted Experimental Station (see reference in Problem 5) were studied for production of wheat straw. For a random sample of years, the annual wheat straw production (in pounds) from one plot was as follows:

6.17	6.05	5.89	5.94	7.31	7.18
7.06	5.79	6.24	5.91	6.14	

Use a calculator to verify that, for the preceding data, $s^2 \approx 0.318$.

Another random sample of years for a second plot gave the following annual wheat straw production (in pounds):

6.85	7.71	8.23	6.01	7.22	5.58	5.47	5.86

Use a calculator to verify that, for these data, $s^2 \approx 1.078$.

Test the claim that there is a difference (either way) in the population variance of wheat straw production for these two plots. Use a 5% level of significance.

7. ***Economics: Productivity*** An economist wonders if corporate productivity in some countries is more *volatile* than that in other countries. One measure of a company's productivity is annual percentage yield based on total company assets. Data for this problem are based on information taken from *Forbes Top Companies*, edited by J. T. Davis. A random sample of leading companies in France gave the following percentage yields based on assets:

4.4	5.2	3.7	3.1	2.5	3.5	2.8	4.4	5.7	3.4	4.1
6.8	2.9	3.2	7.2	6.5	5.0	3.3	2.8	2.5	4.5	

Use a calculator to verify that $s^2 \approx 2.044$ for this sample of French companies.

Another random sample of leading companies in Germany gave the following percentage yields based on assets:

3.0	3.6	3.7	4.5	5.1	5.5	5.0	5.4	3.2
3.5	3.7	2.6	2.8	3.0	3.0	2.2	4.7	3.2

Use a calculator to verify that $s^2 \approx 1.038$ for this sample of German companies.

Test the claim that there is a difference (either way) in the population variance of percentage yields for leading companies in France and Germany. Use a 5% level of significance. How could your test conclusion relate to the economist's question regarding *volatility* (data spread) of corporate productivity of large companies in France compared with large companies in Germany?

8. *Economics: Productivity* A random sample of leading companies in South Korea gave the following percentage yields based on assets (see reference in Problem 7):

2.5	2.0	4.5	1.8	0.5	3.6	2.4
0.2	1.7	1.8	1.4	5.4	1.1	

Use a calculator to verify that $s^2 = 2.247$ for these South Korean companies.

Another random sample of leading companies in Sweden gave the following percentage yields based on assets:

2.3	3.2	3.6	1.2	3.6	2.8	2.3	3.5	2.8

Use a calculator to verify that $s^2 = 0.624$ for these Swedish companies.

Test the claim that the population variance of percentage yields on assets for South Korean companies is higher than that for companies in Sweden. Use a 5% level of significance. How could your test conclusion relate to an economist's question regarding *volatility* of corporate productivity of large companies in South Korea compared with that in Sweden?

9. *Investing: Mutual Funds* You don't need to be rich to buy a few shares in a mutual fund. The question is, "How *reliable* are mutual funds as investments?" That depends on the type of fund you buy. The following data are based on information taken from *Morningstar,* a mutual fund guide available in most libraries. A random sample of percentage annual returns for mutual funds holding stocks in aggressive-growth small companies is shown below.

−1.8	14.3	41.5	17.2	−16.8	4.4	32.6	−7.3	16.2	2.8	34.3
−10.6	8.4	−7.0	−2.3	−18.5	25.0	−9.8	−7.8	−24.6	22.8	

Use a calculator to verify that $s^2 \approx 348.43$ for the sample of aggressive-growth small company funds.

Another random sample of percentage annual returns for mutual funds holding value (i.e., market underpriced) stocks in large companies is shown below.

16.2	0.3	7.8	−1.6	−3.8	19.4	−2.5	15.9	32.6	22.1	3.4
−0.5	−8.3	25.8	−4.1	14.6	6.5	18.0	21.0	0.2	−1.6	

Use a calculator to verify that $s^2 \approx 137.31$ for value stocks in large companies.

Test the claim that the population variance for mutual funds holding aggressive-growth small stocks is larger than the population variance for mutual funds holding value stocks in large companies. Use a 5% level of significance. How could your test conclusion relate to the question of *reliability* of returns for each type of mutual fund?

10. *Investing: Mutual Funds* How *reliable* are mutual funds that invest in bonds? Again, this depends on the bond fund you buy (see reference in Problem 9). A random sample of annual percentage returns for mutual funds holding short-term U.S. government bonds is shown below.

4.6	4.7	1.9	9.3	−0.8	4.1	10.5
4.2	3.5	3.9	9.8	−1.2	7.3	

Use a calculator to verify that $s^2 \approx 13.59$ for the preceding data.

A random sample of annual percentage returns for mutual funds holding intermediate-term corporate bonds is shown below.

−0.8	3.6	20.2	7.8	−0.4	18.8	−3.4	10.5
8.0	−0.9	2.6	−6.5	14.9	8.2	18.8	14.2

Use a calculator to verify that $s^2 \approx 72.06$ for returns from mutual funds holding intermediate-term corporate bonds.

Use $\alpha = 0.05$ to test the claim that the population variance for annual percentage returns of mutual funds holding short-term government bonds is different from the population variance for mutual funds holding intermediate-term corporate bonds. How could your test conclusion relate to the question of *reliability* of returns for each type of mutual fund?

11. ***Engineering: Fuel Injection*** A new fuel injection system has been engineered for pickup trucks. The new system and the old system both produce about the same average miles per gallon. However, engineers question which system (old or new) will give better *consistency* in fuel consumption (miles per gallon) under a variety of driving conditions. A random sample of 31 trucks were fitted with the new fuel injection system and driven under different conditions. For these trucks, the sample variance of gasoline consumption was 58.4. Another random sample of 25 trucks were fitted with the old fuel injection system and driven under a variety of different conditions. For these trucks, the sample variance of gasoline consumption was 31.6. Test the claim that there is a difference in population variance of gasoline consumption for the two injection systems. Use a 5% level of significance. How could your test conclusion relate to the question regarding the *consistency* of fuel consumption for the two fuel injection systems?

12. ***Engineering: Thermostats*** A new thermostat has been engineered for the frozen food cases in large supermarkets. Both the old and the new thermostats hold temperatures at an average of 25°F. However, it is hoped that the new thermostat might be more *dependable* in the sense that it will hold temperatures closer to 25°F. One frozen food case was equipped with the new thermostat, and a random sample of 21 temperature readings gave a sample variance of 5.1. Another, similar frozen food case was equipped with the old thermostat, and a random sample of 16 temperature readings gave a sample variance of 12.8. Test the claim that the population variance of the old thermostat temperature readings is larger than that for the new thermostat. Use a 5% level of significance. How could your test conclusion relate to the question regarding the *dependability* of the temperature readings?

SECTION 10.5

One-Way ANOVA: Comparing Several Sample Means

FOCUS POINTS

- Learn about the risk α of a type I error when we test several means at once.
- Learn about the notation and setup for a one-way ANOVA test.
- Compute mean squares between groups and within groups.
- Compute the sample F statistic.
- Use the F distribution to estimate a P-value and conclude the test.

In our past work, to determine the existence (or nonexistence) of a significant difference between population means, we restricted our attention to only two data groups representing the means in question. Many statistical applications in psychology, social science, business administration, and natural science involve many means and many data groups. Questions commonly asked are: Which of *several* alternative methods yields the best results in a particular setting? Which of *several* treatments leads to the highest incidence of patient recovery? Which of *several* teaching methods leads to greatest student retention? Which of *several* investment schemes leads to greatest economic gain?

Using our previous methods (Sections 8.4 and 8.5) of comparing only *two* means would require many tests of significance to answer the preceding questions. For example, even if we had only 5 variables, we would be required to perform 10 tests of significance in order to compare each variable to each of the other variables. If we had the time and patience, we could perform all 10 tests, but what about the risk of accepting a difference where there really is no difference (a type I error)? If the risk of a type I error on each test is $\alpha = 0.05$, then on 10 tests we expect the number of tests with a type I error to be 10(0.05), or 0.5 (see expected value, Section 5.3). This situation may not seem too serious to you, but remember that in a "real-world" problem and with the aid of a high-speed computer, a researcher may want to study the effect of 50 variables on the outcome of an experiment. Using a little mathematics, we can show that the study would require 1225 separate tests to check *each pair* of variables for a significant difference of means. At the $\alpha = 0.05$ level of significance for each test, we could expect (1225)(0.05), or 61.25, of the tests to have a type I error. In other words, these 61.25 tests would say that there are differences between means when there really are no differences.

Introduction

ANOVA

To avoid such problems, statisticians have developed a method called *analysis of variance* (abbreviated *ANOVA*). We will study single-factor analysis of variance (also called *one-way ANOVA*) in this section and two-way ANOVA in Section 10.6. With appropriate modification, methods of single-factor ANOVA generalize to *n*-dimensional ANOVA, but we leave that topic to more advanced studies.

EXAMPLE 8 One-way ANOVA test

Explanation of notation and formulas

A psychologist is studying the effects of dream deprivation on a person's anxiety level during waking hours. Brain waves, heart rate, and eye movements can be used to determine if a sleeping person is about to enter into a dream period. Three groups of subjects were randomly chosen from a large group of college students who volunteered to participate in the study. Group I subjects had their sleep interrupted four times each night, but never during or immediately before a dream. Group II subjects had their sleep interrupted four times also, but on two occasions they were wakened at the onset of a dream. Group III subjects were wakened four times, each time at the onset of a dream. This procedure was repeated for 10 nights, and each day all subjects were given a test to determine their levels of anxiety. The data in Table 10-17 record the total of the test scores for each person over the entire project. Higher totals mean higher anxiety levels.

TABLE 10-17 Dream Deprivation Study

Group I $n_1 = 6$ Subjects		Group II $n_2 = 7$ Subjects		Group III $n_3 = 5$ Subjects	
x_1	x_1^2	x_2	x_2^2	x_3	x_3^2
9	81	10	100	15	225
7	49	9	81	11	121
3	9	11	121	12	144
6	36	10	100	9	81
5	25	7	49	10	100
8	64	6	36		
		8	64		
$\Sigma x_1 = 38$	$\Sigma x_1^2 = 264$	$\Sigma x_2 = 61$	$\Sigma x_2^2 = 551$	$\Sigma x_3 = 57$	$\Sigma x_3^2 = 671$

$N = n_1 + n_2 + n_3 = 18$
$\Sigma x_{TOT} = \Sigma x_1 + \Sigma x_2 + \Sigma x_3 = 156$
$\Sigma x_{TOT}^2 = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 = 1486$


Tanya Constantine/Blend Images/Jupiter Images

From Table 10-17, we see that group I had $n_1 = 6$ subjects, group II had $n_2 = 7$ subjects, and group III had $n_3 = 5$ subjects. For each subject, the anxiety score (x value) and the square of the test score (x^2 value) are also shown. In addition, special sums are shown.

We will outline the procedure for single-factor ANOVA in six steps. Each step will contain general methods and rationale appropriate to all single-factor ANOVA tests. As we proceed, we will use the data of Table 10-17 for a specific reference example.

Basic requirements of ANOVA

Our application of ANOVA has three basic requirements. In a general problem with k groups:

1. We require that each of our k groups of measurements is obtained from a population with a *normal* distribution.
2. Each group is randomly selected and is *independent* of all other groups. In particular, this means that we will not use the same subjects in more than one group and that the scores of one subject will not have an effect on the scores of another subject.
3. We assume that the variables from each group come from distributions with approximately the *same standard deviation.*

STEP 1: Determine the Null and Alternate Hypotheses

The purpose of an ANOVA test is to determine the existence (or nonexistence) of a statistically significant difference *among* the group means. In a general problem with k groups, we call the (population) mean of the first group μ_1, the population mean of the second group μ_2, and so forth. The null hypothesis is simply that *all* the group population means are the same. Since our basic requirements state that each of the k groups of measurements comes from normal, independent distributions with common standard deviation, the null hypothesis states that all the sample groups come from *one and the same* population. The alternate hypothesis is that *not all* the group population means are equal. Therefore, in a problem with k groups, we have

Hypotheses

Hypotheses for one-way ANOVA

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

H_1: At least two of the means $\mu_1, \mu_2, \ldots, \mu_k$ are not equal.

Notice that the alternate hypothesis claims that *at least* two of the means are not equal. If more than two of the means are unequal, the alternate hypothesis is, of course, satisfied.

In our dream problem, we have $k = 3$; μ_1 is the population mean of group I, μ_2 is the population mean of group II, and μ_3 is the population mean of group III. Therefore,

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

H_1: At least two of the means μ_1, μ_2, μ_3 are not equal.

We will test the null hypothesis using an $\alpha = 0.05$ level of significance. Notice that only one test is being performed even though we have $k = 3$ groups and three corresponding means. Using ANOVA avoids the problem mentioned earlier of using multiple tests.

STEP 2: Find SS_{TOT}

Sum of squares, *SS*

The concept of *sum of squares* is very important in statistics. We used a sum of squares in Chapter 3 to compute the sample standard deviation and sample variance.

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} \quad \text{sample standard deviation}$$

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} \quad \text{sample variance}$$

The numerator of the sample variance is a special sum of squares that plays a central role in ANOVA. Since this numerator is so important, we give it the special name *SS* (for "sum of squares").

$$SS = \Sigma(x - \bar{x})^2 \tag{2}$$

Using some college algebra, it can be shown that the following, simpler formula is equivalent to Equation (2) and involves fewer calculations:

$$SS = \Sigma x^2 - \frac{(\Sigma x)^2}{n} \tag{3}$$

where n is the sample size.

In future references to *SS*, we will use Equation (3) because it is easier to use than Equation (2).

The **total sum of squares SS_{TOT}** can be found by using the entire collection of all data values in all groups:

$$SS_{TOT} = \Sigma x_{TOT}^2 - \frac{(\Sigma x_{TOT})^2}{N} \tag{4}$$

where $N = n_1 + n_2 + \cdots + n_k$ is the total sample size from all groups.

$$\Sigma x_{TOT} = \text{sum of all data} = \Sigma x_1 + \Sigma x_2 + \cdots + \Sigma x_k$$
$$\Sigma x_{TOT}^2 = \text{sum of all data squares} = \Sigma x_1^2 + \cdots + \Sigma x_k^2$$

Using the specific data given in Table 10-17 for the dream example, we have

$$k = 3 \quad \text{total number of groups}$$
$$N = n_1 + n_2 + n_3 = 6 + 7 + 5 = 18 \quad \text{total number of subjects}$$
$$\Sigma x_{TOT} = \text{total sum of } x \text{ values} = \Sigma x_1 + \Sigma x_2 + \Sigma x_3 = 38 + 61 + 57 = 156$$
$$\Sigma x_{TOT}^2 = \text{total sum of } x^2 \text{ values} = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 = 264 + 551 + 671 = 1486$$

Therefore, using Equation (4), we have

$$SS_{TOT} = \Sigma x_{TOT}^2 - \frac{(\Sigma x_{TOT})^2}{N} = 1486 - \frac{(156)^2}{18} = 134$$

The numerator for the total variation for all groups in our dream example is $SS_{TOT} = 134$. What interpretation can we give to SS_{TOT}? If we let $\bar{x}_{TOT}$ be the mean of all x values for all groups, then

$$\text{Mean of all } x \text{ values} = \bar{x}_{TOT} = \frac{\Sigma x_{TOT}}{N}$$

Under the null hypothesis (that all groups come from the same normal distribution), $SS_{TOT} = \Sigma(x_{TOT} - \bar{x}_{TOT})^2$ represents the numerator of the sample variance for all groups. Therefore, SS_{TOT} represents total variability of the data. Total variability can occur in two ways:

1. Scores may differ from one another because they belong to *different groups* with different means (recall that the alternate hypothesis states that the means are not all equal). This difference is called **between-group variability** and is denoted SS_{BET}.
2. Inherent differences unique to each subject and differences due to chance may cause a particular score to be different from the mean of its *own group*. This difference is called **within-group variability** and is denoted SS_W.

SS_{BET}

SS_W

Because total variability SS_{TOT} is the sum of between-group variability SS_{BET} and within-group variability SS_W, we may write

$$SS_{TOT} = SS_{BET} + SS_W$$

As we will see, SS_{BET} and SS_W are going to help us decide whether or not to reject the null hypothesis. Therefore, our next two steps are to compute these two quantities.

STEP 3: Find SS_{BET}

Sum of squares between groups

Recall that $\bar{x}_{TOT}$ is the mean of all x values from all groups. Between-group variability (SS_{BET}) measures the variability of group means. Because different groups may have different numbers of subjects, we must "weight" the variability contribution from each group by the group size n_i.

$$SS_{BET} = \sum_{\text{all groups}} n_i(\bar{x}_i - \bar{x}_{TOT})^2$$

where $n_i =$ sample size of group i
$\bar{x}_i =$ sample mean of group i
$\bar{x}_{TOT} =$ mean for values from all group

If we use algebraic manipulations, we can write the formula for SS_{BET} in the following computationally easier form:

Sum of squares between groups

$$SS_{BET} = \sum_{\text{all groups}} \left(\frac{(\Sigma x_i)^2}{n_i}\right) - \frac{(\Sigma x_{TOT})^2}{N} \tag{5}$$

where, as before, $N = n_1 + n_2 + \cdots + n_k$
$\Sigma x_i =$ sum of data in group i
$\Sigma x_{TOT} =$ sum of data from all groups

Using data from Table 10-17 for the dream example, we have

$$\begin{aligned} SS_{BET} &= \sum_{\text{all groups}} \left(\frac{(\Sigma x_i)^2}{n_i}\right) - \frac{(\Sigma x_{TOT})^2}{N} \\ &= \frac{(\Sigma x_1)^2}{n_1} + \frac{(\Sigma x_2)^2}{n_2} + \frac{(\Sigma x_3)^2}{n_3} - \frac{(\Sigma x_{TOT})^2}{N} \\ &= \frac{(38)^2}{6} + \frac{(61)^2}{7} + \frac{(57)^2}{5} - \frac{(156)^2}{18} \\ &= 70.038 \end{aligned}$$

Therefore, the numerator of the between-group variation is

$$SS_{BET} = 70.038$$

STEP 4: Find SS_W

Sum of squares within groups

We could find the value of SS_W by using the formula relating SS_{TOT} to SS_{BET} and SS_W and solving for SS_W:

$$SS_W = SS_{TOT} - SS_{BET}$$

However, we prefer to compute SS_W in a different way and to use the preceding formula as a check on our calculations.

SS_W is the numerator of the variation within groups. Inherent differences unique to each subject and differences due to chance create the variability assigned to SS_W. In a general problem with k groups, the variability within the ith group can be represented by

$$SS_i = \Sigma(x_i - \bar{x}_i)^2$$

or by the mathematically equivalent formula

$$SS_i = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n_i} \qquad (6)$$

Because SS_i represents the variation within the ith group and we are seeking SS_W, the variability within *all* groups, we simply add SS_i for all groups:

Sum of squares within groups

$$SS_W = SS_1 + SS_2 + \cdots + SS_k \qquad (7)$$

Using Equations (6) and (7) and the data of Table 10-17 with $k = 3$, we have

$$SS_1 = \Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n_1} = 264 - \frac{(38)^2}{6} = 23.333$$

$$SS_2 = \Sigma x_2^2 - \frac{(\Sigma x_2)^2}{n_2} = 551 - \frac{(61)^2}{7} = 19.429$$

$$SS_3 = \Sigma x_3^2 - \frac{(\Sigma x_3)^2}{n_3} = 671 - \frac{(57)^2}{5} = 21.200$$

$$SS_W = SS_1 + SS_2 + SS_3 = 23.333 + 19.429 + 21.200 = 63.962$$

Let us check our calculation by using SS_{TOT} and SS_{BET}.

$$SS_{TOT} = SS_{BET} + SS_W$$

$$134 = 70.038 + 63.962 \quad \text{(from steps 2 and 3)}$$

We see that our calculation checks.

STEP 5: Find Variance Estimates (Mean Squares)

Mean squares

In steps 3 and 4, we found SS_{BET} and SS_W. Although these quantities represent variability between groups and within groups, they are not yet the variance estimates we need for our ANOVA test. You may recall our study of the Student's t distribution, in which we introduced the concept of degrees of freedom. Degrees of freedom represent the number of values that are free to vary once we have placed certain restrictions on our data. In ANOVA, there are two types of degrees of freedom: $d.f._{BET}$, representing the degrees of freedom between groups, and $d.f._W$, representing degrees of freedom within groups. A theoretical discussion beyond the scope of this text would show

Degrees of freedom between and within groups

$d.f._{BET} = k - 1$ where k is the number of groups

$d.f._W = N - k$ where N is the total sample size

(*Note:* $d.f._{BET} + d.f._W = N - 1$.)

The variance estimates we are looking for are designated as follows:

MS_{BET}

MS_W

MS_{BET}, the variance between groups (read "mean square between")

MS_W, the variance within groups (read "mean square within")

In the literature of ANOVA, the variances between and within groups are usually referred to as *mean squares between* and *within* groups, respectively. We will use the mean-square notation because it is used so commonly. However, remember that the notations MS_{BET} and MS_W both refer to *variances*, and you might occasionally see the variance notations S^2_{BET} and S^2_W used for these quantities. The formulas for the variances between and within samples follow the pattern of the basic formula for sample variance.

$$\text{Sample variance} = s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{SS}{n - 1}$$

However, instead of using $n - 1$ in the denominator for MS_{BET} and MS_W variances, we use their respective degrees of freedom.

$$\text{Mean square between} = MS_{BET} = \frac{SS_{BET}}{d.f._{BET}} = \frac{SS_{BET}}{k - 1}$$

$$\text{Mean square within} = MS_W = \frac{SS_W}{d.f._W} = \frac{SS_W}{N - k}$$

Using these two formulas and the data of Table 10-17, we find the mean squares within and between variances for the dream deprivation example:

$$MS_{BET} = \frac{SS_{BET}}{k - 1} = \frac{70.038}{3 - 1} = 35.019$$

$$MS_W = \frac{SS_W}{N - k} = \frac{63.962}{18 - 3} = 4.264$$

STEP 6: Find the *F* Ratio and Complete the ANOVA Test

The logic of our ANOVA test rests on the fact that one of the variances, MS_{BET}, *can* be influenced by population differences among means of the several groups, whereas the other variance, MS_W, *cannot* be so influenced. For instance, in the dream deprivation and anxiety study, the variance between groups MS_{BET} will be affected if any of the treatment groups has a population mean anxiety score that is *different* from that of any other group. On the other hand, the variance within groups MS_W compares anxiety scores of each treatment group to its own group anxiety mean, and the fact that group means might differ *does not* affect the MS_W value.

Recall that the null hypothesis claims that all the groups are samples from populations having the *same* (normal) distributions. The alternate hypothesis states that at least two of the sample groups come from populations with *different* (normal) distributions.

If the *null* hypothesis is *true,* MS_{BET} and MS_W should both estimate the *same* quantity. Therefore, if H_0 is true, the *F ratio*

F ratio

Sample *F* statistic

> **Sample *F* ratio**
>
> $$F = \frac{MS_{BET}}{MS_W}$$
>
> The *F* ratio is the *sample test statistic F* for ANOVA tests.

should be approximately 1, and variations away from 1 should occur only because of sampling errors. The variance within groups MS_W is a good estimate of the overall population variance, but the variance between groups MS_{BET} consists of the population variance *plus* an additional variance stemming from the differences between samples. Therefore, if the *null* hypothesis is *false,* MS_{BET} will be larger than MS_W, and the *F* ratio will tend to be *larger* than 1.

The decision of whether or not to reject the null hypothesis is determined by the relative size of the *F* ratio. Table 8 of Appendix II gives *F* values.

For our example about dreams, the computed *F* ratio is

$$F = \frac{MS_{BET}}{MS_W} = \frac{35.019}{4.264} = 8.213$$

ANOVA test is right-tailed

Because large *F* values tend to discredit the null hypothesis, we use a *right-tailed test* with the *F distribution.* To find (or estimate) the *P*-value for the sample *F* statistic, we use the *F*-distribution table, Table 8 of Appendix II. The table requires us to know *degrees of freedom for the numerator* and *degrees of freedom for the denominator.*

Degrees of freedom for sample test statistic *F*

> **Degrees of freedom for sample test statistic *F* in one-way ANOVA**
>
> Degrees of freedom numerator $= d.f._N = d.f._{BET} = k - 1$
>
> Degrees of freedom denominator $= d.f._D = d.f._W = N - k$
>
> where k = number of groups
>
> N = total sample size across all groups

For our example about dreams,

$$d.f._N = k - 1 = 3 - 1 = 2 \qquad d.f._D = N - k = 18 - 3 = 15$$

Finding the *P*-value

Let's use the *F*-distribution table (Table 8, Appendix II) to find the *P*-value of the sample statistic $F = 8.213$. The *P*-value is a *right-tail area,* as shown in Figure 10-16. In Table 8, look in the block headed by column $d.f._N = 2$ and row $d.f._D = 15$. For convenience, the entries are shown in Table 10-18 (Excerpt from Table 8). We see that the sample $F = 8.213$ falls between the entries 6.36 and 11.34, with corresponding right-tail areas 0.010 and 0.001.

Test conclusion

The *P*-value is in the interval $0.001 < P\text{-value} < 0.010$. Since $\alpha = 0.05$, we see that the *P*-value is less than α and we reject H_0.

FIGURE 10-16

P-value

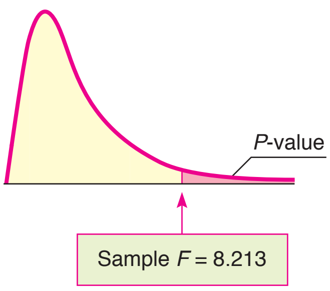


TABLE 10-18 **Excerpt from Table 8, Appendix II**

	Right-tail Area	$d.f._N$ 2
	0.100	2.70
$d.f._D$	0.050	3.68
✓ 15	0.025	4.77
	0.010	6.36
	0.001	11.34

At the 5% level of significance, we reject H_0 and conclude that not all the means are equal. The amount of dream deprivation *does* make a difference in mean anxiety level. *Note:* Technology gives *P*-value ≈ 0.0039.

This completes our single-factor ANOVA test. Before we consider another example, let's summarize the main points.

PROCEDURE

HOW TO CONSTRUCT A ONE-WAY ANOVA TEST

Requirements

You need k independent data groups, with each group belonging to a normal distribution and all groups having (approximately) the same standard deviation. N is the total number of data values across all groups.

Procedure

1. Set the *level of significance* α and the *hypotheses*

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

   H_1: not all of $\mu_1, \mu_2, \ldots, \mu_k$, are equal

   where μ_i is the population mean of group i.

2. Compute the *sample test statistic F* using the following steps or appropriate technology.

   (a) $SS_{TOT} = \Sigma x^2_{TOT} - \dfrac{(\Sigma x_{TOT})^2}{N}$

   where Σx_{TOT} is the sum of all data elements from all groups
   Σx^2_{TOT} is the sum of all data elements squared from all groups
   N is the total sample size

   (b) $SS_{TOT} = SS_{BET} + SS_W$

   where $SS_{BET} = \sum_{\text{all groups}} \left(\dfrac{(\Sigma x_i)^2}{n_i} \right) - \dfrac{(\Sigma x_{TOT})^2}{N}$

   n_i is the number of data elements in group i
   Σx_i is the sum of the data elements in group i

   $SS_W = \sum_{\text{all groups}} \left(\Sigma x_i^2 - \dfrac{(\Sigma x_i)^2}{n_i} \right)$

Continued

(c) $MS_{BET} = \dfrac{SS_{BET}}{d.f._{BET}}$ where $d.f._{BET} = k - 1$

$MS_W = \dfrac{SS_W}{d.f._W}$ where $d.f._W = N - k$

(d) $F = \dfrac{MS_{BET}}{MS_W}$ with $d.f._N = k - 1$ and $d.f._D = N - k$

Because an ANOVA test requires a number of calculations, we recommend that you summarize your results in a table such as Table 10-19. This is the type of table that is often generated by computer software.

3. Find (or estimate) the *P-value* using the *F* distribution (Table 8, Appendix II). The test is a *right-tailed* test.
4. *Conclude* the test. If P-value $\leq \alpha$, then reject H_0. If P-value $> \alpha$, then do not reject H_0.
5. *Interpret your conclusion* in the context of the application.

Summary table for ANOVA

TABLE 10-19 Summary of ANOVA Results

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square (Variance)	*F* Ratio	*P*-value	Test Decision
			Basic Model			
Between groups	SS_{BET}	$d.f._{BET}$	MS_{BET}	$\dfrac{MS_{BET}}{MS_W}$	From table	Reject H_0 or fail to reject H_0
Within groups	SS_W	$d.f._W$	MS_W			
Total	SS_{TOT}	$N - 1$				
			Summary of ANOVA Results from Dream Experiment (Example 8)			
Between groups	70.038	2	35.019	8.213	< 0.010	Reject H_0
Within groups	63.962	15	4.264			
Total	134	17				

GUIDED EXERCISE 11 One-way ANOVA test

A psychologist is studying pattern-recognition skills under four laboratory settings. In each setting, a fourth-grade child is given a pattern-recognition test with 10 patterns to identify. In setting I, the child is given *praise* for each correct answer and no comment about wrong answers. In setting II, the child is given *criticism* for each wrong answer and no comment about correct answers. In setting III, the child is given no praise or criticism, but the observer expresses *interest* in what the child is doing. In setting IV, the observer remains *silent* in an adjacent room watching the child through a one-way mirror. A random sample of fourth-grade children was used, and each child participated in the test only once. The test scores (number correct) for each group follow. (See Table 10-20.)

(a) Fill in the missing entries of Table 10-20.

$\Sigma x_{TOT} =$ ________

$\Sigma x_{TOT} = \Sigma x_1 + \Sigma x_2 + \Sigma x_3 + \Sigma x_4$
$= 41 + 14 + 38 + 28 = 121$

Continued

GUIDED EXERCISE 11 *continued*

$\Sigma x^2_{TOT} =$ ________

$N =$ ________

$k =$ ________

$\Sigma x^2_{TOT} = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 + \Sigma x_4^2$
$= 339 + 54 + 264 + 168 = 825$

$N = n_1 + n_2 + n_3 + n_4 = 5 + 4 + 6 + 5 = 20$

$k = 4$ groups

(b) What assumptions are we making about the data to apply a single-factor ANOVA test?

→ Because each of the groups comes from independent random samples (no child was tested twice), we need assume only that each group of data came from a normal distribution, and that all the groups came from distributions with about the same standard deviation.

Table 10-20 **Pattern-Recognition Experiment**

Group I (Praise) $n_1 = 5$		Group II (Criticism) $n_2 = 4$		Group III (Interest) $n_3 = 6$		Group IV (Silence) $n_4 = 5$	
x_1	x_1^2	x_2	x_2^2	x_3	x_3^2	x_4	x_4^2
9	81	2	4	9	81	5	25
8	64	5	25	3	9	7	49
8	64	4	16	7	49	3	9
9	81	3	9	8	64	6	36
7	49			5	25	7	49
				6	36		
$\Sigma x_1 = 41$		$\Sigma x_2 = 14$		$\Sigma x_3 = 38$		$\Sigma x_4 = 28$	
$\Sigma x_1^2 = 339$		$\Sigma x_2^2 = 54$		$\Sigma x_3^2 = 264$		$\Sigma x_4^2 = 168$	
$\Sigma x_{TOT} =$ _____		$\Sigma x^2_{TOT} =$ _____		$N =$ _____		$k =$ _____	

(c) What are the null and alternate hypotheses?

→ H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

In words, all the groups have the same population mean, and this hypothesis, together with the basic assumptions of part (b), states that all the groups come from the same population.

H_1: not all the means $\mu_1, \mu_2, \mu_3, \mu_4$ are equal.

In words, not all the groups have the same population mean, so at least one group did not come from the same population as the others.

(d) Find the value of SS_{TOT}.

→ $$SS_{TOT} = \Sigma x^2_{TOT} - \frac{(\Sigma x_{TOT})^2}{N} = 825 - \frac{(121)^2}{20}$$
$$= 92.950$$

(e) Find SS_{BET}.

→ $$SS_{BET} = \sum_{\text{all groups}} \left(\frac{(\Sigma x_i)^2}{n_i}\right) - \frac{(\Sigma x_{TOT})^2}{N}$$
$$= \frac{(41)^2}{5} + \frac{(14)^2}{4} + \frac{(38)^2}{6}$$
$$+ \frac{(28)^2}{5} - \frac{(121)^2}{20} = 50.617$$

(f) Find SS_W and check your calculations using the formula

→ $$SS_W = \sum_{\text{all groups}} \left(\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n_i}\right)$$

Continued

GUIDED EXERCISE 11 *continued*

$SS_{TOT} = SS_{BET} + SS_W$

$$SS_W = SS_1 + SS_2 + SS_3 + SS_4$$

$$SS_1 = \Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n_1} = 339 - \frac{(41)^2}{5} = 2.800$$

$$SS_2 = \Sigma x_2^2 - \frac{(\Sigma x_2)^2}{n_2} = 54 - \frac{(14)^2}{4} = 5.000$$

$$SS_3 = \Sigma x_3^2 - \frac{(\Sigma x_3)^2}{n_3} = 264 - \frac{(38)^2}{6} \approx 23.333$$

$$SS_4 = \Sigma x_4^2 - \frac{(\Sigma x_4)^2}{n_4} = 168 - \frac{(28)^2}{5} = 11.200$$

$$SS_W = 42.333$$

Check: $SS_{TOT} = SS_{BET} + SS_W$

$92.950 = 50.617 + 42.333$ checks

(g) Find $d.f._{BET}$ and $d.f._W$.

$$d.f._{BET} = k - 1 = 4 - 1 = 3$$

$$d.f._W = N - k = 20 - 4 = 16$$

Check: $N - 1 = d.f._{BET} + d.f._W$

$20 - 1 = 3 + 16$ checks

(h) Find the mean squares MS_{BET} and MS_W.

$$MS_{BET} = \frac{SS_{BET}}{d.f._{BET}} = \frac{50.617}{3} \approx 16.872$$

$$MS_W = \frac{SS_W}{d.f._W} = \frac{42.333}{16} \approx 2.646$$

(i) Find the F ratio (sample test statistic F).

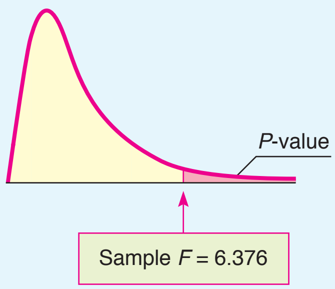
$$F = \frac{MS_{BET}}{MS_W} = \frac{16.872}{2.646} \approx 6.376$$

(j) Estimate the P-value for the sample $F = 6.376$. (Use Table 8 of Appendix II.)

$$d.f._N = d.f._{BET} = k - 1 = 3$$

$$d.f._D = d.f._W = N - k = 16$$

FIGURE 10-17 *P*-value



	Right-tail Area	$d.f._N$ 3
	0.100	2.46
$d.f._D$	0.050	3.24
✓ 16	0.025	4.08
	0.010	5.29
	0.001	9.01

$0.001 < P\text{-value} < 0.010$

(k) Conclude the test using a 1% level of significance. Does the test indicate that we should reject or fail to reject the null hypothesis? Explain.

$\alpha = 0.01$. Technology gives P-value ≈ 0.0048.

Since the P-value is less than 0.01, we reject H_0 and conclude that there is a significant difference in population means among the four groups. The laboratory setting *does* affect the mean scores.

(l) Make a summary table of this ANOVA test.

See Table 10-21.

Continued

GUIDED EXERCISE 11 *continued*

TABLE 10-21 Summary of ANOVA Results for Pattern-Recognition Experiment

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square (Variance)	F Ratio	P-value	Test Decision
Between groups	50.617	3	16.872	6.376	< 0.01	Reject H_0
Within groups	42.333	16	2.646			
Total	92.950	19				

TECH NOTES

After you understand the process of ANOVA, technology tools offer valuable assistance in performing one-way ANOVA. The TI-84Plus/TI-83Plus/TI-*n*spire calculators, Excel 2007, and Minitab all support one-way ANOVA. Both the TI-84Plus/TI-83Plus/TI-*n*spire calculators and Minitab use the terminology

Factor for Between Groups

Error for Within Groups

In all the technologies, enter the data for each group in separate columns. The displays show results for Guided Exercise 11.

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) Use **STAT, TESTS,** and option **F:ANOVA** and enter the lists containing the data.

```
One-way ANOVA
 F=6.376902887
 p=.0047646422
 Factor
  df=3
  SS=50.6166667
↓ MS=16.8722222
```

```
One-way ANOVA
↑  MS=16.8722222
 Error
  df=16
  SS=42.3333333
  MS=2.64583333
 Sxp=1.62660177
```

Excel 2007 Enter the data for each group in separate columns. On the home screen, click the **Data** tab. Then in the Analysis group, click **Data Analysis.** In the dialogue box, select **Anova: Single Factor.**

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
x1	5	41	8.2	0.7		
x2	4	14	3.5	1.666666667		
x3	6	38	6.333333333	4.666666667		
x4	5	28	5.6	2.8		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	50.61666667	3	16.87222222	6.376902887	0.004764642	5.292235983
Within Groups	42.33333333	16	2.645833333			
Total	92.95	19				

Minitab Enter the data for each group in separate columns. Use **Stat ➤ ANOVA ➤ Oneway(unstacked).**

```
One-way Analysis of Variance
Analysis of Variance
Source    DF      SS      MS      F       P
Factor     3   50.62   16.87   6.38   0.005
Error     16   42.33    2.65
Total     19   92.95
                                Individual 95% CIs For Mean
                          Based on Pooled StDev
Level      N    Mean   StDev  ---+---------+---------+---------+---
x1         5   8.200   0.837                     (-----*-----)
x2         4   3.500   1.291  (------*------)
x3         6   6.333   2.160              (-----*-----)
x4         5   5.600   1.673          (------*------)
                             ---+---------+---------+---------+---
Pooled StDev = 1.627           2.5       5.0       7.5      10.0
```

VIEWPOINT Gear Guide!

How much does that backpack cost? Do brand names really make any difference in cost? One way to answer such a question is to construct a one-way ANOVA test to study list prices for various backpacks with similar features but different brand names. For more information, see Backpacker Magazine, *Vol. 25, issue 157.*

SECTION 10.5 PROBLEMS

In each problem, assume that the distributions are normal and have approximately the same population standard deviation.

In each problem, please provide the following information.

(a) What is the level of significance? State the null and alternate hypotheses.
(b) Find SS_{TOT}, SS_{BET}, and SS_W and check that $SS_{TOT} = SS_{BET} + SS_W$. Find $d.f._{BET}$, $d.f._W$, MS_{BET}, and MS_W. Find the value of the sample test statistic F (F ratio). What are the degrees of freedom?
(c) Find (or estimate) the P-value of the sample test statistic.
(d) Based on your answers in parts (a) to (c), will you reject or fail to reject the null hypothesis?
(e) *Interpret* your conclusion in the context of the application.
(f) Make a summary table for your ANOVA test.

1. *Archaeology: Ceramics* Wind Mountain is an archaeological study area located in southwestern New Mexico. Pot sherds are broken pieces of prehistoric Native American clay vessels. One type of painted ceramic vessel is called *Mimbres classic black-on-white*. At three different sites, the number of such sherds was counted in local dwelling excavations (Source: Based on information from *Mimbres Mogollon Archaeology* by A. I. Woosley and A. J. McIntyre, University of New Mexico Press).

Site I	Site II	Site III
61	25	12
34	18	36
25	54	69
12	67	27

data are continued on next page

Site I	Site II	Site III
79		18
55		14
20		

Shall we reject or not reject the claim that there is no difference in population mean Mimbres classic black-on-white sherd counts for the three sites? Use a 1% level of significance.

2. ***Archaeology: Ceramics*** Another type of painted ceramic vessel is called *three-circle red-on-white* (see reference in Problem 1). At four different sites in the Wind Mountain archaeological region, the number of such sherds was counted in local dwelling excavations.

Site I	Site II	Site III	Site IV
17	18	32	13
23	4	19	19
6	33	18	14
19	8	43	34
11	25		12
	16		15

Shall we reject or not reject the claim that there is no difference in the population mean three-circle red-on-white sherd counts for the four sites? Use a 5% level of significance.

3. ***Economics: Profits per Employee*** How productive are U.S. workers? One way to answer this question is to study annual profits per employee. A random sample of companies in computers (I), aerospace (II), heavy equipment (III), and broadcasting (IV) gave the following data regarding annual profits per employee (units in thousands of dollars) (Source: *Forbes Top Companies,* edited by J. T. Davis, John Wiley and Sons).

I	II	III	IV
27.8	13.3	22.3	17.1
23.8	9.9	20.9	16.9
14.1	11.7	7.2	14.3
8.8	8.6	12.8	15.2
11.9	6.6	7.0	10.1
	19.3		9.0

Shall we reject or not reject the claim that there is no difference in population mean annual profits per employee in each of the four types of companies? Use a 5% level of significance.

4. ***Economics: Profits per Employee*** A random sample of companies in electric utilities (I), financial services (II), and food processing (III) gave the following information regarding annual profits per employee (units in thousands of dollars). (See reference in Problem 3.)

I	II	III
49.1	55.6	39.0
43.4	25.0	37.3
32.9	41.3	10.8
27.8	29.9	32.5
38.3	39.5	15.8
36.1		42.6
20.2		

Shall we reject or not reject the claim that there is no difference in population mean annual profits per employee in each of the three types of companies? Use a 1% level of significance.

5. *Ecology: Deer* Where are the deer? Random samples of square-kilometer plots were taken in different ecological locations of Mesa Verde National Park. The deer counts per square kilometer were recorded and are shown in the following table (Source: *The Mule Deer of Mesa Verde National Park,* edited by G. W. Mierau and J. L. Schmidt, Mesa Verde Museum Association).

Mountain Brush	Sagebrush Grassland	Pinon Juniper
30	20	5
29	58	7
20	18	4
29	22	9

Shall we reject or accept the claim that there is no difference in the mean number of deer per square kilometer in these different ecological locations? Use a 5% level of significance.

6. *Ecology: Vegetation* Wild irises are beautiful flowers found throughout the United States, Canada, and northern Europe. This problem concerns the length of the sepal (leaf-like part covering the flower) of different species of wild iris. Data are based on information taken from an article by R. A. Fisher in *Annals of Eugenics* (Vol. 7, part 2, pp. 179–188). Measurements of sepal length in centimeters from random samples of *Iris setosa* (I), *Iris versicolor* (II), and *Iris virginica* (III) are as follows:

I	II	III
5.4	5.5	6.3
4.9	6.5	5.8
5.0	6.3	4.9
5.4	4.9	7.2
4.4	5.2	5.7
5.8	6.7	6.4
5.7	5.5	
	6.1	

Shall we reject or not reject the claim that there are no differences among the population means of sepal length for the different species of iris? Use a 5% level of significance.

7. *Insurance: Sales* An executive at the home office of Big Rock Life Insurance is considering three branch managers as candidates for promotion to vice president. The branch reports include records showing sales volume for each salesperson in the branch (in hundreds of thousands of dollars). A random sample of these records was selected for salespersons in each branch. All three branches are located in cities in which per capita income is the same. The executive wishes to compare these samples to see if there is a significant difference in performance of salespersons in the three different branches. If so, the information will be used to determine which of the managers to promote.

Branch Managed by Adams	Branch Managed by McDale	Branch Managed by Vasquez
7.2	8.8	6.9
6.4	10.7	8.7
10.1	11.1	10.5
11.0	9.8	11.4
9.9		
10.6		

Use an $\alpha = 0.01$ level of significance. Shall we reject or not reject the claim that there are no differences among the performances of the salespersons in the different branches?

8. *Ecology: Pollution* The quantity of dissolved oxygen is a measure of water pollution in lakes, rivers, and streams. Water samples were taken at four different locations in a river in an effort to determine if water pollution varied from location to location. Location I was 500 meters above an industrial plant water discharge point and near the shore. Location II was 200 meters above the discharge point and in midstream. Location III was 50 meters downstream from the discharge point and near the shore. Location IV was 200 meters downstream from the discharge point and in midstream. The following table shows the results. Lower dissolved oxygen readings mean more pollution. Because of the difficulty in getting midstream samples, ecology students collecting the data had fewer of these samples. Use an $\alpha = 0.05$ level of significance. Do we reject or not reject the claim that the quantity of dissolved oxygen does not vary from one location to another?

Location I	Location II	Location III	Location IV
7.3	6.6	4.2	4.4
6.9	7.1	5.9	5.1
7.5	7.7	4.9	6.2
6.8	8.0	5.1	
6.2		4.5	

9. *Sociology: Ethnic Groups* A sociologist studying New York City ethnic groups wants to determine if there is a difference in income for immigrants from four different countries during their first year in the city. She obtained the data in the following table from a random sample of immigrants from these countries (incomes in thousands of dollars). Use a 0.05 level of significance to test the claim that there is no difference in the earnings of immigrants from the four different countries.

Country I	Country II	Country III	Country IV
12.7	8.3	20.3	17.2
9.2	17.2	16.6	8.8
10.9	19.1	22.7	14.7
8.9	10.3	25.2	21.3
16.4		19.9	19.8

SECTION 10.6

Introduction to Two-Way ANOVA

FOCUS POINTS

- Learn the notation and setup for two-way ANOVA tests.
- Learn about the three main types of deviations and how they break into additional effects.
- Use mean-square values to compute different sample *F* statistics.
- Use the *F* distribution to estimate *P*-values and conclude the test.
- Summarize experimental design features using a completely randomized design flow chart.

Suppose that Friendly Bank is interested in average customer satisfaction regarding the issue of obtaining bank balances and a summary of recent account transactions. Friendly Bank uses two systems, the first being a completely automated voice mail information system requiring customers to enter account numbers and passwords using the telephone keypad, and the second being the use of bank tellers or bank representatives to give the account information personally to customers. In addition, Friendly Bank wants to learn if average customer satisfaction is the same regardless of the time of day of contact. Three times of day are under study: morning, afternoon, and evening.

Friendly Bank could do two studies: one regarding average customer satisfaction with regard to type of contact (automated or bank representative) and one regarding average customer satisfaction with regard to time of day. The first study could be done using a difference-of-means test because there are only two types of contact being studied. The second study could be accomplished using one-way ANOVA.

Two-way ANOVA

Interaction

However, Friendly Bank could use just *one* study and the technique of *two-way analysis of variance* (known as *two-way ANOVA*) to simultaneously study average customer satisfaction with regard to the variable type of contact and the variable time of day, and *also* with regard to the *interaction* between the two variables. An interaction is present if, for instance, the difference in average customer satisfaction regarding type of contact is much more pronounced during the evening hours than, say, during the afternoon hours or the morning hours.

Factor

Levels

Let's begin our study of two-way ANOVA by considering the organization of data appropriate to two-way ANOVA. Two-way ANOVA involves *two* variables. These variables are called *factors*. The *levels* of a factor are the different values the factor can assume. Example 9 demonstrates the use of this terminology for the information Friendly Bank is seeking.

EXAMPLE 9 FACTORS AND LEVELS

For the Friendly Bank study discussed earlier, identify the factors and their levels, and create a table displaying the information.

SOLUTION: There are two factors. Call factor 1 *time of day*. This factor has three levels: morning, afternoon, and evening. Factor 2 is *type of contact*. This factor has two levels: automated contact and personal contact through a bank representative. Table 10-22 shows how the information regarding customer satisfaction can be organized with respect to the two factors.

TABLE 10-22 Table for Recording Average Customer Response

Factor 1: Time of Day	Factor 2: Type of Contact	
	Automated	Bank Representative
Morning	Morning Automated	Morning Bank Representative
Afternoon	Afternoon Automated	Afternoon Bank Representative
Evening	Evening Automated	Evening Bank Representative

Cell

When we look at Table 10-22, we see six contact–time-of-day combinations. Each such combination is called a *cell* in the table. The number of cells in any two-way ANOVA data table equals the product of the number of levels of the row factor times the number of levels of the column factor. In the case illustrated by Table 10-22, we see that the number of cells is 3×2, or 6.

Basic requirements of two-way ANOVA

Just as for one-way ANOVA, our application of two-way ANOVA has some basic requirements:

1. The measurements in each cell of a two-way ANOVA model are assumed to be drawn from a population with a normal distribution.
2. The measurements in each cell of a two-way ANOVA model are assumed to come from distributions with approximately the same variance.
3. The measurements in each cell come from *independent* random samples.
4. There are the *same number of measurements* in each cell.

Procedure to Conduct a Two-Way ANOVA Test (More Than One Measurement per Cell)

We will outline the procedure for two-way ANOVA in five steps. Each step will contain general methods and rationale appropriate to all two-way ANOVA tests with more than one data value in each cell. As we proceed, we will see how the method is applied to the Friendly Bank study.

Let's assume that Friendly Bank has taken random samples of customers fitting the criteria of each of the six cells described in Table 10-22. This means that a random sample of four customers fitting the morning-automated cell were surveyed. Another random sample of four customers fitting the afternoon-automated cell were surveyed, and so on. The bank measured customer satisfaction on a scale of 0 to 10 (10 representing highest customer satisfaction). The data appear in Table 10-23. Table 10-23 also shows cell means, row means, column means, and the total mean $\bar{\bar{x}}$ computed for all 24 data values. We will use these means as we conduct the two-way ANOVA test.

Overview

As in any statistical test, the first task is to establish the hypotheses for the test. Then, as in one-way ANOVA, the *F* distribution is used to determine the test conclusion. To compute the sample *F* value for a given null hypothesis, many of the same kinds of computations are done as are done in one-way ANOVA. In particular, we will use degrees of freedom $d.f. = N - 1$ (where N is the total sample size) allocated among the row factor, the column factor, the interaction, and the error (corresponding to "within groups" of one-way ANOVA). We look at the sum of squares *SS* (which measures variation) for the row factor, the column factor, the interaction, and the error. Then we compute the mean square *MS* for each category by taking the *SS* value and dividing by the corresponding degrees of freedom. Finally, we compute the sample *F* statistic for each factor and for the interaction by dividing the appropriate *MS* value by the *MS* value of the error.

TABLE 10-23 Customer Satisfaction at Friendly Bank

Factor 1: Time of Day	Factor 2: Type of Contact — Automated	Bank Representative	Row Means
Morning	6, 5, 8, 4 $\bar{x} = 5.75$	8, 7, 9, 9 $\bar{x} = 8.25$	Row 1 $\bar{x} = 7.00$
Afternoon	3, 5, 6, 5 $\bar{x} = 4.75$	9, 10, 6, 8 $\bar{x} = 8.25$	Row 2 $\bar{x} = 6.50$
Evening	5, 5, 7, 5 $\bar{x} = 5.50$	9, 10, 10, 9 $\bar{x} = 9.50$	Row 3 $\bar{x} = 7.50$
Column means	Column 1 $\bar{x} = 5.33$	Column 2 $\bar{x} = 8.67$	Total $\bar{x} = 7.00$

STEP 1: Establish the Hypotheses

Because we have two factors, we have hypotheses regarding each of the factors separately (called *main effects*) and then hypotheses regarding the interaction between the factors.

These three sets of hypotheses are

Hypotheses

1. H_0: There is no difference in population means among the levels of the row factor.
   H_1: At least two population means are different among the levels of the row factor.
2. H_0: There is no difference in population means among the levels of the column factor.
   H_1: At least two population means are different among the levels of the column factor.
3. H_0: There is no interaction between the factors.
   H_1: There is an interaction between the factors.

In the case of Friendly Bank, the hypotheses regarding the main effects are

H_0: There is no difference in population mean satisfaction depending on time of contact.
H_1: At least two population mean satisfaction measures are different depending on time of contact.
H_0: There is no difference in population mean satisfaction between the two types of customer contact.
H_1: There is a difference in population mean satisfaction between the two types of customer contact.

The hypotheses regarding interaction between factors are

H_0: There is no interaction between type of contact and time of contact.
H_1: There is an interaction between type of contact and time of contact.

STEP 2: Compute Sum of Squares (*SS*) Values

The calculations for the SS values are usually done on a computer. The main questions are whether population means differ according to the factors or the interaction of the factors. As we look at the Friendly Bank data in Table 10-23, we see that sample averages for customer satisfaction differ not only in each cell but also across the rows and across the columns. In addition, the total sample mean (designated $\bar{\bar{x}}$) differs from almost all the means. We know that different samples of the same size from the same population certainly can have different sample means. We need to decide if the differences are simply due to chance (sampling error) or are occurring because the samples have been taken from different populations with means that are not the same.

The tools we use to analyze the differences among the data values, the cell means, the row means, the column means, and the total mean are similar to those we used in Section 10.5 for one-way ANOVA. In particular, we first examine deviations of various measurements from the total mean $\bar{\bar{x}}$, and then we compute the sum of the squares *SS*.

There are basically three types of deviations:

Total deviation	**= Treatment deviation**	**+ Error deviation**
Compare each data value with the total mean $\bar{\bar{x}}$, $(x - \bar{\bar{x}})$.	For each data value, compare the mean of each cell with the total mean $\bar{\bar{x}}$, (cell $\bar{x} - \bar{\bar{x}}$).	Compare each data value with the mean of its cell, ($x -$ cell $\bar{x}$).

The treatment deviation breaks down further as

Treatment deviation	**= Deviation for main effect of factor 1**	**+ Deviation for main effect of factor 2**	**+ Deviation for interaction**
For each data value, (cell $\bar{x} - \bar{\bar{x}}$)	For each data value, compare the row mean with the total mean $\bar{\bar{x}}$, (row $\bar{x} - \bar{\bar{x}}$).	For each data value, compare the column mean with the total mean $\bar{\bar{x}}$, (column $\bar{x} - \bar{\bar{x}}$).	For each data value, (cell $\bar{x}$ − corresponding row $\bar{x}$ − corresponding column $\bar{x} + \bar{\bar{x}}$).

The deviations for each data value, row mean, column mean, or cell mean are then *squared and totaled over all the data*. This results in sums of squares, or variations. The *treatment variations* correspond to *between-group variations* of one-way ANOVA. The *error variation* corresponds to the *within-group variation* of one-way ANOVA.

$$\begin{array}{ccccc}
\textbf{Total varition} & = & \textbf{Treatment variation} & + & \textbf{Error variation} \\
\downarrow & & \downarrow & & \downarrow \\
\displaystyle\sum_{\text{all data}} (x - \bar{\bar{x}})^2 & = & \displaystyle\sum_{\text{all data}} (\text{cell } \bar{x} - \bar{\bar{x}})^2 & + & \displaystyle\sum_{\text{all data}} (x - \text{cell } \bar{x})^2 \\
\downarrow & & \downarrow & & \downarrow \\
SS_{TOT} & = & SS_{TR} & + & SS_E
\end{array}$$

where

$$\begin{array}{ccccccc}
\textbf{Treatment variation} & = & \textbf{Factor 1 variation} & + & \textbf{Factor 2 variation} & + & \textbf{Interaction variation} \\
\downarrow & & \downarrow & & \downarrow & & \downarrow \\
SS_{TR} & = & SS_{F1} & + & SS_{F2} & + & SS_{F1 \times F2} \\
\downarrow & & \downarrow & & \downarrow & & \downarrow \\
\displaystyle\sum_{\text{all data}} (\text{cell } \bar{x} - \bar{\bar{x}})^2 & = & \displaystyle\sum_{\text{all data}} (\text{row } \bar{x} - \bar{\bar{x}})^2 & + & \displaystyle\sum_{\text{all data}} (\text{col } \bar{x} - \bar{\bar{x}})^2 & + & \displaystyle\sum_{\text{all data}} (\text{cell } \bar{x} - \text{row } \bar{x} - \text{col } \bar{x} + \bar{\bar{x}})^2
\end{array}$$

The actual calculation of all the required SS values is quite time-consuming. In most cases, computer packages are used to obtain the results. For the Friendly Bank data, the following table is a Minitab printout giving the sum of squares *SS* for the type-of-contact factor, the time-of-day factor, the interaction between time and type of contact, and the error.

Minitab Printout for Customer Satisfaction at Friendly Bank

Analysis of Variance for Response

Source	DF	SS	MS	F	P
Time	2	4.00	2.00	1.24	0.313
Type	1	66.67	66.67	41.38	0.000
Interaction	2	2.33	1.17	0.72	0.498
Error	18	29.00	1.61		
Total	23	102.00			

We see that $SS_{\text{type}} = 66.77$, $SS_{\text{time}} = 4.00$, $SS_{\text{interaction}} = 2.33$, $SS_{\text{error}} = 29.00$, and $SS_{TOT} = 102$ (the total of the other four sums of squares).

STEP 3: Compute the Mean Square (*MS*) Values

The calculations for the MS values are usually done on a computer. Although the sum of squares computed in step 2 represents variation, we need to compute mean-square (*MS*) values for two-way ANOVA. As in one-way ANOVA, we compute *MS* values by dividing the *SS* values by respective degrees of freedom:

$$\text{Mean square } MS = \frac{\text{Corresponding sum of squares } SS}{\text{Respective degrees of freedom}}$$

For two-way ANOVA with more than one data value per cell, the degrees of freedom are

Degrees of freedom

$d.f.$ of row factor $= r - 1$

$d.f.$ of column factor $= c - 1$

$d.f.$ of total $= nrc - 1$

$d.f.$ of interaction $= (r - 1)(c - 1)$

$d.f.$ of error $= rc(n - 1)$

where r = number of rows, c = number of columns, and n = number of data values in one cell.

The Minitab table shows the degrees of freedom and the *MS* values for the main effect factors, the interaction, and the error for the Friendly Bank study.

STEP 4: Compute the Sample *F* Statistic for Each Factor and for the Interaction

Under the assumption of the respective null hypothesis, we have

$$\text{Sample } F \text{ for row factor} = \frac{MS \text{ for row factor}}{MS \text{ for error}}$$

with degrees of freedom numerator, $d.f._N = d.f.$ of row factor

degrees of freedom denominator, $d.f._D = d.f.$ of error

$$\text{Sample } F \text{ for column factor} = \frac{MS \text{ for column factor}}{MS \text{ for error}}$$

with degrees of freedom numerator, $d.f._N = d.f.$ of column factor

degrees of freedom denominator, $d.f._D = d.f.$ of error

$$\text{Sample } F \text{ for interaction} = \frac{MS \text{ for interaction}}{MS \text{ for error}}$$

with degrees of freedom numerator, $d.f._N = d.f.$ of interaction

degrees of freedom denominator, $d.f._D = d.f.$ of error

For the Friendly Bank study, the sample F values are

$$\text{Sample } F \text{ for time:} \quad F = \frac{MS_{\text{time}}}{MS_{\text{error}}} = \frac{2.00}{1.61} = 1.24$$

$$d.f._N = 2 \quad \text{and} \quad d.f._D = 18$$

Sample F for type of contact: $F = \dfrac{MS_{\text{type}}}{MS_{\text{error}}} = \dfrac{66.67}{1.61} = 41.41$

$d.f._N = 1$ and $d.f._D = 18$

Sample F for interaction: $F = \dfrac{MS_{\text{interaction}}}{MS_{\text{error}}} = \dfrac{1.17}{1.61} = 0.73$

$d.f._N = 2$ and $d.f._D = 18$

Due to rounding, the sample F values we just computed differ slightly from those shown in the Minitab printout.

STEP 5: Conclude the Test

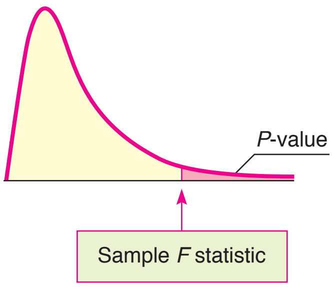
As with one-way ANOVA, larger values of the sample F statistic discredit the null hypothesis that there is no difference in population means across a given factor. The smaller the area to the right of the sample F statistic, the more likely there is an actual difference in some population means across the different factors. Smaller areas to the right of the sample F for interaction indicate greater likelihood of interaction between factors. Consequently, the P-value of a sample F statistic is the area of the F distribution to the *right* of the sample F statistic. Figure 10-18 shows the P-value associated with a sample F statistic.

Finding the P-value

Most statistical computer software packages provide P-values for the sample test statistic. You can also use the F distribution (Table 8 of Appendix II) to estimate the P-value. Once you have the P-value, compare it to the preset level of significance α. If the P-value is less than or equal to α, then reject H_0. Otherwise, do not reject H_0.

Be sure to test for interaction between the factors *first*. If you *reject* the null hypothesis of no interaction, then you should *not* test for a difference of means in the levels of the row factors or for a difference of means in the levels of the column factors because the interaction of the factors makes interpretation of the results of the main effects more complicated. A more extensive study of two-way ANOVA beyond the scope of this book shows how to interpret the results of the test of the main factors when there is interaction. For our purposes, we will simply stop the analysis rather than draw misleading conclusions.

FIGURE 10-18



For two-way ANOVA, the P-value of the sample statistic is the area of the F distribution to the *right* of the sample F statistic.

In two-way ANOVA, test for *interaction* first. If you reject the null hypothesis of no interaction, then do not continue with any tests of differences of means among other factors (unless you know techniques more advanced than those presented in this section).

If the test for interaction between the factors indicates that there is no evidence of interaction, then proceed to test the hypotheses regarding the levels of the row factor and the hypotheses regarding the levels of the column factor.

For the Friendly Bank study, we proceed as follows:

1. First, we determine if there is any evidence of interaction between the factors. The sample test statistic for interaction is $F = 0.73$, with P-value ≈ 0.498. Since the P-value is greater than $\alpha = 0.05$, we do not reject H_0. There is no evidence of interaction. Because there is no evidence of interaction between the main effects of type of contact and time of day, we proceed to test each factor for a difference in population mean satisfaction among the respective levels of the factors.
2. Next, we determine if there is a difference in mean satisfaction according to type of contact. The sample test statistic for type of contact is $F = 41.41$, with P-value ≈ 0.000 (to three places after the decimal). Since the P-value is less than $\alpha = 0.05$, we reject H_0. At the 5% level of significance, we conclude that there is a difference in average customer satisfaction between contact with an automated system and contact with a bank representative.
3. Finally, we determine if there is a difference in mean satisfaction according to time of day. The sample test statistic for time of day is $F = 1.24$, with P-value ≈ 0.313. Because the P-value is greater than $\alpha = 0.05$, we do not reject H_0. We conclude that at the 5% level of significance, there is no evidence that population mean customer satisfaction is different according to time of day.

Special Case: One Observation in Each Cell with No Interaction

In the case where our data consist of only one value in each cell, there are no measures for sum of squares *SS* interaction or mean-square *MS* interaction, and we cannot test for interaction of factors using two-way ANOVA. If it *seems reasonable* (based on other information) to assume that there is *no* interaction between the factors, then we can use two-way ANOVA techniques to test for average response differences due to the main effects. In Guided Exercise 12, we look at two-way ANOVA applied to the special case of only one measurement per cell and no interactions.

GUIDED EXERCISE 12 Special-case two-way ANOVA

Let's use two-way ANOVA to test if the average fat content (grams of fat per 3-oz serving) of potato chips is different according to the brand or according to which laboratory made the measurement. Use $\alpha = 0.05$. (See the following Minitab tables.)

Average Grams of Fat in a 3-oz Serving of Potato Chips

	Laboratory		
Brand	Lab I	Lab II	Lab III
Texas Chips	32.4	33.1	32.9
Great Chips	37.9	37.7	37.8
Chip Ooh	29.1	29.4	29.5

Minitab Printout for Potato Chip Data

Analysis of Variance for Fat

Source	DF	SS	MS	F	P
Brand	2	108.7022	54.3511	968.63	0.000
Lab	2	0.1422	0.0711	1.27	0.375
Error	4	0.2244	0.0561		
Total	8	109.0689			

Continued

GUIDED EXERCISE 12 *continued*

(a) List the factors and the number of levels for each.

The factors are brand and laboratory. Each factor has three levels.

(b) Assuming there is no interaction, list the hypotheses for each factor.

For brand,
H_0: There is no difference in population mean fat by brands.
H_1: At least two brands have different population mean fat contents.
For laboratory,
H_0: There is no difference in mean fat content as measured by the labs.
H_1: At least two of the labs give different mean fat measurements.

(c) Calculate the sample F statistic for brands and compare it to the value given in the Minitab printout. Look at the P-value in the printout. What is your conclusion regarding average fat content among brands?

For brand,

$$\text{Sample } F = \frac{MS_{\text{brand}}}{MS_{\text{error}}} \approx \frac{54.35}{0.056} \approx 970$$

Using the Minitab printout, we see P-value ≈ 0.000 (to three places after the decimal). Using Table 8 of Appendix II, we see P-value < 0.001. Since the P-value is less than $\alpha = 0.05$, we reject H_0 and conclude that at the 5% level of significance, at least two of the brands have different mean fat contents.

(d) Calculate the sample F statistic for laboratories and compare it to the value given in the Minitab printout. What is your conclusion regarding average fat content as measured by the different laboratories?

For laboratories,

$$\text{Sample } F = \frac{MS_{\text{lab}}}{MS_{\text{error}}} \approx \frac{0.0711}{0.056} \approx 1.27$$

Using the Minitab printout, we see P-value ≈ 0.375. Using Table 8 of Appendix II, we see P-value > 0.100. Because the P-value is greater than $\alpha = 0.05$, we conclude that at the 5% level of significance, there is no evidence of differences in average measurements of fat content as determined by the different laboratories.

TECH NOTES The calculations involved in two-way ANOVA are usually done using statistical or spreadsheet software. Basic printouts from various software packages are similar. Specific instructions for using Excel 2007, Minitab, and SPSS are given in the Using Technology section at the end of the chapter.

Experimental Design

In the preceding section and in this section, we have seen aspects of one-way and two-way ANOVA, respectively. Now let's take a brief look at some experimental design features that are appropriate for the use of these techniques.

Completely randomized design

For one-way ANOVA, we have one factor. Different levels for the factor form the treatment groups under study. In a *completely randomized design*, independent random samples of experimental subjects or objects are selected for each treatment group. For example, suppose a researcher wants to study the effects of different treatments for the condition of slightly high blood pressure. Three treatments are under study: diet, exercise, and medication. In a completely randomized design, the people participating in the experiment are *randomly* assigned to each treatment group. Table 10-24 shows the process.

TABLE 10-24 Completely Randomized Design Flow Chart

Subjects with slightly high blood pressure	→	Random assignment	→ Treatment 1: Diet → Treatment 2: Exercise → Treatment 3: Medication

TABLE 10-25 **Randomized Block Design Flow Chart**

Blocks			Treatments	
Subjects with slightly high blood pressure	Under 30	Random assignment	Treatment 1:	Diet
			Treatment 2:	Exercise
			Treatment 3:	Medication
	Ages 31–50	Random assignment	Treatment 1:	Diet
			Treatment 2:	Exercise
			Treatment 3:	Medication
	Over 50	Random assignment	Treatment 1:	Diet
			Treatment 2:	Exercise
			Treatment 3:	Medication

For two-way ANOVA, there are *two* factors. When we *block* experimental subjects or objects together based on a similar characteristic that might affect responses to treatments, we have a *block design*. For example, suppose the researcher studying treatments for slightly high blood pressure believes that the age of subjects might affect the response to the three treatments. In such a case, blocks of subjects in specified age groups are used. The factor "age" is used to form blocks. Suppose age has three levels: under age 30, ages 31–50, and over age 50. The same number of subjects is assigned to each block. Then the subjects in each block are randomly assigned to the different treatments of diet, exercise, or medication. Table 10-25 shows the *randomized block design*.

Randomized block design

Experimental design is an essential component of good statistical research. The design of experiments can be quite complicated, and if the experiment is complex, the services of a professional statistician may be required. The use of blocks helps the researcher account for some of the most important sources of variability among the experimental subjects or objects. Then, randomized assignments to different treatment groups help average out the effects of other remaining sources of variability. In this way, differences among the treatment groups are more likely to be caused by the treatments themselves rather than by other sources of variability.

VIEWPOINT Who Watches Cable TV?

Consider the following claim: Average cable TV viewers are, generally speaking, as affluent as newspaper readers and better off than radio or magazine audiences. How do we know whether this claim is true? One way to answer such a question is to construct a two-way ANOVA test in which the rows represent household income levels and the columns represent cable TV viewers, news paper readers, magazine audiences, and radio listeners. The response variable is an index that represents the ratio of the specific medium compared with U.S. averages. For more information and data, see American Demographics *(Vol. 17, No. 6).*

SECTION 10.6 PROBLEMS

1. *Statistical Literacy: Physical Therapy* Does talking while walking slow you down? A study reported in the journal *Physical Therapy* (Vol. 72, No. 4) considered mean cadence (steps per minute) for subjects using no walking device, a standard walker, and a rolling walker. In addition, the cadence was measured when the subjects had to perform dual tasks. The second task was to respond vocally to a signal while walking. Cadence was measured for subjects who were

just walking (using no device, a standard walker, or a rolling walker) and for subjects required to respond to a signal while walking. List the factors and the number of levels of each factor. How many cells are there in the data table?

2. *Statistical Literacy: Salary Survey* *Academe, Bulletin of the American Association of University Professors* (Vol. 83, No. 2) presents results of salary surveys (average salary) by rank of the faculty member (professor, associate, assistant, instructor) and by type of institution (public, private). List the factors and the number of levels of each factor. How many cells are there in the data table?

3. *Critical Thinking: Physical Therapy* For the study regarding mean cadence (see Problem 1), two-way ANOVA was used. Recall that the two factors were walking device (none, standard walker, rolling walker) and dual task (being required to respond vocally to a signal or no dual task required). Results of two-way ANOVA showed that there was no evidence of interaction between the factors. However, according to the article, "The ANOVA conducted on the cadence data revealed a main effect of walking device." When the hypothesis regarding no difference in mean cadence according to which, if any, walking device was used, the sample F was 30.94, with $d.f._N = 2$ and $d.f._D = 18$. Further, the P-value for the result was reported to be less than 0.01. From this information, what is the conclusion regarding any difference in mean cadence according to the factor "walking device used"?

4. *Education: Media Usage* In a study of media usage versus education level (*American Demographics,* Vol. 17, No. 6), an index was used to measure media usage, where a measurement of 100 represents the U.S. average. Values above 100 represent above-average media usage.

	Media				
Education Level	Cable Network	Prime-Time TV	Radio	Newspaper	Magazine
Less than high school	80	112	87	76	85
High school graduate	103	105	100	99	101
Some college	107	94	106	105	107
College graduate	108	90	106	116	108

Source: American Demographics by Staff. Copyright 1995 by CRAIN COMMUNICATIONS INC. Reproduced with permission of CRAIN COMMUNICATIONS INC in the formats Textbook and Other book via Copyright Clearance Center.

(a) List the factors and the number of levels of each factor.
(b) Assume there is no interaction between the factors. Use two-way ANOVA and the following Minitab printout to determine if there is a difference in population mean index based on education. Use $\alpha = 0.05$.
(c) Determine if there is a difference in population mean index based on media. Use $\alpha = 0.05$.

Minitab Printout for Media/Education Data

```
Analysis of Variance for Index
Source     DF      SS      MS       F        P
Edu         3     961     320     2.96     0.075
Media       4       5       1     0.01     1.000
Error      12    1299     108
Total      19    2264
```

5. *Income: Media Usage* In the same study described in Problem 4, media usage versus household income also was considered. The media usage indices for the various media and income levels follow.

Income Level	Media				
	Cable Network	Prime-Time TV	Radio	Newspaper	Magazine
Less than $20,000	78	112	89	80	91
$20,000–$39,999	97	105	100	97	100
$40,000–$74,999	113	92	106	111	105
$75,000 or more	121	94	107	121	105

Source: *American Demographics* by Staff. Copyright 1995 by CRAIN COMMUNICATIONS INC. Reproduced with permission of CRAIN COMMUNICATIONS INC in the formats Textbook and Other book via Copyright Clearance Center.

(a) List the factors and the number of levels of each factor.
(b) Assume there is no interaction between the factors. Use two-way ANOVA and the following Minitab printout to determine if there is a difference in population mean index based on income. Use $\alpha = 0.05$.
(c) Determine if there is a difference in population mean index based on media. Use $\alpha = 0.05$.

Minitab Printout for Media/Income Data

Analysis of Variance for Index

Source	DF	SS	MS	F	P
Income	3	1078	359	2.77	0.088
Media	4	15	4	0.03	0.998
Error	12	1558	130		
Total	19	2651			

6. *Gender: Grade Point Average* Does college grade point average (GPA) depend on gender? Does it depend on class (freshman, sophomore, junior, senior)? In a study, the following GPA data were obtained for random samples of college students in each of the cells.

Gender	Class							
	Freshman		Sophomore		Junior		Senior	
Male	2.8	2.1	2.5	2.3	3.1	2.9	3.8	3.6
	2.7	3.0	2.9	3.5	3.2	3.8	3.5	3.1
Female	2.3	2.9	2.6	2.4	2.6	3.6	3.2	3.5
	3.5	3.9	3.3	3.6	3.3	3.7	3.8	3.6

(a) List the factors and the number of levels of each factor.
(b) Use two-way ANOVA and the following Minitab printout to determine if there is any evidence of interaction between the two factors at a level of significance of 0.05.

Minitab Printout of GPA Based on Gender and Class

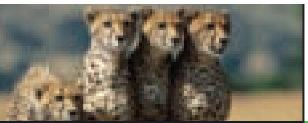
Analysis of Variance for GPA

Source	DF	SS	MS	F	P
Gender	1	0.281	0.281	1.26	0.273
Class	3	2.226	0.742	3.32	0.037
Interaction	3	0.286	0.095	0.43	0.736
Error	24	5.365	0.224		
Total	31	8.159			

(c) If there is no evidence of interaction, use two-way ANOVA and the Minitab printout to determine if there is a difference in mean GPA based on class. Use $\alpha = 0.05$.

(d) If there is no evidence of interaction, use two-way ANOVA and the Minitab printout to determine if there is a difference in mean GPA based on gender. Use $\alpha = 0.05$.

7. *Experimental Design: Teaching Style* A researcher forms three blocks of students interested in taking a history course. The groups are based on grade point average (GPA). The first group consists of students with a GPA less than 2.5, the second group consists of students with a GPA between 2.5 and 3.1, and the last group consists of students with a GPA greater than 3.1. History courses are taught in three ways: traditional lecture, small-group collaborative method, and independent study. The researcher randomly assigns 10 students from each block to sections of history taught each of the three ways. Sections for each teaching style then have 10 students from each block. The researcher records the scores on a common course final examination administered to each student. Draw a flow chart showing the design of this experiment. Does the design fit the model for randomized block design?



Chapter Review

SUMMARY

In this chapter, we introduced applications of two probability distributions: the chi-square distribution and the F distribution.

- The chi-square distribution is used for tests of independence or homogeneity, tests of goodness of fit, tests of variance σ^2, and tests to estimate a variance σ^2.
- The F distribution is used for tests of two variances, one-way ANOVA, and two-way ANOVA.

ANOVA tests are used to determine whether there are differences among means for several groups.

- If groups are based on the value of only one variable, we have one-way ANOVA.
- If groups are formed using two variables, we use two-way ANOVA to test for differences of means based on either variable or on an interaction between the variables.

IMPORTANT WORDS & SYMBOLS

Section 10.1

Independence test 593
Chi-square distribution overview 592
Observed frequency of a cell, O 593
Contingency table with cells 593
Expected frequency of a cell, E 594
Row total 595
Column total 595
Degrees of freedom, $d.f. = (R - 1)(C - 1)$ for χ^2 distribution and tests of independence 597
Test of homogeneity 600
Multinomial experiments 603

Section 10.2

Goodness-of-fit test 609
Degrees of freedom $d.f.$ for χ^2 distribution and goodness-of-fit tests 611

Section 10.3

Hypotheses tests about σ^2 624
Confidence interval for σ^2 625
χ^2_U, χ^2_L 625

Section 10.4

Hypothesis tests about two variances 631
F distribution overview 630

Section 10.5

ANOVA 641
Sum of squares, SS 643
SS_{TOT}, SS_{BET}, SS_W 645

VIEWPOINT Movies and Money!

Young adults are the movie industry's best customers. However, going to the movies is expensive, which may explain why attendance rates increase with household income. Using what you have learned in this chapter, you can create appropriate chi-square tests to determine how good a fit exists between national percentage rates of attendance by household income and attendance rates in your demographic area. For more information and national data, see American Demographics *(Vol. 18, No. 12).*

CHAPTER REVIEW PROBLEMS

1. *Statistical Literacy* Of the following random variables, which have only non-negative values: z, t, chi-square, F?
2. *Statistical Literacy* Of the following probability distributions, which are always symmetric: normal, Student's t, chi-square, F?
3. *Critical Thinking* Suppose you took random samples from three distinct age groups. Through a survey, you determined how many respondents from each age group preferred to get news from TV, newspapers, the Internet, or another source (respondents could select only one mode). What type of test would be appropriate to determine if there is sufficient statistical evidence to claim that the proportions of each age group preferring the different modes of obtaining news are not the same? Select from tests of independence, homogeneity, goodness of fit, and ANOVA.
4. *Critical Thinking* Suppose you take a random sample from a normal population and you want to determine whether there is sufficient statistical evidence to claim that the population variance differs from a corresponding variance specified in a government contract. Which type of test is appropriate, a test of one variance or a test of two variances?

Before you solve Problems 5–14, first classify the problem as one of the following:

Chi-square test of independence or homogeneity
Chi-square goodness of fit
Chi-square for testing or estimating σ^2 or σ
F test for two variances
One-way ANOVA
Two-way ANOVA

Then, in each of the problems when a test is to be performed, do the following:

(i) Give the value of the level of significance. State the null and alternate hypotheses.
(ii) Find the sample test statistic.
(iii) Find or estimate the P-value of the sample test statistic.
(iv) Conclude the test.

(v) *Interpret* the conclusion in the context of the application.
(vi) In the case of one-way ANOVA, make a summary table.

5. *Sales: Packaging* The makers of Country Boy Corn Flakes are thinking about changing the packaging of the cereal in the hope of improving sales. In an experiment, five stores of similar size in the same region sold Country Boy Corn Flakes in different-shaped containers for 2 weeks. Total packages sold are given in the following table. Using a 0.05 level of significance, shall we reject or fail to reject the hypothesis that the mean sales are the same, no matter which box shape is used?

Cube	Cylinder	Pyramid	Rectangle
120	110	74	165
88	115	62	98
65	180	110	125
95	96	66	87
71	85	83	118

6. *Education: Exams* Professor Fair believes that extra time does not improve grades on exams. He randomly divided a group of 300 students into two groups and gave them all the same test. One group had exactly 1 hour in which to finish the test, and the other group could stay as long as desired. The results are shown in the following table. Test at the 0.01 level of significance that time to complete a test and test results are independent.

Time	A	B	C	F	Row Total
1 h	23	42	65	12	142
Unlimited	17	48	85	8	158
Column Total	40	90	150	20	300

7. *Tires: Blowouts* A consumer agency is investigating the blowout pressures of Soap Stone tires. A Soap Stone tire is said to blow out when it separates from the wheel rim due to impact forces usually caused by hitting a rock or a pothole in the road. A random sample of 30 Soap Stone tires were inflated to the recommended pressure, and then forces measured in foot-pounds were applied to each tire (1 foot-pound is the force of 1 pound dropped from a height of 1 foot). The customer complaint is that some Soap Stone tires blow out under small-impact forces, while other tires seem to be well made and don't have this fault. For the 30 test tires, the sample standard deviation of blowout forces was 1353 foot-pounds.
(a) Soap Stone claims its tires will blow out at an average pressure of 20,000 foot-pounds, with a standard deviation of 1020 foot-pounds. The average blowout force is not in question, but the variability of blowout forces is in question. Using a 0.01 level of significance, test the claim that the variance of blowout pressures is more than Soap Stone claims it is.
(b) Find a 95% confidence interval for the variance of blowout pressures, using the information from the random sample.

8. *Computer Science: Data Processing* Anela is a computer scientist who is formulating a large and complicated program for a type of data processing. She has three ways of storing and retrieving data: CD, tape, and disk. As an experiment, she sets up her program in three different ways: one using CDs, one using tapes, and the other using disks. Then she makes four test runs of this type of data processing on each program. The time required to execute each program is shown in the following table (in minutes). Use a 0.01 level of significance to test the hypothesis that the mean processing time is the same for each method.

CD	Tape	Disks
8.7	7.2	7.0
9.3	9.1	6.4
7.9	7.5	9.8
8.0	7.7	8.2

9. ***Teacher Ratings: Grades*** Professor Stone complains that students' teacher ratings depend on the grade students receive. In other words, according to Professor Stone, a teacher who gives good grades gets good ratings, and a teacher who gives bad grades gets bad ratings. To test this claim, the Student Assembly took a random sample of 300 teacher ratings on which the students' grades for the course also were indicated. The results are given in the following table. Test the hypothesis that teacher ratings and student grades are independent at the 0.01 level of significance.

Rating	A	B	C	F (or withdrawal)	Row Total
Excellent	14	18	15	3	50
Average	25	35	75	15	150
Poor	21	27	40	12	100
Column Total	60	80	130	30	300

10. ***Packaging: Corn Flakes*** A machine that puts corn flakes into boxes is adjusted to put an average of 15 ounces into each box, with standard deviation of 0.25 ounce. If a random sample of 12 boxes gave a sample standard deviation of 0.38 ounce, do these data support the claim that the variance has increased and the machine needs to be brought back into adjustment? (Use a 0.01 level of significance.)

11. ***Sociology: Age Distribution*** A sociologist is studying the age of the population in Blue Valley. Ten years ago, the population was such that 20% were under 20 years old, 15% were in the 20- to 35-year-old bracket, 30% were between 36 and 50, 25% were between 51 and 65, and 10% were over 65. A study done this year used a random sample of 210 residents. This sample showed

Under 20	20–35	36–50	51–65	Over 65
26	27	69	68	20

At the 0.01 level of significance, has the age distribution of the population of Blue Valley changed?

12. ***Engineering: Roller Bearings*** Two processes for manufacturing large roller bearings are under study. In both cases, the diameters (in centimeters) are being examined. A random sample of 21 roller bearings from the old manufacturing process showed the sample variance of diameters to be $s^2 = 0.235$. Another random sample of 26 roller bearings from the new manufacturing process showed the sample variance of their diameters to be $s^2 = 0.128$. Use a 5% level of significance to test the claim that there is a difference (either way) in the population variances between the old and new manufacturing processes.

13. ***Engineering: Light Bulbs*** Two processes for manufacturing 60-watt light bulbs are under study. In both cases, the life (in hours) of the bulb before it burns out is being examined. A random sample of 18 light bulbs manufactured using the old process showed the sample variance of lifetimes to be $s^2 = 51.87$. Another random sample of 16 light bulbs manufactured using the new process showed the sample variance of the lifetimes to be $s^2 = 135.24$. Use a 5% level of significance to test the claim that the population variance of lifetimes for the new manufacturing process is larger than that of the old process.

14. *Advertising: Newspapers* Does the newspaper section in which an ad is placed make a difference in the average daily number of people responding to the ad? Is there a difference in average daily number of people responding to the ad if it is placed in the Sunday newspaper as compared with the Wednesday newspaper? Video Entertainment is a video club that sells videotapes featuring movies of all kinds—instructional videos, videos of TV specials, etc. To attract new customers, Video Entertainment ran ads in the Wednesday newspaper and in the Sunday newspaper offering six free videotapes or DVDs to new members. In addition, the ads were placed in the sports section, the entertainment section, and the business section of the local newspaper. Ads running in the different sections carried different promotion codes. Different codes also were used for Wednesday ads compared with Sunday ads. The numbers of people responding to the ads for days selected at random were recorded according to the promotion codes. The results follow:

Day	Section of Newspaper								
	Sports			Entertainment			Business		
Wed.	12	15	11	22	14	17	2	4	3
	12	15	20	22	18	12	5	0	1
Sun.	20	23	25	32	26	28	13	16	13
	33	15	17	31	25	41	15	14	10

(a) List the factors and the number of levels for each factor.

(b) Use the following Minitab printout and the *F* distribution table (Table 8 of Appendix II) to test for interaction between the variables. Use a 1% level of significance.

```
Minitab Printout for Mean Number of Responses per Day
Analysis of Variance for Response
Source          DF       SS        MS        F        P
Day              1    1024.0    1024.0    55.28    0.000
Section          2    1573.6     786.8    42.48    0.000
Interaction      2      38.0      19.0     1.03    0.371
Error           30     555.7      18.5
Total           35    3191.2
```

(c) If there is no evidence of interaction between the factors, test for a difference in mean number of daily responses for the levels of the day factor. Use $\alpha = 0.01$.

(d) If there is no evidence of interaction between the factors, test for a difference in mean number of daily responses for the levels of the section factor. Use $\alpha = 0.01$.

DATA HIGHLIGHTS: GROUP PROJECTS

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

The *Statistical Abstract of the United States* reported information about the percentage of arrests of all drunk drivers according to age group. In the following table, the entry 3.7 in the first row means that in the entire United States, about 3.7% of all people arrested for drunk driving were in the age group 16–17 years. The Freemont County Sheriff's Office obtained data about the number of drunk drivers arrested in each age group over the past several years. In the following table, the entry 8 in the first row means that eight people in the age group 16–17 years were arrested for drunk driving in Freemont County.

Use a chi-square test with 5% level of significance to test the claim that the age distribution of drunk drivers arrested in Freemont County is the same as the national age distribution of drunk drivers arrested.

(a) State the null and alternate hypotheses.

(b) Find the value of the chi-square test statistic from the sample.

(c) Find the degrees of freedom and the *P*-value of the test statistic.

(d) Decide whether you should reject or not reject the null hypothesis.

(e) State your conclusion in the context of the problem.

(f) How could you gather data and conduct a similar test for the city or county in which you live? Explain.

Distribution of Drunk Driver Arrests by Age

Age	National Percentage	Number in Freemont County
16–17	3.7	8
18–24	18.9	35
25–29	12.9	23
30–34	10.3	19
35–39	8.5	12
40–44	7.9	14
45–49	8.0	16
50–54	7.9	13
55–59	6.8	10
60–64	5.7	9
65 and over	9.4	15
	100%	174

LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

1. In this chapter, you studied the chi-square distribution and three principal applications of the distribution.

   (a) Outline the basic ideas behind the chi-square test of independence. What is a contingency table? What are the null and alternate hypotheses? How is the test statistic constructed? What basic assumptions underlie this application of the chi-square distribution?

   (b) Outline the basic ideas behind the chi-square test of goodness of fit. What are the null and alternate hypotheses? How is the test statistic constructed? There are a number of direct similarities between tests of independence and tests of goodness of fit. Discuss and summarize these similarities.

   (c) Outline the basic ideas behind the chi-square method of testing and estimating a standard deviation. What basic assumptions underlie this process?

   (d) Outline the basic ideas behind the chi-square test of homogeneity. What are the null and alternate hypotheses? How is the test statistic contructed? What basic assumptions underlie the application of the chi-square distribution?

2. The *F* distribution is used to construct a one-way ANOVA test for comparing several sample means.

   (a) Outline the basic purpose of ANOVA. How does ANOVA avoid high risk due to multiple type I errors?

   (b) Outline the basic assumption for ANOVA.

   (c) What are the null and alternate hypotheses in an ANOVA test? If the test conclusion is to reject the null hypothesis, do we know which of the population means are different from each other?

   (d) What is the *F* distribution? How are the degrees of freedom for numerator and denominator determined?

   (e) What do we mean by a summary table of ANOVA results? What are the main components of such a table? How is the final decision made?

Using Technology

Application

Analysis of Variance (One-Way ANOVA)

The following data comprise a winter mildness/severity index for three European locations near 50° north latitude. For each decade, the number of unmistakably mild months minus the number of unmistakably severe months for December, January, and February is given.

Decade	Britain	Germany	Russia
1800	−2	−1	+1
1810	−2	−3	−1
1820	0	0	0
1830	−3	−2	−1
1840	−3	−2	+1
1850	−1	−2	+3
1860	+8	+6	+1
1870	0	0	−3
1880	−2	0	+1
1890	−3	−1	+1
1900	+2	0	+2
1910	+5	+6	+1
1920	+8	+6	+2
1930	+4	+4	+5
1940	+1	−1	−1
1950	0	+1	+2

Table is based on data from Exchanging Climate by H. H. Lamb; copyright © 1966. Reprinted by permission of Routledge, UK.

1. We wish to test the null hypothesis that the mean winter indices for Britain, Germany, and Russia are all equal against the alternate hypothesis that they are not all equal. Use a 5% level of significance.
2. What is the sum of squares between groups? Within groups? What is the sample F ratio? What is the P-value? Shall we reject or fail to reject the statement that the mean winter indices for these locations in Britain, Germany, and Russia are the same?
3. What is the smallest level of significance at which we could conclude that the mean winter indices for these locations are not all equal?

Technology Hints (One-Way ANOVA)

SPSS

Enter all the data in one column. In another column, use integers to designate the group to which each data value belongs. Use the menu selections **Analyze ➤ Compare Means ➤ One-Way ANOVA.** Move the column containing the data to the dependent list. Move the column containing group designation to the factor list.

Technology Hints (Two-Way ANOVA)

Excel 2007

Excel has two commands for two-way ANOVA, depending on how many data values are in each cell. On the home screen, click the **Data** tab. Then in the Analysis group, click **Data Analysis**. Then use **ANOVA: Two-Factor with Replication** if there are two or more sample measurements for each factor combination cell. Again, there must be the same number of data in each cell.

ANOVA: Two-Factor without Replication if there is only one data value in each factor combination cell.

Data entry is fairly straightforward. For example, look at the Excel spreadsheet for the data of Guided Exercise 12 regarding the fat content of different brands of potato chips as measured by different labs.

	A	B	C	D
1		**Lab 1**	**Lab 2**	**Lab 3**
2	Texas	32.4	33.1	32.9
3	Great	37.9	37.7	37.8
4	Chip Ooh	29.1	29.4	29.5

Minitab

For Minitab, all the data for the response variable (in this case, fat content) are entered into a single column. Create two more columns, one for the row number of the cell containing the data value and one for the column number of the cell containing the data value. For the potato chip example, the rows correspond to the brand and the columns to the lab doing the analysis. Use the menu choices **Stat ➤** ANOVA **➤ Two-Way.**

	C1	C2	C3
	Brand	Lab	Fat
1	1	1	32.4
2	1	2	33.1
3	1	3	32.9
4	2	1	37.9
5	2	2	37.7
6	2	3	37.8
7	3	1	29.1
8	3	2	29.4
9	3	3	29.5

SPSS

Data entry for SPSS is similar to that for Minitab. Enter all the data in one column. Use a separate column for each factor and use a label or integer to designate the group in the particular factor. Under Variable View, type appropriate labels for the columns of data. Use the menu selections **Analyze ➤ General Linear Model ➤ Univariate. . . .** In the dialogue box, the dependent variable is the quantity represented by the data. The factors are those found in each factor column. For the special case of only one datum per cell, click the Model button. Select Custom and move the desired factors into Model.

SPSS Data Editor

File Edit View Data Transform Analyze Graphs

	fat	lab	brand
1	32.40	1	Texas
2	37.90	1	Great
3	29.10	1	Chip
4	33.10	2	Texas
5	37.70	2	Great
6	29.40	2	Chip
7	32.90	3	Texas
8	37.80	3	Great
9	29.50	3	Chip

11

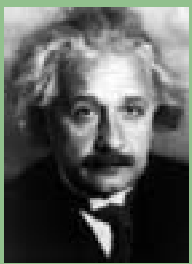


Photo Researchers / Alamy

Make everything as simple as possible, but no simpler.

—Albert Einstein

Business Wire/Getty Images Publicity/Getty Images

The brilliant German-born American physicist Albert Einstein (1879–1955) formulated the theory of relativity.

For online student resources, visit the Brase/Brase, *Understandable Statistics,* 10th edition web site at **http://www.cengage.com/statistics/brase**.

Nonparametric Statistics

James Shafer/PhotoEdit

PREVIEW QUESTIONS

What if you cannot make assumptions about a population distribution? Can you still use statistical methods? What are the advantages and disadvantages? (Section 11.1)

What are nonparametric tests? How do you handle a "before and after" situation? (Section 11.1)

If you can't make assumptions about the population, and you have independent samples, how do you set up a nonparametric test? (Section 11.2)

Suppose you are interested only in rank data (ordinal-type data). If you have ordered pairs (x, y) of ranked data, is there a way to measure and test correlation? (Section 11.3)

Is a sequence random or is there a pattern associated with the sequence? (Section 11.4)

FOCUS PROBLEM

How Cold? Compared to What?

Juneau is the capital of Alaska. The terrain surrounding Juneau is very rugged, and storms that sweep across the Gulf of Alaska usually hit Juneau. However, Juneau is located in southern Alaska, near the ocean, and temperatures are often comparable with those found in the lower 48 states. Madison is the capital of Wisconsin. The city is located between two large lakes. The climate of Madison is described as the typical continental climate of interior North America. Consider the long-term average temperatures (in degrees Fahrenheit) paired by month for the two cities (Source: National Weather Bureau). Use a sign test with a 5% level of significance to test the claim that the overall temperature distribution of Madison is different (either way) from that of Juneau. (See Problem 12 of Section 11.1.)

Month	Madison	Juneau
January	17.5	22.2
February	21.1	27.3
March	31.5	31.9
April	46.1	38.4
May	57.0	46.4
June	67.0	52.8
July	71.3	55.5
August	69.8	54.1
September	60.7	49.0
October	51.0	41.5
November	35.7	32.0
December	22.8	26.9

Medio Images / Photolibrary

SECTION 11.1

The Sign Test for Matched Pairs

FOCUS POINTS

- State the criteria for setting up a matched pair sign test.
- Complete a matched pair sign test.
- Interpret the results in the context of the application.

Nonparametric statistics

There are many situations in which very little is known about the population from which samples are drawn. Therefore, we cannot make assumptions about the population distribution, such as assuming the distribution is normal or binomial. In this chapter, we will study methods that come under the heading of *nonparametric statistics*. These methods are called *nonparametric* because they require no assumptions about the population distributions from which samples are drawn. The obvious advantages of these tests are that they are quite general and (as we shall see) not difficult to apply. The disadvantages are that they tend to waste information and tend to result in acceptance of the null hypothesis more often than they should. As such, nonparametric tests are sometimes *less sensitive* than other tests.

Sign test

The easiest of all the nonparametric tests is probably the *sign test*. The sign test is used when we compare sample distributions from two populations that are *not independent*. This occurs when we measure the sample twice, as in "before and after" studies. The following example shows how the sign test is constructed and used:

Criteria for sign test

As part of their training, 15 police cadets took a special course on identification awareness. To determine how the course affects a cadet's ability to identify a suspect, the 15 cadets were first given an identification-awareness exam and then, after the course, were tested again. The police school would like to use the results of the two tests to see if the identification-awareness course *improves* a cadet's score. Table 11-1 gives the scores for each exam.

The sign of the difference is obtained by subtracting the precourse score from the postcourse score. If the difference is positive, we say that the sign of the difference is +, and if the difference is negative, we indicate it with −. No sign is indicated if the scores are identical; in essence, such scores are ignored when using the sign test. To use the sign test, we need to compute the *proportion x of plus signs* to all signs. We ignore the pairs with no difference of signs. This is demonstrated in Guided Exercise 1.

Photo Researchers, Inc.

TABLE 11-1 **Scores for 15 Police Cadets**

Cadet	Postcourse Score	Precourse Score	Sign of Difference
1	93	76	+
2	70	72	−
3	81	75	+
4	65	68	−
5	79	65	+
6	54	54	No difference
7	94	88	+
8	91	81	+
9	77	65	+
10	65	57	+
11	95	86	+
12	89	87	+
13	78	78	No difference
14	80	77	+
15	76	76	No difference

GUIDED EXERCISE 1 **Proportion of plus signs**

Look at Table 11-1 under the "Sign of Difference" column.

(a) How many plus signs do you see? → 10

(b) How many plus and minus signs do you see? → 12

(c) The *proportion of plus signs* is → $x = \frac{10}{12} = \frac{5}{6} \approx 0.833$

$$x = \frac{\text{Number of plus signs}}{\text{Total number of plus and minus signs}}$$

Use parts (a) and (b) to find x.

We observe that x is the sample proportion of plus signs, and we use p to represent the population proportion of plus signs (if *all* possible police cadets were tested).

Null hypothesis

The null hypothesis is

H_0: $p = 0.5$ (the distributions of scores before and after the course are the same)

The null hypothesis states that the identification-awareness course does *not* affect the distribution of scores. Under the null hypothesis, we expect the number of plus signs and minus signs to be about equal. This means that the proportion of plus signs should be approximately 0.5.

Alternate hypothesis

The police department wants to see if the course *improves* a cadet's score. Therefore, the alternate hypothesis will be

H_1: $p > 0.5$ (the distribution of scores after the course is shifted higher than the distribution before the course)

The alternate hypothesis states that the identification-awareness course tends to improve scores. This means that the proportion of plus signs should be greater than 0.5.

Sampling distribution

To test the null hypothesis H_0: $p = 0.5$ against the alternate hypothesis H_1: $p > 0.5$, we use methods of Section 8.3 for tests of proportions. As in Section 8.3, we will assume that all our samples are sufficiently large to permit a normal approximation to the binomial distribution. For most practical work, this will be the case if the total number of plus and minus signs is 12 or more ($n \geq 12$).

When the total number of plus and minus signs is 12 or more, the sample statistic x (proportion of plus signs) has a distribution that is approximately normal, with mean p and standard deviation $\sqrt{pq/n}$ (See Section 6.6.)

Sample test statistic

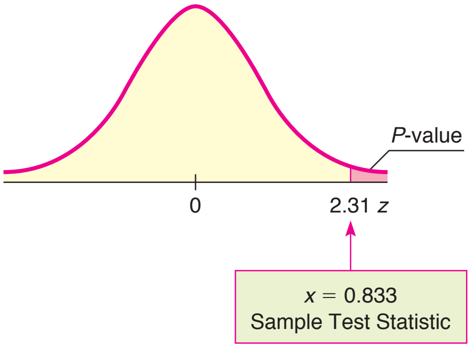
Under the null hypothesis H_0: $p = 0.5$, we assume that the population proportion p of plus signs is 0.5. Therefore, the z value corresponding to the sample test statistic x is

$$z = \frac{x - p}{\sqrt{\frac{pq}{n}}} = \frac{x - 0.5}{\sqrt{\frac{(0.5)(0.5)}{n}}} = \frac{x - 0.5}{\sqrt{\frac{0.25}{n}}}$$

where n is the total number of plus and minus signs, and x is the total number of plus signs divided by n.

FIGURE 11-1

P-value



For the police cadet example, we found $x \approx 0.833$ in Guided Exercise 1. The value of n is 12. (Note that of the 15 cadets in the sample, 3 had no difference in precourse and postcourse test scores, so there are no signs for these 3.) The z value corresponding to $x = 0.833$ is then

$$z \approx \frac{0.833 - 0.5}{\sqrt{\frac{0.25}{12}}} \approx 2.31$$

P-value

We use the standard normal distribution table (Table 5 of Appendix II) to find P-values for the sign test. This table gives areas to the left of z. Recall from Section 8.2 that Table 5 of Appendix II can be used directly to find P-values of one-tailed tests. For *two-tailed* tests, we must *double* the value given in the table. To review the process of finding areas to the right or left of z using Table 5, see Section 6.2.

The alternate hypothesis for the police cadet example is H_1: $p > 0.5$. The P-value for the sample test statistic $z = 2.31$ is shown in Figure 11-1. For a right-tailed test, the P-value is the area to the right of the sample test statistic $z = 2.31$. From Table 5 of Appendix II, $P(z > 2.31) = 0.0104$.

Conclude the test and interpret the results

In our example, the police department wishes to use a 5% level of significance to test the claim that the identification-awareness course improves a cadet's score. Since the P-value of 0.0104 is less than $\alpha = 0.05$, we reject the null hypothesis H_0 that the course makes no difference. Instead, at the 5% level of significance, we say the results are significant. The evidence is sufficient to claim that the identification-awareness course improves cadets' scores.

The steps used to construct a sign test for matched pairs are summarized in the next procedure.

PROCEDURE

HOW TO CONSTRUCT A SIGN TEST FOR MATCHED PAIRS

Setup and Requirements

You first need a random sample of data pairs (A, B). Next, you take the differences $A - B$ and record the sign change for each difference: plus, minus, or no change. The number of data pairs should be large enough that the total number of plus and minus signs is at least 12. The sample proportion of plus signs is

$$x = \frac{\text{number of plus signs}}{\text{total number of plus and minus signs}}$$

Let p represent the population proportion of plus signs if the entire population of all possible data pairs (A, B) were to be used.

Continued

Procedure

1. Set the *level of significance* α. The *null hypothesis* is H_0: $p = 0.5$. In the context of the application, set the *alternate hypothesis:* H_1: $p > 0.5$, H_1: $p < 0.5$, or H_1: $p \neq 0.5$.
2. The *sample test statistic* is

$$z = \frac{x - 0.5}{\sqrt{\frac{0.25}{n}}}$$

where $n \geq 12$ is the total number of plus and minus signs.

3. Use the standard normal distribution and the type of test, one-tailed or two-tailed, to find the *P-value* corresponding to the test statistic.
4. *Conclude* the test. If P-value $\leq \alpha$, then reject H_0. If P-value $> \alpha$, then do not reject H_0.
5. *Interpret your conclusion* in the context of the application.

GUIDED EXERCISE 2 Sign test

Dr. Kick-a-poo's Traveling Circus made a stop at Middlebury, Vermont, where the doctor opened a booth and sold bottles of Dr. Kick-a-poo's Magic Gasoline Additive. The additive is supposed to increase gas mileage when used according to instructions. Twenty local people purchased bottles of the additive and used it according to instructions. These people carefully recorded their mileage with and without the additive. The results are shown in Table 11-2.

TABLE 11-2 **Mileage Before and After Kick-a-poo's Additive**

Car	With Additive	Without Additive	Sign of Difference
1	17.1	16.8	+
2	21.2	20.1	+
3	12.3	12.3	No difference (N.D.)
4	19.6	21.0	−
5	22.5	20.9	+
6	17.0	17.9	___
7	24.2	25.4	___
8	22.2	20.1	___
9	18.3	19.1	___
10	11.0	12.3	___
11	17.6	14.2	___
12	22.1	23.7	___
13	29.9	30.2	___
14	27.6	27.6	___
15	28.4	27.7	___
16	16.1	16.1	___
17	19.0	19.5	___
18	38.7	37.9	___
19	17.6	19.7	___
20	21.6	22.2	___

TABLE 11-3 **Completion of Table 11-2**

Car	Sign of Difference
6	−
7	−
8	+
9	−
10	−
11	+
12	−
13	−
14	N.D.
15	+
16	N.D.
17	−
18	+
19	−
20	−

Continued

GUIDED EXERCISE 2 *continued*

(a) In Table 11-2, complete the column headed "Sign of Difference." How many plus signs are there? How many total plus and minus signs are there? What is the value of x, the proportion of plus signs?

→ There are 7 plus signs and 17 total plus and minus signs. The proportion of plus signs is

$$x = \frac{7}{17} \approx 0.412$$

(b) Most people claim that the additive has no effect. Let's use a 0.05 level of significance to test this claim against the alternate hypothesis that the additive did have an effect (one way or the other). State the null and alternate hypotheses.

→ We use
H_0: $p = 0.5$ (mileage distributions are the same)
H_1: $p \neq 0.5$ (mileage distributions are different)

(c) Convert the sample x value, $x = 0.412$, to a z value.

→ To find the z value corresponding to $x = 0.412$, we use $n = 17$ (total number of signs).

$$z = \frac{x - 0.5}{\sqrt{0.25/n}} \approx \frac{0.412 - 0.5}{\sqrt{0.25/17}} \approx -0.73$$

(d) Find the corresponding P-value.

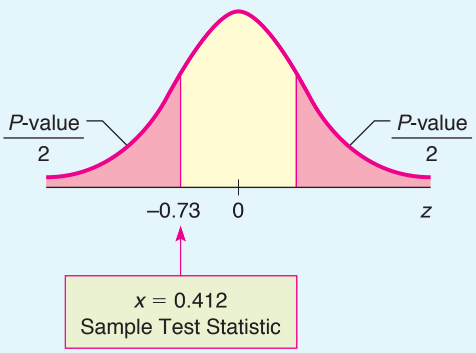
→ Table 5 of Appendix II gives the area to the left of $z = -0.73$.

$P(z < -0.73) = 0.2327$

Because this is a two-tailed test, the P-value is double this area.

$P\text{-value} = 2(0.2327) = 0.4654$

FIGURE 11-2 *P*-value



(e) Conclude the test.

→ For $\alpha = 0.05$, we see that the P-value $= 0.4654$ is greater than α. We fail to reject H_0.

(f) ***Interpret*** the results.

→ At the 5% level of significance, the data are not statistically significant, and we cannot reject the hypothesis that the mileage distribution is the same with or without the additive.

VIEWPOINT *Yukon News*

The Yukon News *featured an article entitled "Resurgence of the Dreaded White Plague," about the resurgence of tuberculosis (TB) in the far north. TB, also known as the white plague, has been present in Canada since it was brought in by European immigrants in the 17th century. Although antibiotics are widely used today, the disease has never been eradicated. Canadian National Health data suggest that TB is spreading faster in the Yukon than elsewhere in Canada. Because of this, the Canadian government has established many new TB clinics in remote Yukon villages. Using what you have learned in this section and Canadian National Health data, can you think of a way to use a sign test to study the claim that in these villages, the rate of TB in the population dropped after the clinics were activated?*

SECTION 11.1 PROBLEMS

1. *Statistical Literacy* To apply the sign test, do you need independent or dependent (matched pair) data?

2. *Statistical Literacy* For the sign test of matched pairs, do pairs for which the difference in values is zero enter into any calculations?

For Problems 3–12, please provide the following information.
(a) What is the level of significance? State the null and alternate hypotheses.
(b) Compute the sample test statistic. What is the sampling distribution?
(c) Find the P-value of the sample test statistic.
(d) Conclude the test.
(e) *Interpret* the conclusion in the context of the application.

3. *Economic Growth: Asia* Asian economies impact some of the world's largest populations. The growth of an economy has a big influence on the everyday lives of ordinary people. Are Asian economies changing? A random sample of 15 Asian economies gave the following information about annual percentage growth rate (Reference: *Handbook of International Economic Statistics,* U.S. Government Documents).

Region	1	2	3	4	5	6	7	8
Modern Growth Rate %	4.0	2.3	7.8	2.8	0.7	5.1	2.9	4.2
Historic Growth Rate %	3.3	1.9	7.0	5.5	3.3	6.0	3.2	8.2

Region	9	10	11	12	13	14	15
Modern Growth Rate %	4.9	5.8	6.8	3.6	3.2	0.8	7.3
Historic Growth Rate %	6.4	7.2	6.1	1.5	1.0	2.1	5.1

Does this information indicate a change (either way) in the growth rate of Asian economies? Use a 5% level of significance.

4. *Debt: Developing Countries* Borrowing money may be necessary for business expansion. However, too much borrowed money can also mean trouble. Are developing countries tending to borrow more? A random sample of 20 developing countries gave the following information regarding foreign debt per capita (in U.S. dollars, inflation adjusted) (Reference: *Handbook of International Economic Statistics,* U.S. Government Documents).

Country	1	2	3	4	5	6	7	8	9	10
Modern Debt per Capita	179	157	129	125	91	80	31	25	29	85
Historic Debt per Capita	144	132	88	112	53	66	31	30	40	75

Country	11	12	13	14	15	16	17	18	19	20
Modern Debt per Capita	27	20	17	21	195	189	143	126	106	76
Historic Debt per Capita	21	19	15	24	104	150	142	118	117	79

Does this information indicate that foreign debt per capita is increasing in developing countries? Use a 1% level of significance.

5. *Education: Exams* A high school science teacher decided to give a series of lectures on current events. To determine if the lectures had any effect on student awareness of current events, an exam was given to the class before the lectures, and a similar exam was given after the lectures. The scores follow.

Use a 0.05 level of significance to test the claim that the lectures made no difference against the claim that the lectures did make some difference (one way or the other).

Student	1	2	3	4	5	6	7	8	9
After Lectures	107	115	120	78	83	56	71	89	77
Before Lectures	111	110	93	75	88	56	75	73	83

Student	10	11	12	13	14	15	16	17	18
After Lectures	44	119	130	91	99	96	83	100	118
Before Lectures	40	115	101	110	90	98	76	100	109

6. ***Grain Yields: Feeding the World*** With an ever-increasing world population, grain yields are extremely important. A random sample of 16 large grain-producing regions in the world gave the following information about grain production (in kg/hectare) (Reference: *Handbook of International Economic Statistics,* U.S. Government Documents).

Region	1	2	3	4	5	6	7	8
Modern Production	1610	2230	5270	6990	2010	4560	780	6510
Historic Production	1590	2360	5161	7170	1920	4760	660	6320

Region	9	10	11	12	13	14	15	16
Modern Production	2850	3550	1710	2050	2750	2550	6750	3670
Historic Production	2920	2440	1340	2180	3110	2070	7330	2980

Does this information indicate that modern grain production is higher? Use a 5% level of significance.

7. ***Identical Twins: Reading Skills*** To compare two elementary schools regarding teaching of reading skills, 12 sets of identical twins were used. In each case, one child was selected at random and sent to school A, and his or her twin was sent to school B. Near the end of fifth grade, an achievement test was given to each child. The results follow:

Twin Pair	1	2	3	4	5	6
School A	177	150	112	95	120	117
School B	86	135	115	110	116	84

Twin Pair	7	8	9	10	11	12
School A	86	111	110	142	125	89
School B	93	77	96	130	147	101

Use a 0.05 level of significance to test the hypothesis that the two schools have the same effectiveness in teaching reading skills against the alternate hypothesis that the schools are not equally effective.

8. ***Incomes: Electricians and Carpenters*** How do the average weekly incomes of electricians and carpenters compare? A random sample of 17 regions in the United States gave the following information about average weekly income (in dollars) (Reference: U.S. Department of Labor, Bureau of Labor Statistics).

Region	1	2	3	4	5	6	7	8	9
Electricians	461	713	593	468	730	690	740	572	805
Carpenters	540	812	512	473	686	507	785	657	475

Region	10	11	12	13	14	15	16	17
Electricians	593	593	700	572	863	599	596	653
Carpenters	485	646	675	382	819	600	559	501

Does this information indicate a difference (either way) in the average weekly incomes of electricians compared to those of carpenters? Use a 5% level of significance.

9. ***Quitting Smoking: Hypnosis*** One program to help people stop smoking cigarettes uses the method of posthypnotic suggestion to remind subjects to avoid smoking. A random sample of 18 subjects agreed to test the program. All subjects counted the number of cigarettes they usually smoke a day; then they counted the number of cigarettes smoked the day after hypnosis. (*Note:* It usually takes several weeks for a subject to stop smoking completely, and the method does not work for everyone.) The results follow.

	Cigarettes Smoked per Day			Cigarettes Smoked per Day	
Subject	After Hypnosis	Before Hypnosis	Subject	After Hypnosis	Before Hypnosis
1	28	28	10	5	19
2	15	35	11	12	32
3	2	14	12	20	42
4	20	20	13	30	26
5	31	25	14	19	37
6	19	40	15	0	19
7	6	18	16	16	38
8	17	15	17	4	23
9	1	21	18	19	24

Using a 1% level of significance, test the claim that the number of cigarettes smoked per day was less after hypnosis.

10. ***Incomes: Lawyers and Architects*** How do the average weekly incomes of lawyers and architects compare? A random sample of 18 regions in the United States gave the following information about average weekly incomes (in dollars) (Reference: U.S. Department of Labor, Bureau of Labor Statistics).

Region	1	2	3	4	5	6	7	8	9
Lawyers	709	898	848	1041	1326	1165	1127	866	1033
Architects	859	936	887	1100	1378	1295	1039	888	1012

Region	10	11	12	13	14	15	16	17	18
Lawyers	718	835	1192	992	1138	920	1397	872	1142
Architects	794	900	1150	1038	1197	939	1124	911	1171

Does this information indicate that architects tend to have a larger average weekly income? Use $\alpha = 0.05$.

11. *High School Dropouts: Male versus Female* Is the high school dropout rate higher for males or females? A random sample of population regions gave the following information about percentage of 15- to 19-year-olds who are high school dropouts (Reference: *Statistical Abstract of the United States,* 121st Edition).

Region	1	2	3	4	5	6	7	8	9	10
Male	7.3	7.5	7.7	21.8	4.2	12.2	3.5	4.2	8.0	9.7
Female	7.5	6.4	6.0	20.0	2.6	5.2	3.1	4.9	12.1	10.8

Region	11	12	13	14	15	16	17	18	19	20
Male	14.1	3.6	3.6	4.0	5.2	6.9	15.6	6.3	8.0	6.5
Female	15.6	6.3	4.0	3.9	9.8	9.8	12.0	3.3	7.1	8.2

Does this information indicate that the dropout rates for males and females are different (either way)? Use $\alpha = 0.01$.

12. *Focus Problem: Meteorology* The Focus Problem at the beginning of this chapter asks you to use a sign test with a 5% level of significance to test the claim that the overall temperature distribution of Madison, Wisconsin, is different (either way) from that of Juneau, Alaska. The monthly average data (in °F) are as follows.

Month	Jan.	Feb.	March	April	May	June
Madison	17.5	21.1	31.5	46.1	57.0	67.0
Juneau	22.2	27.3	31.9	38.4	46.4	52.8

Month	July	Aug.	Sept.	Oct.	Nov.	Dec.
Madison	71.3	69.8	60.7	51.0	35.7	22.8
Juneau	55.5	54.1	49.0	41.5	32.0	26.9

What is your conclusion?

SECTION 11.2

The Rank-Sum Test

FOCUS POINTS

- State the criteria for setting up a rank-sum test.
- Use the distribution of ranks to complete the test.
- Interpret the results in the context of the application.

The sign test is used when we have paired data values coming from dependent samples, as in "before and after" studies. However, if the data values are *not paired,* the sign test should *not* be used.

Criteria for rank-sum test

For the situation in which we draw *independent random samples* from two populations, there is another nonparametric method for testing the difference between sample means; it is called the *rank-sum test* (also called the *Mann–Whitney test*). The rank-sum test can be used when assumptions about *normal* populations are not satisfied. To fix our thoughts on a definite problem, let's consider the following example:

When a scuba diver makes a deep dive, nitrogen builds up in the diver's blood. After returning to the surface, the diver must wait in a decompression

TABLE 11-4 Decompression Times for 23 Navy Divers (in min)

Group A (had pill)	41	56	64	42	50	70	44	57	63	65	52	
	Mean time = 54.91 min											
Group B (no pill)	66	43	72	62	55	80	74	75	77	78	47	60
	Mean time = 65.75 min											

chamber until the nitrogen level of the blood returns to normal. A physiologist working with the Navy has invented a pill that a diver takes 1 hour before diving. The pill is supposed to reduce the waiting time spent in the decompression chamber. Twenty-three Navy divers volunteered to help the physiologist determine if the pill has any effect. The divers were randomly divided into two groups: group A had 11 divers who took the pill, and group B had 12 divers who did not take the pill. All the divers worked the same length of time on a deep salvage operation and returned to the decompression chamber. A monitoring device in the decompression chamber measured the waiting time for each diver's nitrogen level to return to normal. These times are recorded in Table 11-4.

Rank the data

The means of our two samples are 54.91 and 65.75 minutes. We will use the rank-sum test to decide whether the difference between the means is significant. First, we arrange the two samples jointly in order of increasing time. To do this, we use the data of groups A and B as if they were one sample. The times (in minutes), groups, and ranks are shown in Table 11-5.

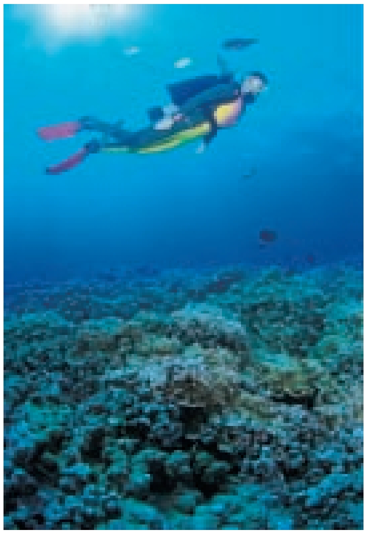
Group A occupies the ranks 1, 2, 4, 6, 7, 9, 10, 13, 14, 15, and 17, while group B occupies the ranks 3, 5, 8, 11, 12, 16, 18, 19, 20, 21, 22, and 23. We add up the ranks of the group with the *smaller* sample size, in this case, group A.

Sum the ranks of the smaller group

The sum of the ranks is denoted by R:

$$R = 1 + 2 + 4 + 6 + 7 + 9 + 10 + 13 + 14 + 15 + 17 = 98$$

Let n_1 be the size of the *smaller sample* and n_2 be the size of the *larger sample*. In the case of the divers, $n_1 = 11$ and $n_2 = 12$. So, R is the sum of the ranks from the smaller sample. If both samples are of the same size, then $n_1 = n_2$ and R is the sum of the ranks of either group (but not both groups).



Andrew G. Wood/Photo Researchers, Inc.

TABLE 11-5 Ranks for Decompression Time

Time	Group	Rank	Time	Group	Rank
41	A	1	63	A	13
42	A	2	64	A	14
43	B	3	65	A	15
44	A	4	66	B	16
47	B	5	70	A	17
50	A	6	72	B	18
52	A	7	74	B	19
55	B	8	75	B	20
56	A	9	77	B	21
57	A	10	78	B	22
60	B	11	80	B	23
62	B	12			

Distribution of ranks

When both n_1 and n_2 are sufficiently large (each greater than 10), advanced mathematical statistics can be used to show that R is approximately normally distributed, with mean

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2}$$

and standard deviation

$$\sigma_R = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$$

GUIDED EXERCISE 3 **Mean and standard deviation of ranks**

For the Navy divers, compute μ_R and σ_R. (Recall that $n_1 = 11$ and $n_2 = 12$.)

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{11(11 + 12 + 1)}{2} = 132$$

$$\sigma_R = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$$

$$= \sqrt{\frac{11 \cdot 12(11 + 12 + 1)}{12}} \approx 16.25$$

Sample test statistic

Since $n_1 = 11$ and $n_2 = 12$, the samples are large enough to assume that the rank R is approximately normally distributed. We convert the sample test statistic R to a z value using the following formula, with $R = 98$, $\mu_R = 132$, and $\sigma_R \approx 16.25$:

$$z = \frac{R - \mu_R}{\sigma_R} \approx \frac{98 - 132}{16.25} \approx -2.09$$

Hypotheses

When using the rank-sum test, the null hypothesis is that the distributions are the same, while the alternate hypothesis is that the distributions are different. In the case of the Navy divers, we have

H_0: Decompression time distributions are the same.

H_1: Decompression time distributions are different.

We'll test the decompression time distributions using level of significance 5%.

P-value

To find the P-value of the sample test statistic $z = -2.09$, we use the normal distribution (Table 5 of Appendix II) and the fact that we have a two-tailed test. Figure 11-3 shows the P-value.

The area to the left of -2.09 is 0.0183. This is a two-tailed test, so

$$P\text{-value} = 2(0.0183) = 0.0366$$

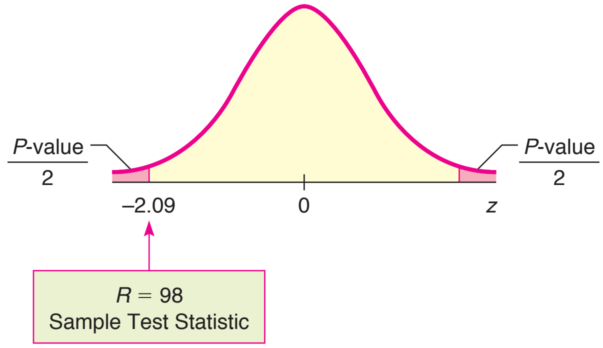
Conclusion

Since the P-value is less than $\alpha = 0.05$, we reject H_0. At the 5% level of significance, we have sufficient evidence to conclude that the pill changes decompression times for divers.

The steps necessary for a rank-sum test are summarized by the procedure on the next page.

FIGURE 11-3

P-value



PROCEDURE

How to construct a rank-sum test

Setup and Requirements

You first need independent random samples (both of size 11 or more) from two populations A and B. Let n_1 be the sample size of the *smaller* sample and let n_2 be the sample size of the larger sample. If the sample sizes are equal, then simply use the common value for n_1 and n_2. Next, you need to rank-order the data as if they were one big sample. Label each rank A or B according to the population from which it came. Let R be a random variable that represents the sum of ranks from the sample of size n_1. If $n_1 = n_2$, then R is the sum of ranks from either group (but not both).

Procedure

1. Set the *level of significance* α. The *null* and *alternate hypotheses* are

   H_0: The two samples come from populations with the same distribution (the two populations are identical).

   H_1: The two samples come from populations with different distributions (the populations differ in some way).

2. The *sample test statistic* is

   $$z = \frac{R - \mu_R}{\sigma_R}$$

   where $R =$ sum of ranks from the sample of size n_1 (smaller sample),

   $$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2}$$

   $$\sigma_R = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$$

   and $n_1 > 10$, $n_2 > 10$.

3. Use the standard normal distribution with a two-tailed test to find the *P-value* corresponding to the test statistic.
4. *Conclude* the test. If P-value $\leq \alpha$, then reject H_0. If P-value $> \alpha$, then do not reject H_0.
5. *Interpret your conclusion* in the context of the application.

Procedure for tied ranks

NOTE For the decompression time data, there were no ties for any rank. If a tie does occur, then each of the tied observations is given the *mean* of the ranks that it occupies. For example, if we rank the numbers

41 42 44 44 44 44

TABLE 11-6

Observation	Rank
41	1
42	2
44	4.5
44	4.5
44	4.5
44	4.5

we see that 44 occupies ranks 3, 4, 5, and 6. Therefore, we give each of the 44's a rank that is the mean of 3, 4, 5, and 6:

$$\text{Mean of ranks} = \frac{3 + 4 + 5 + 6}{4} = 4.5$$

The final ranking would then be that shown in Table 11-6.

For samples where n_1 or n_2 is less than 11, there are statistical tables that give appropriate critical values for the rank-sum test. Most libraries contain such tables, and the interested reader can find such information by looking under the *Mann–Whitney U Test*.

GUIDED EXERCISE 4 Rank-sum test

A biologist is doing research on elk in their natural Colorado habitat. Two regions are under study, both having about the same amount of forage and natural cover. However, region A seems to have fewer predators than region B. To determine if there is a difference in elk life spans between the two regions, a sample of 11 mature elk from each region are tranquilized and have a tooth removed. A laboratory examination of the teeth reveals the ages of the elk. Results for each sample are given in Table 11-7. The biologist uses a 5% level of significance to test for a difference in life spans.

TABLE 11-7 **Ages of Elk**

Group A	4	10	11	2	2	3	9	4	12	6	6
Group B	7	3	8	4	8	5	6	4	2	4	3

(a) Fill in the remaining ranks of Table 11-8. Be sure to use the process of taking the mean of tied ranks.

TABLE 11-8 **Ranks of Elk**

Age	Group	Rank	Age	Group	Rank
2	A	2	5	B	12
2	A	2	6	A	—
2	B	2	6	A	—
3	A	5	6	B	—
3	B	5	7	B	—
3	B	5	8	B	—
4	A	9	8	B	—
4	A	9	9	A	—
4	B	9	10	A	—
4	B	9	11	A	—
4	B	9	12	A	—

Rank
12
14
14
14
16
17.5
17.5
19
20
21
22

(b) What is α? State the null and alternate hypotheses.

$\alpha = 0.05$

H_0: Distributions of life spans are the same.

H_1: Distributions of life spans are different.

Continued

GUIDED EXERCISE 4 *continued*

(c) Find μ_R, σ_R, and R. Convert R to a sample z statistic.

Since $n_1 = 11$ and $n_2 = 11$,

$$\mu_R = \frac{(11)(11 + 11 + 1)}{2} = 126.5$$

$$\sigma_R = \sqrt{\frac{11 \cdot 11(11 + 11 + 1)}{12}} \approx 15.23$$

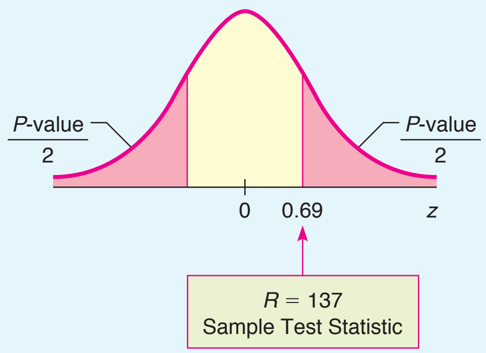
Since $n_1 = n_2 = 11$, we can use the sum of the ranks of either the A group or the B group. Let's use the A group. The A group ranks are 2, 2, 5, 9, 9, 14, 14, 19, 20, 21, and 22. Therefore,

$$R = 2 + 2 + 5 + 9 + 9 + 14 + 14 + 19 + 20 + 21 + 22 = 137$$

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{137 - 126.5}{15.23} \approx 0.69$$

(d) Find the P-value shown in Figure 11-4.

FIGURE 11-4 *P*-value



Using Table 5 of Appendix II, the area to the right of 0.69 is 0.2451. Since this is a two-tailed test,

P-value $= 2(0.2451) = 0.4902$

Comment: If we use the sum of ranks of group B, then $R_B = 116$ and $z = -0.69$. The P-value is again 0.4902, and we have the same conclusion.

(e) ***Interpretation*** What is the conclusion?

The P-value of 0.4902 is greater than $\alpha = 0.05$, so we do not reject H_0. The evidence does not support the claim that the age distribution of elk is different between the two regions.

VIEWPOINT Point Barrow, Alaska

Point Barrow is located very near the northernmost point of land in the United States. In 1935, Will Rogers (an American humorist, social critic, and philosopher) was killed with Wiley Post (a pioneer aviator) at a landing strip near Point Barrow. Since 1920, a weather station at the (now named) Wiley Post–Will Rogers Memorial Landing Strip has recorded daily high and low temperatures. From these readings, annual mean maximum and minimum temperatures have been computed. Is Point Barrow warming up, cooling down, or neither? Can you think of a way to gather data and construct a nonparametric test to investigate long-term temperature highs and lows at Point Barrow? For weather-related data, visit the Brase/Brase statistics site at **http://www.cengage.com/statistics/brase** *and find a link to the Geophysical Institute at the University of Alaska in Fairbanks. Then follow the links to Point Barrow.*

SECTION 11.2 PROBLEMS

1. *Statistical Literacy* When applying the rank-sum test, do you need independent or dependent samples?

2. *Statistical Literacy* If two or more data values are the same, how is the rank of each of the tied data computed?

For Problems 3–11, please provide the following information.
(a) What is the level of significance? State the null and alternate hypotheses.
(b) Compute the sample test statistic. What is the sampling distribution? What conditions are necessary to use this distribution?
(c) Find the P-value of the sample test statistic.
(d) Conclude the test.
(e) ***Interpret*** the conclusion in the context of the application.

3. ***Agriculture: Lima Beans*** Are yields for organic farming different from conventional farming yields? Independent random samples from method A (organic farming) and method B (conventional farming) gave the following information about yield of lima beans (in tons/acre) (Reference: *Agricultural Statistics,* U.S. Department of Agriculture).

Method A	1.83	2.34	1.61	1.99	1.78	2.01	2.12	1.15	1.41	1.95	1.25	
Method B	2.15	2.17	2.11	1.89	1.34	1.88	1.96	1.10	1.75	1.80	1.53	2.21

Use a 5% level of significance to test the hypothesis that there is no difference between the yield distributions.

4. ***Agriculture: Sweet Corn*** Are yields for organic farming different from conventional farming yields? Independent random samples from method A (organic farming) and method B (conventional farming) gave the following information about yield of sweet corn (in tons/acre) (Reference: *Agricultural Statistics,* U.S. Department of Agriculture).

Method A	6.88	6.86	7.12	5.91	6.80	6.92	6.25	6.98	7.21	7.33	5.85	6.72
Method B	5.71	6.93	7.05	7.15	6.79	6.87	6.45	7.34	5.68	6.78	6.95	

Use a 5% level of significance to test the claim that there is no difference between the yield distributions.

5. ***Horse Trainer: Jumps*** A horse trainer teaches horses to jump by using two methods of instruction. Horses being taught by method A have a lead horse that accompanies each jump. Horses being taught by method B have no lead horse. The table shows the number of training sessions required before each horse performed the jumps properly.

Method A	28	35	19	41	37	31	38	40	25	27	36	43
Method B	42	33	26	24	44	46	34	20	48	39	45	

Use a 5% level of significance to test the claim that there is no difference between the training session distributions.

6. ***Violent Crime: FBI Report*** Is the crime rate in New York different from the crime rate in New Jersey? Independent random samples from region A (cities in New York) and region B (cities in New Jersey) gave the following information about violent crime rate (number of violent crimes per 100,000 population) (Reference: U.S. Department of Justice, Federal Bureau of Investigation).

Region A	554	517	492	561	577	621	512	580	543	605	531	
Region B	475	419	505	575	395	433	521	388	375	411	586	415

Use a 5% level of significance to test the claim that there is no difference in the crime rate distributions of the two states.

7. ***Psychology: Testing*** A cognitive aptitude test consists of putting together a puzzle. Eleven people in group A took the test in a competitive setting (first and second to finish received a prize). Twelve people in group B took the test in a noncompetitive setting. The results follow (in minutes required to complete the puzzle).

Group A	7	12	10	15	22	17	18	13	8	16	11	
Group B	9	16	30	11	33	28	19	14	24	27	31	29

Use a 5% level of significance to test the claim that there is no difference in the distributions of time to complete the test.

8. ***Psychology: Testing*** A psychologist has developed a mental alertness test. She wishes to study the effects (if any) of type of food consumed on mental alertness. Twenty-one volunteers were randomly divided into two groups. Both groups were told to eat the amount they usually eat for lunch at noon. At 2:00 P.M., all subjects were given the alertness test. Group A had a low-fat lunch with no red meat, lots of vegetables, carbohydrates, and fiber. Group B had a high-fat lunch with red meat, vegetable oils, and low fiber. The only drink for both groups was water. The test scores are shown below.

Group A	76	93	52	81	68	79	88	90	67	85	60	
Group B	44	57	60	91	62	86	82	65	96	42	68	98

Use a 1% level of significance to test the claim that there is no difference in mental alertness distributions based on type of lunch.

9. ***Lifestyles: Exercise*** Is there a link between exercise and level of education? Independent random samples of adults from group A (college graduates) and group B (no high school diploma) gave the following information about percentage who exercise regularly (Reference: Center for Disease Control and Prevention).

A(%)	63.3	55.1	50.0	47.1	58.2	60.0	44.3	49.1	68.7	57.3	59.9	
B(%)	33.7	40.1	53.3	36.9	29.1	59.6	35.7	44.2	38.2	46.6	45.2	60.2

Use a 1% level of significance to test the claim that there is no difference in the exercise rate distributions according to education level.

10. ***Doctor's Degree: Years of Study*** Does the average length of time to earn a doctorate differ from one field to another? Independent random samples from large graduate schools gave the following averages for length of registered time (in years) from bachelor's degree to doctorate. Sample A was taken from the humanities field, and sample B from the social sciences field (Reference: *Education Statistics*, U.S. Department of Education).

Field A	8.9	8.3	7.2	6.4	8.0	7.5	7.1	6.0	9.2	8.7	7.5	
Field B	7.6	7.9	6.2	5.8	7.8	8.3	8.5	7.0	6.3	5.4	5.9	7.7

Use a 1% level of significance to test the claim that there is no difference in the distributions of time to complete a doctorate for the two fields.

11. ***Education: Spelling*** Twenty-two fourth-grade children were randomly divided into two groups. Group A was taught spelling by a phonetic method. Group B was taught spelling by a memorization method. At the end of the fourth grade, all children were given a standard spelling exam. The scores are as follows.

Group A	77	95	83	69	85	92	61	79	87	93	65	78
Group B	62	90	70	81	63	75	80	72	82	94	65	79

Use a 1% level of significance to test the claim that there is no difference in the test score distributions based on instruction method.

SECTION 11.3

Spearman Rank Correlation

FOCUS POINTS

- Learn about monotone relations and the Spearman rank correlation coefficient.
- Compute the Spearman correlation coefficient and conduct statistical tests for significance.
- Interpret the results in the context of the application.

Data given in ranked form (ordinal type) are different from data given in measurement form (interval or ratio type). For instance, if we compared the test performances of three students and, say, Elizabeth did the best, Joel did next best, and Sally did the worst, we are giving the information in ranked form. We cannot say how much better Elizabeth did than Sally or Joel, but we do know how the three scores compare. If the actual test scores for the three tests were given, we would have data in measurement form and could tell exactly how much better Elizabeth did than Joel or Sally. In Chapter 9, we studied linear correlation of data in measurement form. In this section, we will study correlation of data in ranked form.

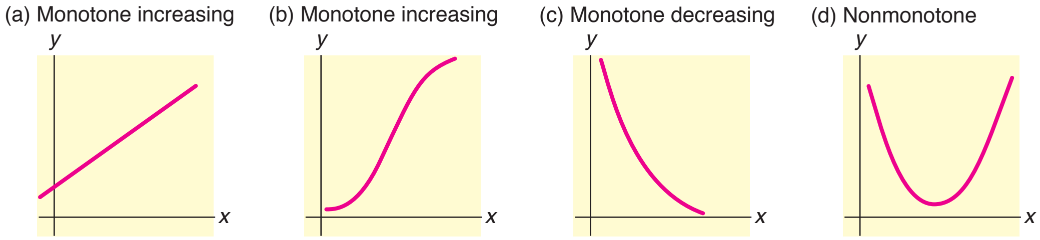
As a specific example of a situation in which we might want to compare ranked data from two sources, consider the following. Hendricks College has a new faculty position in its political science department. A national search to fill this position has resulted in a large number of qualified candidates. The political science faculty reserves the right to make the final hiring decision. However, the faculty is interested in comparing its opinion with student opinion about the teaching ability of the candidates. A random sample of nine equally qualified candidates were asked to give a classroom presentation to a large class of students. Both faculty and students attended the lectures. At the end of each lecture, both faculty and students filled out a questionnaire about the teaching performance of the candidate. Based on these questionnaires, each candidate was given an overall rank from the faculty and an overall rank from the students. The results are shown in Table 11-9. Higher ranks mean better teaching performance.

TABLE 11-9 Faculty and Student Ranks of Candidates

Candidate	Faculty Rank	Student Rank
1	3	5
2	7	7
3	5	6
4	9	8
5	2	3
6	8	9
7	1	1
8	6	4
9	4	2

FIGURE 11-5

Examples of Monotone Relations



Using data in ranked form, we answer the following questions:

1. Do candidates getting higher ranks from faculty tend to get higher ranks from students?
2. Is there any relation between faculty rankings and student rankings?
3. Do candidates getting higher ranks from faculty tend to get lower ranks from students?

We will use the Spearman rank correlation to answer such questions. In the early 1900s, Charles Spearman of the University of London developed the techniques that now bear his name. The Spearman test of rank correlation requires us to use *ranked variables*. Because we are using only ranks, we cannot use the Spearman test to check for the existence of a linear relationship between the variables as we did with the Pearson correlation coefficient (Section 9.1). The Spearman test checks only for the existence of a *monotone* relationship between the variables. (See Figure 11-5.) By a *monotone relationship** between variables x and y, we mean a relationship in which

Monotone relationship

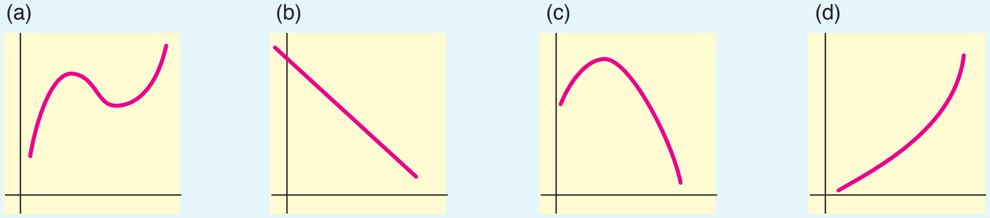
1. as x increases, y also increases, or
2. as x increases, y decreases.

The relationship shown in Figure 11-5(d) is a nonmonotone relationship because as x increases, y at first decreases, but later starts to increase. Remember, for a relation to be monotone, as x increases, y must *always* increase or *always* decrease. In a nonmonotone relationship, as x increases, y sometimes increases and sometimes decreases or stays unchanged.

GUIDED EXERCISE 5 *Monotonic behavior*

Identify each of the relations in Figure 11-6 as monotone increasing, monotone decreasing, or nonmonotone.

FIGURE 11-6



Answers: (a) nonmonotone, (b) monotone decreasing, (c) nonmonotone, (d) monotone increasing

*Some advanced texts call the monotone relationship we describe *strictly monotone*.

Before we can complete the solution of our problem about the political science department at Hendricks College, we need the following information.

Suppose we have a sample of size n of randomly obtained ordered pairs (x, y), where both the x and y values are from *ranked variables*. If there are no ties in the ranks, then the Pearson product-moment correlation coefficient (Section 9.1) can be reduced to a simpler equation. The new equation produces the *Spearman rank correlation coefficient*, r_s.

Spearman rank correlation coefficient r_s

Spearman rank correlation coefficient

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} \quad \text{where } d = x - y$$

The Spearman rank correlation coefficient has the following properties.

Properties of the Spearman rank correlation coefficient

1. $-1 \le r_s \le 1$. If $r_s = -1$, the relation between x and y is perfectly mono tone decreasing. If $r_s = 0$, there is no monotone relation between x and y. If $r_s = 1$, the relation between x and y is perfectly monotone increasing. Values of r_s close to 1 or -1 indicate a strong tendency for x and y to have a monotone relationship (increasing or decreasing). Values of r_s close to 0 indicate a very weak (or perhaps nonexistent) monotone relationship.
2. The probability distribution of r_s depends on the sample size n. It is symmetric about $r_s = 0$. Table 9 of Appendix II gives critical values for certain specified one-tail and two-tail areas. Use of the table requires no assumptions that x and y are normally distributed variables. In addition, we make no assumption about the x and y relationship being linear.
3. The Spearman rank correlation coefficient r_s is the *sample* estimate for the *population* Spearman rank correlation coefficient ρ_s.

Population Spearman rank correlation coefficient ρ_s

Hypotheses

We construct a test of significance for the Spearman rank correlation coefficient in much the same way that we tested the Pearson correlation coefficient (Section 9.3). The null hypothesis states that there is no monotone relation between x and y (either increasing or decreasing).

H_0: $\rho_s = 0$

The alternate hypothesis is one of the following:

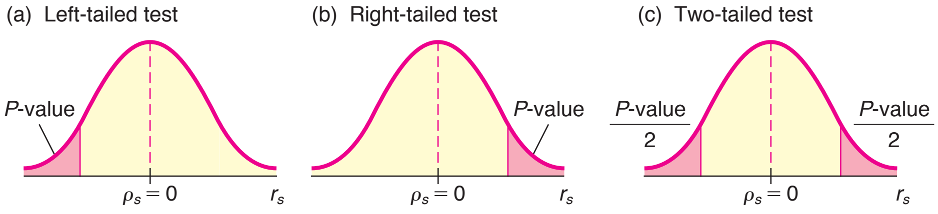
H_1: $\rho_s < 0$	H_1: $\rho_s > 0$	H_1: $\rho_s \neq 0$
(left-tailed)	(right-tailed)	(two-tailed)

A left-tailed alternate hypothesis claims there is a monotone-decreasing relation between x and y. A right-tailed alternate hypothesis claims there is a monotone-increasing relation between x and y, while a two-tailed alternate hypothesis claims there is a monotone relation (either increasing or decreasing) between x and y.

Figure 11-7 shows the type of test and corresponding P-value region.

FIGURE 11-7

Type of Test and *P*-value Region



EXAMPLE 1 Testing the Spearman rank correlation coefficient

Using the information about the Spearman rank correlation coefficient, let's finish our problem about the search for a new member of the political science department at Hendricks College. Our work is organized in Table 11-10, where the rankings given by students and faculty are listed for each of the nine candidates.

Zigy Kaluzny-Charles Thatcher/Stone/Getty Images

P-value

(a) Using a 1% level of significance, let's test the claim that the faculty and students tend to agree about a candidate's teaching ability. This means that the x and y variables should be monotone increasing (as x increases, y increases). Since ρ_s is the population Spearman rank correlation coefficient, we have

H_0: $\rho_s = 0$ (There is no monotone relation.)

H_1:$\rho_s > 0$ (There is a monotone-increasing relation.)

(b) Compute the sample test statistic.

SOLUTION: Since the sample size is $n = 9$, and from Table 11-10 we see that $\Sigma d^2 = 16$, the Spearman rank correlation coefficient is

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6(16)}{9(81-1)} \approx 0.867$$

(c) Find or estimate the *P*-value.

SOLUTION: To estimate the *P*-value for the sample test statistic $r_s = 0.867$, we use Table 9 of Appendix II. The sample size is $n = 9$ and the test is a one-tailed test. We find the location of the sample test statistic in row 9, and then read the corresponding one-tail area. From the Table 9, Appendix II excerpt, we see that the sample test statistic $r_s = 0.867$ falls between the entries 0.834 and 0.917 in the $n = 9$ row. These values correspond to *one-tail areas* between 0.005 and 0.001.

$$0.001 < P\text{-value} < 0.005$$

TABLE 11-10 **Student and Faculty Ranks of Candidates and Calculations for the Spearman Rank Correlation Test**

Candidate	Faculty Rank x	Student Rank y	$d = x - y$	d^2
1	3	5	−2	4
2	7	7	0	0
3	5	6	−1	1
4	9	8	1	1
5	2	3	−1	1
6	8	9	−1	1
7	1	1	0	0
8	6	4	2	4
9	4	2	2	4
				$\Sigma d^2 = 16$

(d) Conclude the test. *Interpret* the results.

SOLUTION:

✓ One-tail area	0.005	0.001
$n = 9$	0.834	0.917
	↑ Sample $r_s = 0.867$	

Since the *P*-value is less than $\alpha = 0.01$, we reject H_0. At the 1% level of significance, we conclude that the relation between faculty and student ratings is monotone increasing. This means that faculty and students tend to rank the teaching performance of candidates in a similar way: Higher student ratings of a candidate correspond with higher faculty ratings of the same candidate.

The following procedure summarizes the steps involved in testing the population Spearman rank correlation coefficient.

PROCEDURE

HOW TO TEST THE SPEARMAN RANK CORRELATION COEFFICIENT ρ_s

Setup

You first need a random sample (of size n) of data pairs (x, y), where both the x and y values are *ranked* variables. Let ρ_s represent the population Spearman rank correlation coefficient, which is in theory computed from the population of all possible (x, y) data pairs.

Procedure

1. Set the *level of significance* α. The *null hypothesis* is H_0: $\rho_s = 0$. In the context of the application, choose the *alternate hypothesis* to be H_1: $\rho_s > 0$ or H_1: $\rho_s < 0$ or H_1: $\rho_s \neq 0$.
2. If there are no ties in the ranks, or if the number of ties is small compared to the number of data pairs n, then compute the *sample test statistic*

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

   where $d = x - y$ is the difference in ranks

   $n =$ number of data pairs

   and the sum is over all sample data pairs.
3. Use Table 9 of Appendix II to find or estimate the *P-value* corresponding to r_s and $n =$ number of data pairs.
4. *Conclude* the test. If *P*-value $\leq \alpha$, then reject H_0. If *P*-value $> \alpha$, then do not reject H_0.
5. *Interpret your conclusion* in the context of the application.

GUIDED EXERCISE 6 **Testing the Spearman rank correlation coefficient**

Fishermen in the Adirondack Mountains are complaining that acid rain caused by air pollution is killing fish in their region. To research this claim, a team of biologists studied a random sample of 12 lakes in the region. For each lake, they measured the level of acidity of rain in the drainage leading into the lake and the density of fish in the lake (number of fish per acre-foot of water). They then did a ranking of $x =$ acidity and $y =$ density of fish. The results are shown in Table 11-11. Higher x ranks mean more acidity, and higher y ranks mean higher density of fish.

Continued

GUIDED EXERCISE 6 *continued*

Table 11-11 Acid Rain and Density of Fish

Lake	Acidity x	Fish Density y	$d = x - y$	d^2
1	5	8	−3	9
2	8	6	2	4
3	3	9	−6	36
4	2	12	−10	100
5	6	7	−1	1
6	1	10	−9	81
7	10	2	8	64
8	12	1	___	___
9	7	5	___	___
10	4	11	___	___
11	9	4	___	___
12	11	3	___	___
			$\Sigma d^2 =$	___

(a) Complete the entries in the d and d^2 columns of Table 11-11, and find Σd^2.

Lake	x	y	d	d^2
8	12	1	11	121
9	7	5	2	4
10	4	11	−7	49
11	9	4	5	25
12	11	3	8	64
				$\Sigma d^2 = 558$

(b) Compute r_s.

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6(558)}{12(144 - 1)} \approx -0.951$$

(c) The fishermen claim that more acidity means lower density of fish. Does this claim state that x and y have a monotone-increasing relation, a monotone-decreasing relation, or no monotone relation?

The claim states that as x increases, y decreases, so the relation between x and y is monotone decreasing.

(d) To test the fishermen's claim, what should we use for the null hypothesis and for the alternate hypothesis? Use $\alpha = 0.01$.

H_0: $\rho_s = 0$ (no monotone relation)
H_1: $\rho_s < 0$ monotone-decreasing relation)

(e) Find or estimate the P-value of the sample test statistic $r_s = -0.951$.

Use Table 9 of Appendix II. There are $n = 12$ data pairs. The sample statistic r_s is negative. Because the r_s distribution is symmetric about 0, we look up the corresponding positive value 0.951 in the row headed by $n = 12$. Use one-tail areas, since this is a left-tailed test.

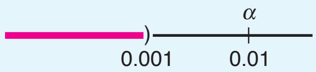
✓ One-tail area	0.001
$n = 12$	0.826
	↑ $-r_s = 0.951$

As positive r_s values increase, corresponding right-tail areas decrease. Therefore,

P-value < 0.001

GUIDED EXERCISE 6 *continued*

(f) Use $\alpha = 0.01$ and conclude the test.



Since the P-value is less than $\alpha = 0.01$, we reject H_0 and conclude that there is a monotone-decreasing relationship between the acidity of the water and the number of fish.

(g) ***Interpretation*** Do the data support the claim that higher acidity means fewer fish?

At the 1% level of significance, we conclude that higher acidity means fewer fish.

Ties of ranks

If ties occur in the assignment of ranks, we follow the usual method of averaging tied ranks. This method was discussed in Section 11.2 (The Rank-Sum Test). The next example illustrates the method.

COMMENT Technically, the use of the given formula for r_s requires that there be no ties in rank. However, if the number of ties in rank is small relative to the number of ranks, the formula can be used with quite a bit of reliability.

EXAMPLE 2 TIED RANKS

Do people who smoke more tend to drink more cups of coffee? The following data were obtained from a random sample of $n = 10$ cigarette smokers who also drink coffee.

Person	Cigarettes per Day	Cups of Coffee per Day
1	8	4
2	15	7
3	20	10
4	5	3
5	22	9
6	15	5
7	15	8
8	25	11
9	30	18
10	35	18

(a) To use the Spearman rank correlation test, we need to rank the data. It does not matter if we rank from smallest to largest or from largest to smallest. The only requirement is that we be consistent in our rankings. Let us rank from smallest to largest.

First, we rank the data for each variable as though there were no ties; then we average the ties as shown in Tables 11-12 and 11-13.

(b) Using 0.01 as the level of significance, we test the claim that x and y have a monotone-increasing relationship. In other words, we test the claim that people who tend to smoke more tend to drink more cups of coffee (Table 11-14).

H_0: $\rho_s = 0$ (There is no montone relation.)

H_1: $\rho_s > 0$ (Right-tailed test)

TABLE 11-12 Rankings of Cigarettes Smoked per Day

Person	Cigarettes per Day	Rank	Average Rank x
4	5	1	1
1	8	2	2
2	15	3	4
6	15 Ties	4 Average rank is 4.	4 Use the average rank for tied data.
7	15	5	4
3	20	6	6
5	22	7	7
8	25	8	8
9	30	9	9
10	35	10	10

TABLE 11-13 Rankings of Cups of Coffee per Day

Person	Cups of Coffee per Day	Rank	Average Rank y
4	3	1	1
1	4	2	2
6	5	3	3
2	7	4	4
7	8	5	5
5	9	6	6
3	10	7	7
8	11	8	8
9	18 Ties	9 Average rank is 9.5.	9.5 Use the average rank for tied data.
10	18	10	9.5

TABLE 11-14 Ranks to Be Used for a Spearman Rank Correlation Test

Person	Cigarette Rank x	Coffee Rank y	$d = x - y$	d^2
1	2	2	0	0
2	4	4	0	0
3	6	7	−1	1
4	1	1	0	0
5	7	6	1	1
6	4	3	1	1
7	4	5	−1	1
8	8	8	0	0
9	9	9.5	−0.5	0.25
10	10	9.5	0.5	0.25
				$\Sigma d^2 = 4.5$

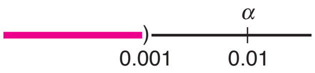
(c) Next, we compute the observed sample test statistic r_s using the results shown in Table 11-14.

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6(4.5)}{10(100 - 1)} \approx 0.973$$

(d) Find or estimate the P-value for the sample test statistic $r_s = 0.973$.
We use Table 9 of Appendix II to estimate the P-value. Using $n = 10$ and a one-tailed test, we see that $r_s = 0.973$ is to the right of the entry 0.879. Therefore, the P-value is smaller than 0.001.

✓ One-tail area	0.001
$n = 9$	0.879
	↑ Sample $r_s = 0.973$

(e) Conclude the test and *interpret* the results.



Since the P-value is less than $\alpha = 0.01$, we reject H_0. At the 1% level of significance, it appears that there is a monotone-increasing relationship between the number of cigarettes smoked and the amount of coffee consumed. People who smoke more cigarettes tend to drink more coffee.

VIEWPOINT Rug Rats!

When do babies start to crawl? Janette Benson, in her article "Infant Behavior and Development," claims that crawling age is related to temperature during the month in which babies first try to crawl. To find a data file for this subject, visit the Brase/Brase statistics site at **http://www.cengage.com/statistics/brase** *and find the link to DASL, the Carnegie Mellon University Data and Story Library. Then look under Psychology in the Data Subjects and select the Crawling Datafile. Can you think of a way to gather data and construct a nonparametric test to study this claim?*

SECTION 11.3 PROBLEMS

1. *Statistical Literacy* For data pairs (x, y), if y always increases as x increases, is the relationship monotone increasing, monotone decreasing, or nonmonotone?

2. *Statistical Literacy* Consider the Spearman rank correlation coefficient r_s for data pairs (x, y). What is the monotone relationship, if any, between x and y implied by a value of
(a) $r_s = 0$?
(b) r_s close to 1?
(c) r_s close to -1?

For Problems 3–11, please provide the following information.
(a) What is the level of significance? State the null and alternate hypotheses.
(b) Compute the sample test statistic.
(c) Find or estimate the P-value of the sample test statistic.
(d) Conclude the test.
(e) *Interpret* the conclusion in the context of the application.

3. ***Training Program: Sales*** A data-processing company has a training program for new salespeople. After completing the training program, each trainee is ranked by his or her instructor. After a year of sales, the same class of trainees is again ranked by a company supervisor according to net value of the contracts they have acquired for the company. The results for a random sample of 11 salespeople trained in the previous year follow, where x is rank in training class and y is rank in sales after 1 year. Lower ranks mean higher standing in class and higher net sales.

Person	1	2	3	4	5	6	7	8	9	10	11
x rank	6	8	11	2	5	7	3	9	1	10	4
y rank	4	9	10	1	6	7	8	11	3	5	2

Using a 0.05 level of significance, test the claim that the relation between x and y is monotone (either increasing or decreasing).

4. ***Economics: Stocks*** As an economics class project, Debbie studied a random sample of 14 stocks. For each of these stocks, she found the cost per share (in dollars) and ranked each of the stocks according to cost. After 3 months, she found the earnings per share for each stock (in dollars). Again, Debbie ranked each of the stocks according to earnings. The way Debbie ranked, higher ranks mean higher cost and higher earnings. The results follow, where x is the rank in cost and y is the rank in earnings.

Stock	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x rank	5	2	4	7	11	8	12	3	13	14	10	1	9	6
y rank	5	13	1	10	7	3	14	6	4	12	8	2	11	9

Using a 0.01 level of significance, test the claim that there is a monotone relation, either way, between the ranks of cost and earnings.

5. ***Psychology: Rat Colonies*** A psychology professor is studying the relation between overcrowding and violent behavior in a rat colony. Eight colonies with different degrees of overcrowding are being studied. By using a television monitor, lab assistants record incidents of violence. Each colony has been ranked for crowding and violence. A rank of 1 means most crowded or most violent. The results for the eight colonies are given in the following table, with x being the population density rank and y the violence rank.

Colony	1	2	3	4	5	6	7	8
x rank	3	5	6	1	8	7	4	2
y rank	1	3	5	2	8	6	4	7

Using a 0.05 level of significance, test the claim that lower crowding ranks mean lower violence ranks (i.e., the variables have a monotone-increasing relationship).

6. ***FBI Report: Murder and Arson*** Is there a relation between murder and arson? A random sample of 15 Midwest cities (over 10,000 population) gave the following information about annual number of murder and arson cases (Reference: Federal Bureau of Investigation, U.S. Department of Justice).

City	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Murder	12	7	25	4	10	15	9	8	11	18	23	19	21	17	6
Arson	62	12	153	2	63	93	31	29	47	131	175	129	162	115	4

(i) Rank-order murder using 1 as the largest data value. Also rank-order arson using 1 as the largest data value. Then construct a table of ranks to be used for a Spearman rank correlation test.
(ii) Use a 1% level of significance to test the claim that there is a monotone-increasing relationship between the ranks of murder and arson.

7. *Psychology: Testing* An army psychologist gave a random sample of seven soldiers a test to measure sense of humor and another test to measure aggressiveness. Higher scores mean greater sense of humor or more aggressiveness.

Soldier	1	2	3	4	5	6	7
Score on humor test	60	85	78	90	93	45	51
Score on aggressiveness test	78	42	68	53	62	50	76

(i) Ranking the data with rank 1 for highest score on a test, make a table of ranks to be used in a Spearman rank correlation test.
(ii) Using a 0.05 level of significance, test the claim that rank in humor has a monotone-decreasing relation to rank in aggressiveness.

8. *FBI Report: Child Abuse and Runaway Children* Is there a relation between incidents of child abuse and number of runaway children? A random sample of 15 cities (over 10,000 population) gave the following information about the number of reported incidents of child abuse and the number of runaway children (Reference: Federal Bureau of Investigation, U.S. Department of Justice).

City	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Abuse cases	49	74	87	10	26	119	35	13	89	45	53	22	65	38	29
Runaways	382	510	581	163	210	791	275	153	491	351	402	209	410	312	210

(i) Rank-order abuse using 1 as the largest data value. Also rank-order runaways using 1 as the largest data value. Then construct a table of ranks to be used for a Spearman rank correlation test.
(ii) Use a 1% level of significance to test the claim that there is a monotone-increasing relationship between the ranks of incidents of abuse and number of runaway children.

9. *Demographics: Police and Fire Protection* Is there a relation between police protection and fire protection? A random sample of large population areas gave the following information about the number of local police and the number of local firefighters (units in thousands) (Reference: *Statistical Abstract of the United States*).

Area	1	2	3	4	5	6	7	8	9	10	11	12	13
Police	11.1	6.6	8.5	4.2	3.5	2.8	5.9	7.9	2.9	18.0	9.7	7.4	1.8
Firefighters	5.5	2.4	4.5	1.6	1.7	1.0	1.7	5.1	1.3	12.6	2.1	3.1	0.6

(i) Rank-order police using 1 as the largest data value. Also rank-order firefighters using 1 as the largest data value. Then construct a table of ranks to be used for a Spearman rank correlation test.
(ii) Use a 5% level of significance to test the claim that there is a monotone relationship (either way) between the ranks of number of police and number of firefighters.

10. *Ecology: Wetlands* Turbid water is muddy or cloudy water. Sunlight is necessary for most life forms; thus turbid water is considered a threat to wetland ecosystems. Passive filtration systems are commonly used to reduce turbidity in wetlands. Suspended solids are measured in mg/l. Is there a relation between input and output turbidity for a passive filtration system and, if so, is

it statistically significant? At a wetlands environment in Illinois, the inlet and outlet turbidity of a passive filtration system have been measured. A random sample of measurements is shown below (Reference: *EPA Wetland Case Studies*).

Reading	1	2	3	4	5	6	7	8	9	10	11	12
Inlet (mg/l)	8.0	7.1	24.2	47.7	50.1	63.9	66.0	15.1	37.2	93.1	53.7	73.3
Outlet (mg/l)	2.4	3.6	4.5	14.9	7.4	7.4	6.7	3.6	5.9	8.2	6.2	18.1

(i) Rank-order the inlet readings using 1 as the largest data value. Also rank-order the outlet readings using 1 as the largest data value. Then construct a table of ranks to be used for a Spearman rank correlation test.
(ii) Use a 1% level of significance to test the claim that there is a monotone relation ship (either way) between the ranks of the inlet readings and outlet readings.

11. *Insurance: Sales* Big Rock Insurance Company did a study of per capita income and volume of insurance sales in eight Midwest cities. The volume of sales in each city was ranked, with 1 being the largest volume. The per capita income was rounded to the nearest thousand dollars.

City	1	2	3	4	5	6	7	8
Rank of insurance sales volume	6	7	1	8	3	2	5	4
Per capita income in $1000	17	18	19	11	16	20	15	19

(i) Using a rank of 1 for the highest per capita income, make a table of ranks to be used for a Spearman rank correlation test.
(ii) Using a 0.01 level of significance, test the claim that there is a monotone relation (either way) between rank of sales volume and rank of per capita income.

SECTION 11.4

Runs Test for Randomness

FOCUS POINTS

- Test a sequence of *symbols* for randomness.
- Test a sequence of *numbers* for randomness about the median.

Astronomers have made an extensive study of galaxies that are $\pm 16°$ above and below the celestial equator. Of special interest is the flux, or change in radio signals, that originates from large electromagnetic disturbances deep in space. The flux units (10^{-26} watts/m^2/Hz) are very small. However, modern radio astronomy can detect and analyze these signals using large antennas (Reference: *Journal of Astrophysics,* Vol. 148, pp. 321–365).

A very important question is the following: Are changes in flux simply random, or is there some kind of nonrandom pattern? Let us use the symbol S to represent a strong or moderate flux and the symbol W to represent a faint or weak flux. Astronomers have received the following signals in order of occurrence.

S S W W W S W W S S S W W W S S W W W S S

Is there a statistical test to help us decide whether or not this sequence of radio signals is random? Well, we're glad you asked, because that is the topic of this section.

We consider applications in which *two* symbols are used (e.g., S or W). Applications using more than two symbols are left to specialized studies in mathematical combinatorics.

Sequence

Run

A **sequence** is an *ordered set* of consecutive symbols.

A **run** is a sequence of one or more occurrences of the *same* symbol.

n_1 = number of times the first symbol occurs in a sequence

n_2 = number of times the second symbol occurs in a sequence

R is a random variable that represents the **number of runs in a sequence.**

EXAMPLE 3 BASIC TERMINOLOGY

In this example, we use the symbols S and W, where S is the first symbol and W is the second symbol, to demonstrate sequences and runs. Identify the runs.

(a) S S W W W is a sequence.

SOLUTION: Table 11-15 shows the sequence of runs. There are $R = 2$ runs in the sequence. The first symbol S occurs $n_1 = 2$ times. The second symbol W occurs $n_2 = 3$ times.

TABLE 11-15 Runs

Run 1	Run 2
S S	W W W

(b) S S W W W S W W S S S S W is a sequence.

SOLUTION: The sequence of runs are shown in Table 11-16. There are $R = 6$ runs in the sequence. The first symbol S occurs $n_1 = 7$ times. The second symbol W occurs $n_2 = 6$ times.

TABLE 11-16 Runs

Run 1	Run 2	Run 3	Run 4	Run 5	Run 6
S S	W W W	S	W W	S S S S	W

Hypotheses

To test a sequence of two symbols for randomness, we use the following hypotheses.

Hypotheses for runs test for randomness

Hypotheses for runs test for randomness

H_0: The symbols are randomly mixed in the sequence.

H_1: The symbols are not randomly mixed in the sequence.

The decision procedure will reject H_0 if either R is too small (too few runs) or R is too large (too many runs).

Sample test statistic

The number of runs R is a *sample test statistic* with its own sampling distribution. Table 10 of Appendix II gives critical values of R for a significance level $\alpha = 0.05$. There are two parameters associated with R. They are n_1 and n_2, the numbers of times the first and second symbols appear in the sequence, respectively. If either $n_1 > 20$ or $n_2 > 20$, you can apply the normal approximation to construct the test. This will be discussed in Problems 11 and 12 at the end of this section. For now, we assume that $n_1 \leq 20$ and $n_2 \leq 20$.

Critical values

For each pair of n_1 and n_2 values, Table 10 of Appendix II provides two critical values: a smaller value denoted c_1 and a larger value denoted c_2. These two values are used to decide whether or not to reject the null hypothesis H_0 that the symbols are randomly mixed in the sequence.

Decision process when $n_1 \le 20$ and $n_2 \le 20$

Use Table 10 of Appendix II with n_1 and n_2 to find the critical values c_1 and c_2. At the $\alpha = 5\%$ level of significance, use the following decision process, where R is the number of runs: If either $R \le c_1$ (too few runs) or $R \ge c_2$ (too many runs), then *reject* H_0. Otherwise, *do not reject* H_0.

COMMENT If either n_1 or n_2 is larger than 20, a normal approximation can be used. See Problems 11 and 12 at the end of this section.

Let's apply this decision process to the astronomy example regarding the sequence of strong and weak electromagnetic radio signals coming from a distant galaxy.

EXAMPLE 4 **Runs test**

Recall that our astronomers had received the following sequence of electromagnetic signals, where S represents a strong flux and W represents a weak flux.

S S W W W S W W S S S W W W S S W W W S S

Is this a random sequence or not? Use a 5% level of significance.

(a) What is the level of significance α? State the null and alternate hypotheses.

SOLUTION: $\alpha = 0.05$

H_0: The symbols S and W are randomly mixed in the sequence.

H_1: The symbols S and W are not randomly mixed in the sequence.

(b) Find the sample test statistic R and the parameters n_1 and n_2.

SOLUTION: We break the sequence according to runs.

Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9
SS	WWW	S	WW	SSS	WWW	SS	WWW	SS

We see that there are $n_1 = 10$ S symbols and $n_2 = 11$ W symbols. The number of runs is $R = 9$.

(c) Use Table 10 of Appendix II to find the critical values c_1 and c_2.

SOLUTION: Since $n_1 = 10$ and $n_2 = 11$, then $c_1 = 6$ and $c_2 = 17$.

Eastcott-Momatiuk/The Image Works

(d) Conclude the test.

SOLUTION:

$R \le 6$	✓ $7 \le R \le 16$	$R \ge 17$
Reject H_0.	✓ Fail to reject H_0.	Reject H_0.

Since $R = 9$, we fail to reject H_0 at the 5% level of significance.

(e) *Interpret* the conclusion in the context of the problem.

SOLUTION: At the 5% level of significance, there is insufficient evidence to conclude that the sequence of electromagnetic signals is not random.

Randomness about the median

An important application of the runs test is to help us decide if a sequence of numbers is a random sequence about the median. This is done using the *median* of the sequence of numbers. The process is explained in the next example.

EXAMPLE 5 RUNS TEST ABOUT THE MEDIAN

Silver iodide seeding of summer clouds was done over the Santa Catalina mountains of Arizona. Of great importance is the direction of the wind during the seeding process. A sequence of consecutive days gave the following compass readings for wind direction at seeding level at 5 A.M. (0° represents true north) (Reference: *Proceedings of the National Academy of Science*, Vol. 68, pp. 649–652).

174	160	175	288	195	140	124	219	197	184
183	224	33	49	175	74	103	166	27	302
61	72	93	172						

University Corporation for Atmospheric Research

We will test this sequence for randomness above and below the median using a 5% level of significance.

Part I: Adjust the sequence so that it has only two symbols, A and B.

SOLUTION: First rank-order the data and find the median (see Section 3.1). Doing this, we find the median to be 169. Next, give each data value in the original sequence the label A if it is *above* the median and the label B if it is *below* the median. Using the original sequence, we get

A | B | AAA | BB | AAAAA | BB | A | BBBB | A | BBB | A

We see that

$n_1 = 12$ (number of A's) $\quad n_2 = 12$ (number of B's) $\quad R = 11$ (number of runs)

Note: In this example, none of the data values actually equals the median. If a data value *equals the median*, we put neither A nor B in the sequence. This eliminates from the sequence any data values that equal the median.

Part II: Test the sequence of A and B symbols for randomness.

(a) What is the level of significance α? State the null and alternate hypotheses.

SOLUTION: $\alpha = 0.05$

H_0: The symbols A and B are randomly mixed in the sequence.

H_1: The symbols A and B are not randomly mixed in the sequence.

(b) Find the sample test statistic R and the parameters n_1 and n_2.

SOLUTION: As shown in Part I, for the sequence of A's and B's,

$n_1 = 12$; $n_2 = 12$; $R = 11$

(c) Use Table 10 of Appendix II to find the critical values c_1 and c_2.

SOLUTION: Since $n_1 = 12$ and $n_2 = 12$, we find $c_1 = 7$ and $c_2 = 19$.

(d) Conclude the test.

SOLUTION:

$R \le 7$	✓ $8 \le R \le 18$	$R \ge 19$
Reject H_0.	✓ Fail to reject H_0.	Reject H_0.

Since $R = 11$, we fail to reject H_0 at the 5% level of significance.

(e) *Interpret* the conclusion in the context of the problem.

SOLUTION: At the 5% level of significance, there is insufficient evidence to conclude that the sequence of wind directions above and below the median direction is not random.

PROCEDURE

How to construct a runs test for randomness

Setup

You need a sequence (ordered set) consisting of two symbols. If your sequence consists of measurements of some type, then convert it to a sequence of two symbols in the following way:

(a) Find the median of the entries in the sequence.

(b) Label an entry A if it is above the median and B if it is below the median. If an entry equals the median, then put neither A nor B in the sequence.

Now you have a sequence with two symbols.

Let n_1 = number of times the first symbol occurs in the sequence.

n_2 = number of times the second symbol occurs in the sequence.

Note: Either symbol can be called the "first" symbol.

Let R = number of runs in the sequence.

Procedure

1. The *level of significance is* $\alpha = 0.05$. The *null and alternate hypotheses* are:

   H_0: The two symbols are randomly mixed in the sequence.

   H_1: The two symbols are not randomly mixed in the sequence.

2. The *sample test statistic* is the number of runs R.
3. Use Table 10, Appendix II, with parameters n_1 and n_2 to find the *lower and upper critical values* c_1 and c_2.
4. Use the *critical values* c_1 and c_2 in the following *decision process*.

$R \le c_1$	$c_1 + 1 \le R \le c_2 - 1$	$R \le c_2$
Reject H_0.	Fail to reject H_0.	Reject H_0.

5. *Interpret your conclusion* in the context of the application.

   Note: If your original sequence consisted of measurements (not just symbols), it is important to remember that you are testing for randomness about the median of these measurements. In any case, you are testing for randomness regarding a mix of two symbols in a given sequence.

Problem 11 describes how to use a normal approximation for the sample test statistic. Problem 12 gives additional practice.

COMMENT In many applications, $n_1 \le 20$ and $n_2 \le 20$. What happens if either $n_1 > 20$ or $n_2 > 20$? In this case, you can use the normal approximation, which is presented in Problems 11 and 12 at the end of this section.

GUIDED EXERCISE 7 **Runs test for randomness of two symbols**

The majority party of the U.S. Senate for each year from 1973 to 2003 is shown below, where D and R represent Democrat and Republican, respectively (Reference: *Statistical Abstract of the United States*).

D D D D R R R D D D D R R R R D D R

Test the sequence for randomness. Use a 5% level of significance.

Continued

GUIDED EXERCISE 7 *continued*

(a) What is α? State the null and alternate hypotheses.

$\alpha = 0.05$

H_0: The two symbols are randomly mixed.

H_1: The two symbols are not randomly mixed.

(b) Block the sequence into runs. Find the values of n_1, n_2, and R.

DDDD | RRR | DDDD | RRRR | DD | R

Letting D be the first symbol, we have

$n_1 = 10$; $n_2 = 8$; $R = 6$

(c) Use Table 10 of Appendix II to find the critical values c_1 and c_2.

Lower critical value $c_1 = 5$
Upper critical value $c_2 = 15$

(d) Using critical values, do you reject or fail to reject H_0?

$R \le 5$	✓ $6 \le R \le 14$	$R \ge 15$
Reject H_0.	✓ Fail to reject H_0.	Reject H_0.

Since $R = 6$, we fail to reject H_0.

(e) ***Interpret*** the conclusion in the context of the application.

The sequence of party control of the U.S. Senate appears to be random. At the 5% level of significance, the evidence is insufficient to reject H_0, that the sequence is random.

GUIDED EXERCISE 8 Runs test for randomness about the median

The national percentage distribution of burglaries is shown by month, starting in January (Reference: *FBI Crime Report*, U.S. Department of Justice).

7.8	6.7	7.6	7.7	8.3	8.2	9.0	9.1	8.6	9.3	8.8	8.9

Test the sequence for randomness about the median. Use a 5% level of significance.

(a) What is α? State the null and alternate hypotheses.

$\alpha = 0.05$

H_0: The sequence of values above and below the median is random.

H_1: The sequence of values above and below the median is not random.

(b) Find the median. Assign the symbol A to values above the median and the symbol B to values below the median. Next block the sequence of A's and B's into runs. Find n_1, n_2, and R.

First order the numbers. Then find the median. Median = 8.45. The original sequence translates to

BBBBBB | AAAAAA

$n_1 = 6$; $n_2 = 6$; $R = 2$

(c) Use Table 10 of Appendix II to find the critical values c_1 and c_2.

Lower critical value $c_1 = 3$
Upper critical value $c_2 = 11$

(d) Using the critical values, do you reject or fail to reject H_0?

✓ $R \le 3$	$4 \le R \le 10$	$R \le 11$
✓ Reject H_0.	Fail to reject H_0.	Reject H_0.

Since $R = 2$, we reject H_0.

(e) ***Interpret*** the conclusion in the context of the application.

At the 5% level of significance, there is sufficient evidence to claim that the sequence of burglaries is not random about the median. It appears that from January to June, there tend to be fewer burglaries.

TECH NOTES

Minitab Enter your sequence of numbers in a column. Use the menu choices **Stat ➤ Nonparametrics ➤ Runs.** In the dialogue box, select the column containing the sequence. The default is to test the sequence for randomness above and below the mean. Otherwise, you can test for randomness above and below any other value, such as the median.

SECTION 11.4 PROBLEMS

1. *Statistical Literacy* To apply a runs test for randomness as described in this section to a sequence of symbols, how many different symbols are required?

2. *Statistical Literacy* Suppose your data consist of a sequence of numbers. To apply a runs test for randomness about the median, what process do you use to convert the numbers into two distinct symbols?

For Problems 3–10, please provide the following information.
(a) What is the level of significance? State the null and alternate hypotheses.
(b) Find the sample test statistic R, the number of runs.
(c) Find the upper and lower critical values in Table 10 of Appendix II.
(d) Conclude the test.
(e) ***Interpret*** the conclusion in the context of the application.

3. ***Presidents: Party Affiliation*** For each successive presidential term from Teddy Roosevelt to George W. Bush (first term), the party affiliation controlling the White House is shown below, where R designates Republican and D designates Democrat (Reference: *The New York Times Almanac*).

   R R R D D R R D D D D D R D R R D R R R D D R

   Historical Note: In cases in which a president died in office or resigned, the period during which the vice president finished the term is not counted as a new term. Test the sequence for randomness. Use $\alpha = 0.05$.

4. ***Congress: Party Affiliation*** The majority party of the U.S. House of Representatives for each year from 1973 to 2003 is shown below, where D and R represent Democrat and Republican, respectively (Reference: *Statistical Abstract of the United States*).

   D D D D D D D D D D D R R R R R R R

   Test the sequence for randomness. Use $\alpha = 0.05$.

5. ***Cloud Seeding: Arizona*** Researchers experimenting with cloud seeding in Arizona want a random sequence of days for their experiments (Reference: *Proceedings of the National Academy of Science*, Vol. 68, pp. 649–652). Suppose they have the following itinerary for consecutive days, where S indicates a day for cloud seeding and N indicates a day for no cloud seeding.

   S S S N S N S S S S N N S N S S S N N S S S S

   Test this sequence for randomness. Use $\alpha = 0.05$.

6. ***Astronomy: Earth's Rotation*** Changes in the earth's rotation are exceedingly small. However, a very long-term trend could be important. (Reference: *Journal of Astronomy*, Vol. 57, pp. 125–146). Let I represent an increase and D a decrease in the rate of the earth's rotation. The following sequence represents historical increases and decreases measured every consecutive fifth year.

   D D D D D I I I D D D D D I I I I I I I I I D I I I I

   Test the sequence for randomness. Use $\alpha = 0.05$.

7. ***Random Walk: Stocks*** Many economists and financial experts claim that the *price level* of a stock or bond is not random; rather, the *price changes* tend to follow a random sequence over time. The following data represent annual percentage returns on Vanguard Total Stock Index for a sequence of recent years. This fund represents nearly all publicly traded U.S. stocks (Reference: *Morningstar Mutual Fund Analysis*).

   10.4 10.6 −0.2 35.8 21.0 31.0 23.3 23.8 −10.6
   −11.0 −21.0 12.8

   (i) Convert this sequence of numbers to a sequence of symbols A and B, where A indicates a value above the median and B a value below the median.
   (ii) Test the sequence for randomness about the median. Use $\alpha = 0.05$.

8. ***Random Walk: Bonds*** The following data represent annual percentage returns on Vanguard Total Bond Index for a sequence of recent years. This fund represents nearly all publicly traded U.S. bonds (Reference: *Morningstar Mutual Fund Analysis*).

   7.1 9.7 −2.7 18.2 3.6 9.4 8.6 −0.8 11.4 8.4 8.3 0.8

   (i) Convert this sequence of numbers to a sequence of symbols A and B, where A indicates a value above the median and B a value below the median.
   (ii) Test the sequence for randomness about the median. Use $\alpha = 0.05$.

9. ***Civil Engineering: Soil Profiles*** Sand and clay studies were conducted at the West Side Field Station of the University of California (Reference: Professor D. R. Nielsen, University of California, Davis). Twelve consecutive depths, each about 15 cm deep, were studied and the following percentages of sand in the soil were recorded.

   19.0 27.0 30.0 24.3 33.2 27.5 24.2 18.0 16.2 8.3 1.0 0.0

   (i) Convert this sequence of numbers to a sequence of symbols A and B, where A indicates a value above the median and B a value below the median.
   (ii) Test the sequence for randomness about the median. Use $\alpha = 0.05$.

10. ***Civil Engineering: Soil Profiles*** Sand and clay studies were conducted at the West Side Field Station of the University of California (Reference: Professor D. R. Nielsen, University of California, Davis). Twelve consecutive depths, each about 15 cm deep, were studied and the following percentages of clay in the soil were recorded.

    47.4 43.4 48.4 42.6 41.4 40.7 46.4 44.8 36.5 35.7 33.7 42.6

    (i) Convert this sequence of numbers to a sequence of symbols A and B, where A indicates a value above the median and B a value below the median.
    (ii) Test the sequence for randomness about the median. Use $\alpha = 0.05$.

11. ***Expand Your Knowledge: Either $n_1 > 20$ or $n_2 > 20$*** For each successive presidential term from Franklin Pierce (the 14th president, elected in 1853) to George W. Bush (43rd president), the party affiliation controlling the White House is shown below, where R designates Republican and D designates Democrat (Reference: *The New York Times Almanac*).

    Historical Note: We start this sequence with the 14th president because earlier presidents belonged to political parties such as the Federalist or Wigg (not Democratic or Republican) party. In cases in which a president died in office or resigned, the period during which the vice president finished the term is not counted as a new term. The one exception is the case in which Lincoln (a Republican) was assassinated and the vice president Johnson (a Democrat) finished the term.

D D R R D R R R R D R D R R R R D D R R
D D D D D R R D D R R D R R R D D R

Test the sequence for randomness at the 5% level of significance. Use the following outline.

(a) State the null and alternate hypotheses.

(b) Find the number of runs R, n_1, and n_2. Let $n_1 =$ number of Republicans and $n_2 =$ number of Democrats.

(c) In this case, $n_1 = 21$, so we cannot use Table 10 of Appendix II to find the critical values. Whenever either n_1 or n_2 exceeds 20, the number of runs R has a distribution that is approximately normal, with

$$\mu_R = \frac{2n_1n_2}{n_1 + n_2} + 1 \quad \text{and} \quad \sigma_R = \sqrt{\frac{(2n_1n_2)(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}$$

We convert the number of runs R to a z value, and then use the normal distribution to find the critical values. Convert the sample test statistic R to z using the formula

$$z = \frac{R - \mu_R}{\sigma_R}$$

(d) The critical values of a normal distribution for a two-tailed test with level of significance $\alpha = 0.05$ are -1.96 and 1.96 (see Table 5(c) of Appendix II). Reject H_0 if the sample test statistic $z \leq -1.96$ or if the sample test statistic $z \geq 1.96$. Otherwise, do not reject H_0.

Sample $z \leq -1.96$	$-1.96 <$ sample $z < 1.96$	Sample $z \geq 1.96$
Reject H_0.	Fail to reject H_0.	Reject H_0.

Using this decision process, do you reject or fail to reject H_0 at the 5% level of significance? What is the P-value for this two-tailed test? At the 5% level of significance, do you reach the same conclusion using the P-value that you reach using critical values? Explain.

(e) ***Interpret*** your results in the context of the application.

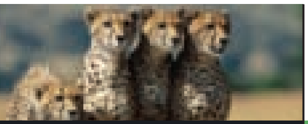
12. ***Expand Your Knowledge: Either $n_1 > 20$ or $n_2 > 20$*** Professor Cornish studied rainfall cycles and sunspot cycles (Reference: *Australian Journal of Physics*, Vol. 7, pp. 334–346). Part of the data include amount of rain (in mm) for 6-day intervals. The following data give rain amounts for consecutive 6-day intervals at Adelaide, South Australia.

6	29	6	0	68	0	0	2	23	5	18	0	50	163
64	72	26	0	0	3	8	142	108	3	90	43	2	5
0	21	2	57	117	51	3	157	43	20	14	40	0	23
18	73	25	64	114	38	31	72	54	38	9	1	17	0
13	6	2	0	1	5	9	11						

Verify that the median is 17.5.

(a) Convert this sequence of numbers to a sequence of symbols A and B, where A indicates a value above the median and B a value below the median.

(b) Test the sequence for randomness about the median at the 5% level of significance. Use the large sample theory outlined in Problem 11.



Chapter Review

SUMMARY

When we cannot assume that data come from a normal, binomial, or Student's t distribution, we can employ tests that make no assumptions about data distribution. Such tests are called nonparametric tests. We studied four widely used tests: the sign test, the rank-sum test, the Spearman rank correlation coefficient test, and the runs test for randomness. Nonparametric tests have both advantages and disadvantages:

- Advantages of nonparametric tests

  No requirements concerning the distributions of populations under investigation.

  Easy to use.

- Disadvantages of nonparametric tests

  Waste information.

  Are less sensitive.

It is usually good advice to use standard tests when possible, keeping nonparametric tests for situations wherein assumptions about the data distribution cannot be made.

IMPORTANT WORDS & SYMBOLS

Section 11.1
Nonparametric statistics 678
Sign test 678

Section 11.2
Rank-sum test 686

Section 11.3
Monotone relationship 695
Spearman rank correlation coefficient r_s 696
Population Spearman rank correlation coefficient ρ_s 696

Section 11.4
Sequence 706
Run 706
Runs test for randomness 705

VIEWPOINT Lending a Hand

Whom would you ask for help if you were sick? in need of money? upset with your spouse? depressed? Consider the following claims: People look to sisters for emotional help and brothers for physical help. After that, people look to parents, clergy, or friends. Can you think of nonparametric tests to study such claims? For more information, see American Demographics, *Vol. 18, No. 8.*

CHAPTER REVIEW PROBLEMS

1. *Statistical Literacy* For nonparametric tests, what assumptions, if any, need to be made concerning the distributions of the populations under investigation?

2. *Critical Thinking* Suppose you want to test whether there is a difference in means in a matched pair, "before and after" situation. If you know that the populations under investigation are at least mound-shaped and symmetrical and you have a large sample, is it better to use the parametric paired differences test or the nonparametric sign test for matched pairs? Explain.

For Problems 3–10, please provide the following information.
(a) State the test used.
(b) Give α. State the null and alternate hypotheses.
(c) Find the sample test statistic.

(d) For the sign test, rank-sum test, and Spearman correlation coefficient test, find the P-value of the sample test statistic. For the runs test of randomness, find the critical values from Table 10 of Appendix II.
(e) Conclude the test and ***interpret*** the results in the context of the application.

3. ***Chemistry: Lubricant*** In the production of synthetic motor lubricant from coal, a new catalyst has been discovered that seems to affect the viscosity index of the lubricant. In an experiment consisting of 23 production runs, 11 used the new catalyst and 12 did not. After each production run, the viscosity index of the lubricant was determined to be as follows.

With catalyst	1.6	3.2	2.9	4.4	3.7	2.5	1.1	1.8	3.8	4.2	4.1	
Without catalyst	3.9	4.6	1.5	2.2	2.8	3.6	2.4	3.3	1.9	4.0	3.5	3.1

The two samples are independent. Use a 0.05 level of significance to test the null hypothesis that the viscosity index is unchanged by the catalyst against the alternate hypothesis that the viscosity index has changed.

4. ***Self-Improvement: Memory*** Professor Adams wrote a book called *Improving Your Memory*. The professor claims that if you follow the program outlined in the book, your memory will definitely improve. Fifteen people took the professor's course, in which the book and its program were used. On the first day of class, everyone took a memory exam; and on the last day, everyone took a similar exam. The paired scores for each person follow.

Last exam	225	120	115	275	85	76	114	200	99	135	170	110	216	280	78
First exam	175	110	115	200	60	85	160	190	70	110	140	10	190	200	92

Use a 0.05 level of significance to test the null hypothesis that the scores are the same whether or not people have taken the course against the alternate hypothesis that the scores of people who have taken the course are higher.

5. ***Sales: Paint*** A chain of hardware stores is trying to sell more paint by mailing pamphlets describing the paint. In 15 communities containing one of these hardware stores, the paint sales (in dollars) were recorded for the months before and after the ads were sent out. The paired results for each store follow.

Sales after	610	150	790	288	715	465	280	640	500	118	265	365	93	217	280
Sales before	460	216	640	250	685	430	220	470	370	118	117	360	93	291	430

Use a 0.01 level of significance to test the null hypothesis that the advertising had no effect on sales against the alternate hypothesis that it improved sales.

6. ***Dogs: Obedience School*** An obedience school for dogs experimented with two methods of training. One method involved rewards (food, praise); the other involved no rewards. The dogs were randomly placed into two independent groups of 11 each. The number of sessions required to train each of 22 dogs follows.

With rewards	12	17	15	10	16	20	9	23	8	14	10
No rewards	19	22	11	18	13	25	24	28	21	20	21

Use a 0.05 level of significance to test the hypothesis that the number of sessions was the same for the two groups against the alternate hypothesis that the number of sessions was not the same.

7. *Training Program: Fast Food* At McDouglas Hamburger stands, each employee must undergo a training program before he or she is assigned. A group of nine people went through the training program and were assigned to work at the Teton Park McDouglas Hamburger stand. Rankings in performance after the training program and after one month on the job are shown (a rank of 1 is for best performance).

Employee	1	2	3	4	5	6	7	8	9
Rank, training program	8	9	7	3	6	4	1	2	5
Rank on job	9	8	6	7	5	1	3	4	2

Using a 0.05 level of significance, test the claim that there is a monotone-increasing relation between rank from the training program and rank in performance on the job.

8. *Cooking School: Chocolate Mousse* Two expert French chefs judged chocolate mousse made by students in a Paris cooking school. Each chef ranked the best chocolate mousse as 1.

Student	1	2	3	4	5
Rank by Chef Pierre	4	2	3	1	5
Rank by Chef André	4	1	2	3	5

Use a 0.10 level of significance to test the claim that there is a monotone relation (either way) between ranks given by Chef Pierre and by Chef André.

9. *Education: True–False Questions* Dr. Gill wants to arrange the answers to a true–false exam in random order. The answers in order of occurrence are shown below.

T T T T F T T F F T T T T F F F F F F T T T T T T

Test the sequence for randomness using $\alpha = 0.05$.

10. *Agriculture: Wheat* For the past 16 years, the yields of wheat (in tons) grown on a plot at Rothamsted Experimental Station (England) are shown below. The sequence is by year.

3.8 1.9 0.6 1.7 2.0 3.5 3.0 1.4 2.7 2.3 2.6 2.1
2.4 2.7 1.8 1.9

Use level of significance 5% to test for randomness about the median.

DATA HIGHLIGHTS: GROUP PROJECTS

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

In the world of business and economics, to what extent do assets determine profits? Do the big companies with large assets always make more profits? Is there a rank correlation between assets and profits? The following table is based on information taken from

Company	Asset Rank	Profit Rank
Pepsico	4	2
McDonald's	1	1
Aramark	6	4
Darden Restaurants	7	5
Flagstar	11	11
VIAD	10	8
Wendy's International	2	3
Host Marriott Services	9	10
Brinker International	5	7
Shoney's	3	6
Food Maker	8	9

Fortune (Vol. 135, No. 8). A rank of 1 means highest profits or highest assets. The companies are food service companies.

(a) Compute the Spearman rank correlation coefficient for these data.

(b) Using a 5% level of significance, test the claim that there is a monotone-increasing relation between the ranks of earnings and growth.

(c) Decide whether you should reject or not reject the null hypothesis. Interpret your conclusion in the context of the problem.

(d) As an investor, what are some other features of food companies that you might be interested in ranking? Identify any such features that you think might have a monotone relation.

LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

1. (a) What do we mean by the term *nonparametric statistics*? What do we mean by the term *parametric statistics*? How do nonparametric methods differ from the methods we studied earlier?

   (b) What are the advantages of nonparametric statistical methods? How can they be used in problems to which other methods we have learned would not apply?

   (c) Are there disadvantages to nonparametric statistical methods? What do we mean when we say that nonparametric methods tend to waste information? Why do we say that nonparametric methods are not as *sensitive* as parametric methods?

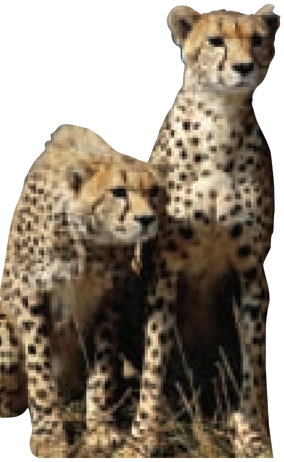
   (d) List three random variables from ordinary experience to which you think nonparametric methods would definitely apply and the application of parametric methods would be questionable.

2. Outline the basic logic and ideas behind the sign test. Describe how the binomial probability distribution was used in the construction of the sign test. What assumptions must be made about the sign test? Why is the sign test so extremely general in its possible applications? Why is it a special test for "before and after" studies?

3. Outline the basic logic and ideas behind the rank-sum test. Under what conditions would you use the rank-sum test and *not* the sign test? What assumptions must be made in order to use the rank-sum test? List two advantages the rank-sum test has that the methods of Section 8.5 do not have. List some advantages the methods of Section 8.5 have that the rank-sum test does not have.

4. What do we mean by a monotone relationship between two variables x and y? What do we mean by ranked variables? Give a graphic example of two variables x and y that have a monotone relationship but do *not* have a linear relationship. Does the Spearman test check for a monotone relationship or a linear relationship? Under what conditions does the Pearson product-moment correlation coefficient reduce to the Spearman rank correlation coefficient? Summarize the basic logic and ideas behind the test for Spearman rank correlation. List variables x and y from daily experience for which you think a strong Spearman rank correlation coefficient exists even though the variables are *not* linearly related.

5. What do we mean by a runs test for randomness? What is a run in a sequence? How can we test for randomness about the median? Why is this an important concept? List at least three applications from your own experience.



Cumulative Review Problems

CHAPTERS 10–11

1. *Goodness-of-Fit Test: Rare Events* This cumulative review problem uses material from Chapters 3, 5, and 10. Recall that the Poisson distribution deals with rare events. Death from the kick of a horse is a rare event, even in the Prussian army. The following data are a classic example of a Poisson application to rare events. A reproduction of the original data can be found in C. P. Winsor, *Human Biology,* Vol. 19, pp. 154–161. The data represent the number of deaths from the kick of a horse per army corps per year for 10 Prussian army corps for 20 years (1875–1894). Let x represent the number of deaths and f the frequency of x deaths.

x	0	1	2	3 or more
f	109	65	22	4

(a) First, we fit the data to a Poisson distribution (see Section 5.4).

Poission distribution: $P(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$

where $\lambda \approx \bar{x}$ (sample mean of x values)
From our study of weighted averages (see Section 3.1),

$$\bar{x} = \frac{\Sigma xf}{\Sigma f}$$

Verify that $\bar{x} \approx 0.61$ *Hint:* For the category 3 or more, use 3.

(b) Now we have $P(x) = \dfrac{e^{-0.61}(0.61)^x}{x!}$ for $x = 0$, 1, 2, 3. . . .

Find $P(0)$, $P(1)$, $P(2)$, and $P(3 \leq x)$. Round to three places after the decimal.

(c) The total number of observations is $\Sigma f = 200$. For a given x, the expected frequency of x deaths is $200P(x)$. The following table gives the observed frequencies O and the expected frequencies $E = 200P(x)$.

x	$O = f$	$E = 200P(x)$
0	109	$200(0.543) = 108.6$
1	65	$200(0.331) = 66.2$
2	22	$200(0.101) = 20.2$
3 or more	4	$200(0.025) = 5$

Compute $\chi^2 = \Sigma \dfrac{(O - E)^2}{E}$

(d) State the null and alternate hypotheses for a chi-square goodness-of-fit test. Set the level of significance to be $\alpha = 0.01$. Find the P-value for a goodness-of-fit test. Interpret your conclusion in the context of this application. Is there reason to believe that the Poisson distribution fits the raw data provided by the Prussian army? Explain.

2. *Test of Independence: Agriculture* Three types of fertilizer were used on 132 identical plots of maize. Each plot was harvested and the yield (in kg) was recorded (Reference: Caribbean Agricultural Research and Development Institute).

	Type of Fertilizer			
Yield (kg)	I	II	III	Row Total
0–2.9	12	10	15	37
3.0–5.9	18	21	11	50
6.0–8.9	16	19	10	45
Column Total	46	50	36	132

Use a 5% level of significance to test the hypothesis that type of fertilizer and yield of maize are independent. Interpret the results.

3. *Testing and Estimating Variances: Iris* Random samples of two species of iris gave the following petal lengths (in cm) (Reference: R. A. Fisher, *Annals of Eugenics*, Vol. 7).

x_1, *Iris virginica*	5.1 5.9 4.5 4.9 5.7 4.8 5.8 6.4 5.6 5.9
x_2, *Iris versicolor*	4.5 4.8 4.7 5.0 3.8 5.1 4.4 4.2

(a) Use a 5% level of significance to test the claim that the population standard deviation of x_1 is larger than 0.55.
(b) Find a 90% confidence interval for the population standard deviation of x_1.
(c) Use a 1% level of significance to test the claim that the population variance of x_1 is larger than that of x_2. Interpret the results.

4. *Sign Test: Wind Direction* The following data are paired by date. Let x and y be random variables representing wind direction at 5 A.M. and 5 P.M., respectively (units are degrees on a compass, with 0° representing true north). The readings were taken at seeding level in a cloud seeding experiment. (Reference: *Proceedings of the National Academy of Science,* Vol. 68, pp. 649–652.) A random sample of days gave the following information.

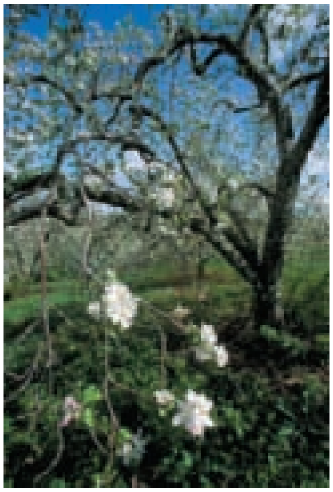
x	177	140	197	224	49	175	257	72	172
y	142	142	217	125	53	245	218	35	147
x	214	265	110	193	180	190	94	8	93
y	205	218	100	170	245	117	140	99	60

Use the sign test with a 5% level of significance to test the claim that the distributions of wind directions at 5 A.M. and 5 P.M. are different. Interpret the results.

5. *Rank-Sum Test: Apple Trees* Commercial apple trees usually consist of two parts grafted together. The upper part, or graft, determines the character of the fruit, while the root stock determines the size of the tree. (Reference: East Malling Research Station, England.) The following data are from two root stocks A and B. The data represent total extension growth (in meters) of the grafts after 4 years.

Stock A	2.81	2.26	1.94	2.37	3.11	2.58	2.74	2.10	3.41	2.94	2.88
Stock B	**2.52**	3.02	2.86	2.91	2.78	2.71	1.96	2.44	2.13	1.58	2.77

Use a 1% level of significance and the rank-sum test to test the claim that the distributions of growths are different for root stocks A and B. Interpret the results.



Eastcott-Momatiuk/The Image Works

6. *Spearman Rank Correlation: Calcium Tests* Random collections of nine different solutions of a calcium compound were given to two laboratories A and B. Each laboratory measured the calcium content (in mmol. per liter) and reported the results. The data are paired by calcium compound (Reference: *Journal of Clinical Chemistry and Clinical Biochemistry,* Vol. 19, pp. 395–426).

Compound	1	2	3	4	5	6	7	8	9
Lab A	13.33	15.79	14.78	11.29	12.59	9.65	8.69	10.06	11.58
Lab B	13.17	15.72	14.66	11.47	12.65	9.60	8.75	10.25	11.56

(a) Rank-order the data using 1 for the lowest calcium reading. Make a table of ranks to be used in a Spearman rank correlation test.
(b) Use a 5% level of significance to test for a monotone relation (either way) between ranks. Interpret the results.

7. *Runs Test for Randomness: Sunspots* The January mean number of sunspots is recorded for a sequence of recent Januaries (Reference: *International Astronomical Union Quarterly Bulletin on Solar Activity*).

57.9	38.7	19.8	15.3	17.5	28.2
110.9	121.8	104.4	111.5	9.13	61.5
43.4	27.6	18.9	8.1	16.4	51.9

Use level of significance 5% to test for randomness about the median. Interpret the results.

Appendix I Additional Topics

PART I

Bayes's Theorem

The Reverend Thomas Bayes (1702–1761) was an English mathematician who discovered an important relation for conditional probabilities. This relation is referred to as *Bayes's rule* or *Bayes's theorem*. It uses conditional probabilities to adjust calculations so that we can accommodate new, relevant information. We will restrict our attention to a special case of Bayes's theorem in which an event B is partitioned into only *two* mutually exclusive events (see Figure AI-1). The general formula is a bit complicated but is a straightforward extension of the basic ideas we will present here. Most advanced texts contain such an extension.

Note: We use the following compact notation in the statement of Bayes's theorem:

Notation	Meaning
A^c	complement of A; *not* A
$P(B\|A)$	probability of event B, *given* event A; $P(B$, *given* $A)$
$P(B\|A^c)$	probability of event B, *given* the complement of A; $P(B$, *given not* $A)$

We will use Figure AI-1 to motivate Bayes's theorem. Let A and B be events in a sample space that have probabilities not equal to 0 or 1. Let A^c be the complement of A.

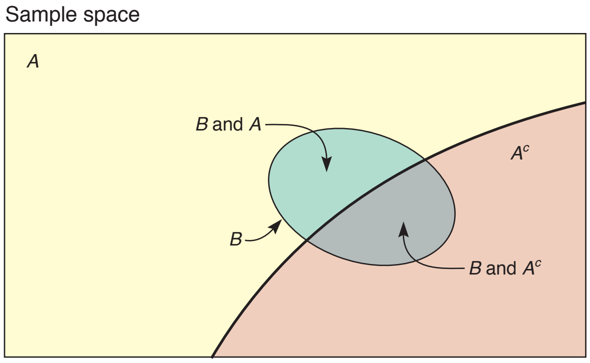
$$\text{Here is Bayes's theorem: } P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \tag{1}$$

Overview of Bayes's Theorem

Suppose we have an event A and we calculate $P(A)$, the unconditional probability of A standing by itself. Now suppose we have a "new" event B and we know the probability of B given that A occurs $P(B|A)$, as well as the probability of B given that A does not occur $P(B|A^c)$. Where does such an event B come from? The event B can be constructed in many possible ways. For example, B can be constructed as

FIGURE A1-1

A Typical Setup for Bayes's Theorem



the result of a consulting service, a testing procedure, or a sorting activity. In the examples and problems, you will find more ways to construct such an event B.

How can we use this "new" information concerning the event B to adjust our calculation of the probability of event A, given B? That is, how can we make our calculation of the probability of A more realistic by including information about the event B? The answer is that we will use Equation (1) of Bayes's theorem.

Let's look at some examples that use Equation (1) of Bayes's theorem. We are grateful to personal friends in the oil and natural gas business in Colorado who provided the basic information in the following example.

EXAMPLE 1 BAYES'S THEOREM

A geologist has examined seismic data and other geologic formations in the vicinity of a proposed site for an oil well. Based on this information, the geologist reports a 65% chance of finding oil. The oil company decides to go ahead and start drilling. As the drilling progresses, sample cores are taken from the well and studied by the geologist. These sample cores have a history of predicting oil *when there is oil* about 85% of the time. However, about 6% of the time the sample cores will predict oil *when there is no oil*. (Note that these probabilities need not add up to 1.) Our geologist is delighted because the sample cores predict oil for this well.

Use the "new" information from the sample cores to revise the geologist's original probability that the well will hit oil. What is the new probability?

SOLUTION: To use Bayes's theorem, we need to identify the events A and B. Then we need to find $P(A)$, $P(A^c)$, $P(B|A)$, and $P(B|A^c)$. From the description of the problem, we have

A is the event that the well strikes oil.
A^c is the event that the well is dry (no oil).
B is the event that the core samples indicate oil.

Again, from the description, we have

$$P(A) = 0.65, \qquad \text{so} \qquad P(A^c) = 1 - 0.65 = 0.35$$

These are our *prior* (before new information) probabilities. New information comes from the sample cores. Probabilities associated with the new information are

$$P(B|A) = 0.85$$

This is the probability that core samples indicate oil when there actually is oil.

$$P(B|A^c) = 0.06$$

This is the probability that core samples indicate oil when there is no oil (dry well).

Now we use Bayes's theorem to revise the probability that the well will hit oil based on the "new" information from core samples. The revised probability is the *posterior* probability we compute that uses the new information from the sample cores:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{(0.85)(0.65)}{(0.85)(0.65) + (0.06)(0.35)} = 0.9634$$

We see that the revised (*posterior*) probability indicates about a 96% chance for the well to hit oil. This is why sample cores that are good can attract money in the form of venture capital (for independent drillers) on a big, expensive well.

GUIDED EXERCISE 1 Bayes's theorem

The Anasazi were prehistoric pueblo people who lived in what is now the southwestern United States. Mesa Verde, Pecos Pueblo, and Chaco Canyon are beautiful national parks and monuments, but long ago they were home to many Anasazi. In prehistoric times, there were several Anasazi migrations, until finally their pueblo homes were completely abandoned. The delightful book *Proceedings of the Anasazi Symposium, 1981*, published by Mesa Verde Museum Association, contains a very interesting discussion about methods anthropologists use to (approximately) date Anasazi objects. There are two popular ways. One is to compare environmental data to other objects of known dates. The other is radioactive carbon dating.

Carbon dating has some variability in its accuracy, depending on how far back in time the age estimate goes and also on the condition of the specimen itself. Suppose experience has shown that the carbon method is correct 75% of the time it is used on an object from a known (given) time period. However, there is a 10% chance that the carbon method will predict that an object is from a certain period even when we already know the object is not from that period.

Using environmental data, an anthropologist reported the probability to be 40% that a fossilized deer bone bracelet was from a certain Anasazi migration period. Then, as a follow-up study, the carbon method also indicated that the bracelet was from this migration period. How can the anthropologist adjust her estimated probability to include the "new" information from the carbon dating?

(a) To use Bayes's theorem, we must identify the events A and B. From the description of the problem, what are A and B?

→ A is the event that the bracelet is from the given migration period. B is the event that carbon dating indicates that the bracelet is from the given migration period.

(b) Find $P(A)$, $P(A^c)$, $P(B|A)$, and $P(B|A^c)$.

→ From the description,

$$\begin{aligned} P(A) &= 0.40 \\ P(A^c) &= 0.60 \\ P(B|A) &= 0.75 \\ P(B|A^c) &= 0.10 \end{aligned}$$

(c) Compute $P(A|B)$, and explain the meaning of this number.

→ Using Bayes's theorem and the results of part (b), we have

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\ &= \frac{(0.75)(0.40)}{(0.75)(0.40) + (0.10)(0.60)} = 0.8333 \end{aligned}$$

The prior (before carbon dating) probability was only 40%. However, the carbon dating enabled us to revise this probability to 83%. Thus, we are about 83% sure that the bracelet came from the given migration period. Perhaps additional research at the site will uncover more information to which Bayes's theorem could be applied again.

The next example is a classic application of Bayes's theorem. Suppose we are faced with two competing hypotheses. Each hypothesis claims to explain the same phenomenon; however, only one hypothesis can be correct. Which hypothesis should we accept? This situation occurs in the natural sciences, the social sciences, medicine, finance, and many other areas of life. Bayes's theorem will help us compute the probabilities that one or the other hypothesis is correct. Then what do we do? Well, the great mathematician and philosopher René Descartes can guide us. Descartes once said, "When it is not in our power to determine what is true, we ought to follow what is most probable." Just knowing probabilities does not allow us with absolute certainty to choose the correct hypothesis, but it does permit us to identify which hypothesis is *most likely* to be correct.

EXAMPLE 2 **Competing hypotheses**

A large hospital uses two medical labs for blood work, biopsies, throat cultures, and other medical tests. Lab I does 60% of the reports. The other 40% of the reports are done by Lab II. Based on long experience, it is known that about 10% of the reports from Lab I contain errors and that about 7% of the reports from Lab II contain errors. The hospital recently received a lab report that, through additional medical work, was revealed to be incorrect. One hypothesis is that the report with the mistake came from Lab I. The competing hypothesis is that the report with the mistake came from Lab II. Which lab do you suspect is the culprit? Why?

SOLUTION: Let's use the following notation.

A = event report is from Lab I
A^c = event report is from Lab II
B = event report contains a mistake

From the information given,

$$P(A) = 0.60 \qquad P(A^c) = 0.40$$
$$P(B|A) = 0.10 \qquad P(B|A^c) = 0.07$$

The probability that the report is from Lab I *given* we have a mistake is $P(A|B)$. Using Bayes's theorem, we get

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\ &= \frac{(0.10)(0.60)}{(0.10)(0.60) + (0.07)(0.40)} \\ &= \frac{0.06}{0.088} \approx 0.682 \approx 68\% \end{aligned}$$

So, the probability is about 68% that Lab I supplied the report with the error. It follows that the probability is about 100% − 68% = 32% that the erroneous report came from Lab II.

PROBLEM

Bayes's theorem applied to quality control

A company that makes steel bolts knows from long experience that about 12% of its bolts are defective. If the company simply ships all bolts that it produces, then 12% of the shipment the customer receives will be defective. To decrease the percentage of defective bolts shipped to customers, an electronic scanner is installed. The scanner is positioned over the production line and is supposed to pick out the good bolts. However, the scanner itself is not perfect. To test the scanner, a large number of (pretested) "good" bolts were run under the scanner, and it accepted 90% of the bolts as good.

Continued

Then a large number of (pretested) defective bolts were run under the scanner, and it accepted 3% of these as good bolts.

(a) If the company does not use the scanner, what percentage of a shipment is expected to be good? What percentage is expected to be defective?

(b) The scanner itself makes mistakes, and the company is questioning the value of using it. Suppose the company does use the scanner and ships only what the scanner passes as "good" bolts. In this case, what percentage of the shipment is expected to be good? What percentage is expected to be defective?

Partial Answer

To solve this problem, we use Bayes's theorem. The result of using the scanner is a dramatic improvement in the quality of the shipped product. If the scanner is not used, only 88% of the shipped bolts will be good. However, if the scanner is used and only the bolts it passes as good are shipped, then 99.6% of the shipment is expected to be good. Even though the scanner itself makes a considerable number of mistakes, it is definitely worth using. Not only does it increase the quality of a shipment, the bolts it rejects can also be recycled into new bolts.

PART II The Hypergeometric Probability Distribution

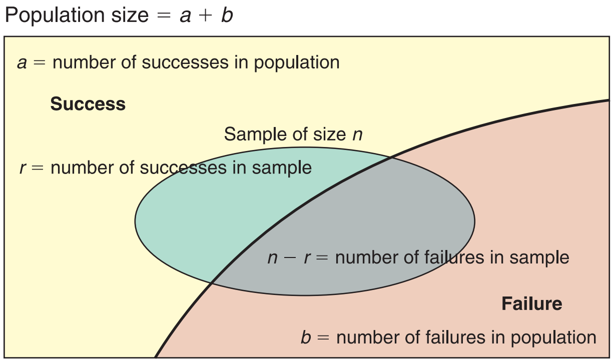
In Chapter 5, we examined the binomial distribution. The binomial probability distribution assumes *independent trials*. If the trials are constructed by drawing samples from a population, then we have two possibilities: We sample either *with replacement* or *without replacement*. If we draw random samples with replacement, the trials can be taken to be independent. If we draw random samples without replacement and the population is very large, then it is reasonable to say that the trials are approximately independent. In this case, we go ahead and use the binomial distribution. However, if the population is relatively small and we draw samples without replacement, the assumption of independent trials is not valid, and we should not use the binomial distribution.

The *hypergeometric distribution* is a probability distribution of a random variable that has two outcomes when sampling is done *without replacement*.

Consider the following notational setup (see Figure AI-2). Suppose we have a population with only *two* distinct types of objects. Such a population might be made up of females and males, students and faculty, residents and nonresidents, defective and nondefective items, and so on. For simplicity of reference, let us call one type of object (your choice) "success" and the other "failure." Let's use the

FIGURE A1-2

Notational Setup for Hypergeometric Distribution



letter a to designate the number of successes in the population and the letter b to designate the number of failures in the population. Thus, the total population size is $a + b$. Next, we draw a random sample (without replacement) of size n from this population. Let r be the number of successes in this sample. Then $n - r$ is the number of failures in the sample. The hypergeometric distribution gives us the probability of r successes in the sample of size n.

Recall from Section 4.3 that the number of combinations of k objects taken j at a time can be computed as

$$C_{k,j} = \frac{k!}{j!(k - j)!}$$

Using the notation of Figure AI-2 and the formula for combinations, the hypergeometric distribution can be calculated.

Hypergeometric distribution

Given that a population has two distinct types of objects, success and failure,

a counts the number of successes in the population.

b counts the number of failures in the population.

For a random sample of size n taken *without replacement* from this population, the probability $P(r)$ of getting r successes in the *sample* is

$$P(r) = \frac{C_{a,r}C_{b,n-r}}{C_{(a+b),n}} \qquad (2)$$

The expected value and standard deviation are

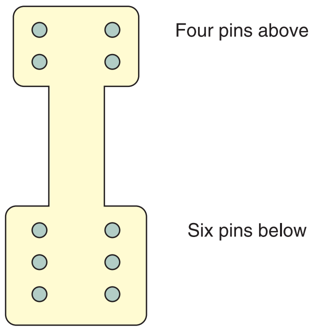
$$\mu = \frac{na}{a + b} \quad \text{and} \quad \sigma = \sqrt{n\left(\frac{a}{a + b}\right)\left(\frac{b}{a + b}\right)\left(\frac{a + b - n}{a + b - 1}\right)}$$

EXAMPLE 3 HYPERGEOMETRIC DISTRIBUTION

A section of an Interstate 95 bridge across the Mianus River in Connecticut collapsed suddenly on the morning of June 28, 1983. (See *To Engineer Is Human: The Role of Failure in Successful Designs* by Henry Petroski.) Three people were killed when their vehicles fell off the bridge. It was determined that the collapse was caused by the failure of a metal hanger design that left a section of the bridge with no support when something went wrong with the pins. Subsequent inspection revealed many cracked pins and hangers in bridges across the United States.

FIGURE A1-3

Steel Hanger Design for Bridge Support



(a) Suppose a hanger design uses four pins in the upper part and six pins in the lower part, as shown in Figure AI-3. The hangers come in a kit consisting of the hanger and 10 pins. When a work crew installs a hanger, they start with the top part and randomly select a pin, which is put into place. This is repeated until all four pins are in the top. Then they finish the lower part.

Assume that three pins in the kit are faulty. The other seven are all right. What is the probability that all three faulty pins get put in the top part of the hanger? This means that the support is held up, in effect, by only one good pin.

SOLUTION: The population consists of 10 pins identical in appearance. However, three are faulty and seven are good. The sampling of four pins for the top part of the hanger is done *without replacement*. Since we are interested in the faulty pins, let us label them "success" (only a convenient label). Using the notation of Figure AI-2 and the hypergeometric distribution, we have

a = number of successes in the population (bad pins) = 3
b = number of failures in the population (good pins) = 7
n = sample size (number of pins put in top) = 4
r = number of successes in sample (number of bad pins in top) = 3

The hypergeometric distribution applies because the population is relatively small (10 pins) and sampling is done without replacement. By Equation (2), we compute $P(r)$:

$$P(r) = \frac{C_{a,r}C_{b,n-r}}{C_{(a+b),n}}$$

Using the preceding information about a, b, n, and r, we get

$$P(r = 3) = \frac{C_{3,3}C_{7,1}}{C_{10,4}}$$

Using the formula for $C_{k,j}$, Table 2 of Appendix II, or the combinations key on a calculator, we get

$$P(r = 3) = \frac{1 \cdot 7}{210} = 0.0333$$

We see that there is a better than 3.3% chance of getting three out of four bad pins in the top part of the hanger.

(b) Suppose that all the hanger kits are like the one described in part (a). On a long bridge that uses 200 such hangers, how many do you expect are held up by only one good pin? How might this affect the safety of the bridge?

SOLUTION: We would expect

$$200(0.0333) \approx 6.7$$

That is, between six and seven hangers are expected to be held up by only one good pin. As time goes on, this pin will corrode and show signs of wear as the bridge vibrates. With only one good pin, there is much less margin of safety.

Professor Petroski discusses the bridge on I-95 across the Mianus River in his book mentioned earlier. He points out that this dramatic accidental collapse resulted in better quality control (for hangers and pins) as well as better overall design of bridges. In addition to this, the government has greatly increased programs for maintenance and inspection of bridges.

GUIDED EXERCISE 2 Hypergeometric distribution

The biology club weekend outing has two groups. One group with seven people will camp at Diamond Lake. The other group with 10 people will camp at Arapahoe Pass. Seventeen duffels were prepacked by the outing committee, but six of these had the tents accidentally left out of the duffel. The group going to Diamond Lake picked up their duffels at random from the collection and started off on the trail. The group going to Arapahoe Pass used the remaining duffels. What is the probability that all six duffels without tents were picked up by the group going to Diamond Lake?

Continued

GUIDED EXERCISE 2 *continued*

(a) What is success? Are the duffels selected with or without replacement? Which probability distribution applies?

Success is taking a duffel without a tent. The duffels are selected without replacement. The hypergeometric distribution applies.

(b) Use the hypergeometric distribution to compute the probability of $r = 6$ successes in the sample of seven people going to Diamond Lake.

To use the hypergeometric distribution, we need to know the values of

a = number of successes in population = 6

b = number of failures in population = 11

n = sample size = 7, since seven people are going to Diamond Lake

r = number of successes in sample = 6

Then, $P(r = 6) = \dfrac{C_{6,6}\, C_{11,1}}{C_{17,7}} = \dfrac{1 \cdot 11}{19448} = 0.0006$

The probability that all six duffels without tents are taken by the seven hikers to Diamond Lake is 0.0006.

Appendix II Tables

1. Random Numbers
2. Binomial Coefficients $C_{n,r}$
3. Binomial Probability Distribution $C_{n,r}p^r q^{n-r}$
4. Poisson Probability Distribution
5. Areas of a Standard Normal Distribution
6. Critical Values for Student's t Distribution
7. The χ^2 Distribution
8. Critical Values for F Distribution
9. Critical Values for Spearman Rank Correlation, r_s
10. Critical Values for Number of Runs R

TABLE 1 Random Numbers

92630	78240	19267	95457	53497	23894	37708	79862	76471	66418
79445	78735	71549	44843	26104	67318	00701	34986	66751	99723
59654	71966	27386	50004	05358	94031	29281	18544	52429	06080
31524	49587	76612	39789	13537	48086	59483	60680	84675	53014
06348	76938	90379	51392	55887	71015	09209	79157	24440	30244
28703	51709	94456	48396	73780	06436	86641	69239	57662	80181
68108	89266	94730	95761	75023	48464	65544	96583	18911	16391
99938	90704	93621	66330	33393	95261	95349	51769	91616	33238
91543	73196	34449	63513	83834	99411	58826	40456	69268	48562
42103	02781	73920	56297	72678	12249	25270	36678	21313	75767
17138	27584	25296	28387	51350	61664	37893	05363	44143	42677
28297	14280	54524	21618	95320	38174	60579	08089	94999	78460
09331	56712	51333	06289	75345	08811	82711	57392	25252	30333
31295	04204	93712	51287	05754	79396	87399	51773	33075	97061
36146	15560	27592	42089	99281	59640	15221	96079	09961	05371
29553	18432	13630	05529	02791	81017	49027	79031	50912	09399
23501	22642	63081	08191	89420	67800	55137	54707	32945	64522
57888	85846	67967	07835	11314	01545	48535	17142	08552	67457
55336	71264	88472	04334	63919	36394	11196	92470	70543	29776
10087	10072	55980	64688	68239	20461	89381	93809	00796	95945
34101	81277	66090	88872	37818	72142	67140	50785	21380	16703
53362	44940	60430	22834	14130	96593	23298	56203	92671	15925
82975	66158	84731	19436	55790	69229	28661	13675	99318	76873
54827	84673	22898	08094	14326	87038	42892	21127	30712	48489
25464	59098	27436	89421	80754	89924	19097	67737	80368	08795
67609	60214	41475	84950	40133	02546	09570	45682	50165	15609
44921	70924	61295	51137	47596	86735	35561	76649	18217	63446
33170	30972	98130	95828	49786	13301	36081	80761	33985	68621
84687	85445	06208	17654	51333	02878	35010	67578	61574	20749
71886	56450	36567	09395	96951	35507	17555	35212	69106	01679

TABLE 1 *continued*

00475	02224	74722	14721	40215	21351	08596	45625	83981	63748
25993	38881	68361	59560	41274	69742	40703	37993	03435	18873
92882	53178	99195	93803	56985	53089	15305	50522	55900	43026
25138	26810	07093	15677	60688	04410	24505	37890	67186	62829
84631	71882	12991	83028	82484	90339	91950	74579	03539	90122
34003	92326	12793	61453	48121	74271	28363	66561	75220	35908
53775	45749	05734	86169	42762	70175	97310	73894	88606	19994
59316	97885	72807	54966	60859	11932	35265	71601	55577	67715
20479	66557	50705	26999	09854	52591	14063	30214	19890	19292
86180	84931	25455	26044	02227	52015	21820	50599	51671	65411
21451	68001	72710	40261	61281	13172	63819	48970	51732	54113
98062	68375	80089	24135	72355	95428	11808	29740	81644	86610
01788	64429	14430	94575	75153	94576	61393	96192	03227	32258
62465	04841	43272	68702	01274	05437	22953	18946	99053	41690
94324	31089	84159	92933	99989	89500	91586	02802	69471	68274
05797	43984	21575	09908	70221	19791	51578	36432	33494	79888
10395	14289	52185	09721	25789	38562	54794	04897	59012	89251
35177	56986	25549	59730	64718	52630	31100	62384	49483	11409
25633	89619	75882	98256	02126	72099	57183	55887	09320	73463
16464	48280	94254	45777	45150	68865	11382	11782	22695	41988

Source: MILLION RANDOM DIGITS WITH 100,000 NORMAL DEVIATES. By Staff. Copyright 1983 by RAND CORPORATION. Reproduced with permission of RAND CORPORATION in the formats Textbook and Other Book via Copyright Clearance Center.

TABLE 2 **Binomial Coefficients $C_{n,r}$**

n \ r	0	1	2	3	4	5	6	7	8	9	10
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1
11	1	11	55	165	330	462	462	330	165	55	11
12	1	12	66	220	495	792	924	792	495	220	66
13	1	13	78	286	715	1,287	1,716	1,716	1,287	715	286
14	1	14	91	364	1,001	2,002	3,003	3,432	3,003	2,002	1,001
15	1	15	105	455	1,365	3,003	5,005	6,435	6,435	5,005	3,003
16	1	16	120	560	1,820	4,368	8,008	11,440	12,870	11,440	8,008
17	1	17	136	680	2,380	6,188	12,376	19,448	24,310	24,310	19,448
18	1	18	153	816	3,060	8,568	18,564	31,824	43,758	48,620	43,758
19	1	19	171	969	3,876	11,628	27,132	50,388	75,582	92,378	92,378
20	1	20	190	1,140	4,845	15,504	38,760	77,520	125,970	167,960	184,756

TABLE 3 **Binomial Probability Distribution $C_{n,r}p^rq^{n-r}$**

This table shows the probability of *r* successes in *n* independent trials, each with probability of success *p*.

		p																			
n	*r*	.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
2	0	.980	.902	.810	.723	.640	.563	.490	.423	.360	.303	.250	.203	.160	.123	.090	.063	.040	.023	.010	.002
	1	.020	.095	.180	.255	.320	.375	.420	.455	.480	.495	.500	.495	.480	.455	.420	.375	.320	.255	.180	.095
	2	.000	.002	.010	.023	.040	.063	.090	.123	.160	.203	.250	.303	.360	.423	.490	.563	.640	.723	.810	.902
3	0	.970	.857	.729	.614	.512	.422	.343	.275	.216	.166	.125	.091	.064	.043	.027	.016	.008	.003	.001	.000
	1	.029	.135	.243	.325	.384	.422	.441	.444	.432	.408	.375	.334	.288	.239	.189	.141	.096	.057	.027	.007
	2	.000	.007	.027	.057	.096	.141	.189	.239	.288	.334	.375	.408	.432	.444	.441	.422	.384	.325	.243	.135
	3	.000	.000	.001	.003	.008	.016	.027	.043	.064	.091	.125	.166	.216	.275	.343	.422	.512	.614	.729	.857
4	0	.961	.815	.656	.522	.410	.316	.240	.179	.130	.092	.062	.041	.026	.015	.008	.004	.002	.001	.000	.000
	1	.039	.171	.292	.368	.410	.422	.412	.384	.346	.300	.250	.200	.154	.112	.076	.047	.026	.011	.004	.000
	2	.001	.014	.049	.098	.154	.211	.265	.311	.346	.368	.375	.368	.346	.311	.265	.211	.154	.098	.049	.014
	3	.000	.000	.004	.011	.026	.047	.076	.112	.154	.200	.250	.300	.346	.384	.412	.422	.410	.368	.292	.171
	4	.000	.000	.000	.001	.002	.004	.008	.015	.026	.041	.062	.092	.130	.179	.240	.316	.410	.522	.656	.815
5	0	.951	.774	.590	.444	.328	.237	.168	.116	.078	.050	.031	.019	.010	.005	.002	.001	.000	.000	.000	.000
	1	.048	.204	.328	.392	.410	.396	.360	.312	.259	.206	.156	.113	.077	.049	.028	.015	.006	.002	.000	.000
	2	.001	.021	.073	.138	.205	.264	.309	.336	.346	.337	.312	.276	.230	.181	.132	.088	.051	.024	.008	.001
	3	.000	.001	.008	.024	.051	.088	.132	.181	.230	.276	.312	.337	.346	.336	.309	.264	.205	.138	.073	.021
	4	.000	.000	.000	.002	.006	.015	.028	.049	.077	.113	.156	.206	.259	.312	.360	.396	.410	.392	.328	.204
	5	.000	.000	.000	.000	.000	.001	.002	.005	.010	.019	.031	.050	.078	.116	.168	.237	.328	.444	.590	.774
6	0	.941	.735	.531	.377	.262	.178	.118	.075	.047	.028	.016	.008	.004	.002	.001	.000	.000	.000	.000	.000
	1	.057	.232	.354	.399	.393	.356	.303	.244	.187	.136	.094	.061	.037	.020	.010	.004	.002	.000	.000	.000
	2	.001	.031	.098	.176	.246	.297	.324	.328	.311	.278	.234	.186	.138	.095	.060	.033	.015	.006	.001	.000
	3	.000	.002	.015	.042	.082	.132	.185	.236	.276	.303	.312	.303	.276	.236	.185	.132	.082	.042	.015	.002
	4	.000	.000	.001	.006	.015	.033	.060	.095	.138	.186	.234	.278	.311	.328	.324	.297	.246	.176	.098	.031
	5	.000	.000	.000	.000	.002	.004	.010	.020	.037	.061	.094	.136	.187	.244	.303	.356	.393	.399	.354	.232
	6	.000	.000	.000	.000	.000	.000	.001	.002	.004	.008	.016	.028	.047	.075	.118	.178	.262	.377	.531	.735
7	0	.932	.698	.478	.321	.210	.133	.082	.049	.028	.015	.008	.004	.002	.001	.000	.000	.000	.000	.000	.000
	1	.066	.257	.372	.396	.367	.311	.247	.185	.131	.087	.055	.032	.017	.008	.004	.001	.000	.000	.000	.000
	2	.002	.041	.124	.210	.275	.311	.318	.299	.261	.214	.164	.117	.077	.047	.025	.012	.004	.001	.000	.000
	3	.000	.004	.023	.062	.115	.173	.227	.268	.290	.292	.273	.239	.194	.144	.097	.058	.029	.011	.003	.000
	4	.000	.000	.003	.011	.029	.058	.097	.144	.194	.239	.273	.292	.290	.268	.227	.173	.115	.062	.023	.004
	5	.000	.000	.000	.001	.004	.012	.025	.047	.077	.117	.164	.214	.261	.299	.318	.311	.275	.210	.124	.041
	6	.000	.000	.000	.000	.000	.001	.004	.008	.017	.032	.055	.087	.131	.185	.247	.311	.367	.396	.372	.257
	7	.000	.000	.000	.000	.000	.000	.000	.001	.002	.004	.008	.015	.028	.049	.082	.133	.210	.321	.478	.698

TABLE 3 *continued*

		p																			
n	*r*	.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
8	0	.923	.663	.430	.272	.168	.100	.058	.032	.017	.008	.004	.002	.001	.000	.000	.000	.000	.000	.000	.000
	1	.075	.279	.383	.385	.336	.267	.198	.137	.090	.055	.031	.016	.008	.003	.001	.000	.000	.000	.000	.000
	2	.003	.051	.149	.238	.294	.311	.296	.259	.209	.157	.109	.070	.041	.022	.010	.004	.001	.000	.000	.000
	3	.000	.005	.033	.084	.147	.208	.254	.279	.279	.257	.219	.172	.124	.081	.047	.023	.009	.003	.000	.000
	4	.000	.000	.005	.018	.046	.087	.136	.188	.232	.263	.273	.263	.232	.188	.136	.087	.046	.018	.005	.000
	5	.000	.000	.000	.003	.009	.023	.047	.081	.124	.172	.219	.257	.279	.279	.254	.208	.147	.084	.033	.005
	6	.000	.000	.000	.000	.001	.004	.010	.022	.041	.070	.109	.157	.209	.259	.296	.311	.294	.238	.149	.051
	7	.000	.000	.000	.000	.000	.000	.001	.003	.008	.016	.031	.055	.090	.137	.198	.267	.336	.385	.383	.279
	8	.000	.000	.000	.000	.000	.000	.000	.000	.001	.002	.004	.008	.017	.032	.058	.100	.168	.272	.430	.663
9	0	.914	.630	.387	.232	.134	.075	.040	.021	.010	.005	.002	.001	.000	.000	.000	.000	.000	.000	.000	.000
	1	.083	.299	.387	.368	.302	.225	.156	.100	.060	.034	.018	.008	.004	.001	.000	.000	.000	.000	.000	.000
	2	.003	.063	.172	.260	.302	.300	.267	.216	.161	.111	.070	.041	.021	.010	.004	.001	.000	.000	.000	.000
	3	.000	.008	.045	.107	.176	.234	.267	.272	.251	.212	.164	.116	.074	.042	.021	.009	.003	.001	.000	.000
	4	.000	.001	.007	.028	.066	.117	.172	.219	.251	.260	.246	.213	.167	.118	.074	.039	.017	.005	.001	.000
	5	.000	.000	.001	.005	.017	.039	.074	.118	.167	.213	.246	.260	.251	.219	.172	.117	.066	.028	.007	.001
	6	.000	.000	.000	.001	.003	.009	.021	.042	.074	.116	.164	.212	.251	.272	.267	.234	.176	.107	.045	.008
	7	.000	.000	.000	.000	.000	.001	.004	.010	.021	.041	.070	.111	.161	.216	.267	.300	.302	.260	.172	.063
	8	.000	.000	.000	.000	.000	.000	.000	.001	.004	.008	.018	.034	.060	.100	.156	.225	.302	.368	.387	.299
	9	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.002	.005	.010	.021	.040	.075	.134	.232	.387	.630
10	0	.904	.599	.349	.197	.107	.056	.028	.014	.006	.003	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
	1	.091	.315	.387	.347	.268	.188	.121	.072	.040	.021	.010	.004	.002	.000	.000	.000	.000	.000	.000	.000
	2	.004	.075	.194	.276	.302	.282	.233	.176	.121	.076	.044	.023	.011	.004	.001	.000	.000	.000	.000	.000
	3	.000	.010	.057	.130	.201	.250	.267	.252	.215	.166	.117	.075	.042	.021	.009	.003	.001	.000	.000	.000
	4	.000	.001	.011	.040	.088	.146	.200	.238	.251	.238	.205	.160	.111	.069	.037	.016	.006	.001	.000	.000
	5	.000	.000	.001	.008	.026	.058	.103	.154	.201	.234	.246	.234	.201	.154	.103	.058	.026	.008	.001	.000
	6	.000	.000	.000	.001	.006	.016	.037	.069	.111	.160	.205	.238	.251	.238	.200	.146	.088	.040	.011	.001
	7	.000	.000	.000	.000	.001	.003	.009	.021	.042	.075	.117	.166	.215	.252	.267	.250	.201	.130	.057	.010
	8	.000	.000	.000	.000	.000	.000	.001	.004	.011	.023	.044	.076	.121	.176	.233	.282	.302	.276	.194	.075
	9	.000	.000	.000	.000	.000	.000	.000	.000	.002	.004	.010	.021	.040	.072	.121	.188	.268	.347	.387	.315
	10	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.003	.006	.014	.028	.056	.107	.197	.349	.599
11	0	.895	.569	.314	.167	.086	.042	.020	.009	.004	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	1	.099	.329	.384	.325	.236	.155	.093	.052	.027	.013	.005	.002	.001	.000	.000	.000	.000	.000	.000	.000
	2	.005	.087	.213	.287	.295	.258	.200	.140	.089	.051	.027	.013	.005	.002	.001	.000	.000	.000	.000	.000

TABLE 3 *continued*

		p																			
n	*r*	.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
11	3	.000	.014	.071	.152	.221	.258	.257	.225	.177	.126	.081	.046	.023	.010	.004	.001	.000	.000	.000	.000
	4	.000	.001	.016	.054	.111	.172	.220	.243	.236	.206	.161	.113	.070	.038	.017	.006	.002	.000	.000	.000
	5	.000	.000	.002	.013	.039	.080	.132	.183	.221	.236	.226	.193	.147	.099	.057	.027	.010	.002	.000	.000
	6	.000	.000	.000	.002	.010	.027	.057	.099	.147	.193	.226	.236	.221	.183	.132	.080	.039	.013	.002	.000
	7	.000	.000	.000	.000	.002	.006	.017	.038	.070	.113	.161	.206	.236	.243	.220	.172	.111	.054	.016	.001
	8	.000	.000	.000	.000	.000	.001	.004	.010	.023	.046	.081	.126	.177	.225	.257	.258	.221	.152	.071	.014
	9	.000	.000	.000	.000	.000	.000	.001	.002	.005	.013	.027	.051	.089	.140	.200	.258	.295	.287	.213	.087
	10	.000	.000	.000	.000	.000	.000	.000	.000	.001	.002	.005	.013	.027	.052	.093	.155	.236	.325	.384	.329
	11	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.004	.009	.020	.042	.086	.167	.314	.569
12	0	.886	.540	.282	.142	.069	.032	.014	.006	.002	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	1	.107	.341	.377	.301	.206	.127	.071	.037	.017	.008	.003	.001	.000	.000	.000	.000	.000	.000	.000	.000
	2	.006	.099	.230	.292	.283	.232	.168	.109	.064	.034	.016	.007	.002	.001	.000	.000	.000	.000	.000	.000
	3	.000	.017	.085	.172	.236	.258	.240	.195	.142	.092	.054	.028	.012	.005	.001	.000	.000	.000	.000	.000
	4	.000	.002	.021	.068	.133	.194	.231	.237	.213	.170	.121	.076	.042	.020	.008	.002	.001	.000	.000	.000
	5	.000	.000	.004	.019	.053	.103	.158	.204	.227	.223	.193	.149	.101	.059	.029	.011	.003	.001	.000	.000
	6	.000	.000	.000	.004	.016	.040	.079	.128	.177	.212	.226	.212	.177	.128	.079	.040	.016	.004	.000	.000
	7	.000	.000	.000	.001	.003	.011	.029	.059	.101	.149	.193	.223	.227	.204	.158	.103	.053	.019	.004	.000
	8	.000	.000	.000	.000	.001	.002	.008	.020	.042	.076	.121	.170	.213	.237	.231	.194	.133	.068	.021	.002
	9	.000	.000	.000	.000	.000	.000	.001	.005	.012	.028	.054	.092	.142	.195	.240	.258	.236	.172	.085	.017
	10	.000	.000	.000	.000	.000	.000	.000	.001	.002	.007	.016	.034	.064	.109	.168	.232	.283	.292	.230	.099
	11	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.003	.008	.017	.037	.071	.127	.206	.301	.377	.341
	12	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.002	.006	.014	.032	.069	.142	.282	.540
15	0	.860	.463	.206	.087	.035	.013	.005	.002	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	1	.130	.366	.343	.231	.132	.067	.031	.013	.005	.002	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	2	.009	.135	.267	.286	.231	.156	.092	.048	.022	.009	.003	.001	.000	.000	.000	.000	.000	.000	.000	.000
	3	.000	.031	.129	.218	.250	.225	.170	.111	.063	.032	.014	.005	.002	.000	.000	.000	.000	.000	.000	.000
	4	.000	.005	.043	.116	.188	.225	.219	.179	.127	.078	.042	.019	.007	.002	.001	.000	.000	.000	.000	.000
	5	.000	.001	.010	.045	.103	.165	.206	.212	.186	.140	.092	.051	.024	.010	.003	.001	.000	.000	.000	.000
	6	.000	.000	.002	.013	.043	.092	.147	.191	.207	.191	.153	.105	.061	.030	.012	.003	.001	.000	.000	.000
	7	.000	.000	.000	.003	.014	.039	.081	.132	.177	.201	.196	.165	.118	.071	.035	.013	.003	.001	.000	.000
	8	.000	.000	.000	.001	.003	.013	.035	.071	.118	.165	.196	.201	.177	.132	.081	.039	.014	.003	.000	.000
	9	.000	.000	.000	.000	.001	.003	.012	.030	.061	.105	.153	.191	.207	.191	.147	.092	.043	.013	.002	.000
	10	.000	.000	.000	.000	.000	.001	.003	.010	.024	.051	.092	.140	.186	.212	.206	.165	.103	.045	.010	.001

TABLE 3 *continued*

n	r	.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
15	11	.000	.000	.000	.000	.000	.000	.001	.002	.007	.019	.042	.078	.127	.179	.219	.225	.188	.116	.043	.005
	12	.000	.000	.000	.000	.000	.000	.000	.000	.002	.005	.014	.032	.063	.111	.170	.225	.250	.218	.129	.031
	13	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.003	.009	.022	.048	.092	.156	.231	.286	.267	.135
	14	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.002	.005	.013	.031	.067	.132	.231	.343	.366
	15	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.002	.005	.013	.035	.087	.206	.463
16	0	.851	.440	.185	.074	.028	.010	.003	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	1	.138	.371	.329	.210	.113	.053	.023	.009	.003	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	2	.010	.146	.275	.277	.211	.134	.073	.035	.015	.006	.002	.001	.000	.000	.000	.000	.000	.000	.000	.000
	3	.000	.036	.142	.229	.246	.208	.146	.089	.047	.022	.009	.003	.001	.000	.000	.000	.000	.000	.000	.000
	4	.000	.006	.051	.131	.200	.225	.204	.155	.101	.057	.028	.011	.004	.001	.000	.000	.000	.000	.000	.000
	5	.000	.001	.014	.056	.120	.180	.210	.201	.162	.112	.067	.034	.014	.005	.001	.000	.000	.000	.000	.000
	6	.000	.000	.003	.018	.055	.110	.165	.198	.198	.168	.122	.075	.039	.017	.006	.001	.000	.000	.000	.000
	7	.000	.000	.000	.005	.020	.052	.101	.152	.189	.197	.175	.132	.084	.044	.019	.006	.001	.000	.000	.000
	8	.000	.000	.000	.001	.006	.020	.049	.092	.142	.181	.196	.181	.142	.092	.049	.020	.006	.001	.000	.000
	9	.000	.000	.000	.000	.001	.006	.019	.044	.084	.132	.175	.197	.189	.152	.101	.052	.020	.005	.000	.000
	10	.000	.000	.000	.000	.000	.001	.006	.017	.039	.075	.122	.168	.198	.198	.165	.110	.055	.018	.003	.000
	11	.000	.000	.000	.000	.000	.000	.001	.005	.014	.034	.067	.112	.162	.201	.210	.180	.120	.056	.014	.001
	12	.000	.000	.000	.000	.000	.000	.000	.001	.004	.011	.028	.057	.101	.155	.204	.225	.200	.131	.051	.006
	13	.000	.000	.000	.000	.000	.000	.000	.000	.001	.003	.009	.022	.047	.089	.146	.208	.246	.229	.142	.036
	14	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.002	.006	.015	.035	.073	.134	.211	.277	.275	.146
	15	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.003	.009	.023	.053	.113	.210	.329	.371
	16	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.003	.010	.028	.074	.185	.440
20	0	.818	.358	.122	.039	.012	.003	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	1	.165	.377	.270	.137	.058	.021	.007	.002	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	2	.016	.189	.285	.229	.137	.067	.028	.010	.003	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	3	.001	.060	.190	.243	.205	.134	.072	.032	.012	.004	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
	4	.000	.013	.090	.182	.218	.190	.130	.074	.035	.014	.005	.001	.000	.000	.000	.000	.000	.000	.000	.000
	5	.000	.002	.032	.103	.175	.202	.179	.127	.075	.036	.015	.005	.001	.000	.000	.000	.000	.000	.000	.000
	6	.000	.000	.009	.045	.109	.169	.192	.171	.124	.075	.036	.015	.005	.001	.000	.000	.000	.000	.000	.000
	7	.000	.000	.002	.016	.055	.112	.164	.184	.166	.122	.074	.037	.015	.005	.001	.000	.000	.000	.000	.000
	8	.000	.000	.000	.005	.022	.061	.114	.161	.180	.162	.120	.073	.035	.014	.004	.001	.000	.000	.000	.000
	9	.000	.000	.000	.001	.007	.027	.065	.116	.160	.177	.160	.119	.071	.034	.012	.003	.000	.000	.000	.000

TABLE 3 *continued*

		p																			
n	*r*	.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
20	10	.000	.000	.000	.000	.002	.010	.031	.069	.117	.159	.176	.159	.117	.069	.031	.010	.002	.000	.000	.000
	11	.000	.000	.000	.000	.000	.003	.012	.034	.071	.119	.160	.177	.160	.116	.065	.027	.007	.001	.000	.000
	12	.000	.000	.000	.000	.000	.001	.004	.014	.035	.073	.120	.162	.180	.161	.114	.061	.022	.005	.000	.000
	13	.000	.000	.000	.000	.000	.000	.001	.005	.015	.037	.074	.122	.166	.184	.164	.112	.055	.016	.002	.000
	14	.000	.000	.000	.000	.000	.000	.000	.001	.005	.015	.037	.075	.124	.171	.192	.169	.109	.045	.009	.000
	15	.000	.000	.000	.000	.000	.000	.000	.000	.001	.005	.015	.036	.075	.127	.179	.202	.175	.103	.032	.002
	16	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.005	.014	.035	.074	.130	.190	.218	.182	.090	.013
	17	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.004	.012	.032	.072	.134	.205	.243	.190	.060
	18	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.003	.010	.028	.067	.137	.229	.285	.189
	19	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.002	.007	.021	.058	.137	.270	.377
	20	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.003	.012	.039	.122	.358

TABLE 4 **Poisson Probability Distribution**

For a given value of λ, entry indicates the probability of obtaining a specified value of *r*.

r	λ 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4	.0000	.0001	.0003	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005
7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

r	λ 1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353
1	.3662	.3614	.3543	.3452	.3347	.3230	.3106	.2975	.2842	.2707
2	.2014	.2169	.2303	.2417	.2510	.2584	.2640	.2678	.2700	.2707
3	.0738	.0867	.0998	.1128	.1255	.1378	.1496	.1607	.1710	.1804
4	.0203	.0260	.0324	.0395	.0471	.0551	.0636	.0723	.0812	.0902
5	.0045	.0062	.0084	.0111	.0141	.0176	.0216	.0260	.0309	.0361
6	.0008	.0012	.0018	.0026	.0035	.0047	.0061	.0078	.0098	.0120
7	.0001	.0002	.0003	.0005	.0008	.0011	.0015	.0020	.0027	.0034
8	.0000	.0000	.0001	.0001	.0001	.0002	.0003	.0005	.0006	.0009
9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002

r	λ 2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	.1225	.1108	.1003	.0907	.0821	.0743	.0672	.0608	.0550	.0498
1	.2572	.2438	.2306	.2177	.2052	.1931	.1815	.1703	.1596	.1494
2	.2700	.2681	.2652	.2613	.2565	.2510	.2450	.2384	.2314	.2240
3	.1890	.1966	.2033	.2090	.2138	.2176	.2205	.2225	.2237	.2240
4	.0992	.1082	.1169	.1254	.1336	.1414	.1488	.1557	.1622	.1680
5	.0417	.0476	.0538	.0602	.0668	.0735	.0804	.0872	.0940	.1008
6	.0146	.0174	.0206	.0241	.0278	.0319	.0362	.0407	.0455	.0504
7	.0044	.0055	.0068	.0083	.0099	.0118	.0139	.0163	.0188	.0216
8	.0011	.0015	.0019	.0025	.0031	.0038	.0047	.0057	.0068	.0081
9	.0003	.0004	.0005	.0007	.0009	.0011	.0014	.0018	.0022	.0027
10	.0001	.0001	.0001	.0002	.0002	.0003	.0004	.0005	.0006	.0008
11	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002	.0002
12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

TABLE 4 *continued*

	λ									
r	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	.0450	.0408	.0369	.0334	.0302	.0273	.0247	.0224	.0202	.0183
1	.1397	.1304	.1217	.1135	.1057	.0984	.0915	.0850	.0789	.0733
2	.2165	.2087	.2008	.1929	.1850	.1771	.1692	.1615	.1539	.1465
3	.2237	.2226	.2209	.2186	.2158	.2125	.2087	.2046	.2001	.1954
4	.1734	.1781	.1823	.1858	.1888	.1912	.1931	.1944	.1951	.1954
5	.1075	.1140	.1203	.1264	.1322	.1377	.1429	.1477	.1522	.1563
6	.0555	.0608	.0662	.0716	.0771	.0826	.0881	.0936	.0989	.1042
7	.0246	.2078	.0312	.0348	.0385	.0425	.0466	.0508	.0551	.0595
8	.0095	.0111	.0129	.0148	.0169	.0191	.0215	.0241	.0269	.0298
9	.0033	.0040	.0047	.0056	.0066	.0076	.0089	.0102	.0116	.0132
10	.0010	.0013	.0016	.0019	.0023	.0028	.0033	.0039	.0045	.0053
11	.0003	.0004	.0005	.0006	.0007	.0009	.0011	.0013	.0016	.0019
12	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005	.0006
13	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0002	.0002
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

	λ									
r	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	.0166	.0150	.0136	.0123	.0111	.0101	.0091	.0082	.0074	.0067
1	.0679	.0630	.0583	.0540	.0500	.0462	.0427	.0395	.0365	.0337
2	.1393	.1323	.1254	.1188	.1125	.1063	.1005	.0948	.0894	.0842
3	.1904	.1852	.1798	.1743	.1687	.1631	.1574	.1517	.1460	.1404
4	.1951	.1944	.1933	.1917	.1898	.1875	.1849	.1820	.1789	.1755
5	.1600	.1633	.1662	.1687	.1708	.1725	.1738	.1747	.1753	.1755
6	.1093	.1143	.1191	.1237	.1281	.1323	.1362	.1398	.1432	.1462
7	.0640	.0686	.0732	.0778	.0824	.0869	.0914	.0959	.1002	.1044
8	.0328	.0360	.0393	.0428	.0463	.0500	.0537	.0575	.0614	.0653
9	.0150	.0168	.0188	.0209	.0232	.0255	.0280	.0307	.0334	.0363
10	.0061	.0071	.0081	.0092	.0104	.0118	.0132	.0147	.0164	.0181
11	.0023	.0027	.0032	.0037	.0043	.0049	.0056	.0064	.0073	.0082
12	.0008	.0009	.0011	.0014	.0016	.0019	.0022	.0026	.0030	.0034
13	.0002	.0003	.0004	.0005	.0006	.0007	.0008	.0009	.0011	.0013
14	.0001	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005
15	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0002

TABLE 4 *continued*

r	λ = 5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	.0061	.0055	.0050	.0045	.0041	.0037	.0033	.0030	.0027	.0025
1	.0311	.0287	.0265	.0244	.0225	.0207	.0191	.0176	.0162	.0149
2	.0793	.0746	.0701	.0659	.0618	.0580	.0544	.0509	.0477	.0446
3	.1348	.1293	.1239	.1185	.1133	.1082	.1033	.0985	.0938	.0892
4	.1719	.1681	.1641	.1600	.1558	.1515	.1472	.1428	.1383	.1339
5	.1753	.1748	.1740	.1728	.1714	.1697	.1678	.1656	.1632	.1606
6	.1490	.1515	.1537	.1555	.1571	.1584	.1594	.1601	.1605	.1606
7	.1086	.1125	.1163	.1200	.1234	.1267	.1298	.1326	.1353	.1377
8	.0692	.0731	.0771	.0810	.0849	.0887	.0925	.0962	.0998	.1033
9	.0392	.0423	.0454	.0486	.0519	.0552	.0586	.0620	.0654	.0688
10	.0200	.0220	.0241	.0262	.0285	.0309	.0334	.0359	.0386	.0413
11	.0093	.0104	.0116	.0129	.0143	.0157	.0173	.0190	.0207	.0225
12	.0039	.0045	.0051	.0058	.0065	.0073	.0082	.0092	.0102	.0113
13	.0015	.0018	.0021	.0024	.0028	.0032	.0036	.0041	.0046	.0052
14	.0006	.0007	.0008	.0009	.0011	.0013	.0015	.0017	.0019	.0022
15	.0002	.0002	.0003	.0003	.0004	.0005	.0006	.0007	.0008	.0009
16	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003
17	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001

r	λ = 6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
0	.0022	.0020	.0018	.0017	.0015	.0014	.0012	.0011	.0010	.0009
1	.0137	.0126	.0116	.0106	.0098	.0090	.0082	.0076	.0070	.0064
2	.0417	.0390	.0364	.0340	.0318	.0296	.0276	.0258	.0240	.0223
3	.0848	.0806	.0765	.0726	.0688	.0652	.0617	.0584	.0552	.0521
4	.1294	.1249	.1205	.1162	.1118	.1076	.1034	.0992	.0952	.0912
5	.1579	.1549	.1519	.1487	.1454	.1420	.1385	.1349	.1314	.1277
6	.1605	.1601	.1595	.1586	.1575	.1562	.1546	.1529	.1511	.1490
7	.1399	.1418	.1435	.1450	.1462	.1472	.1480	.1486	.1489	.1490
8	.1066	.1099	.1130	.1160	.1188	.1215	.1240	.1263	.1284	.1304
9	.0723	.0757	.0791	.0825	.0858	.0891	.0923	.0954	.0985	.1014
10	.0441	.0469	.0498	.0528	.0558	.0588	.0618	.0649	.0679	.0710
11	.0245	.0265	.0285	.0307	.0330	.0353	.0377	.0401	.0426	.0452
12	.0124	.0137	.0150	.0164	.0179	.0194	.0210	.0227	.0245	.0264
13	.0058	.0065	.0073	.0081	.0089	.0098	.0108	.0119	.0130	.0142
14	.0025	.0029	.0033	.0037	.0041	.0046	.0052	.0058	.0064	.0071
15	.0010	.0012	.0014	.0016	.0018	.0020	.0023	.0026	.0029	.0033
16	.0004	.0005	.0005	.0006	.0007	.0008	.0010	.0011	.0013	.0014
17	.0001	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0006
18	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002
19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001

TABLE 4 *continued*

	λ									
r	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
0	.0008	.0007	.0007	.0006	.0006	.0005	.0005	.0004	.0004	.0003
1	.0059	.0054	.0049	.0045	.0041	.0038	.0035	.0032	.0029	.0027
2	.0208	.0194	.0180	.0167	.0156	.0145	.0134	.0125	.0116	.0107
3	.0492	.0464	.0438	.0413	.0389	.0366	.0345	.0324	.0305	.0286
4	.0874	.0836	.0799	.0764	.0729	.0696	.0663	.0632	.0602	.0573
5	.1241	.1204	.1167	.1130	.1094	.1057	.1021	.0986	.0951	.0916
6	.1468	.1445	.1420	.1394	.1367	.1339	.1311	.1282	.1252	.1221
7	.1489	.1486	.1481	.1474	.1465	.1454	.1442	.1428	.1413	.1396
8	.1321	.1337	.1351	.1363	.1373	.1382	.1388	.1392	.1395	.1396
9	.1042	.1070	.1096	.1121	.1144	.1167	.1187	.1207	.1224	.1241
10	.0740	.0770	.0800	.0829	.0858	.0887	.0914	.0941	.0967	.0993
11	.0478	.0504	.0531	.0558	.0585	.0613	.0640	.0667	.0695	.0722
12	.0283	.0303	.0323	.0344	.0366	.0388	.0411	.0434	.0457	.0481
13	.0154	.0168	.0181	.0196	.0211	.0227	.0243	.0260	.0278	.0296
14	.0078	.0086	.0095	.0104	.0113	.0123	.0134	.0145	.0157	.0169
15	.0037	.0041	.0046	.0051	.0057	.0062	.0069	.0075	.0083	.0090
16	.0016	.0019	.0021	.0024	.0026	.0030	.0033	.0037	.0041	.0045
17	.0007	.0008	.0009	.0010	.0012	.0013	.0015	.0017	.0019	.0021
18	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
19	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003	.0003	.0004
20	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002
21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001

	λ									
r	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
0	.0003	.0003	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001
1	.0025	.0023	.0021	.0019	.0017	.0016	.0014	.0013	.0012	.0011
2	.0100	.0092	.0086	.0079	.0074	.0068	.0063	.0058	.0054	.0050
3	.0269	.0252	.0237	.0222	.0208	.0195	.0183	.0171	.0160	.0150
4	.0544	.0517	.0491	.0466	.0443	.0420	.0398	.0377	.0357	.0337
5	.0882	.0849	.0816	.0784	.0752	.0722	.0692	.0663	.0635	.0607
6	.1191	.1160	.1128	.1097	.1066	.1034	.1003	.0972	.0941	.0911
7	.1378	.1358	.1338	.1317	.1294	.1271	.1247	.1222	.1197	.1171
8	.1395	.1392	.1388	.1382	.1375	.1366	.1356	.1344	.1332	.1318
9	.1256	.1269	.1280	.1290	.1299	.1306	.1311	.1315	.1317	.1318
10	.1017	.1040	.1063	.1084	.1104	.1123	.1140	.1157	.1172	.1186
11	.0749	.0776	.0802	.0828	.0853	.0878	.0902	.0925	.0948	.0970
12	.0505	.0530	.0555	.0579	.0604	.0629	.0654	.0679	.0703	.0728
13	.0315	.0334	.0354	.0374	.0395	.0416	.0438	.0459	.0481	.0504
14	.0182	.0196	.0210	.0225	.0240	.0256	.0272	.0289	.0306	.0324
15	.0098	.0107	.0116	.0126	.0136	.0147	.0158	.0169	.0182	.0194

TABLE 4 *continued*

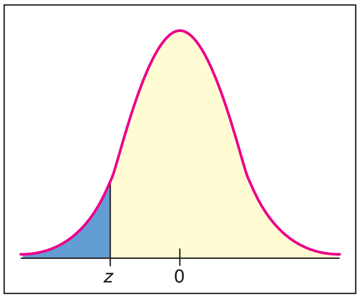
	λ									
r	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
16	.0050	.0055	.0060	.0066	.0072	.0079	.0086	.0093	.0101	.0109
17	.0024	.0026	.0029	.0033	.0036	.0040	.0044	.0048	.0053	.0058
18	.0011	.0012	.0014	.0015	.0017	.0019	.0021	.0024	.0026	.0029
19	.0005	.0005	.0006	.0007	.0008	.0009	.0010	.0011	.0012	.0014
20	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0005	.0006
21	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0002	.0003
22	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001

	λ									
r	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10
0	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0000
1	.0010	.0009	.0009	.0008	.0007	.0007	.0006	.0005	.0005	.0005
2	.0046	.0043	.0040	.0037	.0034	.0031	.0029	.0027	.0025	.0023
3	.0140	.0131	.0123	.0115	.0107	.0100	.0093	.0087	.0081	.0076
4	.0319	.0302	.0285	.0269	.0254	.0240	.0226	.0213	.0201	.0189
5	.0581	.0555	.0530	.0506	.0483	.0460	.0439	.0418	.0398	.0378
6	.0881	.0851	.0822	.0793	.0764	.0736	.0709	.0682	.0656	.0631
7	.1145	.1118	.1091	.1064	.1037	.1010	.0982	.0955	.0928	.0901
8	.1302	.1286	.1269	.1251	.1232	.1212	.1191	.1170	.1148	.1126
9	.1317	.1315	.1311	.1306	.1300	.1293	.1284	.1274	.1263	.1251
10	.1198	.1210	.1219	.1228	.1235	.1241	.1245	.1249	.1250	.1251
11	.0991	.1012	.1031	.1049	.1067	.1083	.1098	.1112	.1125	.1137
12	.0752	.0776	.0799	.0822	.0844	.0866	.0888	.0908	.0928	.0948
13	.0526	.0549	.0572	.0594	.0617	.0640	.0662	.0685	.0707	.0729
14	.0342	.0361	.0380	.0399	.0419	.0439	.0459	.0479	.0500	.0521
15	.0208	.0221	.0235	.0250	.0265	.0281	.0297	.0313	.0330	.0347
16	.0118	.0127	.0137	.0147	.0157	.0168	.0180	.0192	.0204	.0217
17	.0063	.0069	.0075	.0081	.0088	.0095	.0103	.0111	.0119	.0128
18	.0032	.0035	.0039	.0042	.0046	.0051	.0055	.0060	.0065	.0071
19	.0015	.0017	.0019	.0021	.0023	.0026	.0028	.0031	.0034	.0037
20	.0007	.0008	.0009	.0010	.0011	.0012	.0014	.0015	.0017	.0019
21	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
22	.0001	.0001	.0002	.0002	.0002	.0002	.0003	.0003	.0004	.0004
23	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002
24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001

TABLE 4 *continued*

	λ									
r	11	12	13	14	15	16	17	18	19	20
0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	.0010	.0004	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000
3	.0037	.0018	.0008	.0004	.0002	.0001	.0000	.0000	.0000	.0000
4	.0102	.0053	.0027	.0013	.0006	.0003	.0001	.0001	.0000	.0000
5	.0224	.0127	.0070	.0037	.0019	.0010	.0005	.0002	.0001	.0001
6	.0411	.0255	.0152	.0087	.0048	.0026	.0014	.0007	.0004	.0002
7	.0646	.0437	.0281	.0174	.0104	.0060	.0034	.0018	.0010	.0005
8	.0888	.0655	.0457	.0304	.0194	.0120	.0072	.0042	.0024	.0013
9	.1085	.0874	.0661	.0473	.0324	.0213	.0135	.0083	.0050	.0029
10	.1194	.1048	.0859	.0663	.0486	.0341	.0230	.0150	.0095	.0058
11	.1194	.1144	.1015	.0844	.0663	.0496	.0355	.0245	.0164	.0106
12	.1094	.1144	.1099	.0984	.0829	.0661	.0504	.0368	.0259	.0176
13	.0926	.1056	.1099	.1060	.0956	.0814	.0658	.0509	.0378	.0271
14	.0728	.0905	.1021	.1060	.1024	.0930	.0800	.0655	.0514	.0387
15	.0534	.0724	.0885	.0989	.1024	.0992	.0906	.0786	.0650	.0516
16	.0367	.0543	.0719	.0866	.0960	.0992	.0963	.0884	.0772	.0646
17	.0237	.0383	.0550	.0713	.0847	.0934	.0963	.0936	.0863	.0760
18	.0145	.0256	.0397	.0554	.0706	.0830	.0909	.0936	.0911	.0844
19	.0084	.0161	.0272	.0409	.0557	.0699	.0814	.0887	.0911	.0888
20	.0046	.0097	.0177	.0286	.0418	.0559	.0692	.0798	.0866	.0888
21	.0024	.0055	.0109	.0191	.0299	.0426	.0560	.0684	.0783	.0846
22	.0012	.0030	.0065	.0121	.0204	.0310	.0433	.0560	.0676	.0769
23	.0006	.0016	.0037	.0074	.0133	.0216	.0320	.0438	.0559	.0669
24	.0003	.0008	.0020	.0043	.0083	.0144	.0226	.0328	.0442	.0557
25	.0001	.0004	.0010	.0024	.0050	.0092	.0154	.0237	.0336	.0446
26	.0000	.0002	.0005	.0013	.0029	.0057	.0101	.0164	.0246	.0343
27	.0000	.0001	.0002	.0007	.0016	.0034	.0063	.0109	.0173	.0254
28	.0000	.0000	.0001	.0003	.0009	.0019	.0038	.0070	.0117	.0181
29	.0000	.0000	.0001	.0002	.0004	.0011	.0023	.0044	.0077	.0125
30	.0000	.0000	.0000	.0001	.0002	.0006	.0013	.0026	.0049	.0083
31	.0000	.0000	.0000	.0000	.0001	.0003	.0007	.0015	.0030	.0054
32	.0000	.0000	.0000	.0000	.0001	.0001	.0004	.0009	.0018	.0034
33	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0005	.0010	.0020
34	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0012
35	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0007
36	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0004
37	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
38	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
39	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

Source: Biometricka, June 1964, The χ^2 Distribution, H. L. Herter (Table 7). Used by permission of Oxford University Press.



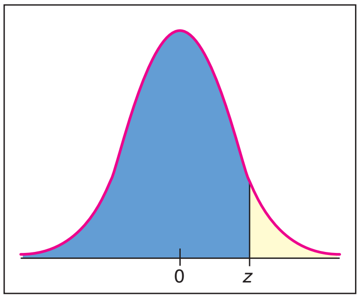
The table entry for z is the area to the left of z.

TABLE 5 Areas of a Standard Normal Distribution

(a) Table of Areas to the Left of z

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

For values of z less than −3.49, use 0.000 to approximate the area.



The table entry for z is the area to the left of z.

TABLE 5(a) *continued*

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

For z values greater than 3.49, use 1.000 to approximate the area.

TABLE 5 *continued*

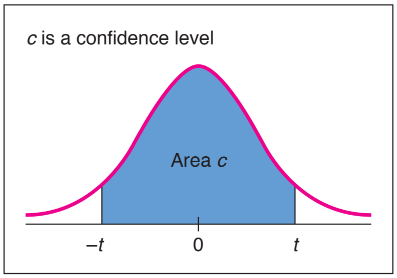
(b) Confidence Interval Critical Values z_c

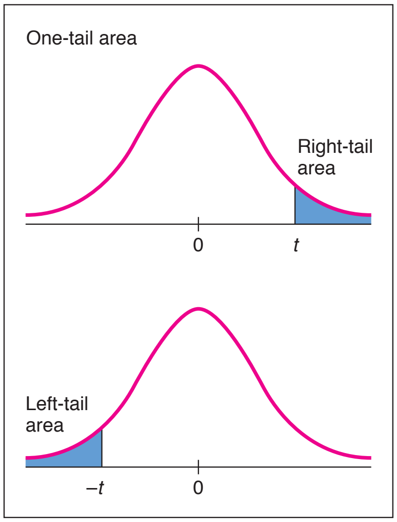
Level of Confidence c	Critical Value z_c
0.70, or 70%	1.04
0.75, or 75%	1.15
0.80, or 80%	1.28
0.85, or 85%	1.44
0.90, or 90%	1.645
0.95, or 95%	1.96
0.98, or 98%	2.33
0.99, or 99%	2.58

TABLE 5 *continued*

(c) Hypothesis Testing, Critical Values z_0

Level of Significance	$\alpha = 0.05$	$\alpha = 0.01$
Critical value z_0 for a left-tailed test	-1.645	-2.33
Critical value z_0 for a right-tailed test	1.645	2.33
Critical values $\pm z_0$ for a two-tailed test	± 1.96	± 2.58





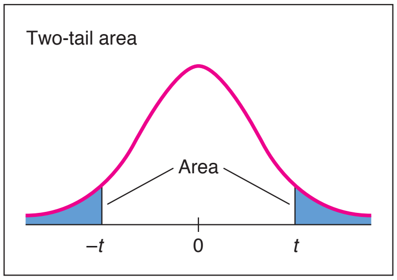
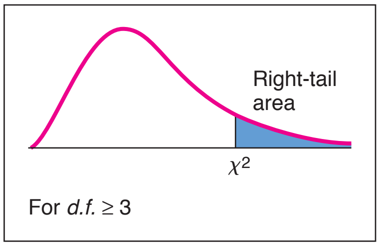


TABLE 6 Critical Values for Student's *t* Distribution

one-tail area	0.250	0.125	0.100	0.075	0.050	0.025	0.010	0.005	0.0005
two-tail area	0.500	0.250	0.200	0.150	0.100	0.050	0.020	0.010	0.0010
d.f. \ *c*	0.500	0.750	0.800	0.850	0.900	0.950	0.980	0.990	0.999
1	1.000	2.414	3.078	4.165	6.314	12.706	31.821	63.657	636.619
2	0.816	1.604	1.886	2.282	2.920	4.303	6.965	9.925	31.599
3	0.765	1.423	1.638	1.924	2.353	3.182	4.541	5.841	12.924
4	0.741	1.344	1.533	1.778	2.132	2.776	3.747	4.604	8.610
5	0.727	1.301	1.476	1.699	2.015	2.571	3.365	4.032	6.869
6	0.718	1.273	1.440	1.650	1.943	2.447	3.143	3.707	5.959
7	0.711	1.254	1.415	1.617	1.895	2.365	2.998	3.499	5.408
8	0.706	1.240	1.397	1.592	1.860	2.306	2.896	3.355	5.041
9	0.703	1.230	1.383	1.574	1.833	2.262	2.821	3.250	4.781
10	0.700	1.221	1.372	1.559	1.812	2.228	2.764	3.169	4.587
11	0.697	1.214	1.363	1.548	1.796	2.201	2.718	3.106	4.437
12	0.695	1.209	1.356	1.538	1.782	2.179	2.681	3.055	4.318
13	0.694	1.204	1.350	1.530	1.771	2.160	2.650	3.012	4.221
14	0.692	1.200	1.345	1.523	1.761	2.145	2.624	2.977	4.140
15	0.691	1.197	1.341	1.517	1.753	2.131	2.602	2.947	4.073
16	0.690	1.194	1.337	1.512	1.746	2.120	2.583	2.921	4.015
17	0.689	1.191	1.333	1.508	1.740	2.110	2.567	2.898	3.965
18	0.688	1.189	1.330	1.504	1.734	2.101	2.552	2.878	3.922
19	0.688	1.187	1.328	1.500	1.729	2.093	2.539	2.861	3.883
20	0.687	1.185	1.325	1.497	1.725	2.086	2.528	2.845	3.850
21	0.686	1.183	1.323	1.494	1.721	2.080	2.518	2.831	3.819
22	0.686	1.182	1.321	1.492	1.717	2.074	2.508	2.819	3.792
23	0.685	1.180	1.319	1.489	1.714	2.069	2.500	2.807	3.768
24	0.685	1.179	1.318	1.487	1.711	2.064	2.492	2.797	3.745
25	0.684	1.198	1.316	1.485	1.708	2.060	2.485	2.787	3.725
26	0.684	1.177	1.315	1.483	1.706	2.056	2.479	2.779	3.707
27	0.684	1.176	1.314	1.482	1.703	2.052	2.473	2.771	3.690
28	0.683	1.175	1.313	1.480	1.701	2.048	2.467	2.763	3.674
29	0.683	1.174	1.311	1.479	1.699	2.045	2.462	2.756	3.659
30	0.683	1.173	1.310	1.477	1.697	2.042	2.457	2.750	3.646
35	0.682	1.170	1.306	1.472	1.690	2.030	2.438	2.724	3.591
40	0.681	1.167	1.303	1.468	1.684	2.021	2.423	2.704	3.551
45	0.680	1.165	1.301	1.465	1.679	2.014	2.412	2.690	3.520
50	0.679	1.164	1.299	1.462	1.676	2.009	2.403	2.678	3.496
60	0.679	1.162	1.296	1.458	1.671	2.000	2.390	2.660	3.460
70	0.678	1.160	1.294	1.456	1.667	1.994	2.381	2.648	3.435
80	0.678	1.159	1.292	1.453	1.664	1.990	2.374	2.639	3.416
100	0.677	1.157	1.290	1.451	1.660	1.984	2.364	2.626	3.390
500	0.675	1.152	1.283	1.442	1.648	1.965	2.334	2.586	3.310
1000	0.675	1.151	1.282	1.441	1.646	1.962	2.330	2.581	3.300
∞	0.674	1.150	1.282	1.440	1.645	1.960	2.326	2.576	3.291

For degrees of freedom *d.f.* not in the table, use the closest *d.f.* that is *smaller.*



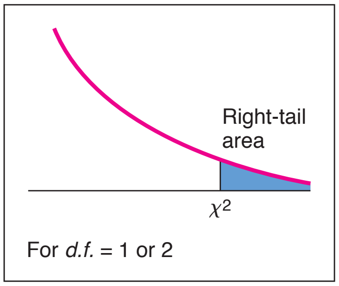
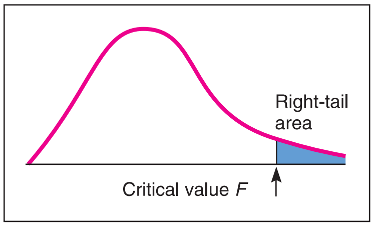


TABLE 7 The χ^2 Distribution

	Right-tail Area									
d.f.	.995	.990	.975	.950	.900	.100	.050	.025	.010	.005
1	0.0^4393	0.0^3157	0.0^3982	0.0^2393	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	8.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.21	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.80	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2

Source: Biometricka, June 1964, The χ^2 Distribution, H. L. Herter (Table 7). Used by permission of Oxford University Press.

TABLE 8 Critical Values for F Distribution



Degrees of freedom denominator, $d.f._D$	Right-tail area	Degrees of freedom numerator, $d.f._N$ 1	2	3	4	5	6	7	8	9
	0.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	0.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
1	0.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	0.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
	0.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
	0.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	0.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
2	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	0.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	0.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
	0.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	0.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
3	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	0.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	0.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86
	0.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	0.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
4	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
	0.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	0.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47
	0.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	0.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
5	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	0.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	0.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24
	0.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	0.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
6	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
	0.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	0.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69
	0.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	0.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
7	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
	0.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	0.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33
	0.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	0.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
8	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	0.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	0.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77

TABLE 8 *continued*

Degrees of freedom denominator, $d.f._D$	Right-tail area	Degrees of freedom numerator, $d.f._N$										
		10	12	15	20	25	30	40	50	60	120	1000
	0.100	60.19	60.71	61.22	61.74	62.05	62.26	62.53	62.69	62.79	63.06	63.30
	0.050	241.88	243.91	245.95	248.01	249.26	250.10	251.14	251.77	252.20	253.25	254.19
1	0.025	968.63	976.71	984.87	993.10	998.08	1001.4	1005.6	1008.1	1009.8	1014.0	1017.7
	0.010	6055.8	6106.3	6157.3	6208.7	6239.8	6260.6	6286.8	6302.5	6313.0	6339.4	6362.7
	0.001	605621	610668	615764	620908	624017	626099	628712	630285	631337	633972	636301
	0.100	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.47	9.48	9.49
	0.050	19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.48	19.49	19.49
2	0.025	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.48	39.49	39.50
	0.010	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.48	99.49	99.50
	0.001	999.40	999.42	999.43	999.45	999.46	999.47	999.47	999.48	999.48	999.49	999.50
	0.100	5.23	5.22	5.20	5.18	5.17	5.17	5.16	5.15	5.15	5.14	5.13
	0.050	8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.58	8.57	8.55	8.53
3	0.025	14.42	14.34	14.25	14.17	14.12	14.08	14.04	14.01	13.99	13.95	13.91
	0.010	27.23	27.05	26.87	26.69	26.58	26.50	26.41	26.35	26.32	26.22	26.14
	0.001	129.25	128.32	127.37	126.42	125.84	125.45	124.96	124.66	124.47	123.97	123.53
	0.100	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.80	3.79	3.78	3.76
	0.050	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.70	5.69	5.66	5.63
4	0.025	8.84	8.75	8.66	8.56	8.50	8.46	8.41	8.38	8.36	8.31	8.26
	0.010	14.55	14.37	14.20	14.02	13.91	13.84	13.75	13.69	13.65	13.56	13.47
	0.001	48.05	47.41	46.76	46.10	45.70	45.43	45.09	44.88	44.75	44.40	44.09
	0.100	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.15	3.14	3.12	3.11
	0.050	4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.44	4.43	4.40	4.37
5	0.025	6.62	6.52	6.43	6.33	6.27	6.23	6.18	6.14	6.12	6.07	6.02
	0.010	10.05	9.89	9.72	9.55	9.45	9.38	9.29	9.24	9.20	9.11	9.03
	0.001	26.92	26.42	25.91	25.39	25.08	24.87	24.60	24.44	24.33	24.06	23.82
	0.100	2.94	2.90	2.87	2.84	2.81	2.80	2.78	2.77	2.76	2.74	2.72
	0.050	4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.75	3.74	3.70	3.67
6	0.025	5.46	5.37	5.27	5.17	5.11	5.07	5.01	4.98	4.96	4.90	4.86
	0.010	7.87	7.72	7.56	7.40	7.30	7.23	7.14	7.09	7.06	6.97	6.89
	0.001	18.41	17.99	17.56	17.12	16.85	16.67	16.44	16.31	16.21	15.98	15.77
	0.100	2.70	2.67	2.63	2.59	2.57	2.56	2.54	2.52	2.51	2.49	2.47
	0.050	3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.32	3.30	3.27	3.23
7	0.025	4.76	4.67	4.57	4.47	4.40	4.36	4.31	4.28	4.25	4.20	4.15
	0.010	6.62	6.47	6.31	6.16	6.06	5.99	5.91	5.86	5.82	5.74	5.66
	0.001	14.08	13.71	13.32	12.93	12.69	12.53	12.33	12.20	12.12	11.91	11.72
	0.100	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.35	2.34	2.32	2.30
	0.050	3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.02	3.01	2.97	2.93
8	0.025	4.30	4.20	4.10	4.00	3.94	3.89	3.84	3.81	3.78	3.73	3.68
	0.010	5.81	5.67	5.52	5.36	5.26	5.20	5.12	5.07	5.03	4.95	4.87
	0.001	11.54	11.19	10.84	10.48	10.26	10.11	9.92	9.80	9.73	9.53	9.36

TABLE 8 *continued*

Degrees of freedom denominator, $d.f._D$	Right-tail area	Degrees of freedom numerator, $d.f._N$ 1	2	3	4	5	6	7	8	9
	0.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	0.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
9	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	0.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	0.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11
	0.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	0.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
10	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
	0.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	0.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96
	0.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
	0.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
11	0.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
	0.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	0.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12
	0.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	0.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
12	0.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	0.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	0.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48
	0.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
	0.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
13	0.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
	0.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	0.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98
	0.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
	0.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
14	0.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21
	0.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	0.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58
	0.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
	0.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
15	0.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12
	0.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	0.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26
	0.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
	0.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
16	0.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05
	0.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
	0.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98

TABLE 8 *continued*

Degrees of freedom denominator, $d.f._D$	Right-tail area	Degrees of freedom numerator, $d.f._N$										
		10	12	15	20	25	30	40	50	60	120	1000
	0.100	2.42	2.38	2.34	2.30	2.27	2.25	2.23	2.22	2.21	2.18	2.16
	0.050	3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.80	2.79	2.75	2.71
9	0.025	3.96	3.87	3.77	3.67	3.60	3.56	3.51	3.47	3.45	3.39	3.34
	0.010	5.26	5.11	4.96	4.81	4.71	4.65	4.57	4.52	4.48	4.40	4.32
	0.001	9.89	9.57	9.24	8.90	8.69	8.55	8.37	8.26	8.19	8.00	7.84
	0.100	2.32	2.28	2.24	2.20	2.17	2.16	2.13	2.12	2.11	2.08	2.06
	0.050	2.98	2.91	2.85	2.77	2.73	2.70	2.66	2.64	2.62	2.58	2.54
10	0.025	3.72	3.62	3.52	3.42	3.35	3.31	3.26	3.22	3.20	3.14	3.09
	0.010	4.85	4.71	4.56	4.41	4.31	4.25	4.17	4.12	4.08	4.00	3.92
	0.001	8.75	8.45	8.13	7.80	7.60	7.47	7.30	7.19	7.12	6.94	6.78
	0.100	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.04	2.03	2.00	1.98
	0.050	2.85	2.79	2.72	2.65	2.60	2.57	2.53	2.51	2.49	2.45	2.41
11	0.025	3.53	3.43	3.33	3.23	3.16	3.12	3.06	3.03	3.00	2.94	2.89
	0.010	4.54	4.40	4.25	4.10	4.01	3.94	3.86	3.81	3.78	3.69	3.61
	0.001	7.92	7.63	7.32	7.01	6.81	6.68	6.52	6.42	6.35	6.18	6.02
	0.100	2.19	2.15	2.10	2.06	2.03	2.01	1.99	1.97	1.96	1.93	1.91
	0.050	2.75	2.69	2.62	2.54	2.50	2.47	2.43	2.40	2.38	2.34	2.30
12	0.025	3.37	3.28	3.18	3.07	3.01	2.96	2.91	2.87	2.85	2.79	2.73
	0.010	4.30	4.16	4.01	3.86	3.76	3.70	3.62	3.57	3.54	3.45	3.37
	0.001	7.29	7.00	6.71	6.40	6.22	6.09	5.93	5.83	5.76	5.59	5.44
	0.100	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.92	1.90	1.88	1.85
	0.050	2.67	2.60	2.53	2.46	2.41	2.38	2.34	2.31	2.30	2.25	2.21
13	0.025	3.25	3.15	3.05	2.95	2.88	2.84	2.78	2.74	2.72	2.66	2.60
	0.010	4.10	3.96	3.82	3.66	3.57	3.51	3.43	3.38	3.34	3.25	3.18
	0.001	6.80	6.52	6.23	5.93	5.75	5.63	5.47	5.37	5.30	5.14	4.99
	0.100	2.10	2.05	2.01	1.96	1.93	1.91	1.89	1.87	1.86	1.83	1.80
	0.050	2.60	2.53	2.46	2.39	2.34	2.31	2.27	2.24	2.22	2.18	2.14
14	0.025	3.15	3.05	2.95	2.84	2.78	2.73	2.67	2.64	2.61	2.55	2.50
	0.010	3.94	3.80	3.66	3.51	3.41	3.35	3.27	3.22	3.18	3.09	3.02
	0.001	6.40	6.13	5.85	5.56	5.38	5.25	5.10	5.00	4.94	4.77	4.62
	0.100	2.06	2.02	1.97	1.92	1.89	1.87	1.85	1.83	1.82	1.79	1.76
	0.050	2.54	2.48	2.40	2.33	2.28	2.25	2.20	2.18	2.16	2.11	2.07
15	0.025	3.06	2.96	2.86	2.76	2.69	2.64	2.59	2.55	2.52	2.46	2.40
	0.010	3.80	3.67	3.52	3.37	3.28	3.21	3.13	3.08	3.05	2.96	2.88
	0.001	6.08	5.81	5.54	5.25	5.07	4.95	4.80	4.70	4.64	4.47	4.33
	0.100	2.03	1.99	1.94	1.89	1.86	1.84	1.81	1.79	1.78	1.75	1.72
	0.050	2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.12	2.11	2.06	2.02
16	0.025	2.99	2.89	2.79	2.68	2.61	2.57	2.51	2.47	2.45	2.38	2.32
	0.010	3.69	3.55	3.41	3.26	3.16	3.10	3.02	2.97	2.93	2.84	2.76
	0.001	5.81	5.55	5.27	4.99	4.82	4.70	4.54	4.45	4.39	4.23	4.08

TABLE 8 *continued*

Degrees of freedom denominator, $d.f._D$	Right-tail area	1	2	3	4	5	6	7	8	9
		Degrees of freedom numerator, $d.f._N$								
17	0.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
	0.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
	0.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98
	0.010	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68
	0.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75
18	0.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	0.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	0.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
	0.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	0.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56
19	0.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
	0.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
	0.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
	0.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	0.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39
20	0.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	0.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	0.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	0.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24
21	0.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	0.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
	0.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
	0.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	0.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11
22	0.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	0.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
	0.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
	0.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	0.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99
23	0.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	0.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
	0.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
	0.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	0.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89
24	0.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	0.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	0.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	0.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	0.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80

TABLE 8 *continued*

Degrees of freedom denominator, $d.f._D$	Right-tail area	Degrees of freedom numerator, $d.f._N$										
		10	12	15	20	25	30	40	50	60	120	1000
	0.100	2.00	1.96	1.91	1.86	1.83	1.81	1.78	1.76	1.75	1.72	1.69
	0.050	2.45	2.38	2.31	2.23	2.18	2.15	2.10	2.08	2.06	2.01	1.97
17	0.025	2.92	2.82	2.72	2.62	2.55	2.50	2.44	2.41	2.38	2.32	2.26
	0.010	3.59	3.46	3.31	3.16	3.07	3.00	2.92	2.87	2.83	2.75	2.66
	0.001	5.58	5.32	5.05	4.78	4.60	4.48	4.33	4.24	4.18	4.02	3.87
	0.100	1.98	1.93	1.89	1.84	1.80	1.78	1.75	1.74	1.72	1.69	1.66
	0.050	2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.04	2.02	1.97	1.92
18	0.025	2.87	2.77	2.67	2.56	2.49	2.44	2.38	2.35	2.32	2.26	2.20
	0.010	3.51	3.37	3.23	3.08	2.98	2.92	2.84	2.78	2.75	2.66	2.58
	0.001	5.39	5.13	4.87	4.59	4.42	4.30	4.15	4.06	4.00	3.84	3.69
	0.100	1.96	1.91	1.86	1.81	1.78	1.76	1.73	1.71	1.70	1.67	1.64
	0.050	2.38	2.31	2.23	2.16	2.11	2.07	2.03	2.00	1.98	1.93	1.88
19	0.025	2.82	2.72	2.62	2.51	2.44	2.39	2.33	2.30	2.27	2.20	2.14
	0.010	3.43	3.30	3.15	3.00	2.91	2.84	2.76	2.71	2.67	2.58	2.50
	0.001	5.22	4.97	4.70	4.43	4.26	4.14	3.99	3.90	3.84	3.68	3.53
	0.100	1.94	1.89	1.84	1.79	1.76	1.74	1.71	1.69	1.68	1.64	1.61
	0.050	2.35	2.28	2.20	2.12	2.07	2.04	1.99	1.97	1.95	1.90	1.85
20	0.025	2.77	2.68	2.57	2.46	2.40	2.35	2.29	2.25	2.22	2.16	2.09
	0.010	3.37	3.23	3.09	2.94	2.84	2.78	2.69	2.64	2.61	2.52	2.43
	0.001	5.08	4.82	4.56	4.29	4.12	4.00	3.86	3.77	3.70	3.54	3.40
	0.100	1.92	1.87	1.83	1.78	1.74	1.72	1.69	1.67	1.66	1.62	1.59
	0.050	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.94	1.92	1.87	1.82
21	0.025	2.73	2.64	2.53	2.42	2.36	2.31	2.25	2.21	2.18	2.11	2.05
	0.010	3.31	3.17	3.03	2.88	2.79	2.72	2.64	2.58	2.55	2.46	2.37
	0.001	4.95	4.70	4.44	4.17	4.00	3.88	3.74	3.64	3.58	3.42	3.28
	0.100	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.65	1.64	1.60	1.57
	0.050	2.30	2.23	2.15	2.07	2.02	1.98	1.94	1.91	1.89	1.84	1.79
22	0.025	2.70	2.60	2.50	2.39	2.32	2.27	2.21	2.17	2.14	2.08	2.01
	0.010	3.26	3.12	2.98	2.83	2.73	2.67	2.58	2.53	2.50	2.40	2.32
	0.001	4.83	4.58	4.33	4.06	3.89	3.78	3.63	3.54	3.48	3.32	3.17
	0.100	1.89	1.84	1.80	1.74	1.71	1.69	1.66	1.64	1.62	1.59	1.55
	0.050	2.27	2.20	2.13	2.05	2.00	1.96	1.91	1.88	1.86	1.81	1.76
23	0.025	2.67	2.57	2.47	2.36	2.29	2.24	2.18	2.14	2.11	2.04	1.98
	0.010	3.21	3.07	2.93	2.78	2.69	2.62	2.54	2.48	2.45	2.35	2.27
	0.001	4.73	4.48	4.23	3.96	3.79	3.68	3.53	3.44	3.38	3.22	3.08
	0.100	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.62	1.61	1.57	1.54
	0.050	2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.86	1.84	1.79	1.74
24	0.025	2.64	2.54	2.44	2.33	2.26	2.21	2.15	2.11	2.08	2.01	1.94
	0.010	3.17	3.03	2.89	2.74	2.64	2.58	2.49	2.44	2.40	2.31	2.22
	0.001	4.64	4.39	4.14	3.87	3.71	3.59	3.45	3.36	3.29	3.14	2.99

TABLE 8 *continued*

Degrees of freedom denominator, $d.f._D$	Right-tail area	1	2	3	4	5	6	7	8	9
		Degrees of freedom numerator, $d.f._N$								
	0.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	0.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
25	0.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
	0.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	0.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71
	0.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	0.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
26	0.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
	0.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
	0.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64
	0.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
	0.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
27	0.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
	0.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
	0.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57
	0.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
	0.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
28	0.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
	0.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
	0.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50
	0.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
	0.050	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
29	0.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59
	0.010	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
	0.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45
	0.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	0.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
30	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
	0.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
	0.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39
	0.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
	0.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
40	0.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45
	0.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
	0.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02
	0.100	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76
	0.050	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
50	0.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38
	0.010	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
	0.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82

TABLE 8 *continued*

Degrees of freedom denominator, $d.f._D$	Right-tail area	Degrees of freedom numerator, $d.f._N$										
		10	12	15	20	25	30	40	50	60	120	1000
	0.100	1.87	1.82	1.77	1.72	1.68	1.66	1.63	1.61	1.59	1.56	1.52
	0.050	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.84	1.82	1.77	1.72
25	0.025	2.61	2.51	2.41	2.30	2.23	2.18	2.12	2.08	2.05	1.98	1.91
	0.010	3.13	2.99	2.85	2.70	2.60	2.54	2.45	2.40	2.36	2.27	2.18
	0.001	4.56	4.31	4.06	3.79	3.63	3.52	3.37	3.28	3.22	3.06	2.91
	0.100	1.86	1.81	1.76	1.71	1.67	1.65	1.61	1.59	1.58	1.54	1.51
	0.050	2.22	2.15	2.07	1.99	1.94	1.90	1.85	1.82	1.80	1.75	1.70
26	0.025	2.59	2.49	2.39	2.28	2.21	2.16	2.09	2.05	2.03	1.95	1.89
	0.010	3.09	2.96	2.81	2.66	2.57	2.50	2.42	2.36	2.33	2.23	2.14
	0.001	4.48	4.24	3.99	3.72	3.56	3.44	3.30	3.21	3.15	2.99	2.84
	0.100	1.85	1.80	1.75	1.70	1.66	1.64	1.60	1.58	1.57	1.53	1.50
	0.050	2.20	2.13	2.06	1.97	1.92	1.88	1.84	1.81	1.79	1.73	1.68
27	0.025	2.57	2.47	2.36	2.25	2.18	2.13	2.07	2.03	2.00	1.93	1.86
	0.010	3.06	2.93	2.78	2.63	2.54	2.47	2.38	2.33	2.29	2.20	2.11
	0.001	4.41	4.17	3.92	3.66	3.49	3.38	3.23	3.14	3.08	2.92	2.78
	0.100	1.84	1.79	1.74	1.69	1.65	1.63	1.59	1.57	1.56	1.52	1.48
	0.050	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.79	1.77	1.71	1.66
28	0.025	2.55	2.45	2.34	2.23	2.16	2.11	2.05	2.01	1.98	1.91	1.84
	0.010	3.03	2.90	2.75	2.60	2.51	2.44	2.35	2.30	2.26	2.17	2.08
	0.001	4.35	4.11	3.86	3.60	3.43	3.32	3.18	3.09	3.02	2.86	2.72
	0.100	1.83	1.78	1.73	1.68	1.64	1.62	1.58	1.56	1.55	1.51	1.47
	0.050	2.18	2.10	2.03	1.94	1.89	1.85	1.81	1.77	1.75	1.70	1.65
29	0.025	2.53	2.43	2.32	2.21	2.14	2.09	2.03	1.99	1.96	1.89	1.82
	0.010	3.00	2.87	2.73	2.57	2.48	2.41	2.33	2.27	2.23	2.14	2.05
	0.001	4.29	4.05	3.80	3.54	3.38	3.27	3.12	3.03	2.97	2.81	2.66
	0.100	1.82	1.77	1.72	1.67	1.63	1.61	1.57	1.55	1.54	1.50	1.46
	0.050	2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.76	1.74	1.68	1.63
30	0.025	2.51	2.41	2.31	2.20	2.12	2.07	2.01	1.97	1.94	1.87	1.80
	0.010	2.98	2.84	2.70	2.55	2.45	2.39	2.30	2.25	2.21	2.11	2.02
	0.001	4.24	4.00	3.75	3.49	3.33	3.22	3.07	2.98	2.92	2.76	2.61
	0.100	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.48	1.47	1.42	1.38
	0.050	2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.66	1.64	1.58	1.52
40	0.025	2.39	2.29	2.18	2.07	1.99	1.94	1.88	1.83	1.80	1.72	1.65
	0.010	2.80	2.66	2.52	2.37	2.27	2.20	2.11	2.06	2.02	1.92	1.82
	0.001	3.87	3.64	3.40	3.14	2.98	2.87	2.73	2.64	2.57	2.41	2.25
	0.100	1.73	1.68	1.63	1.57	1.53	1.50	1.46	1.44	1.42	1.38	1.33
	0.050	2.03	1.95	1.87	1.78	1.73	1.69	1.63	1.60	1.58	1.51	1.45
50	0.025	2.32	2.22	2.11	1.99	1.92	1.87	1.80	1.75	1.72	1.64	1.56
	0.010	2.70	2.56	2.42	2.27	2.17	2.10	2.01	1.95	1.91	1.80	1.70
	0.001	3.67	3.44	3.20	2.95	2.79	2.68	2.53	2.44	2.38	2.21	2.05

TABLE 8 *continued*

Degrees of freedom denominator, $d.f._D$	Right-tail area	1	2	3	4	5	6	7	8	9
		Degrees of freedom numerator, $d.f._N$								
60	0.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	0.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
	0.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
	0.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	0.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69
100	0.100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69
	0.050	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
	0.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24
	0.010	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
	0.001	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44
200	0.100	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66
	0.050	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93
	0.025	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18
	0.010	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50
	0.001	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26
1000	0.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64
	0.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89
	0.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13
	0.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43
	0.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30	3.13

TABLE 8 *continued*

Degrees of freedom denominator, $d.f._D$	Right-tail area	Degrees of freedom numerator, $d.f._N$: 10	12	15	20	25	30	40	50	60	120	1000
	0.100	1.71	1.66	1.60	1.54	1.50	1.48	1.44	1.41	1.40	1.35	1.30
	0.050	1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.56	1.53	1.47	1.40
60	0.025	2.27	2.17	2.06	1.94	1.87	1.82	1.74	1.70	1.67	1.58	1.49
	0.010	2.63	2.50	2.35	2.20	2.10	2.03	1.94	1.88	1.84	1.73	1.62
	0.001	3.54	3.32	3.08	2.83	2.67	2.55	2.41	2.32	2.25	2.08	1.92
	0.100	1.66	1.61	1.56	1.49	1.45	1.42	1.38	1.35	1.34	1.28	1.22
	0.050	1.93	1.85	1.77	1.68	1.62	1.57	1.52	1.48	1.45	1.38	1.30
100	0.025	2.18	2.08	1.97	1.85	1.77	1.71	1.64	1.59	1.56	1.46	1.36
	0.010	2.50	2.37	2.22	2.07	1.97	1.89	1.80	1.74	1.69	1.57	1.45
	0.001	3.30	3.07	2.84	2.59	2.43	2.32	2.17	2.08	2.01	1.83	1.64
	0.100	1.63	1.58	1.52	1.46	1.41	1.38	1.34	1.31	1.29	1.23	1.16
	0.050	1.88	1.80	1.72	1.62	1.56	1.52	1.46	1.41	1.39	1.30	1.21
200	0.025	2.11	2.01	1.90	1.78	1.70	1.64	1.56	1.51	1.47	1.37	1.25
	0.010	2.41	2.27	2.13	1.97	1.87	1.79	1.69	1.63	1.58	1.45	1.30
	0.001	3.12	2.90	2.67	2.42	2.26	2.15	2.00	1.90	1.83	1.64	1.43
	0.100	1.61	1.55	1.49	1.43	1.38	1.35	1.30	1.27	1.25	1.18	1.08
	0.050	1.84	1.76	1.68	1.58	1.52	1.47	1.41	1.36	1.33	1.24	1.11
1000	0.025	2.06	1.96	1.85	1.72	1.64	1.58	1.50	1.45	1.41	1.29	1.13
	0.010	2.34	2.20	2.06	1.90	1.79	1.72	1.61	1.54	1.50	1.35	1.16
	0.001	2.99	2.77	2.54	2.30	2.14	2.02	1.87	1.77	1.69	1.49	1.22

Source: From Biometrika, Tables of Statisticans, Vol. I; Critical Values for *F* Distribution. (Table 8). Reprinted by permission of Oxford University Press.

TABLE 9 Critical Values for Spearman Rank Correlation, r_s

For a right- (left-) tailed test, use the positive (negative) critical value found in the table under One-tail area. For a two-tailed test, use both the positive and the negative of the critical value found in the table under Two-tail area; n = number of pairs.

	One-tail area			
	0.05	0.025	0.005	0.001
	Two-tail area			
n	0.10	0.05	0.01	0.002
5	0.900	1.000		
6	0.829	0.886	1.000	
7	0.715	0.786	0.929	1.000
8	0.620	0.715	0.881	0.953
9	0.600	0.700	0.834	0.917
10	0.564	0.649	0.794	0.879
11	0.537	0.619	0.764	0.855
12	0.504	0.588	0.735	0.826
13	0.484	0.561	0.704	0.797
14	0.464	0.539	0.680	0.772
15	0.447	0.522	0.658	0.750
16	0.430	0.503	0.636	0.730
17	0.415	0.488	0.618	0.711
18	0.402	0.474	0.600	0.693
19	0.392	0.460	0.585	0.676
20	0.381	0.447	0.570	0.661
21	0.371	0.437	0.556	0.647
22	0.361	0.426	0.544	0.633
23	0.353	0.417	0.532	0.620
24	0.345	0.407	0.521	0.608
25	0.337	0.399	0.511	0.597
26	0.331	0.391	0.501	0.587
27	0.325	0.383	0.493	0.577
28	0.319	0.376	0.484	0.567
29	0.312	0.369	0.475	0.558
30	0.307	0.363	0.467	0.549

Source: From G. J. Glasser and R. F. Winter, "Critical Values of the Coefficient of Rank Correlation for Testing the Hypothesis of Independence," *Biometrika, 48,* 444 (1961). Reprinted by permission of Oxford University Press.

TABLE 10 **Critical Values for Number of Runs R (Level of significance $\alpha = 0.05$)**

Value of n_1	Value of n_2: 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
3	1	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
	6	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
4	1	1	1	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
	6	8	9	9	9	10	10	10	10	10	10	10	10	10	10	10	10	10	10
5	1	1	2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
	6	8	9	10	10	11	11	12	12	12	12	12	12	12	12	12	12	12	12
6	1	2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
	6	8	9	10	11	12	12	13	13	13	13	14	14	14	14	14	14	14	14
7	1	2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
	6	8	10	11	12	13	13	14	14	14	14	15	15	15	16	16	16	16	16
8	1	2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
	6	8	10	11	12	13	14	14	15	15	16	16	16	16	17	17	17	17	17
9	1	2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
	6	8	10	12	13	14	14	15	16	16	16	17	17	18	18	18	18	18	18
10	1	2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
	6	8	10	12	13	14	15	16	16	17	17	18	18	18	19	19	19	20	20
11	1	2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
	6	8	10	12	13	14	15	16	17	17	18	19	19	19	20	20	20	21	21
12	2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10
	6	8	10	12	13	14	16	16	17	18	19	19	20	20	21	21	21	22	22
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
	6	8	10	12	14	15	16	17	18	19	19	20	20	21	21	22	22	23	23
14	2	2	3	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11	11
	6	8	10	12	14	15	16	17	18	19	20	20	21	22	22	23	23	23	24
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12
	6	8	10	12	14	15	16	18	18	19	20	21	22	22	23	23	24	24	25
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12
	6	8	10	12	14	16	17	18	19	20	21	21	22	23	23	24	25	25	25
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	13
	6	8	10	12	14	16	17	18	19	20	21	22	23	23	24	25	25	26	26
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
	6	8	10	12	14	16	17	18	19	20	21	22	23	24	25	25	26	26	27
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
	6	8	10	12	14	16	17	18	20	21	22	23	23	24	25	26	26	27	27
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14
	6	8	10	12	14	16	17	18	20	21	22	23	24	25	25	26	27	27	28

Source: The Annals of Mathematical Statistics by Staff. Copyright 1961 by INSTITUTE OF MATHEMATICAL STATISTICS. Reproduced with permission of INSTITUTE OF MATHEMATICAL STATISTICS in the format Textbook via Copyright Clearance Center.

Answers and Key Steps to Odd-Numbered Problems

CHAPTER 1

Section 1.1

1. An individual is a member of the population of interest. A variable is an aspect of an individual subject or object being measured.
3. A parameter is a numerical measurement describing data from a population. A statistic is a numerical measurement describing data from a sample.
5. (a) Nominal level. There is no apparent order relationship among responses.
   (b) Ordinal level. There is an increasing relationship from worst to best level of service. The interval between service levels is not meaningful, nor are ratios.
7. (a) Response regarding meal ordered at fast-food restaurants. (b) Qualitative. (c) Responses for *all* adult fast-food customers in the U.S.
9. (a) Nitrogen concentration (mg nitrogen/l water). (b) Quantitative. (c) Nitrogen concentration (mg nitrogen/l water) in the entire lake.
11. (a) Ratio. (b) Interval. (c) Nominal. (d) Ordinal. (e) Ratio. (f) Ratio.
13. (a) Nominal. (b) Ratio. (c) Interval. (d) Ordinal. (e) Ratio. (f) Interval.
15. Answers vary.
    (a) For example: Use pounds. Round weights to the nearest pound. Since backpacks might weigh as much as 30 pounds, you might use a high-quality bathroom scale. (b) Some students may not allow you to weigh their backpacks for privacy reasons, etc. (c) Possibly. Some students may want to impress you with the heaviness of their backpacks, or they may be embarrassed about the "junk" they have stowed inside and thus may clean out their backpacks.

Section 1.2

1. In a stratified sample, random samples from each stratum are included. In a cluster sample, the clusters to be included are selected at random and then all members of each selected cluster are included.
3. The advice is wrong. A sampling error accounts only for the difference in results based on the use of a sample rather than of the entire population.
5. Use a random-number table to select four distinct numbers corresponding to people in your class.
   (a) Reasons may vary. For instance, the first four students may make a special effort to get to class on time.
   (b) Reasons may vary. For instance, four students who come in late might all be nursing students enrolled in an anatomy and physiology class that meets the hour before in a far-away building. They may be more motivated than other students to complete a degree requirement. (c) Reasons may vary. For instance, four students sitting in the back row might be less inclined to participate in class discussions. (d) Reasons may vary. For instance, the tallest students might all be male.
7. Answers vary. Use groups of two digits.
9. Select a starting place in the table and group the digits in groups of four. Scan the table by rows and include the first six groups with numbers between 0001 and 8615.
11. (a) Yes, when a die is rolled several times, the same number may appear more than once. Outcome on the fourth roll is 2. (b) No, for a fair die, the outcomes are random.
13. Since there are five possible outcomes for each question, read single digits from a random-number table. Select a starting place and proceed until you have 10 digits from 1 to 5. Repetition is required. The correct answer for each question will be the letter choice corresponding to the digit chosen for that question.
15. (a) Simple random sample. (b) Cluster sample. (c) Convenience sample. (d) Systematic sample. (e) Stratified sample.

Section 1.3

1. Answers vary. People with higher incomes are more likely to have high-speed Internet access and to spend more time online. People with high-speed Internet access might spend less time watching TV news or programming. People with higher incomes might have less time to spend watching TV because of access to other entertainment venues.
3. (a) No, those ages 18–29 in 2006 became ages 20–31 in 2008.
   (b) 1977 to 1988 (inclusive).
5. (a) Observational study. (b) Experiment. (c) Experiment. (d) Observational study.
7. (a) Use random selection to pick 10 calves to inoculate; test all calves; no placebo. (b) Use random selection to pick 9 schools to visit; survey all schools; no placebo. (c) Use random selection to pick 40 volunteers for skin patch with drug; survey all volunteers; placebo used.
9. Based on the information given, Scheme A is best because it blocks all plots bordering the river together and all plots not bordering the river together.
   The blocks of Scheme B do not seem to differ from each other.

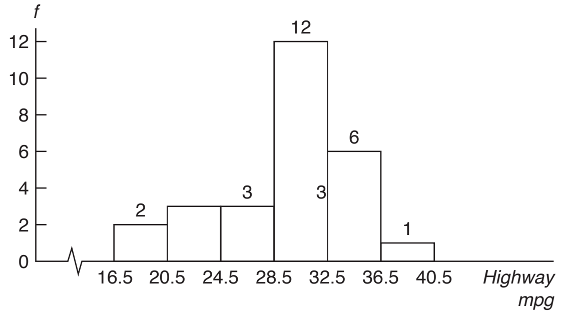
Chapter 1 Review

1. Because of the requirement that each number appear only once in any row, column, or box, it would be very inefficient to use a random-number table to select the numbers. It's better to simply look at existing numbers, list possibilities that meet the requirement, and eliminate numbers that don't work.
3. (a) Stratified. (b) Students on your campus with work-study jobs. (c) Hours scheduled; quantitative; ratio. (d) Rating of applicability of work experience to future employment; qualitative; ordinal. (e) Statistic. (f) 60%; The people choosing not to respond may have some characteristics, such as not working many hours, that would bias the study. (g) No. The sample frame is restricted to one campus.
5. Assign digits so that 3 out of the 10 digits 0 through 9 correspond to the answer "Yes" and 7 of the digits correspond to the answer "No." One assignment is digits 0, 1, and 2 correspond to "Yes," while digits 3, 4, 5, 6, 7, 8, and 9 correspond to "No." Starting with line 1, block 1 of Table 1, this assignment of digits gives the sequence No, Yes, No, No, Yes, No, No.
7. (a) Observational study. (b) Experiment.
9. Possible directions on survey questions: Give height in inches, give age as of last birthday, give GPA to one decimal place, and so forth. Think about the types of responses you wish to have on each question.
11. (a) Experiment, since a treatment is imposed on one colony. (b) The control group receives normal daylight/darkness conditions. The treatment group has light 24 hours per day. (c) The number of fireflies living at the end of 72 hours. (d) Ratio.

CHAPTER 2

Section 2.1

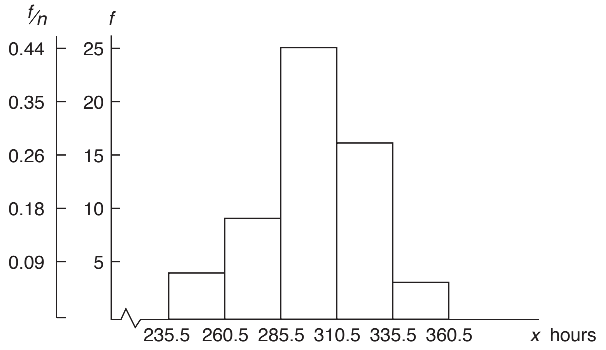
1. Class limits are possible data values. Class limits specify the span of data values that fall within a class. Class boundaries are not possible data values; rather, they are values halfway between the upper class limit of one class and the lower class limit of the next.
3. The classes overlap so that some data values, such as 20, fall within two classes.
5. Class width = 9; class limits: 20–28, 29–37, 38–46, 47–55, 56–64, 65–73, 74–82.
7. (a) Answers vary. Skewed right, if you hope most of the waiting times are low, with only a few times at the higher end of the distribution of waiting times.
(b) A bimodal distribution might reflect the fact that when there are lots of customers, most of the waiting times are longer, especially since the lines are likely to be long. On the other hand, when there are fewer customers, the lines are short or almost nonexistent, and most of the waiting times are briefer.
9. (a) Yes
(b) Histogram of Highway mpg



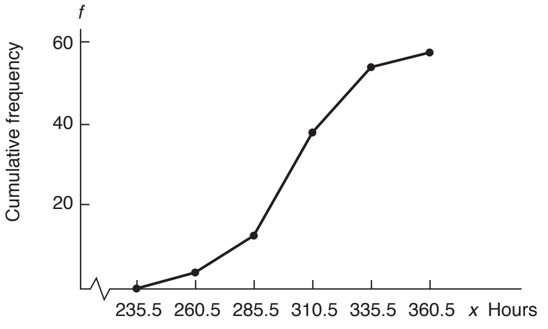
11. (a) Class width = 25.
(b)

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
236–260	235.5–260.5	248	4	0.07	4
261–285	260.5–285.5	273	9	0.16	13
286–310	285.5–310.5	298	25	0.44	38
311–335	310.5–335.5	323	16	0.28	54
336–360	335.5–360.5	348	3	0.05	57

(c, d) Hours to Complete the Iditarod—Histogram, Relative-Frequency Histogram



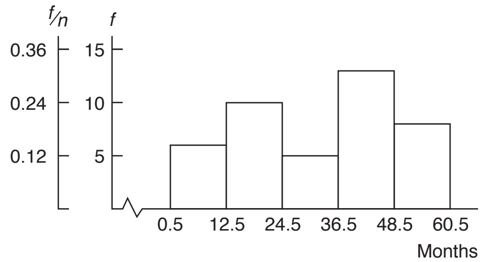
(e) Approximately mound-shaped symmetrical.
(f) Hours to Complete the Iditarod—Ogive



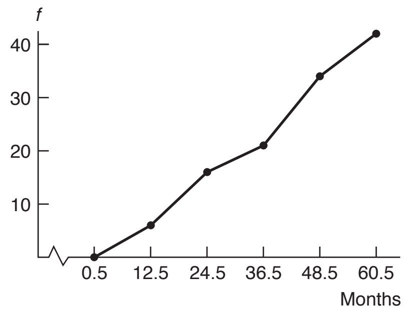
13. (a) Class width = 12.
(b)

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
1–12	0.5–12.5	6.5	6	0.14	6
13–24	12.5–24.5	18.5	10	0.24	16
25–36	24.5–36.5	30.5	5	0.12	21
37–48	36.5–48.5	42.5	13	0.31	34
49–60	48.5–60.5	54.5	8	0.19	42

(c, d) Months Before Tumor Recurrence—Histogram, Relative-Frequency Histogram



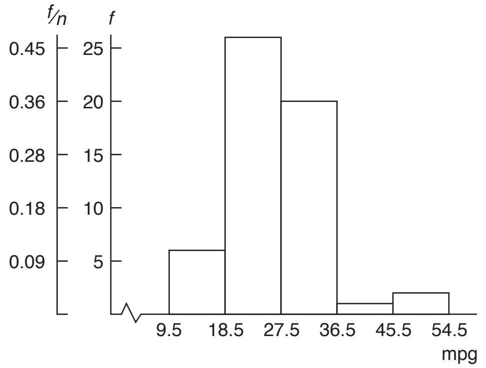
(e) Somewhat bimodal.
(f) Months Before Tumor Recurrence—Ogive



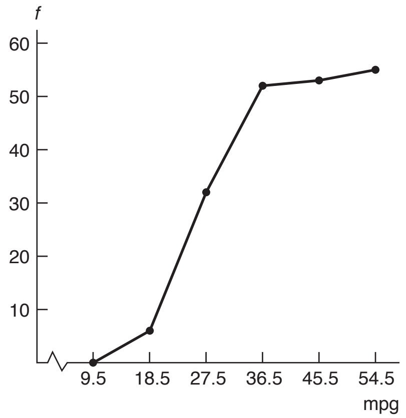
15. (a) Class width = 9.
(b)

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
10–18	9.5–18.5	14	6	0.11	6
19–27	18.5–27.5	23	26	0.47	32
28–36	27.5–36.5	32	20	0.36	52
37–45	36.5–45.5	41	1	0.02	53
46–54	45.5–54.5	50	2	0.04	55

(c, d) Fuel Consumption (mpg)—Histogram, Relative-Frequency Histogram



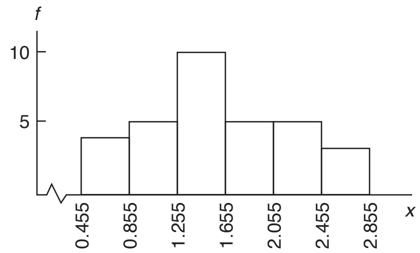
(e) Skewed slightly right.
(f) Fuel Consumption (mpg)—Ogive



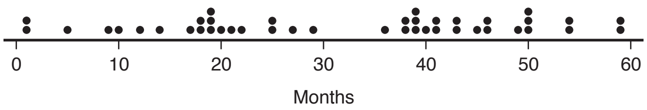
17. (a) Clear the decimals.
(b, c) Class width = 0.40.

Class Limits	Boundaries	Midpoint	Frequency
0.46–0.85	0.455–0.855	0.655	4
0.86–1.25	0.855–1.255	1.055	5
1.26–1.65	1.255–1.655	1.455	10
1.66–2.05	1.655–2.055	1.855	5
2.06–2.45	2.055–2.455	2.255	5
2.46–2.85	2.455–2.855	2.655	3

(c) Tonnes of Wheat—Histogram

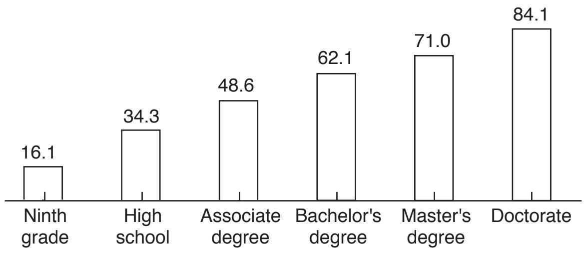


19. (a) One. (b) 5/51 or 9.8%. (c) Interval from 650 to 750.
21. Dotplot for Months Before Tumor Recurrence

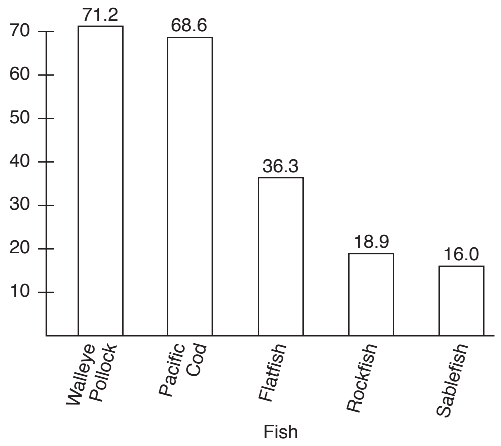


Section 2.2

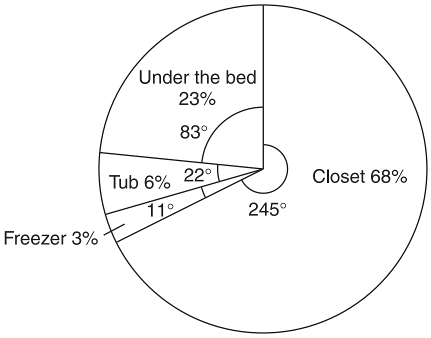
1. (a) Yes, the percentages total more than 100%.
   (b) No, in a circle graph the percentages must total 100% (within rounding error).
   (c) Yes, the graph is organized in order from most frequently selected reason to least.
3. Pareto chart, because it shows the items in order of importance to the greatest number of employees.
5. Highest Level of Education and Average Annual Household Income (in thousands of dollars).



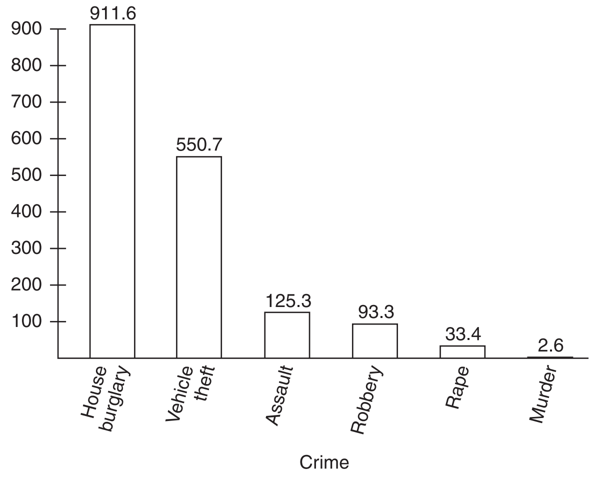
7. Annual Harvest (1000 Metric Tons)—Pareto Chart



9. Where We Hide the Mess

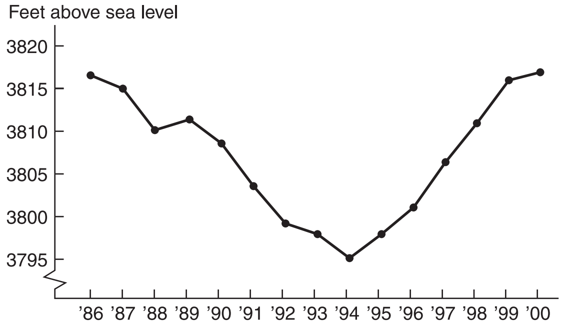


11. (a) Hawaii Crime Rate per 100,000 Population



(b) A circle graph is not appropriate because the data do not reflect all types of crime. Also, the same person may have been the victim of more than one crime.

13. Elevation of Pyramid Lake Surface—Time Plot



Section 2.3

1. (a) Longevity of Cowboys

4	7 = 47 years
4	7
5	2 7 8 8
6	1 6 6 8 8
7	0 2 2 3 3 5 6 7
8	4 4 4 5 6 6 7 9
9	0 1 1 2 3 7

(b) Yes, certainly these cowboys lived long lives.

3. Average Length of Hospital Stay

5	2 = 5.2 years
5	2 3 5 5 6 7
6	0 2 4 6 6 7 7 8 8 8 8 9 9
7	0 0 0 0 0 0 1 1 1 2 2 2 3 3 3 3 4 4 5 5 6 6 8
8	4 5 7
9	4 6 9
10	0 3
11	1

The distribution is skewed right.

5. (a) Minutes Beyond 2 Hours (1961–1980)

0	9 = 9 minutes past 2 hours
0	9 9
1	0 0 2 3 3 4
1	5 5 6 6 7 8 8 9
2	0 2 3 3

(b) Minutes Beyond 2 Hours (1981–2000)

0	7 = 7 minutes past 2 hours
0	7 7 7 8 8 8 8 9 9 9 9 9 9 9 9
1	0 0 1 1 4

(c) In more recent years, the winning times have been closer to 2 hours, with all the times between 7 and 14 minutes over 2 hours. In the earlier period, more than half the times were more than 2 hours and 14 minutes.

7. Milligrams of Tar per Cigarette

1	0 = 1.0 mg tar		
1	0	11	4
2		12	0 4 8
3		13	7
4	1 5	14	1 5 9
5		15	0 1 2 8
6		16	0 6
7	3 8	17	0
8	0 6 8		
9	0		
10		29	8

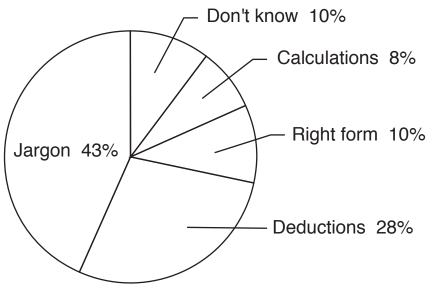
The value 29.8 may be an outlier.

9. Milligrams of Nicotine per Cigarette

0	1 = 0.1 milligram
0	1 4 4
0	5 6 6 6 7 7 7 8 8 9 9 9
1	0 0 0 0 0 0 0 1 2
1	
2	0

Chapter 2 Review

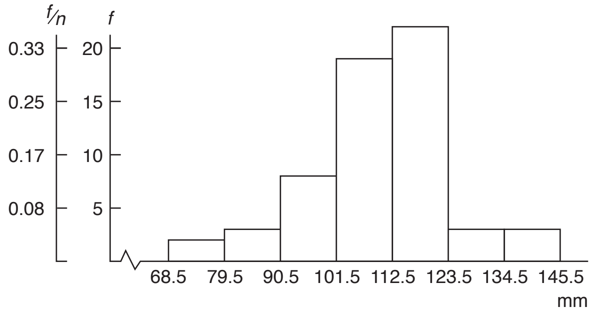
1. (a) Bar graph, Pareto chart, pie chart. (b) All.
3. Any large gaps between bars or stems with leaves at the beginning or end of the data set might indicate that the extreme data values are outliers.
5. (a) Yes, with lines used instead of bars. However, because of the perspective nature of the drawing, the lengths of the bars do not represent the mileages. Thus, the scale for each bar changes. (b) Yes. The scale does not change, and the viewer is not distracted by the graphic of the highway.
7. Problems with Tax Returns



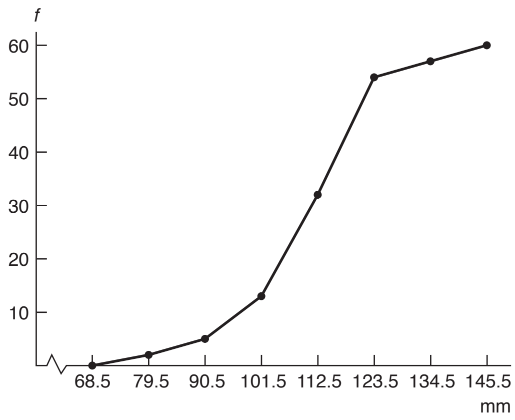
9. (a) Class width = 11.

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
69–79	68.5–79.5	74	2	0.03	2
80–90	79.5–90.5	85	3	0.05	5
91–101	90.5–101.5	96	8	0.13	13
102–112	101.5–112.5	107	19	0.32	32
113–123	112.5–123.5	118	22	0.37	54
124–134	123.5–134.5	129	3	0.05	57
135–145	134.5–145.5	140	3	0.05	60

(b, c) Trunk Circumference (mm)—Histogram, Relative-Frequency Histogram



(d) Skewed slightly left.
(e) Trunk Circumference (mm)—Ogive



11. (a) 1240s had 40 data values. (b) 75. (c) From 1203 to 1212. Little if any repairs or new construction.

CHAPTER 3

Section 3.1

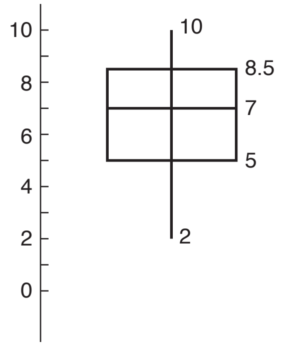
1. Median; mode; mean.
3. $\bar{x} = 5$; median = 6; mode = 2.
5. $\bar{x} = 5$; median = 5.5; mode = 2.
7. Mean, median, and mode are approximately equal.
9. (a) Mode = 5; median = 4; mean = 3.8. (b) Mode. (c) Mean, median, and mode. (d) Mode, median.
11. The supervisor has a legitimate concern because at least half the clients rated the employee below satisfactory. From the information given, it seems that this employee is very inconsistent in her performance.
13. (a) Mode = 2; median = 3; mean = 4.6. (b) Mode = 10; median = 15; mean = 23. (c) Corresponding values are 5 times the original averages. In general, multiplying each data value by a constant c results in the mode, median, and mean changing by a factor of c. (d) Mode = 177.8 cm; median = 172.72 cm; mean = 180.34 cm.
15. $\bar{x} \approx 167.3$°F; median = 171°F; mode = 178°F.
17. (a) $\bar{x} \approx 3.27$; median = 3; mode = 3. (b) $\bar{x} \approx 4.21$; median = 2; mode = 1. (c) Lower Canyon mean is greater; median and mode are less. (d) Trimmed mean = 3.75 and is closer to Upper Canyon mean.
19. (a) $\bar{x}$ = \$136.15; median = \$66.50; mode = \$60. (b) 5% trimmed mean ≈ \$121.28; yes, but still higher than the median. (c) Median. The low and high prices would be useful.
21. 23.
23. $\Sigma wx = 85$; $\Sigma w = 10$; weighted average = 8.5.
25. Approx. 66.67 mph.

Section 3.2

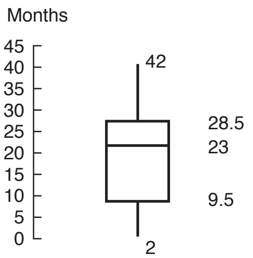
1. Mean.
3. Yes. For the sample standard deviation s, the sum $\Sigma(x - \bar{x})^2$ is divided by $n - 1$, where n is the sample size. For the population standard deviation σ, the sum $\Sigma(x - \mu)^2$ is divided by N, where N is the population size.
5. (a) Range is 4. (b) $s \approx 1.58$. (c) $\sigma \approx 1.41$.
7. For a data set in which not all data values are equal, σ is less than s. The reason is that to compute σ, we divide the sum of the squares by n, and to compute s we divide by the smaller number $n - 1$.
9. (a) (i), (ii), (iii). (b) The data change between data sets (i) and (ii) increased the sum of squared differences $\Sigma(x - \bar{x})^2$ by 10, whereas the data change between data sets (ii) and (iii) increased the sum of squared differences $\Sigma(x - \bar{x})^2$ by only 6.
11. (a) $s \approx 3.6$. (b) $s \approx 18.0$. (c) When each data value is multiplied by 5, the standard deviation is five times greater than that of the original data set. In general, multiplying each data value by the same constant c results in the standard deviation being $|c|$ times as large. (d) No. Multiply 3.1 miles by 1.6 kilometers/mile to obtain $s \approx 4.96$ kilometers.
13. (a) 15. (b) Use a calculator. (c) 37; 6.08. (d) 37; 6.08. (e) $\sigma^2 \approx 29.59$; $\sigma \approx 5.44$.
15. (a) $CV = 10\%$. (b) 14 to 26.
17. (a) 7.87. (b) Use a calculator. (c) $\bar{x} \approx 1.24$; $s^2 \approx 1.78$; $s \approx 1.33$. (d) $CV \approx 107\%$. The standard deviation of the time to failure is just slightly larger than the average time.
19. (a) Use a calculator. (b) $\bar{x} = 49$; $s^2 \approx 687.49$; $s \approx 26.22$. (c) $\bar{y} = 44.8$; $s^2 \approx 508.50$; $s \approx 22.55$. (d) Mallard nests, $CV \approx 53.5\%$; Canada goose nests, $CV \approx 50.3\%$. The CV gives the ratio of the standard deviation to the mean; the CV for mallard nests is slightly higher.
21. Since $CV = s/\bar{x}$, then $s = CV(\bar{x})$; $s = 0.033$.
23. Midpoints: 25.5, 35.5, 45.5; $\bar{x} \approx 35.80$; $s^2 \approx 61.1$; $s \approx 7.82$.
25. Midpoints: 10.55, 14.55, 18.55, 22.55, 26.55; $\bar{x} \approx 15.6$; $s^2 \approx 23.4$; $s \approx 4.8$.

Section 3.3

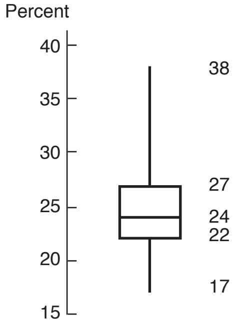
1. 82% or more of the scores were at or below Angela's score; 18% or fewer of the scores were above Angela's score.
3. No, the score 82 might have a percentile rank less than 70.
5. (a) Low = 2; $Q_1 = 5$; median = 7; $Q_3 = 8.5$; high = 10. (b) $IQR = 3.5$. (c) Box-and-Whisker Plot



7. Low = 2; $Q_1 = 9.5$; median = 23; $Q_3 = 28.5$; high = 42; $IQR = 19$.
Nurses' Length of Employment (months)



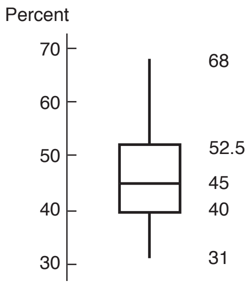
9. (a) Low = 17; $Q_1 = 22$; median = 24; $Q_3 = 27$; high = 38; $IQR = 5$. (b) Third quartile, since it is between the median and Q_3.
Bachelor's Degree Percentage by State



11. (a) California has the lowest premium. Pennsylvania has the highest. (b) Pennsylvania has the highest median premium. (c) California has the smallest range. Texas has the smallest interquartile range. (d) Part (a) is the five-number summary for Texas. It has the smallest IQR. Part (b) is the five-number summary for Pennsylvania. It has the largest minimum. Part (c) is the five-number summary for California. It has the lowest minimum.

Chapter 3 Review

1. (a) Variance and standard deviation. (b) Box-and-whisker plot.
3. (a) For both data sets, mean = 20 and range = 24.
(b) The C1 distribution seems more symmetric because the mean and median are equal, and the median is in the center of the interquartile range. In the C2 distribution, the mean is less than the median.
(c) The C1 distribution has a larger interquartile range that is symmetric around the median. The C2 distribution has a very compressed interquartile range with the median equal to Q_3.
5. (a) Low = 31; $Q_1 = 40$; median = 45; $Q_3 = 52.5$; high = 68; $IQR = 12.5$.
Percentage of Democratic Vote by County



(b) Class width = 8.

Class	Midpoint	f
31–38	34.5	11
39–46	42.5	24
47–54	50.5	15
55–62	58.5	7
63–70	66.5	3

$\bar{x} \approx 46.1$; $s \approx 8.64$; 28.82 to 63.38.
(c) $\bar{x} \approx 46.15$; $s \approx 8.63$.
7. Mean weight = 156.25 pounds.
9. (a) No. (b) \$34,206 to \$68,206. (c) \$10,875.
11. $\Sigma w = 16$, $\Sigma wx = 121$, average = 7.56.

CUMULATIVE REVIEW PROBLEMS

Chapters 1–3

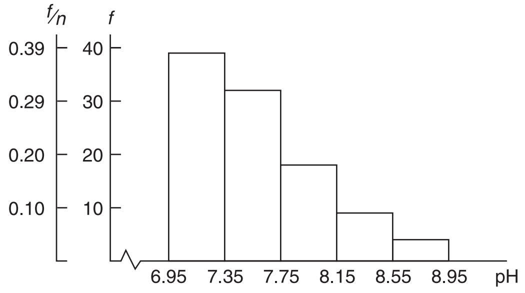
1. (a) Median, percentile. (b) Mean, variance, standard deviation.
2. (a) Gap between first bar and rest of bars or between last bar and rest of bars. (b) Large gap between data on far-left or far-right side and rest of data. (c) Several empty stems after stem including lowest values or before stem including highest values. (d) Data beyond fences placed at $Q_1 - 1.5(IQR)$ and $Q_3 + 1.5(IQR)$.
3. (a) Same. (b) Set B has a higher mean. (c) Set B has a higher standard deviation. (d) Set B has a much longer whisker beyond Q_3.
4. (a) Set A, because 86 is the relatively higher score, since a larger percentage of scores fall below it. (b) Set B, because 86 is more standard deviations above the mean.
5. Assign consecutive numbers to all the wells in the study region. Then use a random-number table, computer, or calculator to select 102 values that are less than or equal to the highest number assigned to a well in the study region. The sample consists of the wells with numbers corresponding to those selected.
6. Ratio.
7. 7 | 0 represents a pH level of 7.0

7	0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1
7	2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3
7	4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5
7	6 6 6 6 6 6 6 6 6 7 7 7 7 7 7
7	8 8 8 8 8 9 9 9 9 9
8	0 1 1 1 1 1 1 1
8	2 2 2 2 2 2 2
8	4 5
8	6 7
8	8 8

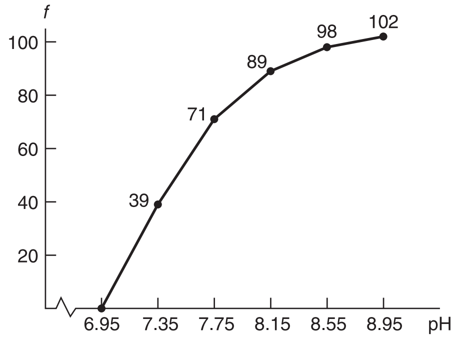
8. Clear the decimals. Then the highest value is 88 and the lowest is 70. The class width for the whole numbers is 4. For the actual data, the class width is 0.4.

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency
7.0–7.3	6.95–7.35	7.15	39	0.38
7.4–7.7	7.35–7.75	7.55	32	0.31
7.8–8.1	7.75–8.15	7.95	18	0.18
8.2–8.5	8.15–8.55	8.35	9	0.09
8.6–8.9	8.55–8.95	8.75	4	0.04

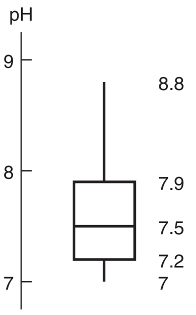
Levels of pH in West Texas Wells—Histogram, Relative-Frequency Histogram



9. Levels of pH in West Texas Wells—Ogive



10. Range = 1.8; $\bar{x} \approx 7.58$; median = 7.5; mode = 7.3.
11. (a) Use a calculator or computer.
    (b) $s^2 \approx 0.20$; $s \approx 0.45$; $CV \approx 5.9\%$.
12. 6.68 to 8.48.
13. Levels of pH in West Texas Wells



$IQR = 0.7$.
14. Skewed right. Lower values are more common.
15. 89%; 50%.
16. No, there are no gaps in the plot, but only 6 out of 102, or about 6%, have pH levels at or above 8.4. Eight wells are neutral.
17. Half the wells have pH levels between 7.2 and 7.9. The data are skewed toward the high values, with the upper half of the pH levels spread out more than the lower half. The upper half ranges between 7.5 and 8.8, while the lower half is clustered between 7 and 7.5.
18. The report should emphasize the relatively low mean, median, and mode, and the fact that half the wells have a pH level less than 7.5. The data are clustered at the low end of the range.

CHAPTER 4

Section 4.1

1. Equally likely outcomes, relative frequency, intuition.
3. (a) 1. (b) 0.
5. $627/1010 \approx 0.62$.
7. Although the probability is high that you will make money, it is not completely certain that you will. In fact, there is a small chance that you could lose your entire investment. If you can afford to lose all of the investment, it might be worthwhile to invest, because there is a high chance of doubling your money.
9. (a) MMM MMF MFM MFF FMM FMF FFM FFF.
   (b) $P(\text{MMM}) = 1/8$. $P(\text{at least one female}) = 1 - P(\text{MMM}) = 7/8$.
11. No. The probability of heads on the second toss is 0.50 regardless of the outcome on the first toss.
13. Answers vary. Probability as a relative frequency. One concern is whether the students in the class are more or less adept at wiggling their ears than people in the general population.
15. (a) $P(0) = 15/375$; $P(1) = 71/375$; $P(2) = 124/375$; $P(3) = 131/375$; $P(4) = 34/375$. (b) Yes, the listed numbers of similar preferences form the sample space.
17. (a) P(best idea 6 A.M.–12 noon) = $290/966 \approx 0.30$; P(best idea 12 noon–6 P.M.) = $135/966 \approx 0.14$; P(best idea 6 P.M.–12 midnight) = $319/966 \approx 0.33$; P(best idea 12 midnight–6 A.M.) = $222/966 \approx 0.23$. (b) The probabilities add up to 1. They should add up to 1 (within rounding errors), provided the intervals do not overlap and each inventor chose only one interval. The sample space is the set of four time intervals.
19. (b) $P(\text{success}) = 2/17 \approx 0.118$. (c) $P(\text{make shot}) = 3/8$ or 0.375.
21. (a) $P(\text{enter if walks by}) = 58/127 \approx 0.46$. (b) $P(\text{buy if entered}) = 25/58 \approx 0.43$. (c) $P(\text{walk in and buy}) = 25/127 \approx 0.20$. (d) $P(\text{not buy}) = 1 - P(\text{buy}) \approx 1 - 0.43 = 0.57$.

Section 4.2

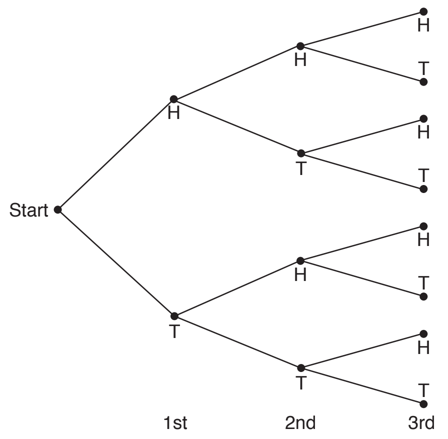
1. No. By definition, mutually exclusive events cannot occur together.
3. (a) 0.7. (b) 0.6.
5. (a) 0.08. (b) 0.04.
7. (a) 0.15. (b) 0.55.
9. (a) Because the events are mutually exclusive, A cannot occur if B occurred. $P(A \mid B) = 0$. (b) Because $P(A \mid B) \neq P(A)$, the events A and B are not independent.
11. (a) $P(A \textit{ and } B)$. (b) $P(B \mid A)$. (c) $P(A^c \mid B)$. (d) $P(A \textit{ or } B)$. (e) $P(B^c \textit{ or } A)$.
13. (a) 0.2; yes. (b) 0.4; yes. (c) $1.0 - 0.2 = 0.8$.
15. (a) Yes. (b) $P(5 \text{ on green } \textit{and } 3 \text{ on red}) = P(5) \cdot P(3) = (1/6)(1/6) = 1/36 \approx 0.028$. (c) $P(3 \text{ on green } \textit{and } 5 \text{ on red}) = P(3) \cdot P(5) = (1/6)(1/6) = 1/36 \approx 0.028$. (d) $P((5 \text{ on green } \textit{and } 3 \text{ on red}) \textit{ or } (3 \text{ on green } \textit{and } 5 \text{ on red})) = (1/36) + (1/36) = 1/18 \approx 0.056$.
17. (a) $P(\text{sum of } 6) = P(1 \textit{ and } 5) + P(2 \textit{ and } 4) + P(3 \textit{ and } 3) + P(4 \textit{ and } 2) + P(5 \textit{ and } 1) = (1/36) + (1/36) + (1/36) + (1/36) + (1/36) = 5/36$. (b) $P(\text{sum of } 4) = P(1 \textit{ and } 3) +$

P(2 *and* 2) + P(3 *and* 1) = (1/36) + (1/36) + (1/36) = 3/36 or 1/12. (c) P(sum of 6 *or* sum of 4) = P(sum of 6) + P(sum of 4) = (5/36) + (3/36) = 8/36 or 2/9; yes.

19. (a) No, after the first draw the sample space becomes smaller and probabilities for events on the second draw change. (b) P(Ace on 1st *and* King on 2nd) = P(Ace) · P(King | Ace) = (4/52)(4/51) = 4/663. (c) P(King on 1st *and* Ace on 2nd) = P(King) · P(Ace | King) = (4/52)(4/51) = 4/663. (d) P(Ace *and* King in either order) = P(Ace on 1st *and* King on 2nd) + P(King on 1st *and* Ace on 2nd) = (4/663) + (4/663) = 8/663.
21. (a) Yes, replacement of the card restores the sample space and all probabilities for the second draw remain unchanged regardless of the outcome of the first card. (b) P(Ace on 1st *and* King on 2nd) = P(Ace) · P(King) = (4/52)(4/52) = 1/169. (c) P(King on 1st *and* Ace on 2nd) = P(King) · P(Ace) = (4/52)(4/52) = 1/169. (d) P(Ace *and* King in either order) = P(Ace on 1st *and* King on 2nd) + P(King on 1st *and* Ace on 2nd) = (1/169) + (1/169) = 2/169.
23. (a) P(6 years old *or* older) = P(6–9) + P(10–12) + P(13 and over) = 0.27 + 0.14 + 0.22 = 0.63. (b) P(12 years old *or* younger) = P(2 and under) + P(3–5) + P(6–9) + P(10–12) = 0.15 + 0.22 + 0.27 + 0.14 = 0.78. (c) P(between 6 and 12) = P(6–9) + P(10–12) = 0.27 + 0.14 = 0.41. (d) P(between 3 and 9) = P(3–5) + P(6–9) = 0.22 + 0.27 = 0.49. The category 13 and over contains far more ages than the group 10–12. It is not surprising that more toys are purchased for this group, since there are more children in this group.
25. The information from James Burke can be viewed as conditional probabilities. P(reports lie | person is lying) = 0.72 and P(reports lie | person is not lying) = 0.07. (a) P(person is not lying) = 0.90; P(person is not lying *and* polygraph reports lie) = P(person is not lying) × P(reports lie | person not lying) = (0.90)(0.07) = 0.063 or 6.3%. (b) P(person is lying) = 0.10; P(person is lying *and* polygraph reports lie) = P(person is lying) × P(reports lie | person is lying) = (0.10)(0.72) = 0.072 or 7.2%. (c) P(person is not lying) = 0.5; P(person is lying) = 0.5; P(person is not lying *and* polygraph reports lie) = P(person is not lying) × P(reports lie | person not lying) = (0.50)(0.07) = 0.035 or 3.5%. P(person is lying *and* polygraph reports lie) = P(person is lying) × P(reports lie | person is lying) = (0.50)(0.72) = 0.36 or 36%. (d) P(person is not lying) = 0.15; P(person is lying) = 0.85; P(person is not lying *and* polygraph reports lie) = P(person is not lying) × P(reports lie | person is not lying) = (0.15)(0.07) = 0.0105 or 1.05%. P(person is lying *and* polygraph reports lie) = P(person is lying) × P(reports lie | person is lying) = (0.85)(0.72) = 0.612 or 61.2%.
27. (a) 686/1160; 270/580; 416/580. (b) No. (c) 270/1160; 416/1160. (d) 474/1160; 310/580. (e) No. (f) 686/1160 + 580/1160 − 270/1160 = 996/1160.
29. (a) 72/154. (b) 82/154. (c) 79/116. (d) 37/116. (e) 72/270. (f) 82/270.
31. (a) $P(A) = 0.65$. (b) $P(B) = 0.71$. (c) $P(B \mid A) = 0.87$. (d) $P(A \textit{ and } B) = P(A) \cdot P(B \mid A) = (0.65)(0.87) \approx 0.57$. (e) $P(A \textit{ or } B) = P(A) + P(B) - P(A \textit{ and } B) \approx 0.65 + 0.71 - 0.57 = 0.79$. (f) P(not close) = P(profit 1st year *or* profit 2nd year) = $P(A \textit{ or } B) \approx 0.79$; P(close) = 1 − P(not close) ≈ 1 − 0.79 = 0.21.
33. (a) P(TB *and* positive) = P(TB)P(positive | TB) = (0.04)(0.82) ≈ 0.033. (b) P(does not have TB) = 1 − P(TB) = 1 − 0.04 = 0.96. (c) P(no TB *and* positive) = P(no TB)P(positive | no TB) = (0.96)(0.09) ≈ 0.086.
35. True. A^c consists of all events not in A.
37. False. If event A^c has occurred, then event A cannot occur.
39. True. $P(A \textit{ and } B) = P(B) \cdot P(A \mid B)$. Since $0 < P(B) < 1$, the product $P(B) \cdot P(A \mid B) \le P(A \mid B)$.
41. True. All the outcomes in event A *and* B are also in event A.
43. True. All the outcomes in event A^c *and* B^c are also in event A^c.
45. False. See Problem 9.
47. True. Since $P(A \textit{ and } B) = P(A) \cdot P(B) = 0$, either $P(A) = 0$ or $P(B) = 0$.
49. True. All simple events of the sample space under the condition "given B" are included in either the event A or the disjoint event A^c

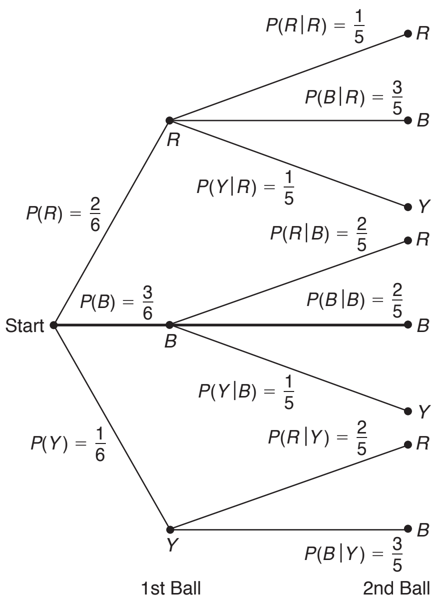
Section 4.3

1. The permutations rule counts the number of different *arrangements* of r items out of n distinct items, whereas the combinations rule counts only the *number* of groups of r items out of n distinct items. The number of permutations is larger than the number of combinations.
3. (a) Use the combinations rule, since only the items in the group and not their arrangement is of concern. (b) Use the permutations rule, since the number of arrangements within each group is of interest.
5. (a) Outcomes for Flipping a Coin Three Times



(b) 3. (c) 3/8.

7. (a) Outcomes for Drawing Two Balls (without replacement)



(b) $P(R \text{ and } R) = 2/6 \cdot 1/5 = 1/15$.
$P(R \text{ 1st and } B \text{ 2nd}) = 2/6 \cdot 3/5 = 1/5$.
$P(R \text{ 1st and } Y \text{ 2nd}) = 2/6 \cdot 1/5 = 1/15$.
$P(B \text{ 1st and } R \text{ 2nd}) = 3/6 \cdot 2/5 = 1/5$.
$P(B \text{ 1st and } B \text{ 2nd}) = 3/6 \cdot 2/5 = 1/5$.
$P(B \text{ 1st and } Y \text{ 2nd}) = 3/6 \cdot 1/5 = 1/10$.
$P(Y \text{ 1st and } R \text{ 2nd}) = 1/6 \cdot 2/5 = 1/15$.
$P(Y \text{ 1st and } B \text{ 2nd}) = 1/6 \cdot 3/5 = 1/10$.

9. $4 \cdot 3 \cdot 2 \cdot 1 = 24$ sequences.
11. $4 \cdot 3 \cdot 3 = 36$.
13. $P_{5,2} = (5!/3!) = 5 \cdot 4 = 20$.
15. $P_{7,7} = (7!/0!) = 7! = 5040$.
17. $C_{5,2} = (5!/(2!3!)) = 10$.
19. $C_{7,7} = (7!/(7!0!)) = 1$.
21. $P_{15,3} = 2730$.
23. $5 \cdot 4 \cdot 3 = 60$.
25. $C_{15,5} = (15!/(5!10!)) = 3003$.
27. (a) $C_{12,6} = (12!/(6!6!)) = 924$.
(b) $C_{7,6} = (7!/(6!1!)) = 7$. (c) $7/924 \approx 0.008$.

Chapter 4 Review

1. (a) The individual does not own a cell phone. (b) The individual owns a cell phone as well as a laptop computer. (c) The individual owns either a cell phone or a laptop computer, and maybe both. (d) The individual owns a cell phone, given he or she owns a laptop computer. (e) The individual owns a laptop computer, given he or she owns a cell phone.
3. For independent events A and B, $P(A) = P(A \mid B)$.
5. (a) $P(\text{offer job 1 } and \text{ offer job 2}) = 0.56$. The probability of getting offers for both jobs is less than the probability of getting each individual job offer.
(b) $P(\text{offer job 1 } or \text{ offer job 2}) = 0.94$. The probability of getting at least one of the job offers is greater than the probability of getting each individual job offer. It seems worthwhile to apply for both jobs since the probability is high of getting at least one offer.
7. (a) No. You need to know that the events are independent or you need to know the value of $P(A \mid B)$ or $P(B \mid A)$. (b) Yes. For independent events, $P(A \text{ and } B) = P(A) \cdot P(B)$.
9. $P(\text{asked}) = 24\%$; $P(\text{received} \mid \text{asked}) = 45\%$; $P(\text{asked } and \text{ received}) = (0.24)(0.45) = 10.8\%$.
11. (a) Drop a fixed number of tacks and count how many land flat side down. Then form the ratio of the number landing flat side down to the total number dropped.
(b) Up, down. (c) $P(\text{up}) = 160/500 = 0.32$; $P(\text{down}) = 340/500 = 0.68$.
13. (a)

Outcomes x	2	3	4	5	6
$P(x)$	0.028	0.056	0.083	0.111	0.139

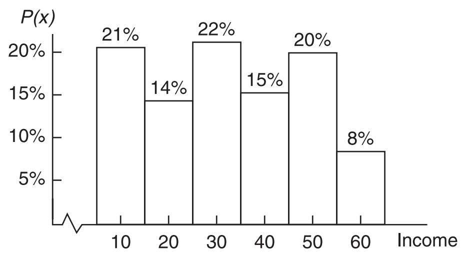
x	7	8	9	10	11	12
$P(x)$	0.167	0.139	0.111	0.083	0.056	0.028

15. $C_{8,2} = (8!/(2!6!)) = (8 \cdot 7/2) = 28$.
17. $4 \cdot 4 \cdot 4 \cdot 4 \cdot 4 = 1024$ choices; $P(\text{all correct}) = 1/1024 \approx 0.00098$.
19. $10 \cdot 10 \cdot 10 = 1000$.

CHAPTER 5

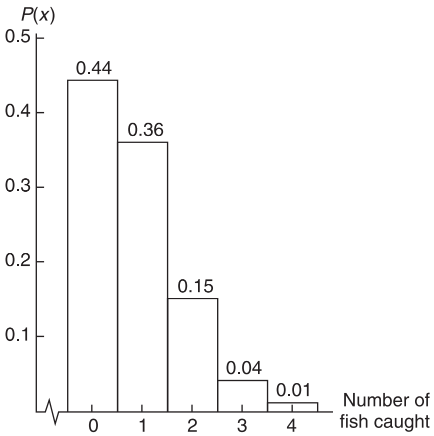
Section 5.1

1. (a) Discrete. (b) Continuous. (c) Continuous. (d) Discrete. (e) Continuous.
3. (a) Yes. (b) No; probabilities total to more than 1.
5. Expected value $= 0.9$. $\sigma \approx 0.6245$.
7. (a) Yes, 7 of the 10 digits represent "making a basket."
(b) Let S represent "making a basket" and F represent "missing the shot." $F, F, S, S, S, F, F, S, S$.
(c) Yes. Again, 7 of the 10 digits represent "making a basket." $S, S, S, S, S, S, S, S, S, S$.
9. (a) Yes, events are distinct and probabilities total to 1.
(b) Income Distribution ($1000)



(c) 32.3 thousand dollars. (d) 16.12 thousand dollars.

11. (a) Number of Fish Caught in a 6-Hour Period at Pyramid Lake, Nevada



(b) 0.56. (c) 0.20. (d) 0.82. (e) 0.899.

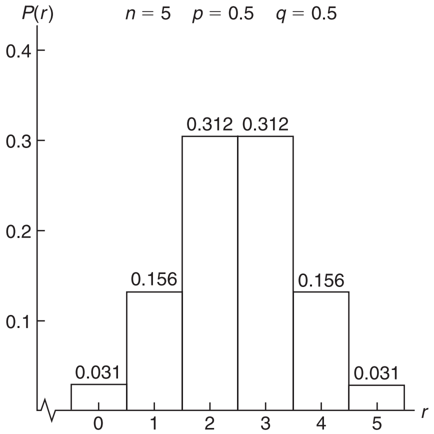
13. (a) 15/719; 704/719. (b) \$0.73; \$14.27.
15. (a) 0.01191; \$595.50. (b) \$646; \$698; \$751.50; \$806.50; \$3497.50 total. (c) \$4197.50. (d) \$1502.50.
17. (a) $\mu_W = 1.5$; $\sigma_W^2 = 208$; $\sigma_W \approx 14.4$.
    (b) $\mu_W = 107.5$; $\sigma_W^2 = 52$; $\sigma_W \approx 7.2$.
    (c) $\mu_L = 90$; $\sigma_L^2 = 92.16$; $\sigma_L \approx 9.6$.
    (d) $\mu_L = 90$; $\sigma_L^2 = 57.76$; $\sigma_L \approx 7.6$.
19. (a) $\mu_W = 50.2$; $\sigma_W^2 = 66.125$; $\sigma_W \approx 8.13$. (b) The means are the same. (c) The standard deviation for two policies is smaller. (d) As we include more policies, the coefficients in W decrease, resulting in smaller σ_W^2 and σ_W. For instance, for three policies, $W = (\mu_1 + \mu_2 + \mu_3)/3 \approx 0.33\mu_1 + 0.33\mu_2 + 0.33\mu_3$ and $\sigma_W^2 \approx (0.33)^2\sigma_1^2 + (0.33)^2\sigma_2^2 + (0.33)^2\sigma_3^2$. Yes, the risk appears to decrease by a factor of $1/\sqrt{n}$.

Section 5.2

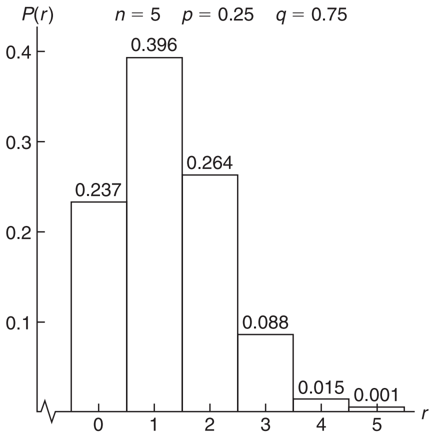
1. The random variable measures the number of successes out of n trials. This text uses the letter r for the random variable.
3. Two outcomes, success or failure.
5. Any monitor failure might endanger patient safely, so you should be concerned about the probability of *at least* one failure, not just exactly one failure.
7. (a) No. A binomial probability model applies to only two outcomes per trial. (b) Yes. Assign outcome A to "success" and outcomes B and C to "failure." $p = 0.40$.
9. (a) A trial consists of looking at the class status of a student enrolled in introductory statistics. Two outcomes are "freshman" and "not freshman." Success is freshman status; failure is any other class status. $P(\text{success}) = 0.40$. (b) Trials are not independent. With a population of only 30 students, in 5 trials without replacement, the probability of success rounded to the nearest hundredth changes for the later trials. Use the hypergeometric distribution for this situation.
11. (a) 0.082. (b) 0.918.
13. (a) 0.000. (b) Yes, the probability of 0 or 1 success is 0.000 to three places after the decimal. It would be a very rare event to get fewer than 2 successes when the probability of success on a single trial is so high.
15. A trial is one flip of a fair quarter. Success = coin shows heads. Failure = coin shows tails. $n = 3$; $p = 0.5$; $q = 0.5$. (a) $P(r = 3 \text{ heads}) = C_{3,3}p^3q^0 = 1(0.5)^3(0.5)^0 = 0.125$. To find this value in Table 3 of Appendix II, use the group in which $n = 3$, the column headed by $p = 0.5$, and the row headed by $r = 3$. (b) $P(r = 2 \text{ heads}) = C_{3,2}p^2q^1 = 3(0.5)^2(0.5)^1 = 0.375$. To find this value in Table 3 of Appendix II, use the group in which $n = 3$, the column headed by $p = 0.5$, and the row headed by $r = 2$. (c) $P(r \text{ is 2 or more}) = P(r = 2 \text{ heads}) + P(r = 3 \text{ heads}) = 0.375 + 0.125 = 0.500$. (d) The probability of getting three tails when you toss a coin three times is the same as getting zero heads. Therefore, $P(3 \text{ tails}) = P(r = 0 \text{ heads}) = C_{3,0}p^0q^3 = 1(0.5)^0(0.5)^3 = 0.125$. To find this value in Table 3 of Appendix II, use the group in which $n = 3$, the column headed by $p = 0.5$, and the row headed by $r = 0$.
17. A trial is recording the gender of one wolf. Success = male. Failure = female. $n = 12$; $p = 0.55$; $q = 0.45$. (a) $P(r \geq 6) = 0.740$. Six or more females means $12 - 6 = 6$ or fewer males; $P(r \leq 6) = 0.473$. Fewer than four females means more than $12 - 4 = 8$ males; $P(r > 8) = 0.135$. (b) A trial is recording the gender of one wolf. Success = male. Failure = female. $n = 12$; $p = 0.70$; $q = 0.30$. $P(r \geq 6) = 0.961$; $P(r \leq 6) = 0.117$; $P(r > 8) = 0.493$.
19. A trial consists of a woman's response regarding her mother-in-law. Success = dislike. Failure = like. $n = 6$; $p = 0.90$; $q = 0.10$. (a) $P(r = 6) = 0.531$. (b) $P(r = 0) = 0.000$ (to three digits). (c) $P(r \geq 4) = P(r = 4) + P(r = 5) + P(r = 6) = 0.098 + 0.354 + 0.531 = 0.983$. (d) $P(r \leq 3) = 1 - P(r \geq 4) \approx 1 - 0.983 = 0.017$ or 0.016 directly from table.
21. A trial is taking a polygraph exam. Success = pass. Failure = fail. $n = 9$; $p = 0.85$; $q = 0.15$. (a) $P(r = 9) = 0.232$. (b) $P(r \geq 5) = P(r = 5) + P(r = 6) + P(r = 7) + P(r = 8) + P(r = 9) = 0.028 + 0.107 + 0.260 + 0.368 + 0.232 = 0.995$. (c) $P(r \leq 4) = 1 - P(r \geq 5) \approx 1 - 0.995 = 0.005$ or 0.006 directly from table. (d) $P(r = 0) = 0.000$ (to three digits).
23. (a) A trial consists of using the Myers–Briggs instrument to determine if a person in marketing is an extrovert. Success = extrovert. Failure = not extrovert. $n = 15$; $p = 0.75$; $q = 0.25$. $P(r \geq 10) = 0.851$; $P(r \geq 5) = 0.999$; $P(r = 15) = 0.013$. (b) A trial consists of using the Myers–Briggs instrument to determine if a computer programmer is an introvert. Success = introvert. Failure = not introvert. $n = 5$; $p = 0.60$; $q = 0.40$. $P(r = 0) = 0.010$; $P(r \geq 3) = 0.683$; $P(r = 5) = 0.078$.
25. $n = 8$; $p = 0.53$; $q = 0.47$. (a) 0.812515; yes, truncated at five digits. (b) 0.187486; 0.18749; yes, rounded to five digits.
27. (a) They are the same. (b) They are the same. (c) $r = 1$. (d) The column headed by $p = 0.80$.
29. (a) $n = 8$; $p = 0.65$; $P(6 \leq r \mid 4 \leq r) = P(6 \leq r)/P(4 \leq r) = 0.428/0.895 \approx 0.478$. (b) $n = 10$; $p = 0.65$; $P(8 \leq r \mid 6 \leq r) = P(8 \leq r)/P(6 \leq r) = 0.262/0.752 \approx 0.348$. (c) Essay. (d) Use event $A = 6 \leq r$ and event $B = 4 \leq r$ in the formula.

Section 5.3

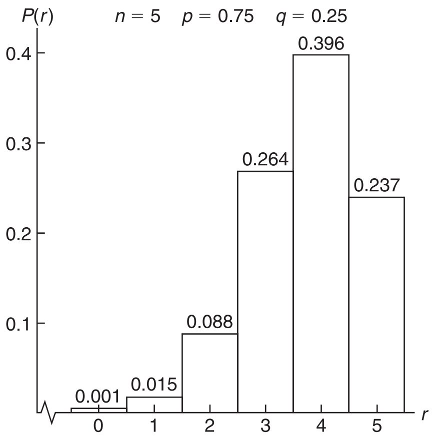
1. The average number of successes.
3. (a) $\mu = 1.6$; $\sigma \approx 1.13$. (b) Yes, 5 successes is more than 2.5σ above the expected value. $P(r \geq 5) = 0.010$.
5. (a) Yes, 120 is more than 2.5 standard deviations above the expected value. (b) Yes, 40 is less than 2.5 standard deviations below the expected value. (c) No, 70 to 90 successes is within 2.5 standard deviations of the expected value.
7. (a) Binomial Distribution
   The distribution is symmetrical.



(b) Binomial Distribution
The distribution is skewed right.

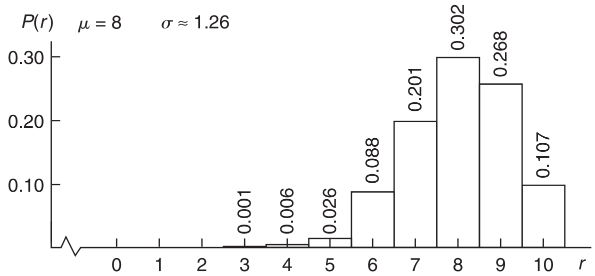


(c) Binomial Distribution
The distribution is skewed left.

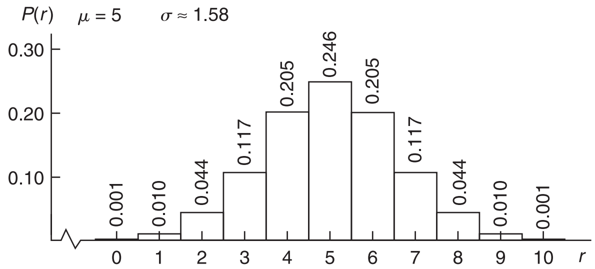


(d) The distributions are mirror images of one another.
(e) The distribution would be skewed left for $p = 0.73$ because the more likely numbers of successes are to the right of the middle.

9. (a) Households with Children Under 2 That Buy Photo Gear

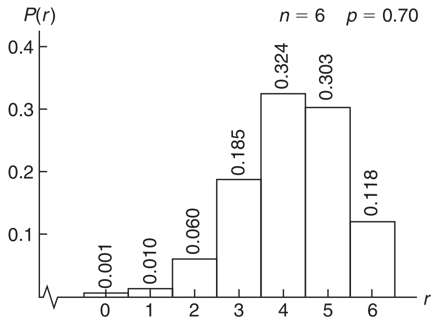


(b) Households with No Children Under 21 That Buy Photo Gear



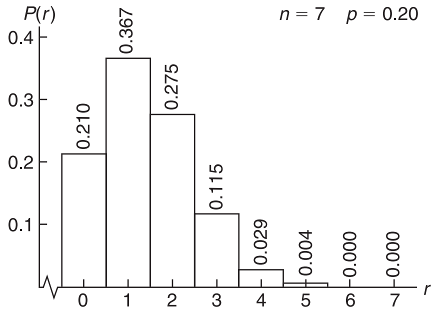
(c) Yes. Adults with children seem to buy more photo gear.

11. (a) Binomial Distribution for Number of Addresses Found



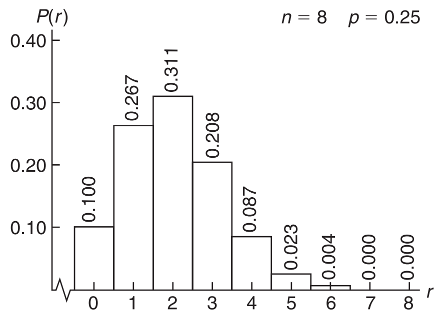
(b) $\mu = 4.2$; $\sigma \approx 1.122$. (c) $n = 5$. Note that $n = 5$ gives $P(r \geq 2) = 0.97$.

13. (a) Binomial Distribution for Number of Illiterate People



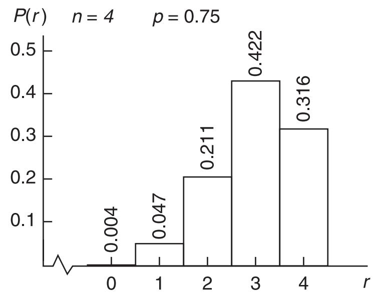
(b) $\mu = 1.4$; $\sigma \approx 1.058$. (c) $n = 12$. Note that $n = 12$ gives $P(r \geq 7) = 0.98$, where success = literate and $p = 0.80$.

15. (a) Binomial Distribution for Number of Gullible Consumers



(b) $\mu = 2$; $\sigma \approx 1.225$. (c) $n = 16$. Note that $n = 16$ gives $P(r \geq 1) = 0.99$.

17. (a) $P(r = 0) = 0.004$; $P(r = 1) = 0.047$; $P(r = 2) = 0.211$; $P(r = 3) = 0.422$; $P(r = 4) = 0.316$.
(b) Binomial Distribution for Number of Parolees Who Do Not Become Repeat Offenders



(c) $\mu = 3$; $\sigma \approx 0.866$. (d) $n = 7$. Note that $n = 7$ gives $P(r \geq 3) = 0.987$.

19. $n = 12$; $p = 0.25$ do not serve; $p = 0.75$ serve.
(a) $P(r = 12 \text{ serve}) = 0.032$. (b) $P(r \geq 6 \text{ do not serve}) = 0.053$. (c) For serving, $\mu = 9$; $\sigma = 1.50$. (d) To be at least 95.9% sure that 12 are available to serve, call 20.

21. $n = 6$; $p = 0.80$ do not solve; $p = 0.20$ solve.
(a) $P(r = 6 \text{ not solved}) = 0.262$. (b) $P(r \geq 1 \text{ solved}) = 0.738$. (c) For solving crime, $\mu = 1.2$; $\sigma \approx 0.98$.
(d) To be 90% sure of solving one or more crimes, investigate $n = 11$ crimes.

23. (a) $P(r = 7 \text{ guilty in U.S.}) = 0.028$; $P(r = 7 \text{ guilty in Japan}) = 0.698$. (b) For guilty in Japan, $\mu = 6.65$; $\sigma \approx 0.58$; for guilty in U.S., $\mu = 4.2$; $\sigma \approx 1.30$. (c) To be 99% sure of at least two guilty convictions in the U.S., look at $n = 8$ trials. To be 99% sure of at least two guilty convictions in Japan, look at $n = 3$ trials.

25. (a) 9. (b) 10.

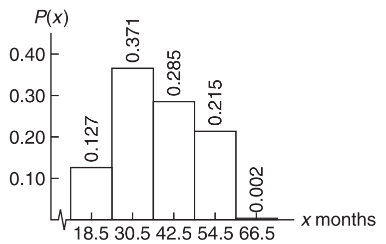
Section 5.4

1. Geometric distribution.
3. No, $n = 50$ is not large enough.
5. 0.144.
7. $\lambda = 8$; 0.1396.
9. (a) $p = 0.77$; $P(n) = (0.77)(0.23)^{n-1}$. (b) $P(1) = 0.77$. (c) $P(2) = 0.1771$. (d) $P(3 \text{ or more tries}) = 1 - P(1) - P(2) = 0.0529$. (e) 1.29, or 1.
11. (a) $P(n) = (0.80)(0.20)^{n-1}$. (b) $P(1) = 0.8$; $P(2) = 0.16$; $P(3) = 0.032$. (c) $P(n \geq 4) = 1 - P(1) - P(2) - P(3) = 1 - 0.8 - 0.16 - 0.032 = 0.008$. (d) $P(n) = (0.04)(0.96)^{n-1}$; $P(1) = 0.04$; $P(2) = 0.0384$; $P(3) = 0.0369$; $P(n \geq 4) = 0.8847$.
13. (a) $P(n) = (0.30)(0.70)^{n-1}$. (b) $P(3) = 0.147$. (c) $P(n > 3) = 1 - P(1) - P(2) - P(3) = 1 - 0.300 - 0.210 - 0.147 = 0.343$. (d) 3.33, or 3.
15. (a) $\lambda = (1.7/10) \times (3/3) = 5.1$ per 30-minute interval; $P(r) = e^{-5.1}(5.1)^r/r!$. (b) Using Table 4 of Appendix II with $\lambda = 5.1$, we find $P(4) = 0.1719$; $P(5) = 0.1753$; $P(6) = 0.1490$. (c) $P(r \geq 4) = 1 - P(0) - P(1) - P(2) - P(3) = 1 - 0.0061 - 0.0311 - 0.0793 - 0.1348 = 0.7487$. (d) $P(r < 4) = 1 - P(r \geq 4) = 1 - 0.7487 = 0.2513$.
17. (a) Births and deaths occur somewhat rarely in a group of 1000 people in a given year. For 1000 people, $\lambda = 16$ births; $\lambda = 8$ deaths. (b) By Table 4 of Appendix II, $P(10 \text{ births}) = 0.0341$; $P(10 \text{ deaths}) = 0.0993$; $P(16 \text{ births}) = 0.0992$; $P(16 \text{ deaths}) = 0.0045$. (c) $\lambda(\text{births}) = (16/1000) \times (1500/1500) = 24$ per 1500 people. $\lambda(\text{deaths}) = (8/1000) \times (1500/1500) = 12$ per 1500 people. By the table, $P(10 \text{ deaths}) = 0.1048$; $P(16 \text{ deaths}) = 0.0543$. Since $\lambda = 24$ is not in the table, use the formula for $P(r)$ to find $P(10 \text{ births}) = 0.00066$; $P(16 \text{ births}) = 0.02186$. (d) $\lambda(\text{births}) = (16/1000) \times (750/750) = 12$ per 750 people. $\lambda(\text{deaths}) = (8/1000) \times (750/750) = 6$ per 750 people. By Table 4 of Appendix II, $P(10 \text{ births}) = 0.1048$; $P(10 \text{ deaths}) = 0.0413$; $P(16 \text{ births}) = 0.0543$; $P(16 \text{ deaths}) = 0.0003$.
19. (a) The Poisson distribution is a good choice for r because gale-force winds occur rather rarely. The occurrences are usually independent. (b) For interval of 108 hours, $\lambda = (1/60) \times (108/108) = 1.8$ per 108 hours. Using Table 4 of Appendix II, we find that $P(2) = 0.2678$; $P(3) = 0.1607$; $P(4) = 0.0723$; $P(r < 2) = P(0) + P(1) = 0.1653 + 0.2975 = 0.4628$. (c) For interval of 180 hours, $\lambda = (1/60) \times (180/180) = 3$ per 180 hours. Table 4 of Appendix II gives $P(3) = 0.2240$; $P(4) = 0.1680$; $P(5) = 0.1008$; $P(r < 3) = P(0) + P(1) + P(2) = 0.0498 + 0.1494 + 0.2240 = 0.4232$.
21. (a) The sales of large buildings are rare events. It is reasonable to assume that they are independent. The variable r = number of sales in a fixed time interval. (b) For a 60-day period, $\lambda = (8/275) \times (60/60) = 1.7$ per 60 days. By Table 4 of Appendix II, $P(0) = 0.1827$; $P(1) = 0.3106$; $P(r \geq 2) = 1 - P(0) - P(1) = 0.5067$. (c) For a 90-day period, $\lambda = (8/275) \times (90/90) = 2.6$ per 90 days. By Table 4 of Appendix II, $P(0) = 0.0743$; $P(2) = 0.2510$; $P(r \geq 3) = 1 - P(0) - P(1) - P(2) = 1 - 0.0743 - 0.1931 - 0.2510 = 0.4816$.
23. (a) The problem satisfies the conditions for a binomial experiment with small $p = 0.0018$ and large $n = 1000$. $np = 1.8$, which is less than 10, so the Poisson approximation to the binomial distribution would be a good choice. $\lambda = np = 1.8$. (b) By Table 4, Appendix II, $P(0) = 0.1653$. (c) $P(r > 1) = 1 - P(0) - P(1) = 1 - 0.1653 - 0.2975 = 0.5372$. (d) $P(r > 2) = 1 - P(0) - P(1) - P(2) = 1 - 0.1653 - 0.2975 - 0.2678 = 0.2694$. (e) $P(r > 3) = 1 - P(0) - P(1) - P(2) - P(3) = 1 - 0.1653 - 0.2975 - 0.2678 - 0.1607 = 0.1087$.

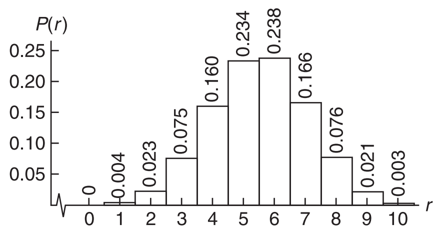
25. (a) The problem satisfies the conditions for a binomial experiment with n large, $n = 175$, and p small. $np = (175)(0.005) = 0.875 < 10$. The Poisson distribution would be a good approximation to the binomial. $n = 175$; $p = 0.005$; $\lambda = np = 0.9$. (b) By Table 4 of Appendix II, $P(0) = 0.4066$. (c) $P(r \geq 1) = 1 - P(0) = 0.5934$. (d) $P(r \geq 2) = 1 - P(0) - P(1) = 0.2275$.
27. (a) $n = 100$; $p = 0.02$; $r = 2$; $P(2) = C_{100,2}(0.02)^2(0.98)^{98} \approx 0.2734$. (b) $\lambda = np = 2$; $P(2) = [e^{-2}(2)^2]/2! \approx 0.2707$. (c) The approximation is correct to two decimal places. (d) $n = 100$; $p = 0.02$; $r = 3$. By the formula for the binomial distribution, $P(3) \approx 0.1823$. By the Poisson approximation, $P(3) \approx 0.1804$. The approximation is correct to two decimal places.
29. (a) $\lambda \approx 3.4$. (b) $P(r \geq 4 \mid r \geq 2) = P(r \geq 4)/P(r \geq 2) \approx 0.4416/0.8531 \approx 0.5176$. (c) $P(r < 6 \mid r \geq 3) = P(3 \leq r < 6)/P(r \geq 3) \approx 0.5308/0.6602 \approx 0.8040$.
31. (a) $P(n) = C_{n-1,\,11}(0.80^{12})(0.20^{n-12})$. (b) $P(12) \approx 0.0687$; $P(13) \approx 0.1649$; $P(14) \approx 0.2144$. (c) 0.4480. (d) 0.5520. (e) $\mu = 15$; $\sigma \approx 1.94$. Susan can expect to get the bonus if she makes 15 contacts, with a standard deviation of about 2 contacts.

Chapter 5 Review

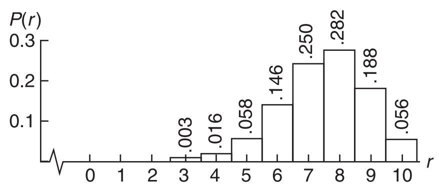
1. A description of all distinct possible values of a random variable x, with a probability assignment $P(x)$ for each value or range of values. $0 \leq P(x) \leq 1$ and $\Sigma P(x) = 1$.
3. (a) Yes. $\mu = 2$ and $\sigma \approx 1.3$. Numbers of successes above 5.25 are unusual. (b) No. It would be unusual to get more than five questions correct.
5. (a) 38; 11.6.
   (b) Duration of Leases in Months



7. (a) Number of Claimants Under 25



   (b) $P(r \geq 6) = 0.504$. (c) $\mu = 5.5$; $\sigma \approx 1.57$.
9. (a) 0.039. (b) 0.403. (c) 8.
11. (a) Number of Good Grapefruit

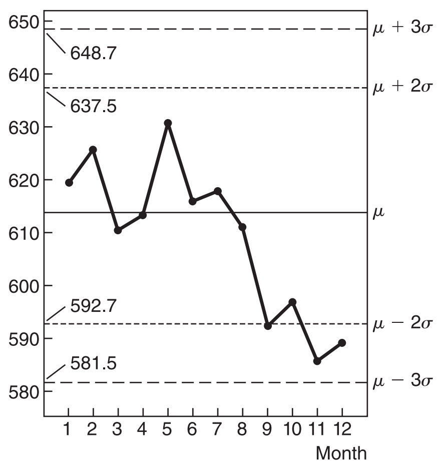


   (b) 0.244, 0.999. (c) 7.5. (d) 1.37.
13. $P(r \leq 2) = 0.000$ (to three digits). The data seem to indicate that the percent favoring the increase in fees is less than 85%.
15. (a) Coughs are a relatively rare occurrence. It is reasonable to assume that they are independent events, and the variable is the number of coughs in a fixed time interval. (b) $\lambda = 11$ coughs per minute; $P(r \leq 3) = P(0) + P(1) + P(2) + P(3) = 0.000 + 0.002 + 0.0010 + 0.0037 = 0.0049$. (c) $\lambda = (11/1) \times (0.5/0.5) = 5.5$ coughs per 30-second period. $P(r \geq 3) = 1 - P(0) - P(1) - P(2) = 1 - 0.0041 - 0.0225 - 0.0618 = 0.9116$.
17. The loan-default problem satisfies the conditions for a binomial experiment. Moreover, p is small, n is large, and $np < 10$. Use of the Poisson approximation to the binomial distribution is appropriate. $n = 300$; $p = 1/350 \approx 0.0029$; and $\lambda = np \approx 300(0.0029) = 0.86 \approx 0.9$; $P(r \geq 2) = 1 - P(0) - P(1) = 1 - 0.4066 - 0.3659 = 0.2275$.
19. (a) Use the geometric distribution with $p = 0.5$. $P(n = 2) = (0.5)(0.5) = 0.25$. As long as you toss the coin at least twice, it does not matter how many more times you toss it. To get the first head on the second toss, you must get a tail on the first and a head on the second. (b) $P(n = 4) = (0.5)(0.5)^3 = 0.0625$; $P(n > 4) = 1 - P(1) - P(2) - P(3) - P(4) = 1 - 0.5 - 0.5^2 - 0.5^3 - 0.5^4 = 0.0625$.

CHAPTER 6

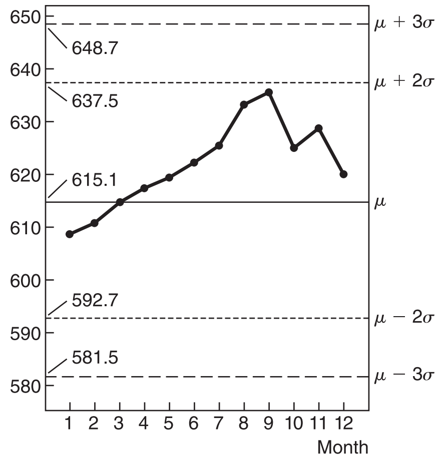
Section 6.1

1. (a) No, it's skewed. (b) No, it crosses the horizontal axis. (c) No, it has three peaks. (d) No, the curve is not smooth.
3. Figure 6-12 has the larger standard deviation. The mean of Figure 6-12 is $\mu = 10$. The mean of Figure 6-13 is $\mu = 4$.
5. (a) 50%. (b) 68%. (c) 99.7%.
7. (a) 50%. (b) 50%. (c) 68%. (d) 95%.
9. (a) From 1207 to 1279. (b) From 1171 to 1315. (c) From 1135 to 1351.
11. (a) From 1.70 mA to 4.60 mA. (b) From 0.25 mA to 6.05 mA.
13. (a) Tri-County Bank Monthly Loan Request—First Year (thousands of dollars)



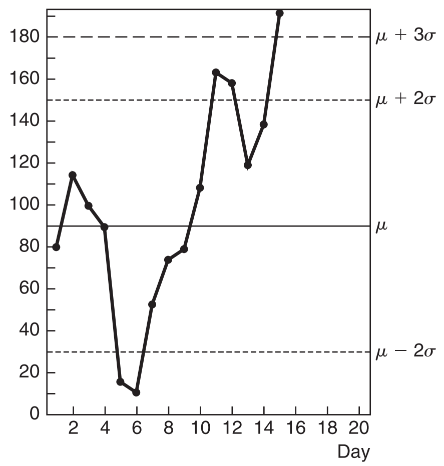
The process is out of control with a type III warning signal, since two of three consecutive points are more than 2 standard deviations below the mean. The trend is down.

(b) Tri-County Bank Monthly Loan Requests—Second Year (thousands of dollars)



The process shows warning signal II, a run of nine consecutive points above the mean. The economy is probably heating up.

15. Visibility Standard Index



There is one point above $\mu + 3\sigma$. Thus control signal I indicates "out of control." Control signal III is present. There are two consecutive points below $\mu - 2\sigma$ and two consecutive points above $\mu + 2\sigma$. The out-of-control signals that cause the most concern are those above the mean. Special pollution regulations may be appropriate for those periods.

17. (a) 0.8000. (b) 0.7000. (c) 0.5000. (d) $\mu = 0$; $\sigma \approx 0.289$. Since $\sigma = 0$, the measurements are unbiased.
19. (a) 0.4493. (b) 0.8454. (c) 0.1857. (d) 120.71.

Section 6.2

1. The number of standard deviations from the mean.
3. 0.
5. (a) −1. (b) 2.4. (c) 20. (d) 36.5.
7. They are the same, since both are 1 standard deviation below the mean.
9. (a) Robert, Juan, and Linda each scored above the mean. (b) Joel scored on the mean. (c) Susan and Jan scored below the mean. (d) Robert, 172; Juan, 184; Susan, 110; Joel, 150; Jan, 134; Linda, 182.
11. (a) $-1.00 < z$. (b) $z < -2.00$. (c) $-2.67 < z < 2.33$. (d) $x < 4.4$. (e) $5.2 < x$. (f) $4.1 < x < 4.5$. (g) A red blood cell count of 5.9 or higher corresponds to a standard z score of 3.67. Practically no data values occur this far above the mean. Such a count would be considered unusually high for a healthy female.
13. 0.5000. 15. 0.0934. 17. 0.6736. 19. 0.0643. 21. 0.8888. 23. 0.4993. 25. 0.8953. 27. 0.3471. 29. 0.0306. 31. 0.5000. 33. 0.4483. 35. 0.8849. 37. 0.0885. 39. 0.8849. 41. 0.8808. 43. 0.3226. 45. 0.4474. 47. 0.2939. 49. 0.6704.

Section 6.3

1. 0.50.
3. Negative.
5. $P(3 \le x \le 6) = P(-0.50 \le z \le 1.00) = 0.5328$.
7. $P(50 \le x \le 70) = P(0.67 \le z \le 2.00) = 0.2286$.
9. $P(8 \le x \le 12) = P(-2.19 \le z \le -0.94) = 0.1593$.
11. $P(x \ge 30) = P(z \ge 2.94) = 0.0016$.
13. $P(x \ge 90) = P(z \ge -0.67) = 0.7486$.
15. −1.555. 17. 0.13. 19. 1.41. 21. −0.92.
23. ±2.33.
25. (a) $P(x > 60) = P(z > -1) = 0.8413$. (b) $P(x < 110) = P(z < 1) = 0.8413$. (c) $P(60 \le x \le 110) = P(-1.00 \le z \le 1.00) = 0.8413 - 0.1587 = 0.6826$. (d) $P(x > 140) = P(z > 2.20) = 0.0139$.
27. (a) $P(x < 3.0 \text{ mm}) = P(z < -2.33) = 0.0099$. (b) $P(x > 7.0 \text{ mm}) = P(z > 2.11) = 0.0174$. (c) $P(3.0 \text{ mm} < x < 7.0 \text{ mm}) = P(-2.33 < z < 2.11) = 0.9727$.
29. (a) $P(x < 36 \text{ months}) = P(z < -1.13) = 0.1292$. The company will replace 13% of its batteries. (b) $P(z < z_0) = 10\%$ for $z_0 = -1.28$; $x = -1.28(8) + 45 = 34.76$. Guarantee the batteries for 35 months.
31. (a) According to the empirical rule, about 95% of the data lies between $\mu - 2\sigma$ and $\mu + 2\sigma$. Since this interval is 4σ wide, we have $4\sigma \approx 6$ years, so $\sigma \approx 1.5$ years. (b) $P(x > 5) = P(z > -2.00) = 0.9772$. (c) $P(x < 10) = P(z < 1.33) = 0.9082$. (d) $P(z < z_0) = 0.10$ for $z_0 = -1.28$; $x = -1.28(1.5) + 8 = 6.08$ years. Guarantee the TVs for about 6.1 years.
33. (a) $\sigma \approx 12$ beats/minute. (b) $P(x < 25) = P(z < -1.75) = 0.0401$. (c) $P(x > 60) = P(z > 1.17) = 0.1210$. (d) $P(25 \le x \le 60) = P(-1.75 \le z \le 1.17) = 0.8389$. (e) $P(z \le z_0) = 0.90$ for $z_0 = 1.28$; $x = 1.28(12) + 46 = 61.36$ beats/minute. A heart rate of 61 beats/minute corresponds to the 90% cutoff point of the distribution.
35. (a) $P(z \ge z_0) = 0.99$ for $z_0 = -2.33$; $x = -2.33(3.7) + 90 \approx 81.38$ months. Guarantee the microchips for 81 months. (b) $P(x \le 84) = P(z \le -1.62) = 0.0526$. (c) Expected loss = (50,000,000)(0.0526) = \$2,630,000. (d) Profit = \$370,000.
37. (a) $z = 1.28$; $x \approx 4.9$ hours. (b) $z = -1.04$; $x \approx 2.9$ hours. (c) Yes; work and/or school schedules may be different on Saturday.
39. (a) In general, $P(A \mid B) = P(A \text{ and } B)/P(B)$; $P(x > 20) = P(z > 0.50) = 0.3085$; $P(x > 15) = P(z > -0.75) = 0.7734$; $P(x > 20 \mid x > 15) = 0.3989$. (b) $P(x > 25) =$

$P(z > 1.75) = 0.0401$; $P(x > 18) = P(z > 0.00) = 0.5000$; $P(x > 25 \mid x > 18) = 0.0802$. (c) Use event $A = x > 20$ and event $B = x > 15$ in the formula.

Section 6.4

1. A set of measurements or counts either existing or conceptual. For example, the population of ages of all people in Colorado; the population of weights of all students in your school; the population count of all antelope in Wyoming.
3. A numerical descriptive measure of a population, such as μ, the population mean; σ, the population standard deviation; or σ^2, the population variance.
5. A statistical inference is a conclusion about the value of a population parameter. We will do both estimation and testing.
7. They help us visualize the sampling distribution through tables and graphs that approximately represent the sampling distribution.
9. We studied the sampling distribution of mean trout lengths based on samples of size 5. Other such sampling distributions abound.

Section 6.5

Note: Answers may differ slightly depending on the number of digits carried in the standard deviation.

1. The standard deviation.
3. $\bar{x}$ is an unbiased estimator for μ; $\hat{p}$ is an unbiased estimator for p.
5. (a) Normal; $\mu_{\bar{x}} = 8$; $\sigma_{\bar{x}} = 2$. (b) 0.50. (c) 0.3085. (d) No, about 30% of all such samples have means exceeding 9.
7. (a) 30 or more. (b) No.
9. The second. The standard error of the first is $\sigma/10$, while that of the second is $\sigma/15$, where σ is the standard deviation of the original x distribution.
11. (a) $\mu_{\bar{x}} = 15$; $\sigma_{\bar{x}} = 2.0$; $P(15 \le \bar{x} \le 17) = P(0 \le z \le 1.00) = 0.3413$. (b) $\mu_{\bar{x}} = 15$; $\sigma_{\bar{x}} = 1.75$; $P(15 \le \bar{x} \le 17) = P(0 \le z \le 1.14) = 0.3729$. (c) The standard deviation is smaller in part (b) because of the larger sample size. Therefore, the distribution about $\mu_{\bar{x}}$ is narrower in part (b).
13. (a) $P(x < 74.5) = P(z < -0.63) = 0.2643$. (b) $P(\bar{x} < 74.5) = P(z < -2.79) = 0.0026$. (c) No. If the weight of coal in only one car were less than 74.5 tons, we could not conclude that the loader is out of adjustment. If the mean weight of coal for a sample of 20 cars were less than 74.5 tons, we would suspect that the loader is malfunctioning. As we see in part (b), the probability of this happening is very low if the loader is correctly adjusted.
15. (a) $P(x < 40) = P(z < -1.80) = 0.0359$. (b) Since the x distribution is approximately normal, the $\bar{x}$ distribution is approximately normal, with mean 85 and standard deviation 17.678. $P(\bar{x} < 40) = P(z < -2.55) = 0.0054$. (c) $P(\bar{x} < 40) = P(z < -3.12) = 0.0009$. (d) $P(\bar{x} < 40) = P(z < -4.02) < 0.0002$. (e) Yes; if the average value based on five tests were less than 40, the patient is almost certain to have excess insulin.
17. (a) $P(x < 54) = P(z < -1.27) = 0.1020$. (b) The expected number undernourished is 2200(0.1020), or about 224. (c) $P(\bar{x} \le 60) = P(z \le -2.99) = 0.0014$. (d) $P(\bar{x} < 64.2) = P(z < 1.20) = 0.8849$. Since the sample average is above the mean, it is quite unlikely that the doe population is undernourished.
19. (a) Since x itself represents a sample mean return based on a large (random) sample of stocks, x has a distribution that is approximately normal (central limit theorem). (b) $P(1\% \le \bar{x} \le 2\%) = P(-1.63 \le z \le 1.09) = 0.8105$. (c) $P(1\% \le \bar{x} \le 2\%) = P(-3.27 \le z \le 2.18) = 0.9849$. (d) Yes. The standard deviation decreases as the sample size increases. (e) $P(\bar{x} < 1\%) = P(z < -3.27) = 0.0005$. This is very unlikely if $\mu = 1.6\%$. One would suspect that μ has slipped below 1.6%.
21. (a) The total checkout time for 30 customers is the sum of the checkout times for each individual customer. Thus, $w = x_1 + x_2 + \cdots + x_{30}$, and the probability that the total checkout time for the next 30 customers is less than 90 is $P(w < 90)$. (b) $w < 90$ is equivalent to $x_1 + x_2 + \cdots + x_{30} < 90$. Divide both sides by 30 to get $\bar{x} < 3$ for samples of size 30. Therefore, $P(w < 90) = P(\bar{x} < 3)$. (c) By the central limit theorem, $\bar{x}$ is approximately normal, with $\mu_{\bar{x}} = 2.7$ minutes and $\sigma_{\bar{x}} = 0.1095$ minute. (d) $P(\bar{x} < 3) = P(z < 2.74) = 0.9969$.
23. (a) $P(w > 90) = P(\bar{x} > 18) = P(z > 0.68) = 0.2483$. (b) $P(w < 80) = P(\bar{x} < 16) = P(z < -0.68) = 0.2483$. (c) $P(80 < w < 90) = P(16 < \bar{x} < 18) = P(-0.68 < z < 0.68) = 0.5034$.

Section 6.6

1. $np > 5$ and $nq > 5$, where $q = 1 - p$.
3. (a) Yes, both $np > 5$ and $nq > 5$. (b) $\mu = 20$; $\sigma \approx 3.162$. (c) $r \ge 23$ corresponds to $x \ge 22.5$. (d) $P(r \ge 23) \approx P(x \ge 22.5) \approx P(z \ge 0.79) \approx 0.2148$. (e) No, the probability that this will occur is about 21%.
5. No, $np = 4.3$ and does not satisfy the criterion that $np > 5$.

Note: Answers may differ slightly depending on how many digits are carried in the computation of the standard deviation and z.

7. $np > 5$; $nq > 5$. (a) $P(r \ge 50) = P(x \ge 49.5) = P(z \ge -27.53) \approx 1$, or almost certain. (b) $P(r \ge 50) = P(x \ge 49.5) = P(z \ge 7.78) \approx 0$, or almost impossible for a random sample.
9. $np > 5$; $nq > 5$. (a) $P(r \ge 15) = P(x \ge 14.5) = P(z \ge -2.35) = 0.9906$. (b) $P(r \ge 30) = P(x \ge 29.5) = P(z \ge 0.62) = 0.2676$. (c) $P(25 \le r \le 35) + P(24.5 \le x \le 35.5) = P(-0.37 \le z \le 1.81) = 0.6092$. (d) $P(r > 40) = P(r \ge 41) = P(x \ge 40.5) = P(z \ge 2.80) = 0.0026$.
11. $np > 5$; $nq > 5$. (a) $P(r \ge 47) = P(x \ge 46.5) = P(z \ge -1.94) = 0.9738$. (b) $P(r \le 58) = P(x \le 58.5) = P(z \le 1.75) = 0.9599$. In parts (c) and (d), let r be the number of products that succeed, and use $p = 1 - 0.80 = 0.20$. (c) $P(r \ge 15) = P(x \ge 14.5) = P(z \ge 0.40) = 0.3446$. (d) $P(r < 10) = P(r \le 9) = P(x \le 9.5) = P(z \le -1.14) = 0.1271$.
13. $np > 5$; $nq > 5$. (a) $P(r > 180) = P(x \ge 180.5) = P(z > -1.11) = 0.8665$. (b) $P(r < 200) = P(x \le 199.5) = P(z \le 1.07) = 0.8577$. (c) P(take sample and

buy product) = P(take sample) · P(buy | take sample) = 0.222. (d) $P(60 \le r \le 80) = P(59.5 \le x \le 80.5) = P(-1.47 \le z \le 1.37) = 0.8439$.

15. $np > 5$; $nq > 5$. (a) 0.94. (b) $P(r \le 255)$. (c) $P(r \le 255) = P(x \le 255.5) = P(z \le 1.16) = 0.8770$.
17. $np > 5$ and $nq > 5$.
19. Yes, since the mean of the approximate sampling distribution is $\mu_{\hat{p}} = p$.
21. (a) Yes, both np and nq exceed 5. $\mu_{\hat{p}} = 0.23$; $\sigma_{\hat{p}} \approx 0.042$. (b) No, $np = 4.6$ and does not exceed 5.

Chapter 6 Review

1. Normal probability distributions are distributions of continuous random variables. They are symmetric about the mean and bell-shaped. Most of the data fall within 3 standard deviations of the mean. The mean and median are the same.
3. No.
5. The points lie close to a straight line.
7. $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.
9. (a) A normal distribution. (b) The mean μ of the x distribution. (c) $\sigma/\sqrt{n}$, where σ is the standard deviation of the x distribution. (d) They will both be approximately normal with the same mean, but the standard deviations will be $\sigma/\sqrt{50}$ and $\sigma/\sqrt{100}$, respectively.
11. (a) 0.9821. (b) 0.3156. (c) 0.2977.
13. 1.645.
15. (a) 0.8665. (b) 0.7330.
17. (a) 0.0166. (b) 0.975.
19. (a) 0.9772. (b) 17.3 hours.
21. (a) $P(x \ge 40) = P(z \ge 0.71) = 0.2389$. (b) $P(\bar{x} \ge 40) = P(z \ge 2.14) = 0.0162$.
23. $P(98 \le \bar{x} \le 102) = P(-1.33 \le z \le 1.33) = 0.8164$.
25. (a) Yes, np and nq both exceed 5. (b) $\mu_{\hat{p}} = 0.4$; $\sigma_{\hat{p}} = 0.1$.

CUMULATIVE REVIEW PROBLEMS

1. The specified ranges of readings are disjoint and cover all possible readings.
2. Essay.
3. Yes; the events constitute the entire sample space.
4. (a) 0.85. (b) 0.70. (c) 0.70. (d) 0.30. (e) 0.15. (f) 0.75. (g) 0.30. (h) 0.05.
5. 0.17
6.

x	$P(x)$
5	0.25
15	0.45
25	0.15
35	0.10
45	0.05

$\mu \approx 17.5$; $\sigma \approx 10.9$.

7. (a) $p = 0.10$. (b) $\mu = 1.2$; $\sigma \approx 1.04$. (c) 0.718. (d) 0.889.
8. (a) 0.05. (b) $P(n) = (0.05)(0.95)^{n-1}$; $n \ge 1$. (c) 0.81.
9. (a) Yes; since $n = 100$ and $np = 5$, the criteria $n \ge 100$ and $np < 10$ are satisfied. $\lambda = 5$. (b) 0.7622. (c) 0.0680.
10. (a) Yes; both np and nq exceed 5. (b) 0.9925. (c) np is too large ($np > 10$) and n is too small ($n < 100$).
11. (a) $\sigma \approx 1.7$. (b) 0.1314. (c) 0.1075.
12. Essay based on material from Chapter 6 and Section 1.2.
13. (a) Because of the large sample size, the central limit theorem describes the $\bar{x}$ distribution (approximately). (b) $P(\bar{x} \le 6820) = P(z \le -2.75) = 0.0030$. (c) The probability that the average white blood cell count for 50 healthy adults is as low as or lower than 6820 is very small, 0.0030. Based on this result, it would be reasonable to gather additional facts.
14. (a) Yes, both np and nq exceed 5. (b) $\mu_{\hat{p}} = p = 0.45$; $\sigma_{\hat{p}} \approx 0.09$.
15. Essay.

CHAPTER 7

Section 7.1

1. True. By definition, critical values z_c are values such that c% of the area under the normal curve falls between $-z_c$ and z_c.
3. True. By definition, the margin of error is the magnitude of the difference between $\bar{x}$ and μ.
5. False. The maximal margin of error is $E = z_c \dfrac{\sigma}{\sqrt{n}}$. As the sample size n increases, the maximal error decreases, resulting in a shorter confidence interval for μ.
7. False. The maximal error of estimate E controls the length of the confidence interval regardless of the value of $\bar{x}$.
9. μ is either in the interval 10.1 to 12.2 or not. Therefore, the probability that μ is in this interval is either 0 or 1, not 0.95.
11. (a) Yes, the x distribution is normal and σ is known so the $\bar{x}$ distribution is also normal. (b) 47.53 to 52.47. (c) You are 90% confident that the confidence interval computed is one that contains μ.
13. (a) 217. (b) Yes, by the central limit theorem.
15. (a) 3.04 gm to 3.26 gm; 0.11 gm. (b) Distribution of weights is normal with known σ. (c) There is an 80% chance that the confidence interval is one of the intervals that contain the population average weight of Allen's hummingbirds in this region. (d) $n = 28$.
17. (a) 34.62 ml/kg to 40.38 ml/kg; 2.88 ml/kg. (b) The sample size is large (30 or more) and σ is known. (c) There is a 99% chance that the confidence interval is one of the intervals that contain the population average blood plasma level for male firefighters. (d) $n = 60$.
19. (a) 125.7 to 151.3 larceny cases; 12.8 larceny cases. (b) 123.3 to 153.7 larceny cases; 15.2 larceny cases. (c) 118.4 to 158.6 larceny cases; 20.1 larceny cases. (d) Yes. (e) Yes.
21. (a) 26.64 to 33.36; 3.36. (b) 27.65 to 32.35; 2.35. (c) 28.43 to 31.57; 1.57. (d) Yes. (e) Yes.
23. (a) The mean rounds to the value given. (b) Using the rounded value of part (a), the 75% interval is from 34.19 thousand to 37.81 thousand. (c) Yes; 30 thousand dollars is below the lower bound of the 75% confidence interval. We can say with 75% confidence that the mean lies between 34.19 thousand and 37.81 thousand. (d) Yes; 40 thousand is above the upper bound of the 75% confidence interval. (e) 33.41 thousand to 38.59 thousand. We can say with 90% confidence that the mean

lies between 33.4 thousand and 38.6 thousand dollars. 30 thousand is below the lower bound and 40 thousand is above the upper bound.

25. (a) 92.5°C to 101.5°C. (b) The balloon will go up.

Section 7.2

1. 2.110.
3. 1.721.
5. $t = 0$.
7. $n = 10$, with $d.f. = 9$.
9. Shorter. For $d.f. = 40$, z_c is less than t_c, and the resulting margin of error E is smaller.
11. (a) Yes, since x has a mound-shaped distribution. (b) 9.12 to 10.88. (c) There is a 90% chance that the confidence interval you computed is one of the confidence intervals that contain μ.
13. (a) The mean and standard deviation round to the values given. (b) Using the rounded values for the mean and standard deviation given in part (a), the interval is from 1249 to 1295. (c) We are 90% confident that the computed interval is one that contains the population mean for the tree-ring date.
15. (a) Use a calculator. (b) 74.7 pounds to 107.3 pounds. (c) We are 75% confident that the computed interval is one that contains the population mean weight of adult mountain lions in the region.
17. (a) The mean and standard deviation round to the given values. (b) 8.41 to 11.49. (c) Since all values in the 99.9% confidence interval are above 6, we can be almost certain that this patient no longer has a calcium deficiency.
19. (a) Boxplots differ in length of interquartile box, location of median, and length of whiskers. The boxplots come from different samples. (b) Yes; no; for 95% confidence intervals, we expect about 95% of the samples to generate intervals that contain the mean of the population.
21. (a) The mean and standard deviation round to the given values. (b) 21.6 to 28.8. (c) 19.4 to 31.0. (d) Using both confidence intervals, we can say that the P/E for Bank One is well below the population average. The P/E for AT&T Wireless is well above the population average. The P/E for Disney is within both confidence intervals. It appears that the P/E for Disney is close to the population average P/E. (e) By the central limit theorem, when n is large, the $\bar{x}$ distribution is approximately normal. In general, $n \geq 30$ is considered large.
23. (a) $d.f. = 30$; 43.59 to 46.82; 43.26 to 47.14; 42.58 to 47.81. (b) 43.63 to 46.77; 43.33 to 47.07; 42.74 to 47.66. (c) Yes; the respective intervals based on the Student's t distribution are slightly longer. (d) For Student's t, $d.f. = 80$; 44.22 to 46.18; 44.03 to 46.37; 43.65 to 46.75. For standard normal, 44.23 to 46.17; 44.05 to 46.35; 43.68 to 46.72. The intervals using the t distribution are still slightly longer than the corresponding intervals using the standard normal distribution. However, with a larger sample size, the differences between the two methods are less pronounced.

Section 7.3

1. $\hat{p} = r/n$.
3. (a) No. (b) The difference between $\hat{p}$ and p. In other words, the margin of error is the difference between results based on a random sample and results based on a population.
5. No, Jerry does not have a random sample of all laptops. In fact, he does not even have a random sample of laptops from the computer science class. Also, because all the laptops he tested for spyware are those of students from the same computer class, it could be that students shared software with classmates and spread the infection among the laptops owned by the students of the class.
7. (a) $n\hat{p} = 30$ and $n\hat{q} = 70$, so both products exceed 5. Also, the trials are binomial trials. (b) 0.225 to 0.375. (c) You are 90% confident that the confidence interval you computed is one of the intervals that contain p.
9. (a) 73. (b) 97.
11. (a) $\hat{p} = 39/62 = 0.6290$. (b) 0.51 to 0.75. If this experiment were repeated many times, about 95% of the intervals would contain p. (c) Both np and nq are greater than 5. If either is less than 5, the normal curve will not necessarily give a good approximation to the binomial.
13. (a) $\hat{p} = 1619/5222 = 0.3100$. (b) 0.29 to 0.33. If we repeat the survey with many different samples of 5222 dwellings, about 99% of the intervals will contain p. (c) Both np and nq are greater than 5. If either is less than 5, the normal curve will not necessarily give a good approximation to the binomial.
15. (a) $\hat{p} = 0.5420$. (b) 0.53 to 0.56. (c) Yes. Both np and nq are greater than 5.
17. (a) $\hat{p} = 0.0304$. (b) 0.02 to 0.05. (c) Yes. Both np and nq are greater than 5.
19. (a) $\hat{p} = 0.8603$. (b) 0.84 to 0.89. (c) A recent study shows that 86% of women shoppers remained loyal to their favorite supermarket last year. The margin of error was 2.5 percentage points.
21. (a) $\hat{p} = 0.25$. (b) 0.22 to 0.28. (c) A survey of 1000 large corporations has shown that 25% will choose a nonsmoking job candidate over an equally qualified smoker. The margin of error was 2.7%.
23. (a) Estimate a proportion; 208. (b) 68.
25. (a) Estimate a proportion; 666. (b) 662.
27. (a) $1/4 - (p - 1/2)^2 = 1/4 - (p^2 - p + 1/4) = -p^2 + p = p(1 - p)$. (b) Since $(p - 1/2)^2 \geq 0$, then $1/4 - (p - 1/2)^2 \leq 1/4$ because we are subtracting $(p - 1/2)^2$ from 1/4.

Section 7.4

1. Two random samples are independent if sample data drawn from one population are completely unrelated to the selection of sample data from the other population.
3. Josh's, because the critical value t_c is smaller based on larger $d.f.$; Kendra's, because her value for t_c is larger.
5. $\mu_1 < \mu_2$.
7. (a) Normal distribution by Theorem 7.1 and the fact that the samples are independent and the population standard deviations are known. (b) $E \approx 1.717$; interval from -3.717 to -0.283. (c) Student's t distribution

with $d.f. = 19$, based on the fact that the original distributions are normal and the samples are independent. (d) $t_{0.90} = 1.729$; $E \approx 1.720$; interval from -3.805 to -0.195. (e) $d.f. \approx 42.85$; interval from -3.755 to -0.245. (f) Since the 90% confidence interval contains all negative values, you can be 90% confident that μ_1 is less than μ_2.

9. (a) Yes, $n_1\hat{p}_1$, $n_1\hat{q}_1$, $n_2\hat{p}_2$, $n_2\hat{q}_2$ all exceed 5. (b) $\hat{\sigma} \approx 0.0943$; $E \approx 0.155$; -0.205 to 0.105. (c) No, the 90% confidence interval contains both negative and positive values.
11. (a) Use a calculator. (b) $d.f. \approx 11$; $E \approx 129.9$; interval from -121.3 to 138.5 ppm. (c) Because the interval contains both positive and negative numbers, we cannot say at the 90% confidence level that one region is more interesting than the other. (d) Student's t because σ_1 and σ_2 are unknown.
13. (a) Use a calculator. (b) $d.f. \approx 15$; $E \approx 5.42$; interval from 12.64% to 23.48% foreign revenue. (c) Because the interval contains only positive values, we can say at the 85% confidence level that technology companies have a higher population mean percentage foreign revenue. (d) Student's t because σ_1 and σ_2 are unknown.
15. (a) Use a calculator. (b) $d.f. \approx 39$; to use Table 6, round down to $d.f. \approx 35$; $E \approx 0.125$; interval from -0.399 to -0.149 feet. (c) Since the interval contains all negative numbers, it seems that at the 90% confidence level the population mean height of pro football players is less than that of pro basketball players. (d) Student's t distribution because σ_1 and σ_2 are unknown. Both samples are large, so no assumptions about the original distributions are needed.
17. (a) Yes, the sample sizes, number of successes, and number of failures are sufficiently large. (b) $\hat{\sigma} \approx 0.0232$; $E = 0.0599$; the interval is from 0.67 to 0.79. (c) The confidence interval contains values that are all positive, so we can be 99% sure that $p_1 > p_2$.
19. (a) Normal distribution since the sample sizes are sufficiently large and both σ_1 and σ_2 are known. (b) $E = 0.3201$; the interval is from -9.12 to -8.48. (c) The interval consists of negative values only. At the 99% confidence level, we can conclude that $\mu_1 < \mu_2$.
21. (a) Yes, the sample sizes, number of successes, and number of failures are sufficiently large. (b) $\hat{p}_1 = 0.3095$; $\hat{p}_2 = 0.1184$; $\hat{\sigma} = 0.0413$; interval from 0.085 to 0.297. (c) The interval contains numbers that are all positive. A greater proportion of hogans exist in Fort Defiance.
23. (a) Use a calculator. (b) Student's t distribution because the population standard deviations are unknown. In addition, since the original distributions are not normal, the sample sizes are too small. (c) $d.f. \approx 9$; $E \approx 5.3$; 3.7 to 14.3 pounds. (d) Interval contains all positive values. At the 85% confidence level, it appears that the population mean weight of gray wolves in Chihuahua is greater than that of gray wolves in Durango.
25. (a) -1.35 to 2.39. (b) 0.06 to 3.86. (c) -0.61 to 3.49. (d) At the 85% confidence level, we can say that the mean index of self-esteem based on competence is greater than the mean index of self-esteem based on physical attractiveness. We cannot conclude that there is a difference between the mean index of self-esteem based on competence and that based on social acceptance. We also cannot conclude that there is a difference in the mean indices based on social acceptance and physical attractiveness.
27. (a) Based on the same data, a 99% confidence interval is longer than a 95% confidence interval. Therefore, if the 95% confidence interval has both positive and negative values, so will the 99% confidence interval. However, for the same data, a 90% confidence interval is shorter than a 95% confidence interval. The 90% confidence interval might contain only positive or only negative values even if the 95% interval contains both. (b) Based on the same data, a 99% confidence interval is longer than a 95% confidence interval. Even if the 95% confidence interval contains values that are all positive, the longer 99% interval could contain both positive and negative values. Since, for the same data, a 90% confidence interval is shorter than a 95% confidence interval, if the 95% confidence interval contains only positive values, so will the 90% confidence interval.
29. (a) $n = 896.1$, or 897 couples in each sample. (b) $n = 768.3$, or 769 couples in each sample.
31. (a) Pooled standard deviation $s \approx 8.6836$; interval from 3.9 to 14.1. (b) The pooled standard deviation method has a shorter interval and a larger $d.f.$

Chapter 7 Review

1. See text.
3. (a) No, the probability that μ is in the interval is either 0 or 1. (b) Yes, 99% confidence intervals are constructed in such a way that 99% of all such confidence intervals based on random samples of the designated size will contain μ.
5. Interval for a mean; 176.91 to 180.49.
7. Interval for a mean. (a) Use a calculator. (b) 64.1 to 84.3.
9. Interval for a proportion; 0.50 to 0.54.
11. Interval for a proportion. (a) $\hat{p} = 0.4072$. (b) 0.333 to 0.482.
13. Difference of means. (a) Use a calculator. (b) $d.f. \approx 71$; to use Table 6, round down to $d.f. \approx 70$; $E \approx 0.83$; interval from -0.06 to 1.6. (c) Because the interval contains both positive and negative values, we cannot conclude at the 95% confidence level that there is any difference in soil water content between the two fields. (d) Student's t distribution because σ_1 and σ_2 are unknown. Both samples are large, so no assumptions about the original distributions are needed.
15. Difference of means. (a) $d.f. \approx 17$; $E \approx 2.5$; interval from 5.5 to 10.5 pounds. (b) Yes, the interval contains values that are all positive. At the 75% level of confidence, it appears that the average weight of adult male wolves from the Northwest Territories is greater.
17. Difference of proportions. (a) $\hat{p}_1 = 0.8495$; $\hat{p}_2 = 0.8916$; -0.1409 to 0.0567. (b) The interval contains both negative and positive numbers. We do not detect a difference in the proportions at the 95% confidence level.

19. (a) $P(A_1 < \mu_1 < B_1 \text{ and } A_2 < \mu_2 < B_2) = (0.80)(0.80) = 0.64$. The complement of the event $A_1 < \mu_1 < B_1$ *and* $A_2 < \mu_2 < B_2$ is that either μ_1 is not in the first interval or μ_2 is not in the second interval, or both. Thus, P(at least one interval fails) $= 1 - P(A_1 < \mu_1 < B_1 \text{ and } A_2 < \mu_2 < B_2) = 1 - 0.64 = 0.36$. (b) Suppose $P(A_1 < \mu_1 < B_1) = c$ and $P(A_2 < \mu_2 < B_2) = c$. If we want the probability that both hold to be 90%, and if x_1 and x_2 are independent, then $P(A_1 < \mu_1 < B_1 \text{ and } A_2 < \mu_2 < B_2) = 0.90$ means $P(A_1 < \mu_1 < B_1) \cdot P(A_2 < \mu_2 < B_2) = 0.90$, so $c^2 = 0.90$, or $c = 0.9487$. (c) In order to have a high probability of success for the whole project, the probability that each component will perform as specified must be significantly higher.

CHAPTER 8

Section 8.1

1. See text.
3. No, if we fail to reject the null hypothesis, we have not proved it beyond all doubt. We have failed only to find sufficient evidence to reject it.
5. Level of significance; α; type I.
7. Fail to reject H_0
9. 0.0184.
11. (a) H_0: $\mu = 40$. (b) H_1: $\mu \neq 40$. (c) H_1: $\mu > 40$. (d) H_1: $\mu < 40$.
13. (a) Yes, because x has a normal distribution. (b) $z \approx 1.12$. (c) 0.2628. (d) Fail to reject H_0 because P-value $> \alpha$.
15. (a) H_0: $\mu = 60$ kg. (b) H_1: $\mu < 60$ kg. (c) H_1: $\mu > 60$ kg. (d) H_1: $\mu \neq 60$ kg. (e) For part (b), the P-value area region is on the left. For part (c), the P-value area is on the right. For part (d), the P-value area is on both sides of the mean.
17. (a) H_0: $\mu = 16.4$ feet. (b) H_1: $\mu > 16.4$ feet. (c) H_1: $\mu < 16.4$ feet. (d) H_1: $\mu \neq 16.4$ feet. (e) For part (b), the P-value area is on the right. For part (c), the P-value area is on the left. For part (d), the P-value area is on both sides of the mean.
19. (a) $\alpha = 0.01$; H_0: $\mu = 4.7\%$; H_1: $\mu > 4.7\%$; right-tailed. (b) Normal; $\bar{x} = 5.38$; $z \approx 0.90$. (c) P-value ≈ 0.1841; on standard normal curve, shade area to the right of 0.90. (d) P-value of $0.1841 > 0.01$ for α; fail to reject H_0. (e) Insufficient evidence at the 0.01 level to reject claim that average yield for bank stocks equals average yield for all stocks.
21. (a) $\alpha = 0.01$; H_0: $\mu = 4.55$ grams; H_1: $\mu < 4.55$ grams; left-tailed. (b) Normal; $\bar{x} = 3.75$ grams; $z \approx -2.80$. (c) P-value ≈ 0.0026; on standard normal curve, shade area to the left of -2.80. (d) P-value of $0.0026 \leq 0.01$ for α; reject H_0. (e) The sample evidence is sufficient at the 0.01 level to justify rejecting H_0. It seems that the hummingbirds in the Grand Canyon region have a lower average weight.
23. (a) $\alpha = 0.01$; H_0: $\mu = 11\%$; H_1: $\mu \neq 11\%$; two-tailed. (b) Normal; $\bar{x} = 12.5\%$; $z = 1.20$. (c) P-value $= 2(0.1151) = 0.2302$; on standard normal curve, shade areas to the right of 1.20 and to the left of -1.20. (d) P-value of $0.2302 > 0.01$ for α; fail to reject H_0. (e) There is insufficient evidence at the 0.01 level to reject H_0. It seems that the average hail damage to wheat crops in Weld County matches the national average.

Section 8.2

1. The P-value for a two-tailed test of μ is twice that for a one-tailed test, based on the same sample data and null hypothesis.
3. $d.f. = n - 1$.
5. Yes. When P-value < 0.01, it is also true that P-value < 0.05.
7. (a) $0.010 < P\text{-value} < 0.020$; technology gives P-value ≈ 0.0150. (b) $0.005 < P\text{-value} < 0.010$; technology gives P-value ≈ 0.0075.
9. (a) Yes, since the original distribution is mound-shaped and symmetric and σ is unknown; $d.f. = 24$. (b) H_0: $\mu = 9.5$; H_1: $\mu \neq 9.5$. (c) $t \approx 1.250$. (d) $0.200 < P\text{-value} < 0.250$; technology gives $t \approx 0.2234$. (e) Fail to reject H_0 because the entire interval containing the P-value > 0.05 for α. (f) The sample evidence is insufficient at the 0.05 level to reject H_0.
11. (a) $\alpha = 0.01$; H_0: $\mu = 16.4$ feet; H_1: $\mu > 16.4$ feet. (b) Normal; $z \approx 1.54$. (c) P-value ≈ 0.618; on standard normal curve, shade area to the right of $z \approx 1.54$. (d) P-value of $0.0618 > 0.01$ for α; fail to reject H_0. (e) At the 1% level, there is insufficient evidence to say that the average storm level is increasing.
13. (a) $\alpha = 0.01$; H_0: $\mu = 1.75$ years; H_1: $\mu > 1.75$ years. (b) Student's t, $d.f. = 45$; $t \approx 2.481$. (c) $0.005 < P\text{-value} < 0.010$; on t graph, shade area to the right of 2.481. From TI-84, P-value ≈ 0.0084. (d) Entire P-interval ≤ 0.01 for α; reject H_0. (e) At the 1% level of significance, the sample data indicate that the average age of the Minnesota region coyotes is higher than 1.75 years.
15. (a) $\alpha = 0.05$; H_0: $\mu = 19.4$; H_1: $\mu \neq 19.4$ (b) Student's t, $d.f. = 35$; $t \approx -1.731$. (c) $0.050 < P\text{-value} < 0.100$; on t graph, shade area to the right of 1.731 and to the left of -1.731. From TI-84, P-value ≈ 0.0923. (d) P-value interval > 0.05 for α; fail to reject H_0. (e) At the 5% level of significance, the sample evidence does not support rejecting the claim that the average P/E of socially responsible funds is different from that of the S&P stock index.
17. i. Use a calculator. Rounded values are used in part ii. ii. (a) $\alpha = 0.05$; H_0: $\mu = 4.8$; H_1: $\mu < 4.8$. (b) Student's t, $d.f. = 5$; $t \approx -3.499$. (c) $0.005 < P\text{-value} < 0.010$; on t graph, shade area to the left of -3.499. From TI-84, P-value ≈ 0.0086. (d) P-value interval ≤ 0.05 for α; reject H_0. (e) At the 5% level of significance, sample evidence supports the claim that the average RBC count for this patient is less than 4.8.
19. i. Use a calculator. Rounded values are used in part ii. ii. (a) $\alpha = 0.01$; H_0: $\mu = 67$; H_1: $\mu \neq 67$. (b) Student's t, $d.f. = 15$; $t \approx -1.962$. (c) $0.050 < P\text{-value} < 0.100$; on t graph, shade area to the right of 1.962 and to the left of -1.962. From TI-84, P-value ≈ 0.0686. (d) P-value interval > 0.01; fail to reject H_0. (e) At the 1% level of significance, the sample evidence does not support a claim that the average thickness of slab avalanches in Vail is different from that in Canada.

21. i. Use a calculator. Rounded values are used in part ii. ii. (a) $\alpha = 0.05$; H_0: $\mu = 8.8$; H_1: $\mu \neq 8.8$. (b) Student's t, $d.f. = 13$; $t \approx -1.337$. (c) $0.200 < P\text{-value} < 0.250$; on t graph, shade area to the right of 1.337 and to the left of -1.337. From TI-84, P-value ≈ 0.2042. (d) P-value interval > 0.05; fail to reject H_0. (e) At the 5% level of significance, we cannot conclude that the average catch is different from 8.8 fish per day.
23. (a) The P-value of a one-tailed test is smaller. For a two-tailed test, the P-value is doubled because it includes the area in both tails. (b) Yes; the P-value of a one-tailed test is smaller, so it might be smaller than α, whereas the P-value of a corresponding two-tailed test may be larger than α. (c) Yes; if the two-tailed P-value is less than α, the smaller one-tail area is also less than α. (d) Yes, the conclusions can be different. The conclusion based on the two-tailed test is more conservative in the sense that the sample data must be more extreme (differ more from H_0) in order to reject H_0.
25. (a) For $\alpha = 0.01$, confidence level $c = 0.99$; interval from 20.28 to 23.72; hypothesized $\mu = 20$ is not in the interval; reject H_0. (b) H_0: $\mu = 20$; H_1: $\mu \neq 20$; $z = 3.000$; P-value ≈ 0.0026; P-value of $0.0026 \leq 0.01$ for α; reject H_0; conclusions are the same.
27. Critical value $z_0 = 2.33$; critical region is values to the right of 2.33; since the sample statistic $z = 1.54$ is not in the critical region, fail to reject H_0. At the 1% level, there is insufficient evidence to say that the average storm level is increasing. Conclusion is same as with P-value method.
29. Critical value is $t_0 = 2.412$ for one-tailed test with $d.f. = 45$; critical region is values to the right of 2.412. Since the sample test statistic $t = 2.481$ is in the critical region, reject H_0. At the 1% level, the sample data indicate that the average age of Minnesota region coyotes is higher than 1.75 years. Conclusion is same as with P-value method.

Section 8.3

1. For the conditions $np > 5$ and $nq > 5$, use the value of p from H_0. Note that $q = 1 - p$.
3. Yes. The corresponding P-value for a one-tailed test is half that for a two-tailed test, so the P-value of the one-tailed test is also less than 0.01.
5. (a) Yes, np and nq are both greater than 5. (b) H_0: $p = 0.50$; H_1: $p \neq 0.50$. (c) $\hat{p} = 0.40$; $z \approx -1.10$. (d) 0.2714. (e) Fail to reject H_0 because P-value of $0.2714 > 0.05$ for α. (f) The sample $\hat{p}$ value based on 30 trials is not sufficiently different from 0.50 to justify rejecting H_0 for $\alpha = 0.05$.
7. i. (a) $\alpha = 0.01$; H_0: $p = 0.301$; H_1: $p < 0.301$. (b) Standard normal; yes, $np \approx 64.7 > 5$ and $nq \approx 150.3 > 5$; $\hat{p} \approx 0.214$; $z \approx -2.78$. (c) P-value ≈ 0.0027; on standard normal curve, shade area to the left of -2.78. (d) P-value of $0.0027 \leq 0.01$ for α; reject H_0. (e) At the 1% level of significance, the sample data indicate that the population proportion of numbers with a leading "1" in the revenue file is less than 0.301, predicted by Benford's Law.
   ii. Yes; the revenue data file seems to include more numbers with higher first nonzero digits than Benford's Law predicts.
   iii. We have not proved H_0 to be false. However, because our sample data led us to reject H_0 and to conclude that there are too few numbers with a leading digit of 1, more investigation is merited.
9. (a) $\alpha = 0.01$; H_0: $p = 0.70$; H_1: $p \neq 0.70$. (b) Standard normal; $\hat{p} = 0.75$; $z \approx 0.62$. (c) P-value $= 2(0.2676) = 0.5352$; on standard normal curve, shade areas to the right of 0.62 and to the left of -0.62. (d) P-value of $0.5352 > 0.01$ for α; fail to reject H_0. (e) At the 1% level of significance, we cannot say that the population proportion of arrests of males aged 15 to 34 in Rock Springs is different from 70%.
11. (a) $\alpha = 0.01$; H_0: $p = 0.77$; H_1: $p < 0.77$. (b) Standard normal; $\hat{p} \approx 0.5556$; $z \approx -2.65$. (c) P-value ≈ 0.004; on standard normal curve, shade area to the left of -2.65. (d) P-value of $0.004 \leq 0.01$ for α; reject H_0. (e) At the 1% level of significance, the data show that the population proportion of driver fatalities related to alcohol is less than 77% in Kit Carson County.
13. (a) $\alpha = 0.01$; H_0: $p = 0.50$; H_1: $p < 0.50$. (b) Standard normal; $\hat{p} \approx 0.2941$; $z \approx -2.40$. (c) P-value $= 0.0082$; on standard normal curve, shade region to the left of -2.40. (d) P-value of $0.0082 \leq 0.01$ for α; reject H_0. (e) At the 1% level of significance, the data indicate that the population proportion of female wolves is now less than 50% in the region.
15. (a) $\alpha = 0.01$; H_0: $p = 0.261$; H_1: $p \neq 0.261$. (b) Standard normal; $\hat{p} \approx 0.1924$; $z \approx -2.78$. (c) P-value $= 2(0.0027) = 0.0054$; on standard normal curve, shade area to the right of 2.78 and to the left of -2.78. (d) P-value of $0.0054 \leq 0.01$ for α; reject H_0. (e) At the 1% level of significance, the sample data indicate that the population proportion of the five-syllable sequence is different from that of Plato's *Republic*.
17. (a) $\alpha = 0.01$; H_0: $p = 0.47$; H_1: $p > 0.47$. (b) Standard normal; $\hat{p} \approx 0.4871$; $z \approx 1.09$. (c) P-value $= 0.1379$; on standard normal curve, shade area to the right of 1.09. (d) P-value of $0.1379 > 0.01$ for α; fail to reject H_0. (e) At the 1% level of significance, there is insufficient evidence to support the claim that the population proportion of customers loyal to Chevrolet is more than 47%.
19. (a) $\alpha = 0.05$; H_0: $p = 0.092$; H_1: $p > 0.092$. (b) Standard normal; $\hat{p} \approx 0.1480$; $z \approx 2.71$. (c) P-value $= 0.0034$; on standard normal curve, shade region to the right of 2.71. (d) P-value of $0.0034 \leq 0.05$ for α; reject H_0. (e) At the 5% level of significance, the data indicate that the population proportion of students with hypertension during final exams week is higher than 9.2%.
21. (a) $\alpha = 0.01$; H_0: $p = 0.82$; H_1: $p \neq 0.82$. (b) Standard normal; $\hat{p} \approx 0.7671$; $z \approx -1.18$. (c) P-value $= 2(0.1190) = 0.2380$; on standard normal curve, shade area to the right of 1.18 and to the left of -1.18. (d) P-value of $0.2380 > 0.01$ for α; fail to reject H_0. (e) At the 1% level of significance, the evidence is insufficient to indicate that the population proportion of

extroverts among college student government leaders is different from 82%.

23. Critical value is $z_0 = -2.33$. The critical region consists of values less than -2.33. The sample test statistic $z = -2.65$ is in the critical region, so we reject H_0. This result is consistent with the P-value conclusion.

Section 8.4

1. Paired data are dependent.
3. H_0: $\mu_d = 0$; that is, the mean of the differences is 0, so there is no difference.
5. $d.f. = n - 1$.
7. (a) Yes. The sample size is sufficiently large. Student's t with $d.f. = 35$. (b) H_0: $\mu_d = 0$; H_1: $\mu_d \neq 0$. (c) $t = 2.400$ with $d.f. = 35$. (d) $0.020 < P\text{-value} < 0.050$. TI-84 gives $P\text{-value} \approx 0.0218$. (e) Reject H_0 since the entire interval containing the $P\text{-value} < 0.05$ for α. (f) At the 5% level of significance and for a sample size of 36, the sample mean of the differences is sufficiently different from 0 that we conclude the population mean of the differences is not zero.
9. (a) $\alpha = 0.05$; H_0: $\mu_d = 0$; H_1: $\mu_d \neq 0$. (b) Student's t, $d.f. = 7$; $\bar{d} \approx 2.25$; $t \approx 0.818$. (c) $0.250 < P\text{-value} < 0.500$; on t graph, shade area to the left of -0.818 and to the right of 0.818. From TI-84, $P\text{-value} \approx 0.4402$. (d) P-value interval > 0.05 for α; fail to reject H_0. (e) At the 5% level of significance, the evidence is insufficient to claim a difference in population mean percentage increases for corporate revenue and CEO salary.
11. (a) $\alpha = 0.01$; H_0: $\mu_d = 0$; H_1: $\mu_d > 0$. (b) Student's t, $d.f. = 4$; $\bar{d} \approx 12.6$; $t \approx 1.243$. (c) $0.125 < P\text{-value} < 0.250$; on t graph, shade area to the right of 1.243. From TI-84, $P\text{-value} \approx 0.1408$. (d) P-value interval > 0.01 for α; fail to reject H_0. (e) At the 1% level of significance, the evidence is insufficient to claim that the average peak wind gusts are higher in January.
13. (a) $\alpha = 0.05$; H_0: $\mu_d = 0$; H_1: $\mu_d > 0$. (b) Student's t, $d.f. = 7$; $\bar{d} \approx 6.125$; $t \approx 1.762$. (c) $0.050 < P\text{-value} < 0.075$; on t graph, shade area to the right of 1.762. From TI-84, $P\text{-value} \approx 0.0607$. (d) P-value interval > 0.05 for α; fail to reject H_0. (e) At the 5% level of significance, the evidence is insufficient to indicate that the population average percentage of male wolves in winter is higher.
15. (a) $\alpha = 0.05$; H_0: $\mu_d = 0$; H_1: $\mu_d > 0$. (b) Student's t, $d.f. = 7$; $\bar{d} \approx 6.0$; $t \approx 0.788$. (c) $0.125 < P\text{-value} < 0.250$; on t graph, shade area to the right of 0.788. From TI-84, $P\text{-value} \approx 0.2282$. (d) P-value interval > 0.05 for α; fail to reject H_0. (e) At the 5% level of significance, the evidence is insufficient to show that the population mean number of inhabited houses is greater than that of hogans.
17. i. Use a calculator. Nonrounded results are used in part ii. ii. (a) $\alpha = 0.05$; H_0: $\mu_d = 0$; H_1: $\mu_d > 0$. (b) Student's t, $d.f. = 35$; $\bar{d} \approx 2.472$; $t \approx 1.223$. (c) $0.100 < P\text{-value} < 0.125$; on t graph, shade area to the right of 1.223. From TI-84, $P\text{-value} \approx 0.1147$. (d) P-value interval > 0.05 for α; fail to reject H_0. (e) At the 5% level of significance, the evidence is insufficient to claim that the population mean cost of living index for housing is higher than that for groceries.
19. (a) $\alpha = 0.05$; H_0: $\mu_d = 0$; H_1: $\mu_d > 0$. (b) Student's t, $d.f. = 8$; $\bar{d} = 2.0$; $t \approx 1.333$. (c) $0.100 < P\text{-value} < 0.125$; on t graph, shade area to the right of 1.333. From TI-84, $P\text{-value} \approx 0.1096$. (d) P-value interval > 0.05 for α; fail to reject H_0. (e) At the 5% level of significance, the evidence is insufficient to claim that the population score on the last round is higher than that on the first.
21. (a) $\alpha = 0.05$; H_0: $\mu_d = 0$; H_1: $\mu_d > 0$. (b) Student's t, $d.f. = 7$; $\bar{d} \approx 0.775$; $t \approx 2.080$. (c) $0.025 < P\text{-value} < 0.050$; on t graph, shade area to the right of 2.080. From TI-84, $P\text{-value} \approx 0.0380$. (d) P-value interval ≤ 0.05 for α; reject H_0. (e) At the 5% level of significance, the evidence is sufficient to claim that the population mean time for rats receiving larger rewards to climb the ladder is less.
23. For a two-tailed test with $\alpha = 0.05$ and $d.f. = 7$, the critical values are $\pm t_0 = \pm 2.365$. The sample test statistic $t = 0.818$ is between -2.365 and 2.365, so we do not reject H_0. This conclusion is the same as that reached by the P-value method.

Section 8.5

1. (a) H_0 says that the population means are equal.
   (b) $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$.
   (c) $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ with $d.f.$ = smaller sample size - 1 or $d.f.$ is from Satterthwaite's formula.
3. H_0: $\mu_1 = \mu_2$ or H_0: $\mu_1 - \mu_2 = 0$.
5. $\bar{p} = \dfrac{r_1 + r_2}{n_1 + n_2}$.
7. H_1: $\mu_1 > \mu_2$; H_1: $\mu_1 - \mu_2 > 0$.
9. (a) Student's t with $d.f. = 48$. Samples are independent, population standard deviations are not known, and sample sizes are sufficiently large. (b) H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$. (c) $\bar{x}_1 - \bar{x}_2 = -2$; $t \approx -3.037$. (d) $0.0010 < P\text{-value} < 0.010$ (using $d.f. = 45$ and Table 6). TI-84 gives $P\text{-value} \approx 0.0030$ with $d.f. \approx 110.96$. (e) Because the entire interval containing the $P\text{-value} < 0.01$ for α, reject H_0. (f) At the 1% level of significance, the sample evidence is sufficiently strong to reject H_0 and conclude that the population means are different.
11. (a) Standard normal. Samples are independent, population standard deviations are known, and sample sizes are sufficiently large. (b) H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$. (c) $\bar{x}_1 - \bar{x}_2 = -2$; $z \approx -3.04$. (d) 0.0024. (e) P-value $0.0024 < 0.01$ for α, reject H_0. (f) At the 1% level of significance, the sample evidence is sufficiently strong to reject H_0 and conclude that the population means are different.
13. (a) $\bar{p} \approx 0.657$. (b) Standard normal distribution because $n_1\bar{p}$, $n_1\bar{q}$, $n_2\bar{p}$, $n_2\bar{q}$ are each greater than 5. (c) H_0: $p_1 = p_2$; H_1: $p_1 \neq p_2$ (d) $\hat{p}_1 - \hat{p}_2 = -0.1$; $z \approx -1.38$. (e) $P\text{-value} \approx 0.1676$. (f) Since P-value of $0.1676 \geq 0.05$ for α, fail to reject H_0. (g) At the 5% level of significance, the difference between the sample probabilities of success for the two binomial

experiments is too small to justify rejecting the hypothesis that the probabilities are equal.

15. (a) $\alpha = 0.01$; H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 > \mu_2$. (b) Standard normal; $\bar{x}_1 - \bar{x}_2 = 0.7$; $z \approx 2.57$. (c) P-value $= P(z > 2.57) \approx 0.0051$; on standard normal curve, shade area to the right of 2.57. (d) P-value of $0.0051 \le 0.01$ for α; reject H_0. (e) At the 1% level of significance, the evidence is sufficient to indicate that the population mean REM sleep time for children is more than that for adults.
17. (a) $\alpha = 0.05$; H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \ne \mu_2$. (b) Standard normal; $\bar{x}_1 - \bar{x}_2 = 0.6$; $z \approx 2.16$. (c) P-value $= 2P(z > 2.16) \approx 2(0.0154) = 0.0308$; on standard normal curve, shade area to the right of 2.16 and to the left of -2.16. (d) P-value of $0.0308 \le 0.05$ for α; reject H_0. (e) At the 5% level of significance, the evidence is sufficient to show that there is a difference between mean responses regarding preference for camping or fishing.
19. i. Use rounded results to compute t.
    ii. (a) $\alpha = 0.01$; H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 < \mu_2$. (b) Student's t, $d.f. = 9$; $\bar{x}_1 - \bar{x}_2 = -0.36$; $t \approx -0.965$. (c) $0.125 < P$-value < 0.250; on t graph, shade area to the left of -0.965. From TI-84, $d.f. \approx 19.96$; P-value ≈ 0.1731. (d) P-value interval > 0.01 for α; do not reject H_0. (e) At the 1% level of significance, the evidence is insufficient to indicate that the violent crime rate in the Rocky Mountain region is higher than that in New England.
21. (a) $\alpha = 0.05$; H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \ne \mu_2$. (b) Student's t, $d.f. = 29$; $\bar{x}_1 - \bar{x}_2 = -9.7$; $t \approx -0.751$. (c) $0.250 < P$-value < 0.500; on t graph, shade area to the right of 0.751 and to the left of -0.751. From TI-84, $d.f. \approx 57.92$; P-value ≈ 0.4556. (d) P-value interval > 0.05 for α; do not reject H_0. (e) At the 5% level of significance, the evidence is insufficient to indicate that there is a difference between the control and experimental groups in the mean score on the vocabulary portion of the test.
23. i. Use rounded results to compute t.
    ii. (a) $\alpha = 0.05$; H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \ne \mu_2$. (b) Student's t, $d.f. = 14$; $\bar{x}_1 - \bar{x}_2 = 0.82$; $t \approx 0.869$. (c) $0.250 < P$-value < 0.500; on t graph, shade area to the right of 0.869 and to the left of -0.869. From TI-84, $d.f. \approx 28.81$; P-value ≈ 0.3940. (d) P-value interval > 0.05 for α; do not reject H_0. (e) At the 5% level of significance, the evidence is insufficient to indicate that there is a difference in the mean number of cases of fox rabies between the two regions.
25. i. Use rounded results to compute t.
    ii. (a) $\alpha = 0.05$; H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \ne \mu_2$. (b) Student's t, $d.f. = 6$; $\bar{x}_1 - \bar{x}_2 = -1.64$; $t \approx -1.041$. (c) $0.250 < P$-value < 0.500; on t graph, shade area to the right of 1.041 and to the left of -1.041. From TI-84, $d.f. \approx 12.28$; P-value ≈ 0.3179. (d) P-value interval > 0.05 for α; do not reject H_0. (e) At the 5% level of significance, the evidence is insufficient to indicate that the mean time lost due to hot tempers is different from that lost due to technical workers' attitudes.
27. (a) $d.f. = 19.96$ (Some software will truncate this to 19.) (b) $d.f. = 9$; the convention of using the smaller of $n_1 - 1$ and $n_2 - 1$ leads to a $d.f.$ that is always less than or equal to that computed by Satterthwaite's formula.
29. (a) $\alpha = 0.05$; H_0: $p_1 = p_2$; H_1: $p_1 \ne p_2$. (b) Standard normal; $\bar{p} \approx 0.2911$; $\hat{p}_1 - \hat{p}_2 \approx -0.052$; $z \approx -1.13$. (c) P-value $\approx 2P(z < -1.13) \approx 2(0.1292) = 0.2584$ on standard normal curve, shade area to the right of 1.13 and to the left of -1.13. (d) P-value of $0.2584 > 0.05$ for α; fail to reject H_0. (e) At the 5% level of significance, there is insufficient evidence to conclude that the population proportion of women favoring more tax dollars for the arts is different from the proportion of men.
31. (a) $\alpha = 0.01$; H_0: $p_1 = p_2$; H_1: $p_1 \ne p_2$. (b) Standard normal; $\bar{p} \approx 0.0676$; $\hat{p}_1 - \hat{p}_2 \approx 0.0237$; $z \approx 0.79$. (c) P-value $\approx 2P(z > 0.79) \approx 2(0.2148) = 0.4296$; on standard normal curve, shade area to the right of 0.79 and to the left of -0.79. (d) P-value of $0.4296 > 0.01$ for α; fail to reject H_0. (e) At the 1% level of significance, there is insufficient evidence to conclude that the population proportion of high school dropouts on Oahu is different from that of Sweetwater County.
33. (a) $\alpha = 0.01$; H_0: $p_1 = p_2$; H_1: $p_1 < p_2$. (b) Standard normal; $\bar{p} = 0.42$; $\hat{p}_1 - \hat{p}_2 = -0.10$; $z \approx -1.43$. (c) P-value $\approx P(z < -1.43) \approx 0.0764$; on standard normal curve, shade area to the left of -1.43. (d) P-value of $0.0764 > 0.01$ for α; fail to reject H_0. (e) At the 1% level of significance, there is insufficient evidence to conclude that the population proportion of adults who believe in extraterrestrials and who attended college is higher than the proportion who believe in extraterrestrials but did not attend college.
35. (a) $\alpha = 0.05$; H_0: $p_1 = p_2$; H_1: $p_1 < p_2$. (b) Standard normal; $\bar{p} \approx 0.2189$; $\hat{p}_1 - \hat{p}_2 \approx -0.074$; $z \approx -2.04$. (c) P-value $\approx P(z < -2.04) \approx 0.0207$; on standard normal curve, shade area to the left of -2.04. (d) P-value of $0.0207 \le 0.05$ for α; reject H_0. (e) At the 5% level of significance, there is sufficient evidence to conclude that the population proportion of trusting people in Chicago is higher for the older group.
37. H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 < \mu_2$; for $d.f. = 9$, $\alpha = 0.01$ in the *one-tail area* row, the critical value is $t_0 = -2.821$; sample test statistic $t = -0.965$ is not in the critical region; fail to reject H_0. This result is consistent with that obtained by the P-value method.

Chapter 8 Review

1. Look at the original x distribution. If it is normal or $n \ge 30$, and σ is known, use the standard normal distribution. If the x distribution is mound-shaped or $n \ge 30$, and σ is unknown, use the Student's t distribution. The $d.f.$ is determined by the application.
3. A larger sample size increases the $|z|$ or $|t|$ value of the sample test statistic.
5. Single mean. (a) $\alpha = 0.05$; H_0: $\mu = 11.1$; H_1: $\mu \ne 11.1$. (b) Standard normal; $z = -3.00$. (c) P-value $= 0.0026$; on standard normal curve, shade area to the right of 3.00 and to the left of -3.00. (d) P-value of $0.0026 \le 0.05$ for α; reject H_0. (e) At the 5% level of significance, the evidence is sufficient to say that the

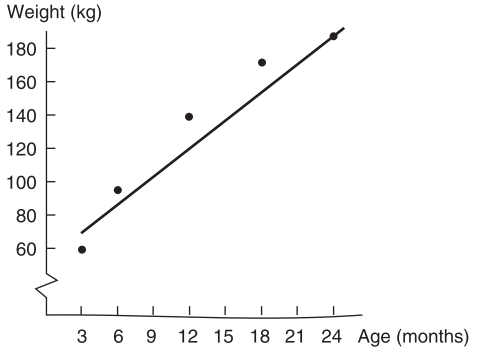
miles driven per vehicle in Chicago is different from the national average.

7. Single mean. (a) $\alpha = 0.01$; H_0: $\mu = 0.8$; H_1: $\mu > 0.8$. (b) Student's t, $d.f. = 8$; $t \approx 4.390$. (c) $0.0005 <$ P-value < 0.005; on t graph, shade area to the right of 4.390. From TI-84, P-value ≈ 0.0012. (d) P-value interval ≤ 0.01 for α; reject H_0. (e) At the 1% level of significance, the evidence is sufficient to say that the Toylot claim of 0.8 A is too low.
9. Single proportion. (a) $\alpha = 0.01$; H_0: $p = 0.60$; H_1: $p < 0.60$. (b) Standard normal; $z = -3.01$. (c) P-value $= 0.0013$; on standard normal curve, shade area to the left of -3.01. (d) P-value of $0.0013 \leq 0.01$ for α; reject H_0. (e) At the 1% level of significance, the evidence is sufficient to show that the mortality rate has dropped.
11. Single mean. (a) $\alpha = 0.01$; H_0: $\mu = 40$; H_1: $\mu > 40$. (b) Standard normal; $z = 3.34$. (c) P-value $= 0.0004$; on standard normal curve, shade area to the right of 3.34. (d) P-value of $0.0004 \leq 0.01$ for α; reject H_0. (e) At the 1% level of significance, the evidence is sufficient to say that the population average number of matches is larger than 40.
13. Difference of means. (a) $\alpha = 0.05$; H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$. (b) Student's t, $d.f. = 50$; $\bar{x}_1 - \bar{x}_2 = 0.3$ cm; $t \approx 1.808$. (c) $0.050 <$ P-value < 0.100; on t graph, shade area to the right of 1.808 and to the left of -1.808. From TI-84, $d.f. \approx 100.27$, P-value ≈ 0.0735. (d) P-value interval > 0.05 for α; do not reject H_0. (e) At the 5% level of significance, the evidence is insufficient to indicate a difference in population mean length between the two types of projectile points.
15. Single mean. (a) $\alpha = 0.05$; H_0: $\mu = 7$ oz; H_1: $\mu \neq 7$ oz. (b) Student's t, $d.f. = 7$; $t \approx 1.697$. (c) $0.100 <$ P-value < 0.150; on t graph, shade area to the right of 1.697 and to the left of -1.697. From TI-84, P-value ≈ 0.1335. (d) P-value interval > 0.05 for α; do not reject H_0. (e) At the 5% level of significance, the evidence is insufficient to show that the population mean amount of coffee per cup is different from 7 oz.
17. Paired difference test. (a) $\alpha = 0.05$; H_0: $\mu_d = 0$; H_1: $\mu_d < 0$. (b) Student's t, $d.f. = 4$; $\bar{d} \approx -4.94$; $t = -2.832$. (c) $0.010 <$ P-value < 0.025; on t graph, shade area to the left of -2.832. From TI-84, P-value ≈ 0.0236. (d) P-value interval ≤ 0.05 for α; reject H_0. (e) At the 5% level of significance, there is sufficient evidence to claim that the population average net sales improved.

CHAPTER 9

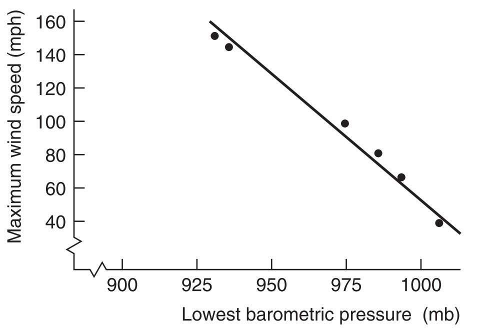
Section 9.1

1. Explanatory variable is placed along horizontal axis, usually x axis. Response variable is placed along vertical axis, usually y axis.
3. Decreases.
5. (a) Moderate. (b) None. (c) High.
7. (a) No. (b) Increasing population might be a lurking variable causing both variables to increase.
9. (a) No. (b) One lurking variable responsible for average annual income increases is inflation. Better training might be a lurking variable responsible for shorter times to run the mile.
11. The correlation coefficient is moderate and negative. It suggests that as gasoline prices increase, consumption decreases, and the relationship is moderately linear. It is risky to apply these results to gasoline prices much higher than \$5.30 per gallon. It could be that many of the discretionary and technical means of reducing consumption have already been applied, so consumers cannot reduce their consumption much more.
13. (a) Ages and Average Weights of Shetland Ponies



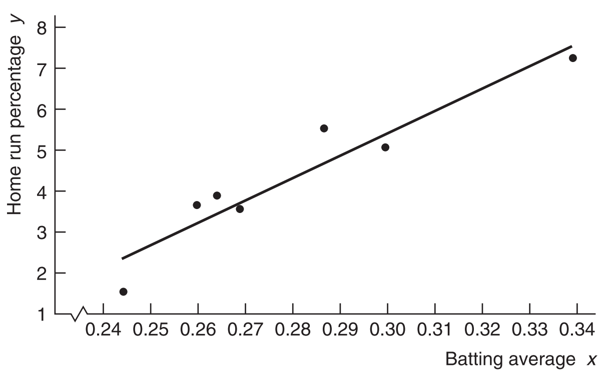
Line slopes upward.
(b) Strong; positive. (c) $r \approx 0.972$; increase.

15. (a) Lowest Barometric Pressure and Maximum Wind Speed for Tropical Cyclones



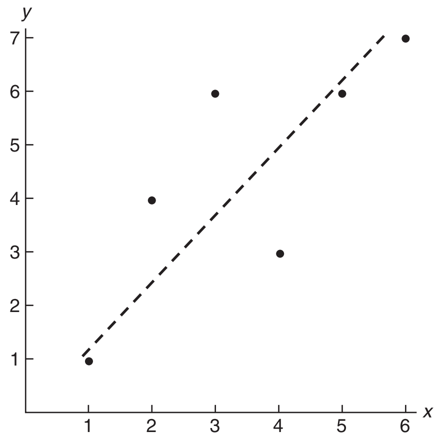
Line slopes downward.
(b) Strong; negative. (c) $r \approx -0.990$; decrease.

17. (a) Batting Average and Home Run Percentage

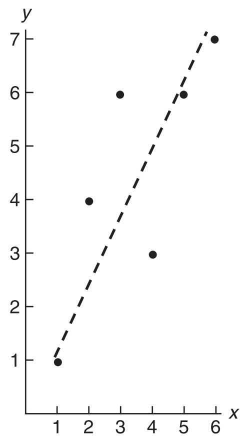


Line slopes upward.
(b) High; positive. (c) $r \approx 0.948$; increase.

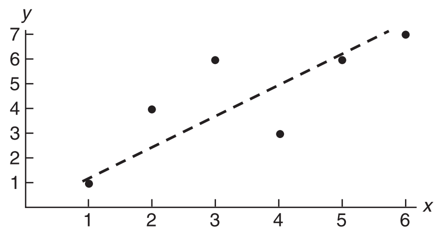
19. (a) Unit Length on y Same as That on x



(b) Unit Length on y Twice That on x



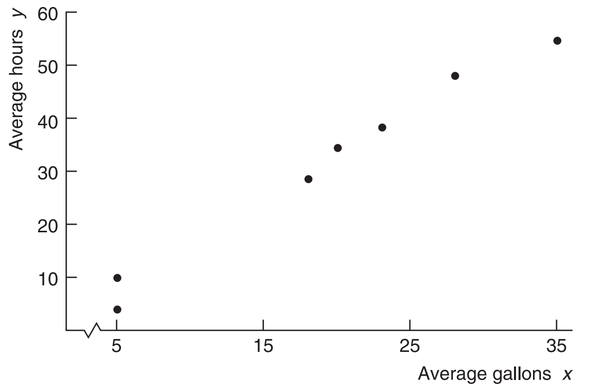
(c) Unit Length on y Half That on x



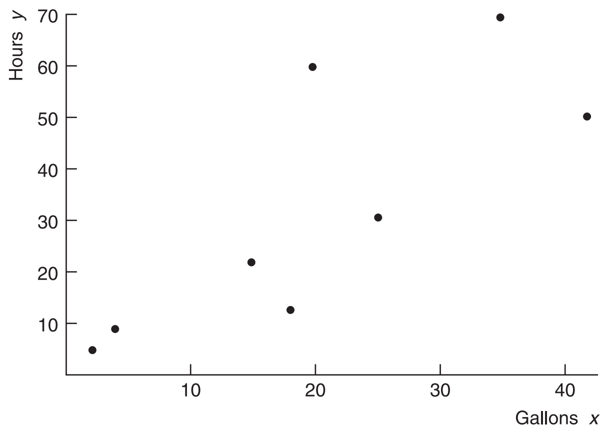
(d) The line in part (b) appears steeper than the line in part (a), whereas the line in part (c) appears flatter than the line in part (a). The slopes actually are all the same, but the lines look different because of the change in unit lengths on the y and x axes.

21. (a) $r \approx 0.972$ with $n = 5$ is significant for $\alpha = 0.05$. For this α, we conclude that age and weight of Shetland ponies are correlated. (b) $r \approx -0.990$ with $n = 6$ is significant for $\alpha = 0.01$. For this α, we conclude that lowest barometric pressure reading and maximum wind speed for cyclones are correlated.

23. (a) Average Hours Lost per Person versus Average Fuel Wasted per Person in Traffic Delays



$r \approx 0.991$.
(b) For variables based on averages, $\overline{x} = 19.25$ hr; $s_x \approx 10.33$ hr; $\overline{y} = 31.13$ gal; $s_y \approx 17.76$ gal. For variables based on single individuals, $\overline{x} = 20.13$ hr; $s_x \approx 13.84$ hr; $\overline{y} = 31.87$ gal; $s_y \approx 25.18$ gal. Dividing by larger numbers results in a smaller value.
(c) Hours Lost per Person versus Fuel Wasted per Person in Traffic Delays

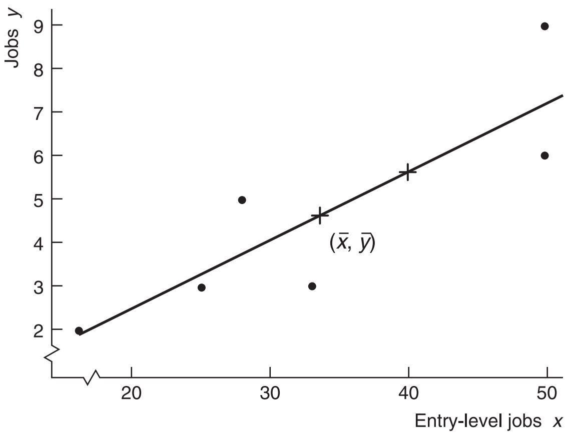


$r \approx 0.794$.
(d) Yes; by the central limit theorem, the $\overline{x}$ distribution has a smaller standard deviation than the corresponding x distribution.

Section 9.2

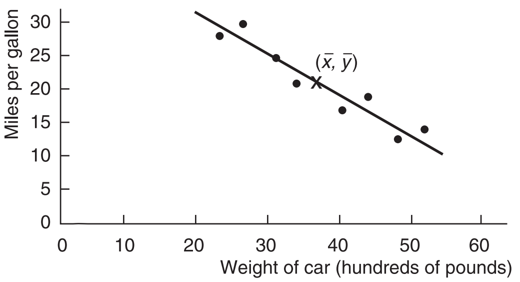
1. $b = -2$. When x changes by 1 unit, y decreases by 2 units.
3. Extrapolating. Extrapolating beyond the range of the data is dangerous because the relationship pattern might change.
5. (a) $\hat{y} \approx 318.16 - 30.878x$. (b) About 31 fewer frost-free days. (c) $r \approx -0.981$. Note that if the slope is negative, r is also negative. (d) 96.3% of variation explained and 3.7% unexplained.

7. (a) Total Number of Jobs and Number of Entry-Level Jobs (Units in 100's)



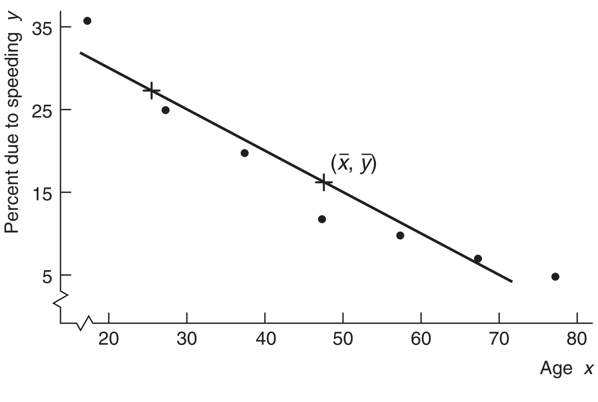
(b) Use a calculator. (c) $\overline{x} \approx 33.67$ jobs; $\overline{y} \approx 4.67$ entry-level jobs; $a \approx -0.748$; $b \approx 0.161$; $\hat{y} \approx -0.748 + 0.161x$ (d) See figure in part (a). (e) $r^2 \approx 0.740$; 74.0% of variation explained and 26.0% unexplained. (f) 5.69 jobs.

9. (a) Weight of Cars and Gasoline Mileage



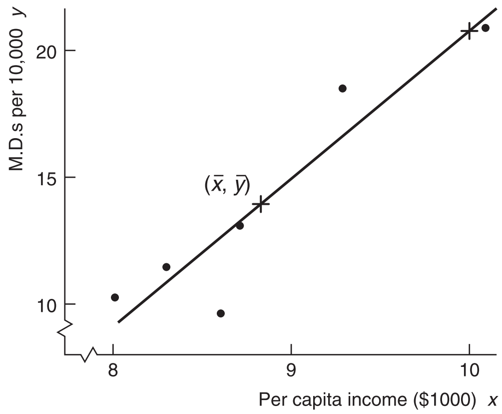
(b) Use a calculator. (c) $\overline{x} \approx 37.375$; $\overline{y} \approx 20.875$ mpg; $a \approx 43.326$; $b \approx -0.6007$; $\hat{y} \approx 43.326 - 0.6007x$. (d) See figure in part (a). (e) $r^2 \approx 0.895$; 89.5% of variation explained and 10.5% unexplained. (f) 20.5 mpg.

11. (a) Age and Percentage of Fatal Accidents Due to Speeding



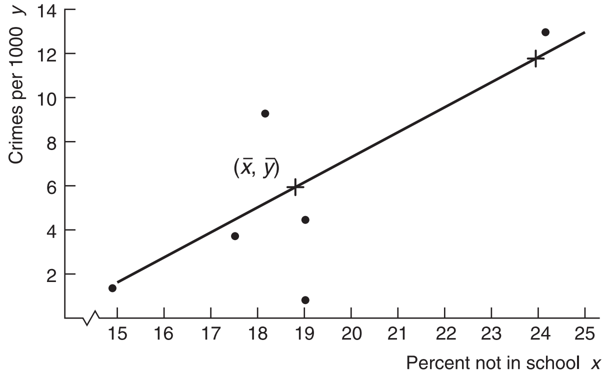
(b) Use a calculator. (c) $\overline{x} \approx 47$ years; $\overline{y} \approx 16.43\%$; $a \approx 39.761$; $b \approx -0.496$; $\hat{y} \approx 39.761 - 0.496x$. (d) See figure in part (a). (e) $r^2 \approx 0.920$; 92.0% of variation explained and 8.0% unexplained. (f) 27.36%.

13. (a) Per Capita Income ($1000) and M.D.s per 10,000 Residents



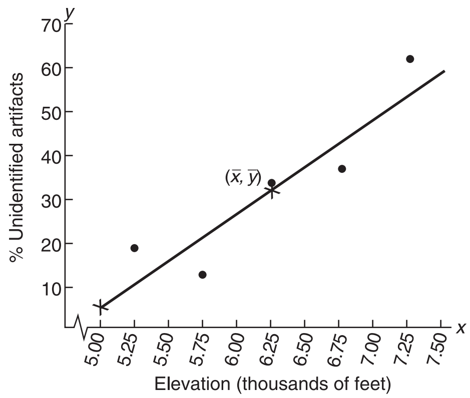
(b) Use a calculator. (c) $\overline{x} = \$8.83$; $\overline{y} \approx 13.95$ M.D.s; $a \approx -36.898$; $b \approx 5.756$; $\hat{y} \approx -36.898 + 5.756x$. (d) See figure in part (a). (e) $r^2 \approx 0.872$; 87.2% of variation explained, 12.8% unexplained. (f) 20.7 M.D.s per 10,000 residents.

15. (a) Percentage of 16- to 19-Year-Olds Not in School and Violent Crime Rate per 1000 Residents



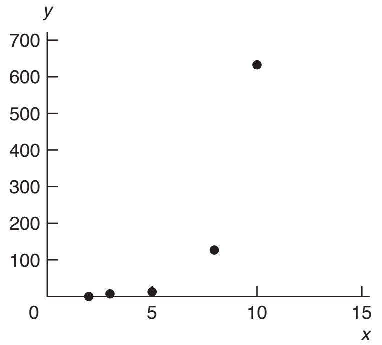
(b) Use a calculator. (c) $\overline{x} = 18.8\%$; $\overline{y} = 5.4$; $a \approx -17.204$; $b \approx 1.202$; $\overline{y} \approx -17.204 + 1.202x$. (d) See figure in part (a). (e) $r^2 \approx 0.584$; 58.4% of variation explained, 41.6% unexplained. (f) 11.6 crimes per 1000 residents.

17. (a) Elevation of Archaeological Sites and Percentage of Unidentified Artifacts

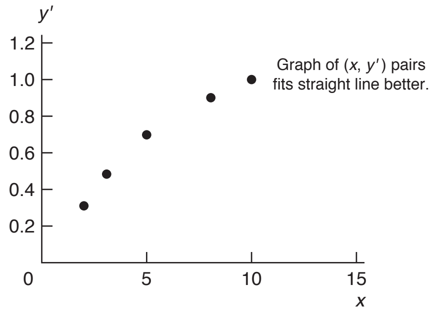


(b) Use a calculator. (c) $\bar{x} = 6.25$; $\bar{y} = 32.8$; $a = -104.7$; $b = 22$; $\hat{y} = -104.7 + 22x$.
(d) See figure in part (a). (e) $r^2 \approx 0.833$; 83.3% of variation explained, 16.7% unexplained. (f) 38.3.

19. (a) Yes. The pattern of residuals appears randomly scattered about the horizontal line at 0. (b) No. There do not appear to be any outliers.
21. (a) Result checks. (b) Result checks. (c) Yes. (d) The equation $x = 0.9337y - 0.1335$ does not match part (b). (e) No. The least-squares equation changes depending on which variable is the explanatory variable and which is the response variable.
23. (a) Model with (x, y) Data Pairs

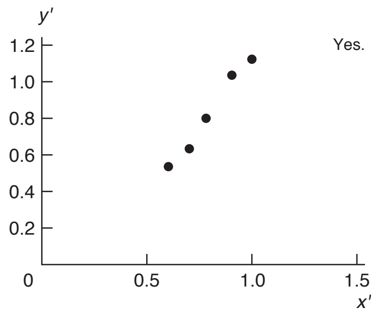


(b) Model with (x, y') Data Pairs



(c) $y' \approx -0.365 + 0.311x$; $r \approx 0.998$.
(d) $\alpha \approx 0.432$; $\beta \approx 2.046$; $y \approx 0.432(2.046)^x$.

25. (a) Model with (x', y') Data Pairs



(b) $y' \approx -0.451 + 1.600x'$; $r \approx 0.991$.
(c) $\alpha \approx 0.354$; $\beta \approx 1.600$; $y \approx 0.354x^{1.600}$.

Section 9.3

1. ρ (Greek letter rho).
3. As x becomes further away from $\bar{x}$, the confidence interval for the predicted y becomes longer.
5. (a) Diameter. (b) $a = -0.223$; $b = 0.7848$; $\hat{y} = -0.223 + 0.7848x$. (c) P-value of b is 0.001. H_0: $\beta = 0$; H_1: $\beta \neq 0$. Since P-value < 0.01, reject H_0 and conclude that the slope is not zero. (d) $r \approx 0.896$. Yes. P-value is 0.001, so we reject H_0 for $\alpha = 0.01$.
7. (a) Use a calculator. (b) $\alpha = 0.05$; H_0: $\rho = 0$; H_1: $\rho > 0$; sample $t \approx 2.522$; $d.f. = 4$; $0.025 < P\text{-value} < 0.050$; reject H_0. There seems to be a positive correlation between x and y. From TI-84, P-value ≈ 0.0326. (c) Use a calculator. (d) 45.36%. (e) Interval from 39.05 to 51.67. (f) $\alpha = 0.05$; H_0: $\beta = 0$; H_1: $\beta > 0$; sample $t \approx 2.522$; $d.f. = 4$; $0.025 < P\text{-value} < 0.050$; reject H_0. There seems to be a positive slope between x and y. From TI-84, P-value ≈ 0.0326. (g) Interval from 0.064 to 0.760. For every percentage increase in successful free throws, the percentage of successful field goals increases by an amount between 0.06 and 0.76.
9. (a) Use a calculator. (b) $\alpha = 0.01$; H_0: $\rho = 0$; H_1: $\rho < 0$; sample $t \approx -10.06$; $d.f. = 5$; P-value < 0.0005; reject H_0. The sample evidence supports a negative correlation. From TI-84, P-value ≈ 0.00008. (c) Use a calculator. (d) 2.39 hours. (e) Interval from 2.12 to 2.66 hours. (f) $\alpha = 0.01$; H_0: $\beta = 0$; H_1: $\beta < 0$; sample $t \approx -10.06$; $d.f. = 5$; P-value < 0.0005; reject H_0. The sample evidence supports a negative slope. From TI-84, P-value ≈ 0.00008. (g) Interval from -0.065 to -0.044. For every additional meter of depth, the optimal time decreases by between 0.04 and 0.07 hour.
11. (a) Use a calculator. (b) $\alpha = 0.01$; H_0: $\rho = 0$; H_1: $\rho > 0$; sample $t \approx 6.534$; $d.f. = 4$; $0.0005 < P\text{-value} < 0.005$; reject H_0. The sample evidence supports a positive correlation. From TI-84, P-value ≈ 0.0014. (c) Use a calculator. (d) \$12.577 thousand. (e) Interval from 12.247 to 12.907 (thousand dollars). (f) $\alpha = 0.01$; H_0: $\beta = 0$; H_1: $\beta > 0$; sample $t \approx 6.534$; $d.f. = 4$; $0.0005 < P\text{-value} < 0.005$; reject H_0. The sample evidence supports a positive slope. From TI-84, P-value ≈ 0.0014. (g) Interval from 0.436 to 1.080. For every \$1000 increase in list price, the dealer price increase is between \$436 and \$1080 higher.
13. (a) H_0: $\rho = 0$; H_1: $\rho \neq 0$; $d.f. = 4$; sample $t = 4.129$; $0.01 < P\text{-value} < 0.02$; do not reject H_0; r is not significant at the 0.01 level of significance. (b) H_0: $\rho = 0$; H_1: $\rho \neq 0$; $d.f. = 8$; sample $t = 5.840$; P-value < 0.001; reject H_0; r is significant at the 0.01 level of significance. (c) As n increases, the t value corresponding to r also increases, resulting in a smaller P-value.

Section 9.4

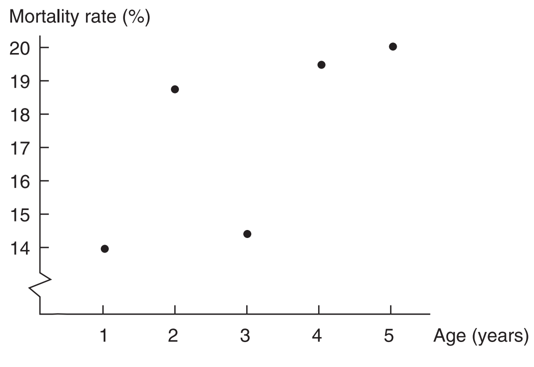
1. (a) Response variable is x_1. Explanatory variables are x_2, x_3, x_4. (b) 1.6 is the constant term; 3.5 is the coefficient of x_2; -7.9 is the coefficient of x_3; and 2.0 is

the coefficient of x_4. (c) $x_1 = 10.7$. (d) 3.5 units; 7 units; -14 units. (e) $d.f. = 8$; $t = 1.860$; 2.72 to 4.28. (f) $\alpha = 0.05$; H_0: $\beta_2 = 0$; H_1: $\beta_2 \neq 0$; $d.f. = 8$; $t = 8.35$; P-value < 0.001; reject H_0.

3. (a) $CVx_1 \approx 9.08$; $CVx_2 \approx 14.59$; $CVx_3 \approx 8.88$; x_2 has greatest spread; x_3 has smallest. (b) $r^2x_1x_2 \approx 0.958$; $r^2x_1x_3 \approx 0.942$; $r^2x_2x_3 \approx 0.895$; x_2; yes; 95.8%; 94.2%. (c) 97.7%. (d) $x_1 = 30.99 + 0.861x_2 + 0.335x_3$; 3.35; 8.61. (e) $\alpha = 0.05$; H_0: coefficient $= 0$; H_1: coefficient $\neq 0$; $d.f. = 8$; for β_2, $t = 3.47$ with P-value $=$ 0.008; for β_3, $t = 2.56$ with P-value $= 0.034$; reject H_0 for each coefficient and conclude that the coefficients of x_2 and x_3 are not zero. (f) $d.f. = 8$; $t = 1.86$; C.I. for β_2 is 0.40 to 1.32; C.I. for β_3 is 0.09 to 0.58. (g) 153.9; 148.3 to 159.4.
5. (a) $CVx_1 \approx 39.64$; $CVx_2 \approx 44.45$; $CVx_3 \approx 50.62$; $CVx_4 \approx 52.15$; x_4; x_1 has a small CV because we divide by a large mean. (b) $r^2x_1x_2 \approx 0.842$; $r^2x_1x_3 \approx 0.865$; $r^2x_1x_4 \approx 0.225$; $r^2x_2x_3 \approx 0.624$; $r^2x_2x_4 \approx 0.184$; $r^2x_3x_4 \approx 0.089$; x_4; 84.2%. (c) 96.7%. (d) $x_1 = 7.68 + 3.66x_2 + 7.62x_3 + 0.83x_4$; 7.62 million dollars. (e) $\alpha = 0.05$; H_0: coefficient $= 0$; H_1: coefficient $\neq 0$; $d.f. = 6$; for β_2, $t = 3.28$ with P-value $= 0.017$; for β_3, $t = 4.60$ with P-value $= 0.004$; for β_4, $t = 1.54$ with P-value $= 0.175$; reject H_0 for β_2 and β_3 and conclude that the coefficients of x_2 and x_3 are not zero. For β_4, fail to reject H_0 and conclude that the coefficient of x_4 could be zero. (f) $d.f. = 6$; $t = 1.943$; C.I. for β_2 is 1.49 to 5.83; C.I. for β_3 is 4.40 to 10.84; C.I. for β_4 is -0.22 to 1.88. (g) 91.95; 77.6 to 106.3. (h) 5.63; 4.21 to 7.04.
7. Depends on data.

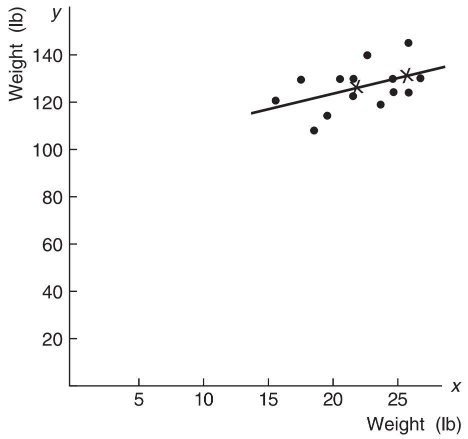
Chapter 9 Review

1. r will be close to 0.
3. Results are more reliable for interpolation.
5. (a) Age and Mortality Rate for Bighorn Sheep

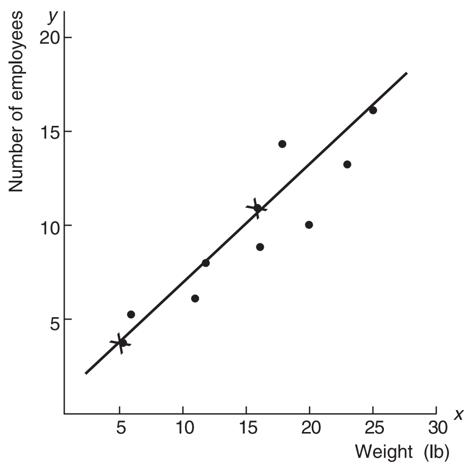


(b) $\bar{x} = 3$; $\bar{y} \approx 17.38$; $b \approx 1.27$; $\hat{y} \approx 13.57 + 1.27x$.
(c) $r \approx 0.685$; $r^2 \approx 0.469$. (d) $\alpha = 0.01$; H_0: $\rho = 0$; H_1: $\rho > 0$; $d.f. = 3$; $t = 1.627$; $0.100 < P\text{-value} < 0.125$; do not reject H_0. There does not seem to be a positive correlation between age and mortality rate of bighorn sheep. From TI-84, P-value ≈ 0.1011. (e) No. Based on these limited data, predictions from the least-squares line model might be misleading. There appear to be other lurking variables that affect the mortality rate of sheep in different age groups.

7. (a) Weight of One-Year-Old versus Weight of Adult



(b) $\bar{x} \approx 21.43$; $\bar{y} \approx 126.79$; $b \approx 1.285$; $\hat{y} \approx 99.25 + 1.285x$. (c) $r \approx 0.468$; $r^2 \approx 0.219$; 21.9% explained. (d) $\alpha = 0.01$; H_0: $\rho = 0$; H_1: $\rho > 0$; $d.f. = 12$; $t = 1.835$; $0.025 < P\text{-value} < 0.050$; do not reject H_0. At the 1% level of significance, there does not seem to be a positive correlation between weight of baby and weight of adult. From TI-84, P-value ≈ 0.0457. (e) 124.95 pounds. However, since r is not significant, this prediction may not be useful. Other lurking variables seem to have an effect on adult weight. (f) Use a calculator. (g) 105.91 to 143.99 pounds. (h) $\alpha = 0.01$; H_0: $\beta = 0$; H_1: $\beta > 0$; $d.f. = 12$; $t = 1.835$; $0.025 < P\text{-value} < 0.050$; do not reject H_0. At the 1% level of significance, there does not seem to be a positive slope between weight of baby x and weight of adult y. From TI-84, P-value ≈ 0.0457. (i) 0.347 to 2.223. At the 80% confidence level, we can say that for each additional pound a female infant weighs at 1 year, the female's adult weight changes by 0.35 to 2.22 pounds.
9. (a) Weight of Mail versus Number of Employees Required



(b) $\bar{x} \approx 16.38$; $\bar{y} \approx 10.13$; $b \approx 0.554$; $\hat{y} \approx 1.051 + 0.554x$.
(c) $r \approx 0.913$; $r^2 \approx 0.833$; 83.3% explained.
(d) $\alpha = 0.01$; H_0: $\rho = 0$; H_1: $\rho > 0$; $d.f. = 6$; $t = 5.467$; $0.0005 < P\text{-value} < 0.005$; reject H_0. At the 1% level of

significance, there is sufficient evidence to show a positive correlation between pounds of mail and number of employees required to process the mail. From TI-84, P-value ≈ 0.0008. (e) 9.36. (f) Use a calculator. (g) 4.86 to 13.86. (h) $\alpha = 0.01$; H_0: $\beta = 0$; H_1: $\beta > 0$; $d.f. = 6$; $t = 5.467$; $0.0005 < P\text{-value} < 0.005$; reject H_0. At the 1% level of significance, there is sufficient evidence to show a positive slope between pounds of mail x and number of employees required to process the mail y. From TI-84, P-value ≈ 0.0008. (i) 0.408 to 0.700. At the 80% confidence level, we can say that for each additional pound of mail, between 0.4 and 0.7 additional employees are needed.

CUMULATIVE REVIEW PROBLEMS

1. (a) i. $\alpha = 0.01$; H_0: $\mu = 2.0$ ug/l; H_1: $\mu > 2.0$ ug/l.
   ii. Standard normal; $z = 2.53$.
   iii. P-value ≈ 0.0057; on standard normal curve, shade area to the right of 2.53.
   iv. P-value of $0.0057 \le 0.01$ for α; reject H_0.
   v. At the 1% level of significance, the evidence is sufficient to say that the population mean discharge level of lead is higher.

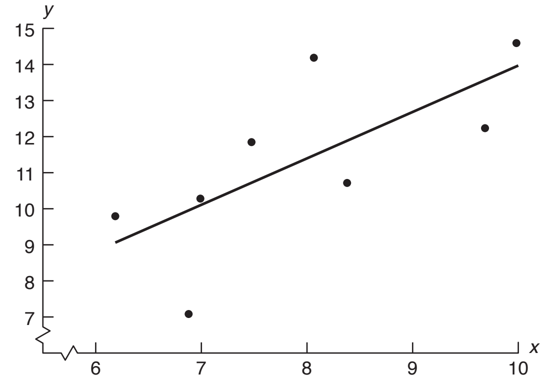
   (b) 2.13 ug/l to 2.99 ug/l. (c) $n = 48$.
2. (a) Use rounded results to compute t in part (b).
   (b) i. $\alpha = 0.05$; H_0: $\mu = 10\%$; H_1: $\mu > 10\%$.
   ii. Student's t, $d.f. = 11$; $t \approx 1.248$.
   iii. $0.100 < P\text{-value} < 0.125$; on t graph, shade area to the right of 1.248. From TI-84, P-value $\approx$ 0.1190.
   iv. P-value interval > 0.05 for α; fail to reject H_0.
   v. At the 5% level of significance, the evidence does not indicate that the patient is asymptomatic.

   (c) 9.27% to 11.71%.
3. (a) i. $\alpha = 0.05$; H_0: $p = 0.10$; H_1: $p \ne 0.10$; yes, $np > 5$ and $nq > 5$; necessary to use normal approximation to the binominal.
   ii. Standard normal; $\hat{p} \approx 0.147$; $z = 1.29$.
   iii. P-value $= 2P(z > 1.29) \approx 0.1970$; on standard normal curve, shade area to the right of 1.29 and to the left of -1.29.
   iv. P-value of $0.1970 > 0.05$ for α; fail to reject H_0.
   v. At the 5% level of significance, the data do not indicate any difference from the national average for the population proportion of crime victims.

   (b) 0.063 to 0.231. (c) From sample, $p \approx \hat{p} \approx 0.147$; $n = 193$.
4. (a) i. $\alpha = 0.05$; H_0: $\mu_d = 0$; H_1: $\mu_d \ne 0$.
   ii. Student's t, $d.f. = 6$; $\bar{d} \approx -0.0039$, $t \approx -0.771$.
   iii. $0.250 < P\text{-value} < 0.500$; on t graph, shade area to the right of 0.771 and to the left of -0.771. From TI-84, P-value ≈ 0.4699.
   iv. P-value interval > 0.05 for α; fail to reject H_0.
   v. At the 5% level of significance, the evidence does not show a population mean difference in phosphorous reduction between the two methods.
5. (a) i. $\alpha = 0.05$; H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \ne \mu_2$.
   ii. Student's t, $d.f. = 15$; $t \approx 1.952$.
   iii. $0.050 < P\text{-value} < 0.100$; on t graph, shade area to the right of 1.952 and to the left of -0.1952. From TI-84, P-value ≈ 0.0609.
   iv. P-value interval > 0.05 for α; fail to reject H_0.
   v. At the 5% level of significance, the evidence does not show any difference in the population mean proportion of on-time arrivals in summer versus winter.

   (b) -0.43% to 9.835%. (c) x_1 and x_2 distributions are approximately normal (mound-shaped and symmetric).
6. (a) i. $\alpha = 0.05$; H_0: $p_1 = p_2$; H_1: $p_1 > p_2$.
   ii. Standard normal; $\hat{p}_1 \approx 0.242$; $\hat{p}_2 \approx 0.207$; $\bar{p} \approx 0.2246$; $z \approx 0.58$.
   iii. P-value ≈ 0.2810; on standard normal curve, shade area to the right of 0.58.
   iv. P-value interval > 0.05 for α; fail to reject H_0.
   v. At the 5% level of significance, the evidence does not indicate that the population proportion of single men who go out dancing occasionally differs from the proportion of single women who do so.

   Since $n_1\bar{p}$, $n_1\bar{q}$, $n_2\bar{p}$, and $n_2\bar{q}$ are all greater than 5, the normal approximation to the binomial is justified. (b) -0.065 to 0.139.
7. (a) Essay. (b) Outline of study.
8. Answers vary.
9. (a) Blood Glucose Level



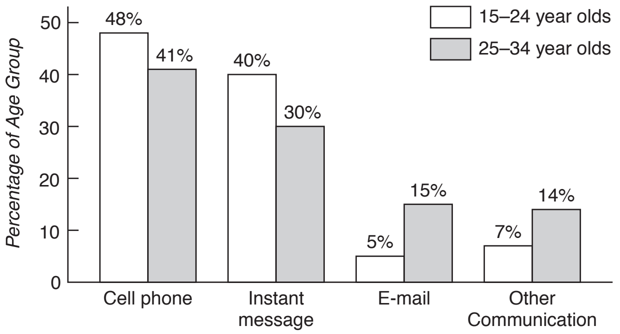
(b) $\hat{y} \approx 1.135 + 1.279x$. (c) $r \approx 0.700$; $r^2 \approx 0.490$; 49% of the variance in y is explained by the model and the variance in x. (d) 12.65; 9.64 to 15.66. (e) $\alpha = 0.01$; H_0: $\rho = 0$; H_1: $\rho \ne 0$; $r \approx 0.700$ with $t \approx 2.40$; $d.f. = 6$; $0.05 < P\text{-value} < 0.10$; do not reject H_0. At the 1% level of significance, the evidence is insufficient to conclude that there is a linear correlation. (f) $S_e \approx 1.901$; $t_c = 1.645$; 0.40 to 2.16.

CHAPTER 10

Section 10.1

1. Skewed right.
3. Right-tailed test.
5. Take random samples from each of the 4 age groups and record the number of people in each age group who recycle each of the 3 product types. Make a contingency table with age groups as labels for rows (or columns) and products as labels for columns (or rows).

7. (a) $d.f. = 6$; $0.005 < P\text{-value} < 0.01$. At the 1% level of significance, we reject H_0 since the P-value is less than 0.01. At the 1% level of significance, we conclude that the age groups differ in the proportions of who recycles each of the specified products.
(b) No. All he can say is that the 4 age groups differ in the proportions of those recycling each specified product. For this study, he cannot determine how the age groups differ regarding the proportions of those recycling the listed products.
9. (a) $\alpha = 0.05$; H_0: Myers–Briggs preference and profession are independent; H_1: Myers–Briggs preference and profession are not independent. (b) $\chi^2 = 8.649$; $d.f. = 2$. (c) $0.010 < P\text{-value} < 0.025$. From TI-84, $P\text{-value} \approx 0.0132$. (d) Reject H_0. (e) At the 5% level of significance, there is sufficient evidence to conclude that Myers–Briggs preference and profession are not independent.
11. (a) $\alpha = 0.01$; H_0: Site type and pottery type are independent; H_1: Site type and pottery type are not independent. (b) $\chi^2 = 0.5552$; $d.f. = 4$. (c) $0.950 < P\text{-value} < 0.975$. From TI-84, $P\text{-value} \approx 0.9679$.
(d) Do not reject H_0. (e) At the 1% level of significance, there is insufficient evidence to conclude that site type and pottery type are not independent.
13. (a) $\alpha = 0.05$; H_0: Age distribution and location are independent; H_1: Age distribution and location are not independent. (b) $\chi^2 = 0.6704$; $d.f. = 4$. (c) $0.950 < P\text{-value} < 0.975$. From TI-84, $P\text{-value} \approx 0.9549$.
(d) Do not reject H_0. (e) At the 5% level of significance, there is insufficient evidence to conclude that age distribution and location are not independent.
15. (a) $\alpha = 0.05$; H_0: Age of young adult and movie preference are independent; H_1: Age of young adult and movie preference are not independent. (b) $\chi^2 = 3.6230$; $d.f. = 4$. (c) $0.100 < P\text{-value} < 0.900$. From TI-84, $P\text{-value} \approx 0.4594$. (d) Do not reject H_0. (e) At the 5% level of significance, there is insufficient evidence to conclude that age of young adult and movie preference are not independent.
17. (a) $\alpha = 0.05$; H_0: Stone tool construction material and site are independent; H_1: Stone tool construction material and site are not independent. (b) $\chi^2 = 11.15$; $d.f. = 3$. (c) $0.010 < P\text{-value} < 0.025$. From TI-84, $P\text{-value} \approx 0.0110$. (d) Reject H_0. (e) At the 5% level of significance, there is sufficient evidence to conclude that stone tool construction material and site are not independent.
19. (i) Communication Preference by Percentage of Age Group



(ii) (a) H_0: The proportions of the different age groups having each communication preference are the same. H_1: The proportions of the different age groups having each communication preference are not the same. (b) $\chi^2 = 9.312$; $d.f. = 3$. (c) $0.025 < P\text{-value} < 0.050$. From TI-84, $P\text{-value} \approx 0.0254$. (d) Reject H_0. (e) At the 5% level of significance, there is sufficient evidence to conclude that the two age groups do not have the same proportions of communications preferences.

Section 10.2

1. $d.f. =$ number of categories $- 1$.
3. The greater the differences between the observed frequencies and the expected frequencies, the higher the sample χ^2 value. Greater χ^2 values lead to the conclusion that the differences between expected and observed frequencies are too large to be explained by chance alone.
5. (a) $\alpha = 0.05$; H_0: The distributions are the same; H_1: The distributions are different. (b) Sample $\chi^2 = 11.788$; $d.f. = 3$. (c) $0.005 < P\text{-value} < 0.010$.
(d) Reject H_0. (e) At the 5% level of significance, the evidence is sufficient to conclude that the age distribution of the Red Lake Village population does not fit the age distribution of the general Canadian population.
7. (a) $\alpha = 0.01$; H_0: The distributions are the same; H_1: The distributions are different. (b) Sample $\chi^2 = 0.1984$; $d.f. = 4$. (c) $P\text{-value} > 0.995$. (Note that as the χ^2 values decrease, the area in the right tail increases, so $\chi^2 < 0.207$ means that the corresponding $P\text{-value} > 0.995$.) (d) Do not reject H_0. (e) At the 1% level of significance, the evidence is insufficient to conclude that the regional distribution of raw materials does not fit the distribution at the current excavation site.
9. (i) Answers vary. (ii) (a) $\alpha = 0.01$; H_0: The distributions are the same; H_1: The distributions are different. (b) Sample $\chi^2 = 1.5693$; $d.f. = 5$.
(c) $0.900 < P\text{-value} < 0.950$. (d) Do not reject H_0.
(e) At the 1% level of significance, the evidence is insufficient to conclude that the average daily July temperature does not follow a normal distribution.
11. (a) $\alpha = 0.05$; H_0: The distributions are the same; H_1: The distributions are different. (b) Sample $\chi^2 = 9.333$; $d.f. = 3$. (c) $0.025 < P\text{-value} < 0.050$. (d) Reject H_0.
(e) At the 5% level of significance, the evidence is sufficient to conclude that the current fish distribution is different than it was 5 years ago.
13. (a) $\alpha = 0.01$; H_0: The distributions are the same; H_1: The distributions are different. (b) Sample $\chi^2 = 13.70$; $d.f. = 5$. (c) $0.010 < P\text{-value} < 0.025$. (d) Do not reject H_0. (e) At the 1% level of significance, the evidence is insufficient to conclude that the census ethnic origin distribution and the ethnic origin distribution of city residents are different.
15. (a) $\alpha = 0.01$; H_0: The distributions are the same; H_1: The distributions are different. (b) Sample $\chi^2 = 3.559$; $d.f. = 8$. (c) $0.100 < P\text{-value} < 0.900$. (d) Do not reject H_0. (e) At the 1% level of significance, the evidence is insufficient to conclude that the distribution of first nonzero digits in the accounting file does not follow Benford's Law.

17. (a) $P(0) \approx 0.179$; $P(1) \approx 0.308$; $P(2) \approx 0.265$; $P(3) \approx 0.152$; $P(r \geq 4) \approx 0.096$. (b) For $r = 0$, $E \approx 16.11$; for $r = 1$, $E \approx 27.72$; for $r = 2$, $E \approx 23.85$; for $r = 3$, $E \approx 13.68$; for $r \geq 4$, $E \approx 8.64$. (c) $\chi^2 \approx 12.55$ with $d.f. = 4$. (d) $\alpha = 0.01$; H_0: The Poisson distribution fits; H_1: The Poisson distribution does not fit; $0.01 < P\text{-value} < 0.025$; do not reject H_0. At the 1% level of significance, we cannot say that the Poisson distribution does not fit the sample data.

Section 10.3

1. Yes. No, the chi-square test of variance requires that the x distribution be a normal distribution.
3. (a) $\alpha = 0.05$; H_0: $\sigma^2 = 42.3$; H_1: $\sigma^2 > 42.3$. (b) $\chi^2 \approx 23.98$; $d.f. = 22$. (c) $0.100 < P\text{-value} < 0.900$. (d) Do not reject H_0. (e) At the 5% level of significance, there is insufficient evidence to conclude that the variance is greater in the new section. (f) $\chi^2_U = 36.78$; $\chi^2_L = 10.98$. Interval for σ^2 is from 27.57 to 92.37.
5. (a) $\alpha = 0.01$; H_0: $\sigma^2 = 136.2$; H_1: $\sigma^2 < 136.2$. (b) $\chi^2 \approx 5.92$; $d.f. = 7$. (c) Right-tailed area between 0.900 and 0.100; $0.100 < P\text{-value} < 0.900$. (d) Do not reject H_0. (e) At the 1% level of significance, there is insufficient evidence to conclude that the variance for number of mountain climber deaths is less than 136.2. (f) $\chi^2_U = 14.07$; $\chi^2_L = 2.17$. Interval for σ^2 is from 57.26 to 371.29.
7. (a) $\alpha = 0.05$; H_0: $\sigma^2 = 9$; H_1: $\sigma^2 < 9$. (b) $\chi^2 \approx 8.82$; $d.f. = 22$. (c) Right-tail area is between 0.995 and 0.990; $0.005 < P\text{-value} < 0.010$. (d) Reject H_0. (e) At the 5% level of significance, there is sufficient evidence to conclude that the variance of protection times for the new typhoid shot is less than 9. (f) $\chi^2_U = 33.92$; $\chi^2_L = 12.34$. Interval for σ is from 1.53 to 2.54.
9. (a) $\alpha = 0.01$; H_0: $\sigma^2 = 0.18$; H_1: $\sigma^2 > 0.18$. (b) $\chi^2 = 90$; $d.f. = 60$. (c) $0.005 < P\text{-value} < 0.010$. (d) Reject H_0. (e) At the 1% level of significance, there is sufficient evidence to conclude that the variance of measurements for the fan blades is higher than the specified amount. The inspector is justified in claiming that the blades must be replaced. (f) $\chi^2_U = 79.08$; $\chi^2_L = 43.19$. Interval for σ is from 0.45 mm to 0.61 mm.
11. (i) (a) $\alpha = 0.05$; H_0: $\sigma^2 = 23$; H_1: $\sigma^2 \neq 23$. (b) $\chi^2 \approx 13.06$; $d.f. = 21$. (c) The area to the left of $\chi^2 = 13.06$ is less than 50%, so we double the left-tail area to find the P-value for the two-tailed test. Right-tail area is between 0.950 and 0.900. Subtracting each value from 1, we find that the left-tail area is between 0.050 and 0.100. Doubling the left-tail area for a two-tailed test gives $0.100 < P\text{-value} < 0.200$. (d) Do not reject H_0. (e) At the 5% level of significance, there is insufficient evidence to conclude that the variance of battery lifetimes is different from 23. (ii) $\chi^2_U = 32.67$; $\chi^2_L = 11.59$. Interval for σ^2 is from 9.19 to 25.91. (iii) Interval for σ is from 3.03 to 5.09.

Section 10.4

1. Independent.
3. F distributions are not symmetrical. Values of the F distribution are all nonnegative.
5. (a) $\alpha = 0.01$; population 1 is annual production from the first plot; H_0: $\sigma_1^2 = \sigma_2^2$; H_1: $\sigma_1^2 > \sigma_2^2$;. (b) $F \approx 3.73$; $d.f._N = 15$; $d.f._D = 15$. (c) $0.001 < P\text{-value} < 0.010$. From TI-84, $P\text{-value} \approx 0.0075$. (d) Reject H_0. (e) At the 1% level of significance, there is sufficient evidence to show that the variance in annual wheat production of the first plot is greater than that of the second plot.
7. (a) $\alpha = 0.05$; population 1 has data from France; H_0: $\sigma_1^2 = \sigma_2^2$; H_1: $\sigma_1^2 \neq \sigma_2^2$. (b) $F \approx 1.97$; $d.f._N = 20$; $d.f._D = 17$. (c) $0.050 < \text{right-tail area} < 0.100$; $0.100 < P\text{-value} < 0.200$. From TI-84, $P\text{-value} \approx 0.1631$. (d) Do not reject H_0. (e) At the 5% level of significance, there is insufficient evidence to show that the variance in corporate productivity of large companies in France and of those in Germany differ. Volatility of corporate productivity does not appear to differ.
9. (a) $\alpha = 0.05$; population 1 has data from aggressive-growth companies; H_0: $\sigma_1^2 = \sigma_2^2$; H_1: $\sigma_1^2 > \sigma_2^2$. (b) $F \approx 2.54$; $d.f._N = 20$; $d.f._D = 20$. (c) $0.010 < P\text{-value} < 0.025$. From TI-84, $P\text{-value} \approx 0.0216$. (d) Reject H_0. (e) At the 5% level of significance, there is sufficient evidence to show that the variance in percentage annual returns for funds holding aggressive-growth small stocks is larger than that for funds holding value stocks.
11. (a) $\alpha = 0.05$; population 1 has data from the new system; H_0: $\sigma_1^2 = \sigma_2^2$; H_1: $\sigma_1^2 \neq \sigma_2^2$. (b) $F \approx 1.85$; $d.f._N = 30$; $d.f._D = 24$. (c) $0.050 < \text{right-tail area} < 0.100$; $0.100 < P\text{-value} < 0.200$. From TI-84, $P\text{-value} \approx 0.1266$. (d) Do not reject H_0. (e) At the 5% level of significance, there is insufficient evidence to show that the variance in gasoline consumption for the two injection systems is different.

Section 10.5

1. (a) $\alpha = 0.01$; H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: Not all the means are equal. (b–f)

Source of Variation	Sum of Squares	Degrees of Freedom	*MS*	*F* Ratio	*P*-value	Test Decision
Between groups	520.280	2	260.14	0.48	> 0.100	Do not reject H_0
Within groups	7544.190	14	538.87			
Total	8064.470	16				

From TI-84, $P\text{-value} \approx 0.6270$.

3. (a) $\alpha = 0.05$; H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$; H_1: Not all the means are equal.
(b–f)

Source of Variation	Sum of Squares	Degrees of Freedom	*MS*	*F* Ratio	*P*-value	Test Decision
Between groups	89.637	3	29.879	0.846	> 0.100	Do not reject H_0
Within groups	635.827	18	35.324			
Total	725.464	21				

From TI-84, P-value ≈ 0.4867.

5. (a) $\alpha = 0.05$; H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: Not all the means are equal.
(b–f)

Source of Variation	Sum of Squares	Degrees of Freedom	*MS*	*F* Ratio	*P*-value	Test Decision
Between groups	1303.167	2	651.58	5.005	between	Reject H_0
Within groups	1171.750	9	130.19		0.025 and 0.050	
Total	2474.917	11				

From TI-84, P-value ≈ 0.0346.

7. (a) $\alpha = 0.01$; H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: Not all the means are equal.
(b–f)

Source of Variation	Sum of Squares	Degrees of Freedom	*MS*	*F* Ratio	*P*-value	Test Decision
Between groups	2.042	2	1.021	0.336	> 0.100	Do not reject H_0
Within groups	33.428	11	3.039			
Total	35.470	13				

From TI-84, P-value ≈ 0.7217.

9. (a) $\alpha = 0.05$; H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$; H_1: Not all the means are equal.
(b–f)

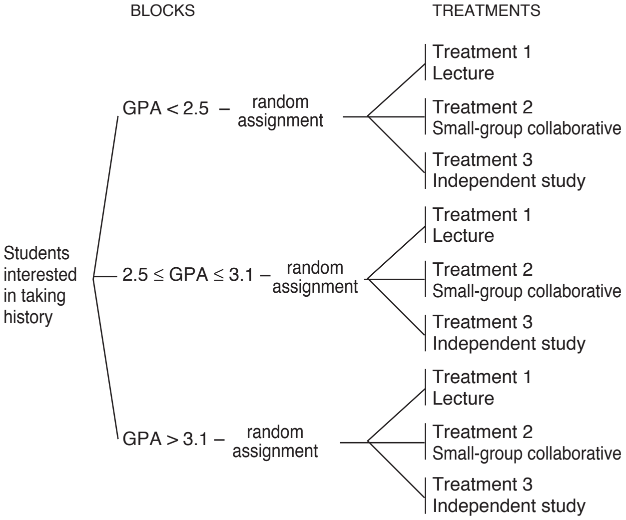
Source of Variation	Sum of Squares	Degrees of Freedom	*MS*	*F* Ratio	*P*-value	Test Decision
Between groups	238.225	3	79.408	4.611	between	Reject H_0
Within groups	258.340	15	17.223		0.010 and 0.025	
Total	496.565	18				

From TI-84, P-value ≈ 0.0177.

Section 10.6

1. Two factors; walking device with 3 levels and task with 2 levels; data table has 6 cells.
3. Since the P-value is less than 0.01, there is a significant difference in mean cadence according to the factor "walking device used."
5. (a) Two factors: income with 4 levels and media type with 5 levels. (b) $\alpha = 0.05$; For income level, H_0: There is no difference in population mean index based on income level; H_1: At least two income levels have different population mean indices; $F_{\text{income}} \approx 2.77$ with P-value ≈ 0.088. At the 5% level of significance, do not reject H_0. The data do not indicate any differences in population mean index according to income level.
(c) $\alpha = 0.05$; For media, H_0: There is no difference in population mean index according to media type; H_1: At least two media types have different population mean indices; $F_{\text{media}} \approx 0.03$ with P-value ≈ 0.998. At the 5% level of significance, do not reject H_0. The data do not indicate any differences in population mean index according to media type.

7. Randomized Block Design



Yes, the design fits the model for randomized block design.

Chapter 10 Review

1. Chi-square, F.
3. Test of homogeneity.
5. One-way ANOVA. $\alpha = 0.05$; H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$; H_1: Not all the means are equal.

Source of Variation	Sum of Squares	Degrees of Freedom	MS
Between groups	6149.75	3	2049.917
Within groups	12,454.80	16	778.425
Total	18,604.55	19	

F Ratio	P-value	Test Decision
2.633	between 0.050 and 0.100	Do not reject H_0

From TI-84, P-value ≈ 0.0854.

7. (a) Chi-square test of σ^2. (i) $\alpha = 0.01$; H_0: $\sigma^2 = 1{,}040{,}400$; H_1: $\sigma^2 > 1{,}040{,}400$. (ii) $\chi^2 \approx 51.03$; $d.f. = 29$. (iii) $0.005 < P$-value < 0.010. (iv) Reject H_0. (v) At the 1% level of significance, there is sufficient evidence to conclude that the variance is greater than claimed. (b) $\chi^2_U = 45.72$; $\chi^2_L = 16.05$; $1{,}161{,}147.4 < \sigma^2 < 3{,}307{,}642.4$.
9. Chi-square test of independence. (i) $\alpha = 0.01$; H_0: Student grade and teacher rating are independent; H_1: Student grade and teacher rating are not independent. (ii) $\chi^2 \approx 9.80$; $d.f. = 6$. (iii) $0.100 < P$-value < 0.900. From TI-84, P-value ≈ 0.1337. (iv) Do not reject H_0. (v) At the 1% level of significance, there is insufficient evidence to claim that student grade and teacher rating are not independent.
11. Chi-square test of goodness of fit. (i) $\alpha = 0.01$; H_0: The distributions are the same; H_1: The distributions are different. (ii) $\chi^2 \approx 11.93$; $d.f. = 4$. (iii) $0.010 < P$-value < 0.025. (iv) Do not reject H_0. (v) At the 1% level of significance, there is insufficient evidence to claim that the age distribution of the population of Blue Valley has changed.
13. F test for two variances. (i) $\alpha = 0.05$; H_0; $\sigma_1^2 = \sigma_2^2$; H_1: $\sigma_1^2 > \sigma_2^2$. (ii) $F \approx 2.61$; $d.f._N = 15$; $d.f._D = 17$. (iii) $0.025 < P$-value < 0.050. From TI-84, P-value ≈ 0.0302. (iv) Reject H_0. (v) At the 5% level of significance, there is sufficient evidence to show that the variance for the lifetimes of bulbs manufactured using the new process is larger than that for bulbs made by the old process.

CHAPTER 11

Section 11.1

1. Dependent (matched pairs).
3. (a) $\alpha = 0.05$; H_0: Distributions are the same; H_1: Distributions are different. (b) $x = 7/15 \approx 0.4667$; $z \approx -0.26$. (c) P-value $= 2(0.3974) = 0.7948$. (d) Do not reject H_0. (e) At the 5% level of significance, the data are not significant. The evidence is insufficient to conclude that the economic growth rates are different.
5. (a) $\alpha = 0.05$; H_0: Distributions are the same; H_1: Distributions are different. (b) $x = 10/16 = 0.625$; $z \approx 1.00$. (c) P-value $= 2(0.1587) = 0.3174$. (d) Do not reject H_0. (e) At the 5% level of significance, the data are not significant. The evidence is insufficient to conclude that the lectures had any effect on student awareness of current events.
7. (a) $\alpha = 0.05$; H_0: Distributions are the same; H_1: Distributions are different. (b) $x = 7/12 \approx 0.5833$; $z \approx 0.58$. (c) P-value $= 2(0.2810) = 0.5620$. (d) Do not reject H_0. (e) At the 5% level of significance, the data are not significant. The evidence is insufficient to conclude that the schools are not equally effective.
9. (a) $\alpha = 0.01$; H_0: Distributions are the same; H_1: Distribution after hypnosis is lower. (b) $x = 3/16 = 0.1875$; $z \approx -2.50$. (c) P-value $= 0.0062$. (d) Reject H_0. (e) At the 1% level of significance, the data are significant. The evidence is sufficient to conclude that the number of cigarettes smoked per day was less after hypnosis.
11. (a) $\alpha = 0.01$; H_0: Distributions are the same; H_1: Distributions are different. (b) $x = 10/20 = 0.5000$; $z = 0$. (c) P-value $= 2(0.5000) = 1$. (d) Do not reject H_0. (e) At the 1% level of significance, the data are not significant. The evidence is insufficient to conclude that the distribution of dropout rates is different for males and females.

Section 11.2

1. Independent.
3. (a) $\alpha = 0.05$; H_0: Distributions are the same; H_1: Distributions are different. (b) $R_A = 126$; $\mu_R = 132$;

$\sigma_R \approx 16.25$; $z \approx -0.37$. (c) P-value $\approx 2(0.3557) = 0.7114$. (d) Do not reject H_0. (e) At the 5% level of significance, the evidence is insufficient to conclude that the yield distributions for organic and conventional farming methods are different.

5. (a) $\alpha = 0.05$; H_0: Distributions are the same; H_1: Distributions are different. (b) $R_B = 148$; $\mu_R = 132$; $\sigma_R \approx 16.25$; $z \approx 0.98$. (c) P-value $\approx 2(0.1635) = 0.3270$. (d) Do not reject H_0. (e) At the 5% level of significance, the evidence is insufficient to conclude that the distributions of the training sessions are different.
7. (a) $\alpha = 0.05$; H_0: Distributions are the same; H_1: Distributions are different. (b) $R_A = 92$; $\mu_R = 132$; $\sigma_R \approx 16.25$; $z \approx -2.46$. (c) P-value $\approx 2(0.0069) = 0.0138$. (d) Reject H_0. (e) At the 5% level of significance, the evidence is sufficient to conclude that the completion time distributions for the two settings are different.
9. (a) $\alpha = 0.01$; H_0: Distributions are the same; H_1: Distributions are different. (b) $R_A = 176$; $\mu_R = 132$; $\sigma_R \approx 16.25$; $z \approx 2.71$. (c) P-value $\approx 2(0.0034) = 0.0068$. (d) Reject H_0. (e) At the 1% level of significance, the evidence is sufficient to conclude that the distributions showing percentage of exercisers differ by education level.
11. (a) $\alpha = 0.01$; H_0: Distributions are the same; H_1: Distributions are different. (b) $R_A = 166$; $\mu_R = 150$; $\sigma_R \approx 17.32$; $z \approx 0.92$. (c) P-value $\approx 2(0.1788) = 0.3576$. (d) Do not reject H_0. (e) At the 1% level of significance, the evidence is insufficient to conclude that the distributions of test scores differ according to instruction method.

Section 11.3

1. Monotone increasing.
3. (a) $\alpha = 0.05$; H_0: $\rho_s = 0$; H_1: $\rho_s \neq 0$. (b) $r_s \approx 0.682$. (c) $n = 11$; $0.01 < P$-value < 0.05. (d) Reject H_0. (e) At the 5% level of significance, we conclude that there is a monotone relationship (either increasing or decreasing) between rank in training class and rank in sales.
5. (a) $\alpha = 0.05$; H_0: $\rho_s = 0$; H_1: $\rho_s > 0$. (b) $r_s \approx 0.571$. (c) $n = 8$; P-value > 0.05. (d) Do not reject H_0. (e) At the 5% level of significance, there is insufficient evidence to indicate a monotone-increasing relationship between crowding and violence.
7. (ii) (a) $\alpha = 0.05$; H_0: $\rho_s = 0$; H_1: $\rho_s < 0$. (b) $r_s \approx -0.214$. (c) $n = 7$; P-value > 0.05. (d) Do not reject H_0. (e) At the 5% level of significance, the evidence is insufficient to conclude that there is a monotone-decreasing relationship between the ranks of humor and aggressiveness.
9. (ii) (a) $\alpha = 0.05$; H_0: $\rho_s = 0$; H_1: $\rho_s \neq 0$. (b) $r_s \approx 0.930$. (c) $n = 13$; P-value < 0.002. (d) Reject H_0. (e) At the 5% level of significance, we conclude that there is a monotone relationship between number of firefighters and number of police.
11. (ii) (a) $\alpha = 0.01$; H_0: $\rho_s = 0$; H_1: $\rho_s \neq 0$. (b) $r_s \approx 0.661$. (c) $n = 8$; $0.05 < P$-value < 0.10. (d) Do not reject H_0. (e) At the 1% level of significance, we conclude that there is insufficient evidence to reject the null hypothesis of no monotone relationship between rank of insurance sales and rank of per capita income.

Section 11.4

1. Exactly two.
3. (a) $\alpha = 0.05$; H_0: The symbols are randomly mixed in the sequence; H_1: The symbols are not randomly mixed in the sequence. (b) $R = 11$. (c) $n_1 = 12$; $n_2 = 11$; $c_1 = 7$; $c_2 = 18$. (d) Do not reject H_0. (e) At the 5% level of significance, the evidence is insufficient to conclude that the sequence of presidential party affiliations is not random.
5. (a) $\alpha = 0.05$; H_0: The symbols are randomly mixed in the sequence; H_1: The symbols are not randomly mixed in the sequence. (b) $R = 11$. (c) $n_1 = 16$; $n_2 = 7$; $c_1 = 6$; $c_2 = 16$. (d) Do not reject H_0. (e) At the 5% level of significance, the evidence is insufficient to conclude that the sequence of days for seeding and not seeding is not random.
7. (i) Median = 11.7; BBBAAAAAABBBA. (ii) (a) $\alpha = 0.05$; H_0: The numbers are randomly mixed about the median; H_1: The numbers are not randomly mixed about the median. (b) $R = 4$. (c) $n_1 = 6$; $n_2 = 6$; $c_1 = 3$; $c_2 = 11$. (d) Do not reject H_0. (e) At the 5% level of significance, the evidence is insufficient to conclude that the sequence of returns is not random about the median.
9. (i) Median = 21.6; BAAAAAABBBBB. (ii) (a) $\alpha = 0.05$; H_0: The numbers are randomly mixed about the median; H_1: The numbers are not randomly mixed about the median. (b) $R = 3$. (c) $n_1 = 6$; $n_2 = 6$; $c_1 = 3$; $c_2 = 11$. (d) Reject H_0. (e) At the 5% level of significance, we can conclude that the sequence of percentages of sand in the soil at successive depths is not random about the median.
11. (a) H_0: The symbols are randomly mixed in the sequence. H_1: The symbols are not randomly mixed in the sequence. (b) $n_1 = 21$; $n_2 = 17$; $R = 18$. (c) $\mu_R \approx 19.80$; $\sigma_R \approx 3.01$; $z \approx -0.60$. (d) Since $-1.96 < z < 1.96$, do not reject H_0; P-value $\approx 2(0.2743) = 0.5486$; at the 5% level of significance, the P-value also tells us not to reject H_0. (e) At the 5% level of significance, the evidence is insufficient to reject the null hypothesis of a random sequence of Democratic and Republican presidential terms.

Chapter 11 Review

1. No assumptions about population distributions are required.
3. (a) Rank-sum test. (b) $\alpha = 0.05$; H_0: Distributions are the same; H_1: Distributions are different. (c) $R_A = 134$; $\mu_R = 132$; $\sigma_R \approx 16.25$; $z \approx 0.12$. (d) P-value $= 2(0.4522) = 0.9044$. (e) Do not reject H_0. At the 5% level of significance, there is insufficient evidence to conclude that the viscosity index distribution has changed with use of the catalyst.
5. (a) Sign test. (b) $\alpha = 0.01$; H_0: Distributions are the same; H_1: Distribution after ads is higher. (c) $x = 0.77$; $z = 1.95$. (d) P-value $= 0.0256$. (e) Do not reject H_0.

At the 1% level of significance, the evidence is insufficient to claim that the distribution is higher after the ads.

7. (a) Spearman rank correlation coefficient test. (b) $\alpha = 0.05$; H_0: $\rho = 0$; H_1: $\rho > 0$. (c) $r_s \approx 0.617$. (d) $n = 9$; $0.025 < P\text{-value} < 0.05$. (e) Reject H_0. At the 5% level of significance, we conclude that there is a monotone-increasing relation between the ranks for the training program and the ranks on the job.
9. (a) Runs test for randomness. (b) $\alpha = 0.05$; H_0: The symbols are randomly mixed in the sequence; H_1: The symbols are not randomly mixed in the sequence. (c) $R = 7$. (d) $n_1 = 16$; $n_2 = 9$; $c_1 = 7$; $c_2 = 18$. (e) Reject H_0. At the 5% level of significance, we can conclude that the sequence of answers is not random.

CUMULATIVE REVIEW PROBLEMS

1. (a) use a calculator. (b) $P(0) \approx 0.543$; $P(1) \approx 0.331$; $P(2) \approx 0.101$; $P(3) \approx 0.025$. (c) 0.3836; $d.f. = 3$. (d) $\alpha = 0.01$; H_0: The distributions are the same; H_1: The distributions are different; $\chi^2 \approx 0.3836$; $0.900 < P\text{-value} < 0.950$; do not reject H_0. At the 1% level of significance, the evidence is insufficient to claim that the distribution does not fit the Poisson distribution.
2. $\alpha = 0.05$; H_0: Yield and fertilizer type are independent; H_1: Yield and fertilizer type are not independent; $\chi^2 \approx 5.005$; $d.f. = 4$; $0.100 < P\text{-value} < 0.900$; do not reject H_0. At the 5% level of significance, the evidence is insufficient to conclude that fertilizer type and yield are not independent.
3. (a) $\alpha = 0.05$; H_0: $\sigma = 0.55$; H_1: $\sigma > 0.55$; $s \approx 0.602$; $d.f. = 9$; $\chi^2 \approx 10.78$; $0.100 < P\text{-value} < 0.900$; do not reject H_0. At the 5% level of significance, there is insufficient evidence to conclude that the standard deviation of petal lengths is greater than 0.55. (b) Interval from 0.44 to 0.99. (c) $\alpha = 0.01$; H_0: $\sigma_1^2 = \sigma_2^2$; H_1: $\sigma_1^2 > \sigma_2^2$; $F \approx 1.95$; $d.f._N = 9$, $d.f._D = 7$; $P\text{-value} > 0.100$; do not reject H_0. At the 1% level of significance, the evidence is insufficient to conclude that the variance of the petal lengths for *Iris virginica* is greater than that for *Iris versicolor*.
4. $\alpha = 0.05$; H_0: $p = 0.5$ (wind direction distributions are the same); H_1: $p \neq 0.5$ (wind direction distributions are different); $x = 11/18$; $z \approx 0.94$; $P\text{-value} = 2(0.1736) = 0.3472$; do not reject H_0. At the 5% level of significance, the evidence is insufficient to conclude that the wind direction distributions are different.
5. $\alpha = 0.01$; H_0: Growth distributions are the same; H_1: Growth distributions are different; $\mu_R = 126.5$; $\sigma_R \approx 15.23$; $R_A = 135$; $z \approx 0.56$; $P\text{-value} = 2(0.2877) = 0.5754$; do not reject H_0. At the 1% level of significance, the evidence is insufficient to conclude that the growth distributions are different for the two root stocks.
6. (b) $\alpha = 0.05$; H_0: $\rho_s = 0$; H_1: $\rho_s \neq 0$; $r_s = 1$; $P\text{-value} < 0.002$; reject H_0. At the 5% level of significance, we can say that there is a monotone relationship between the calcium contents as measured by the labs.
7. Median = 33.45; AABBBBAAAABAABBBBA; $\alpha = 0.05$; H_0: Numbers are random about the median; H_1: Numbers are not random about the median; $R = 7$; $n_1 = n_2 = 9$; $c_1 = 5$; $c_2 = 15$; do not reject H_0. At the 5% level of significance, there is insufficient evidence to conclude that the sunspot activity about the median is not random.

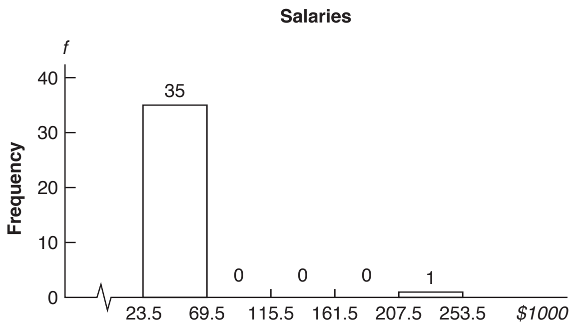
Answers to Selected Even-Numbered Problems

CHAPTER 2

Even-numbered answers not included here appear in the margins of the chapters, next to the problems.

Section 2.1

10. (a) Employee Salaries—Histogram

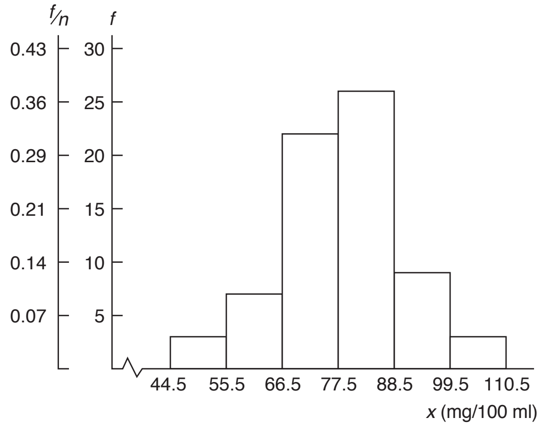


(c) Employee Salaries—Histogram

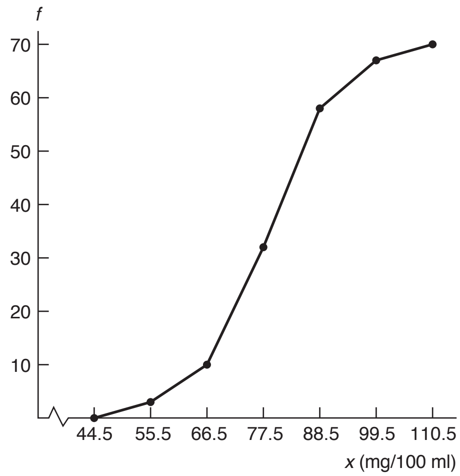
12. (a) Class width = 11.
(b)

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
45–55	44.5–55.5	50	3	0.04	3
56–66	55.5–66.5	61	7	0.10	10
67–77	66.5–77.5	72	22	0.31	32
78–88	77.5–88.5	83	26	0.37	58
89–99	88.5–99.5	94	9	0.13	67
100–110	99.5–110.5	103	3	0.04	70

(c, d) Glucose Level (mg/100 ml)—Histogram, Relative-Frequency Histogram



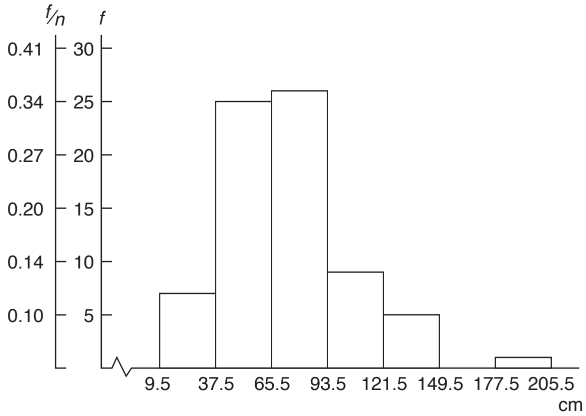
(f) Glucose Level (mg/100 ml)—Ogive



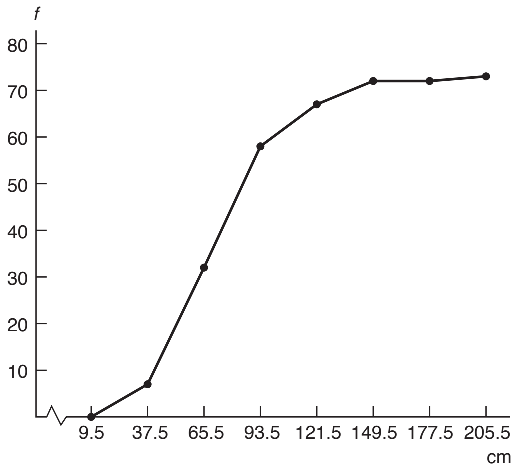
14. (a) Class width = 28.
(b)

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
10–37	9.5–37.5	23.5	7	0.10	7
38–65	37.5–65.5	51.5	25	0.34	32
66–93	65.5–93.5	79.5	26	0.36	58
94–121	93.5–121.5	107.5	9	0.12	67
122–149	121.5–149.5	135.5	5	0.07	72
150–177	149.5–177.5	163.5	0	0.00	72
178–205	177.5–205.5	191.5	1	0.01	73

(c, d) Depth of Artifacts (cm)—Histogram, Relative-Frequency Histogram



(f) Depth of Artifacts (cm)—Ogive

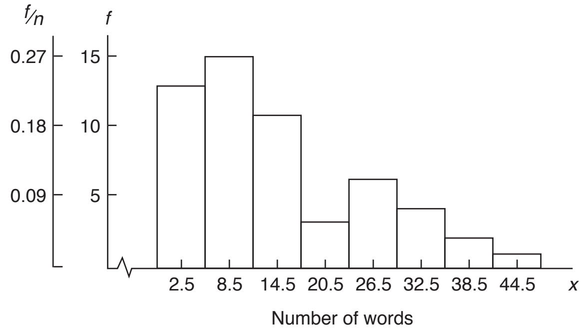


16. (a) Class width = 6.
(b)

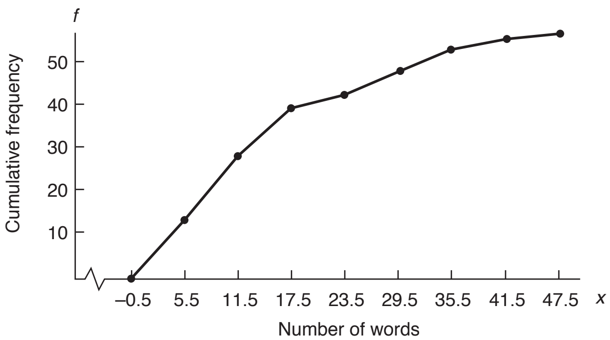
Words of Three Syllables or More

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
0–5	−0.5–5.5	2.5	13	0.24	13
6–11	5.5–11.5	8.5	15	0.27	28
12–17	11.5–17.5	14.5	11	0.20	39
18–23	17.5–23.5	20.5	3	0.05	42
24–29	23.5–29.5	26.5	6	0.11	48
30–35	29.5–35.5	32.5	4	0.07	52
36–41	35.5–41.5	38.5	2	0.04	54
42–47	41.5–47.5	44.5	1	0.02	55

(c, d) Words of Three Syllables or More—Histogram, Relative-Frequency Histogram



(f) Ogive for Words of Three Syllables or More

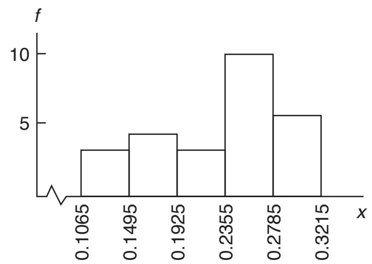


18. (b)

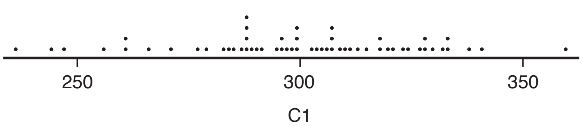
Baseball Batting Averages (class width = 0.043)

Class Limits	Class Boundaries	Midpoint	Frequency
0.107–0.149	0.1065–0.1495	0.128	3
0.150–0.192	0.1495–0.1925	0.171	4
0.193–0.235	0.1925–0.2355	0.214	3
0.236–0.278	0.2355–0.2785	0.257	10
0.279–0.321	0.2785–0.3215	0.3	6

(b, c) Baseball Batting Averages—Histogram

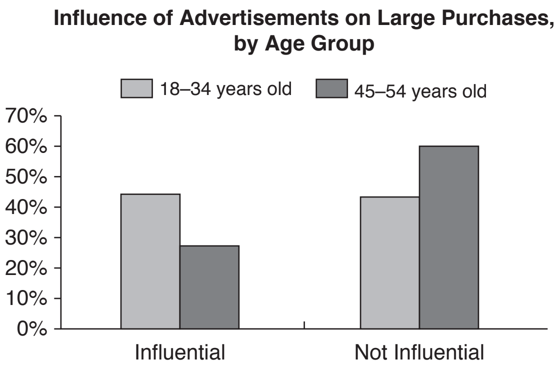


22. Dotplot for Iditarod Finish Time (in hours)

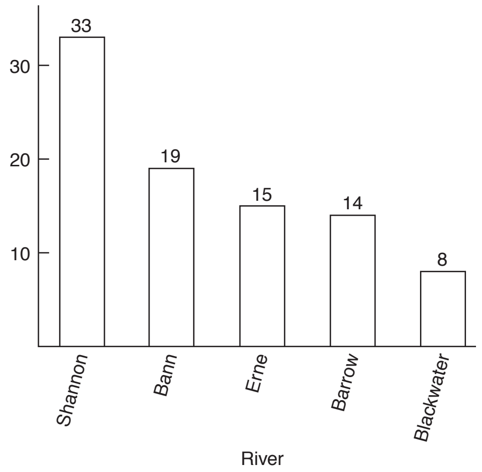


Section 2.2

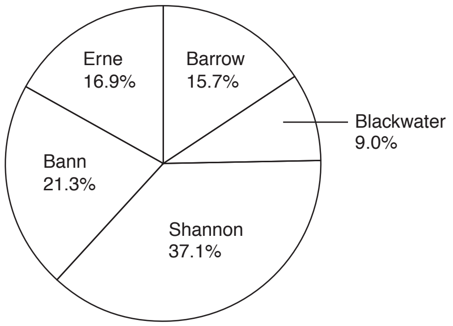
6. (b)



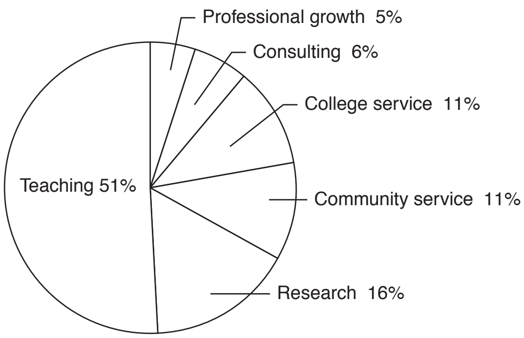
8. (a) Number of Spearheads—Pareto Chart



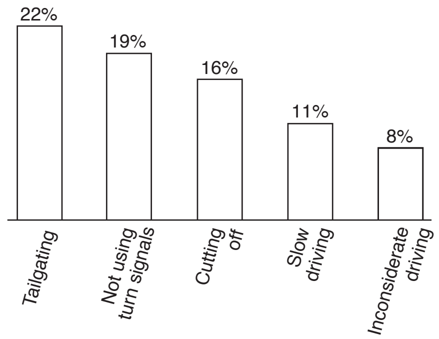
(b) Number of Spearheads—Circle Graph



10. How College Professors Spend Their Time—Circle Graph

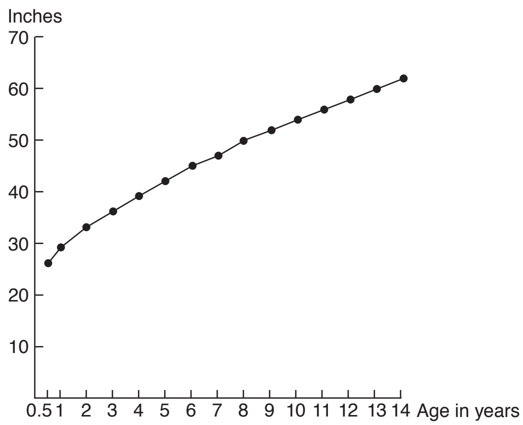


12. Driving Problems—Pareto Chart



No. The total is not 100%, and it is not clear if respondents could mark more than one problem.

14. Changes in Boys' Height with Age—Time-Series Graph



Chapter 2 Review

8. (a)

Age of DUI Arrests

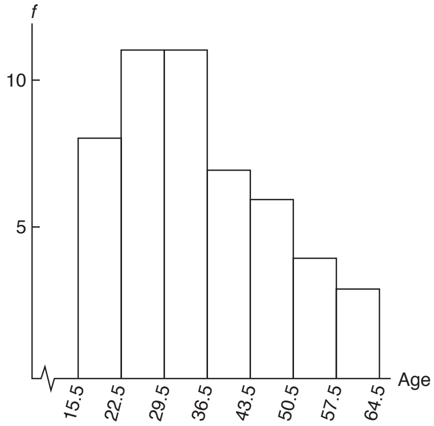
1 |	6 = 16 years
1	6 8
2	0 1 1 2 2 2 3 4 4 5 6 6 6 7 7 7 9
3	0 0 1 1 2 3 4 4 5 5 6 7 8 9
4	0 0 1 3 5 6 7 7 9 9
5	1 3 5 6 8
6	3 4

(b) Class width = 7.

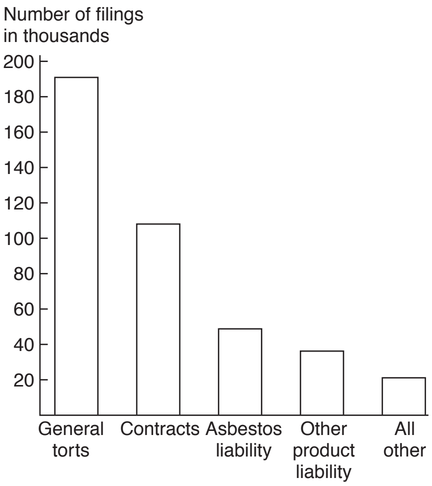
Age Distribution of DUI Arrests

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
16–22	15.5–22.5	19	8	0.16	8
23–29	22.5–29.5	26	11	0.22	19
30–36	29.5–36.5	33	11	0.22	30
37–43	36.5–43.5	40	7	0.14	37
44–50	43.5–50.5	47	6	0.12	43
51–57	50.5–57.5	54	4	0.08	47
58–64	57.5–64.5	61	3	0.06	50

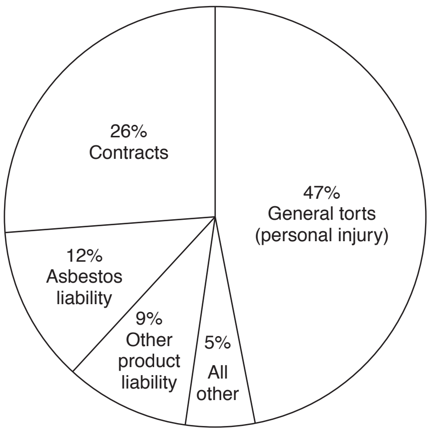
(c) Age Distribution of DUI Arrests—Histogram



10. (a) Distribution of Civil Justice Caseloads Involving Businesses—Pareto Chart



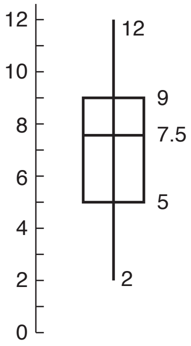
(b) Distribution of Civil Justice Caseloads Involving Businesses—Pie Chart



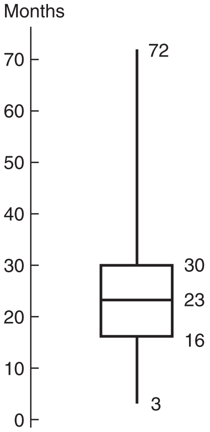
CHAPTER 3

Section 3.3

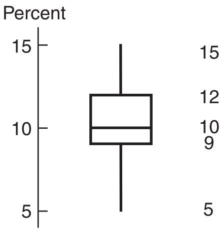
(c) Box-and-Whisker Plot



8. (a) Low = 3; Q_1 = 16; median = 23; Q_3 = 30; high = 72; IQR = 14.
Clerical Staff Length of Employment (months)

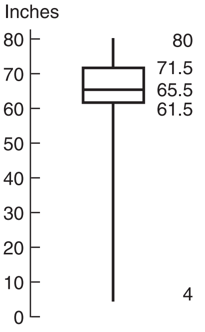


10. (a) Low = 5; Q_1 = 9; median = 10; Q_3 = 12; high = 15; IQR = 3.
(b) First quartile, since it is below Q_1.
High School Dropout Percentage by State



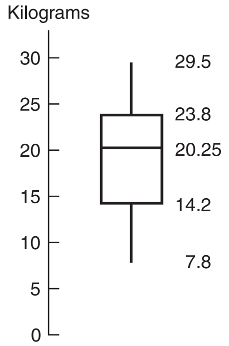
12. (a) Low value = 4; Q_1 = 61.5; median = 65.5; Q_3 = 71.5; high value = 80.
(b) IQR = 10.
(c) Lower limit = 46.5; upper limit = 86.5.

(d) Yes, the value 4 is below the lower limit and is probably an error. Our guess is that one of the students is 4 feet tall and listed height in feet instead of inches. There are no values above the upper limit.
Students' Heights (inches)

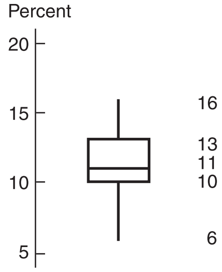


Chapter 3 Review

8. (a) Low = 7.8; Q_1 = 14.2; (kilograms) median = 20.25; Q_3 = 23.8; high = 29.5.
(b) IQR = 9.6 kilograms.
(d) Yes, the lower half shows slightly more spread.
Maize Harvest



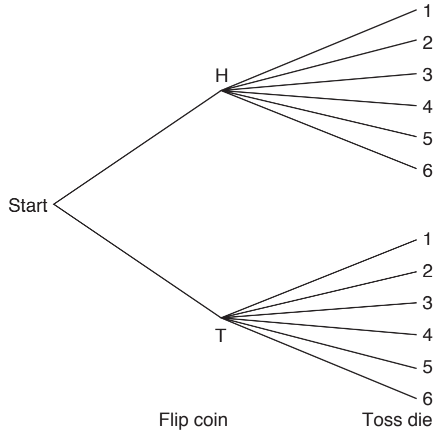
10. (a) Low = 6; Q_1 = 10; median = 11; Q_3 = 13; high = 16; IQR = 3.
Soil Water Content



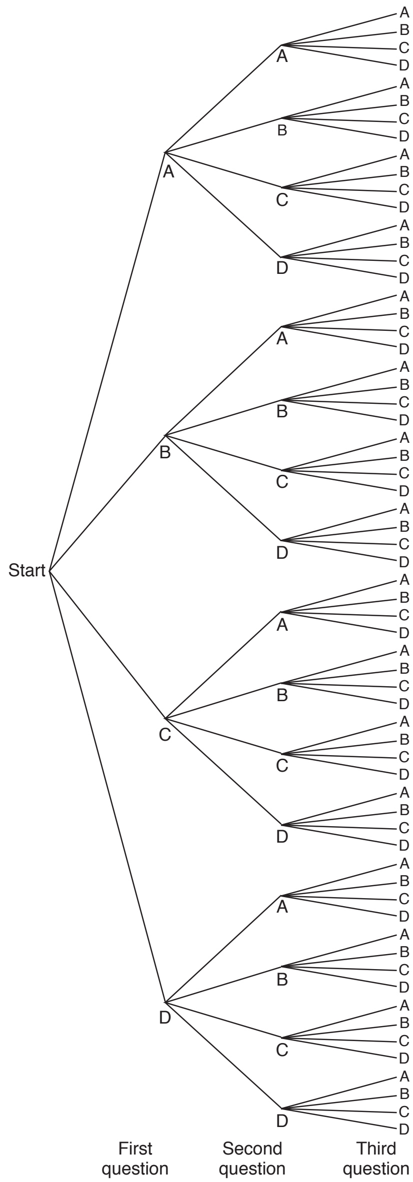
CHAPTER 4

Section 4.3

6. (a) Outcomes of Flipping a Coin and Tossing a Die

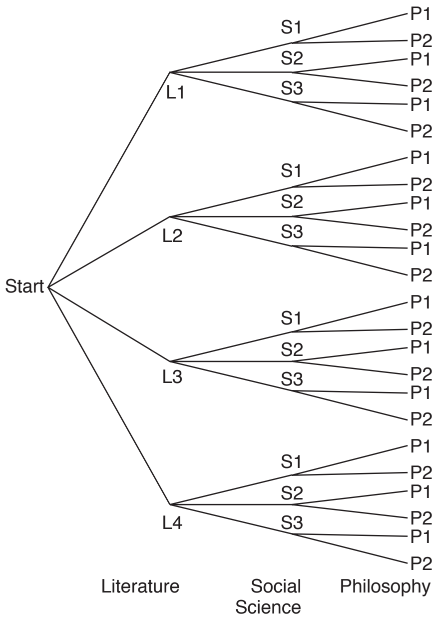


8. (a) Outcomes of Three Multiple-Choice Questions



Chapter 4 Review

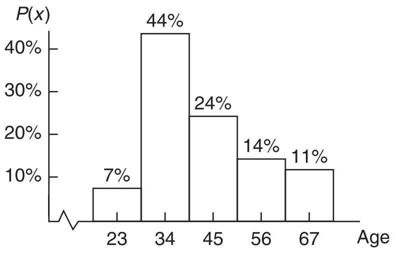
18. Ways to Satisfy Literature, Social Science, and Philosophy Requirements



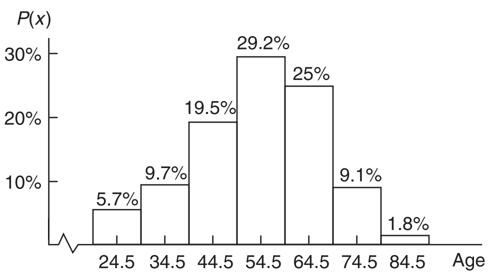
CHAPTER 5

Section 5.1

8. (b) Age of Promotion-Sensitive Shoppers

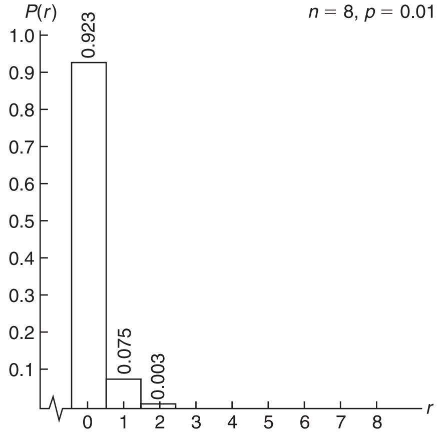


10. (b) Age of Nurses

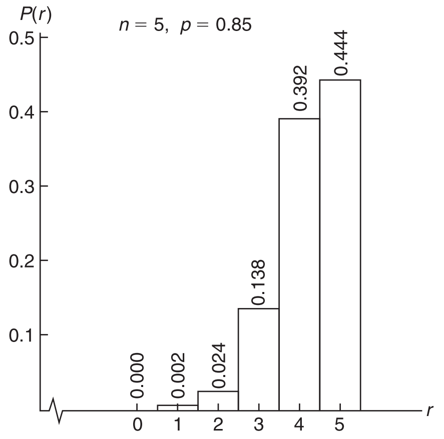


Section 5.3

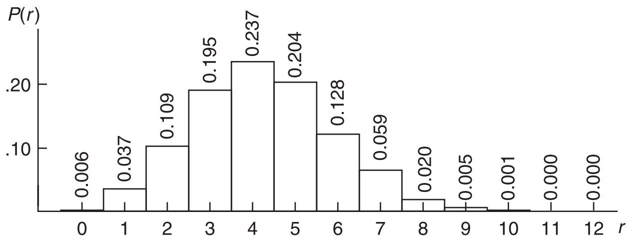
10. (a) Binomial Distribution for Number of Defective Syringes



12. (a) Binomial Distribution for Number of Automobile Damage Claims by People Under Age 25



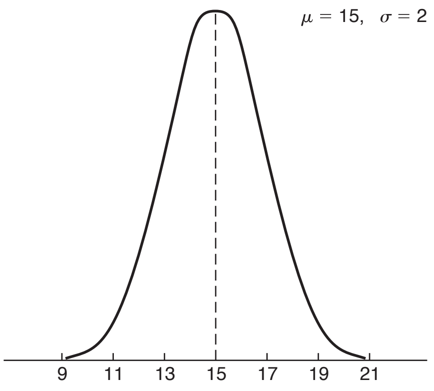
14. (a) Binomial Distribution for Drivers Who Tailgate



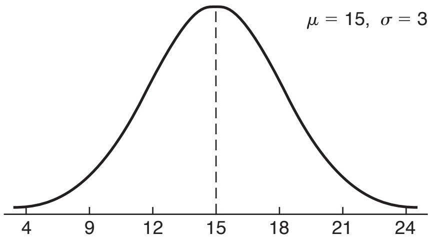
CHAPTER 6

Section 6.1

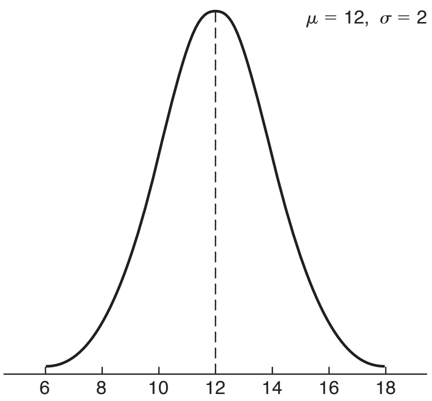
4. (a) Normal Curve



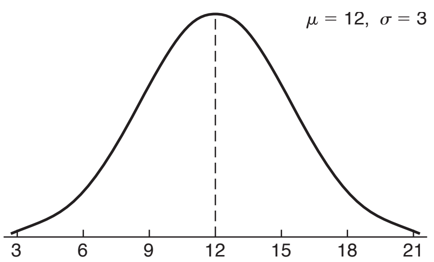
(b) Normal Curve



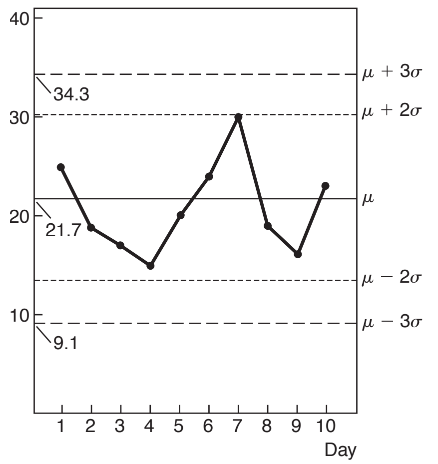
(c) Normal Curve



(d) Normal Curve

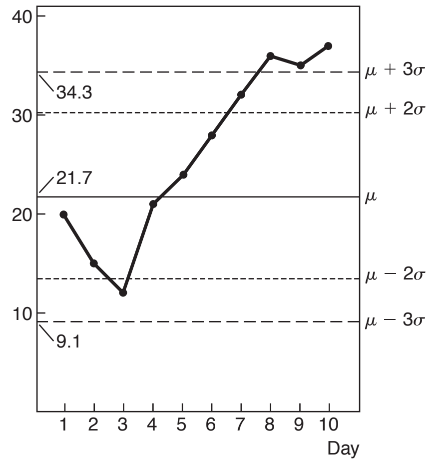


12. (a) Visitors Treated Each Day by YPMS (first period)



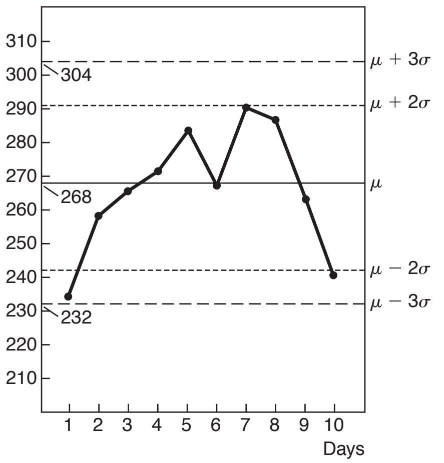
In control.

(b) Visitors Treated Each Day by YPMS (second period)

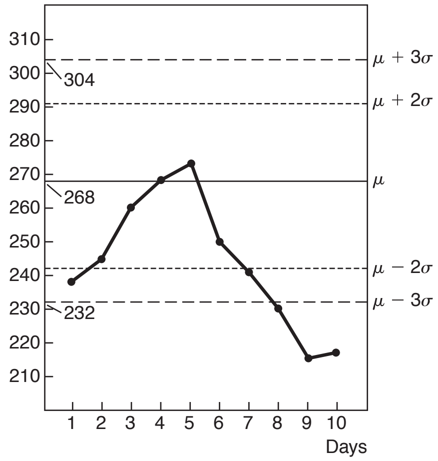


Out-of-control signals I and III are present.

14. (a) Number of Rooms Rented (first period)



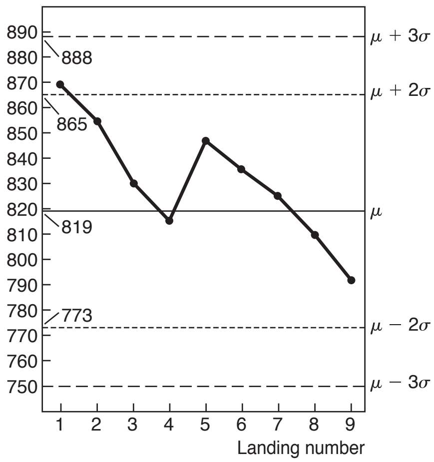
In control.
(b) Number of Rooms Rented (second period)



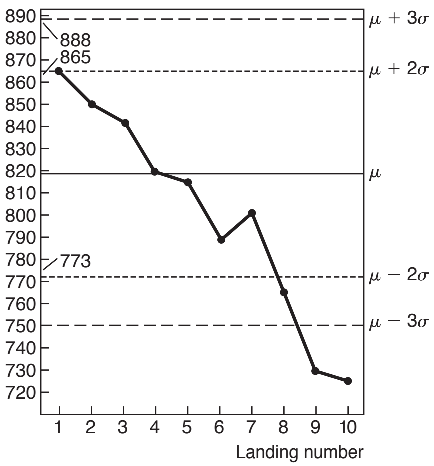
Out-of-control signals I and III are present.

Chapter 6 Review

20. (a) Hydraulic Pressure in Main Cylinder of Landing Gear of Airplanes (psi)—First Data Set



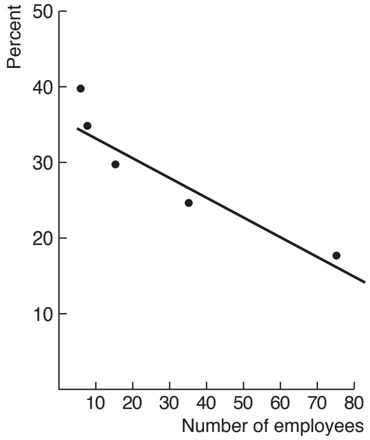
In control.
(b) Hydraulic Pressure in Main Cylinder of Landing Gear of Airplanes (psi)—Second Data Set
Out of control signals I and III are present.



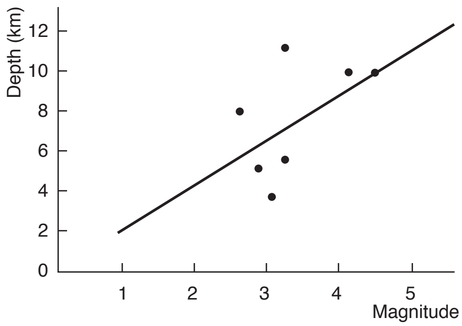
CHAPTER 9

Section 9.1

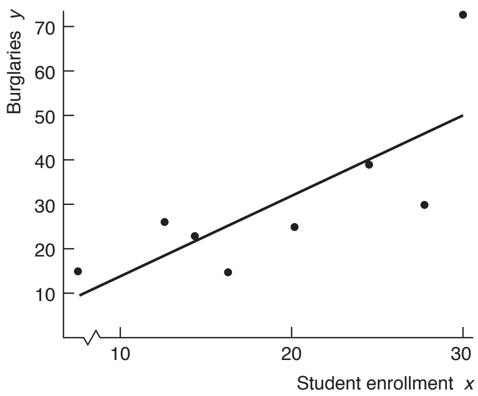
14. (a) Group Health Insurance Plans: Average Number of Employees versus Administrative Costs as a Percentage of Claims



16. (a) Magnitude (Richter Scale) and Depth (km) of Earthquakes

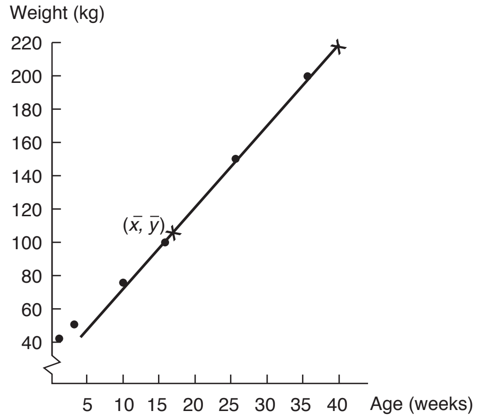


18. (a) Student Enrollment (in thousands) versus Number of Burglaries

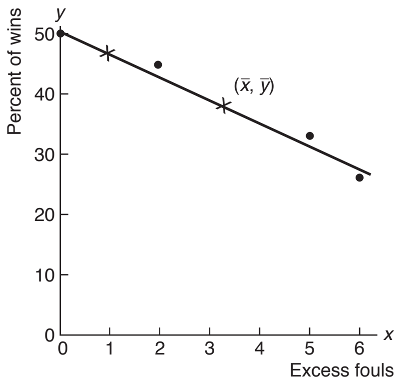


Section 9.2

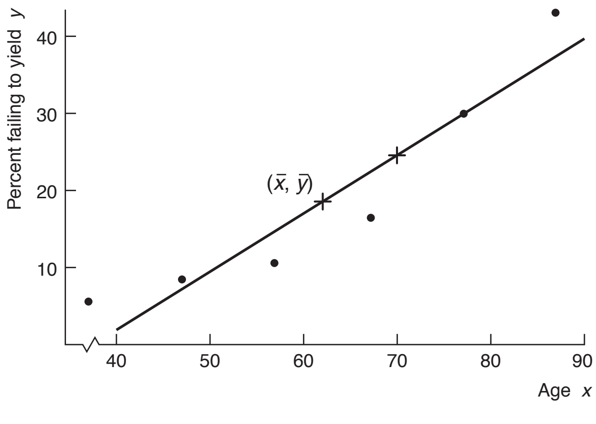
8. (a) Age and Weight of Healthy Calves



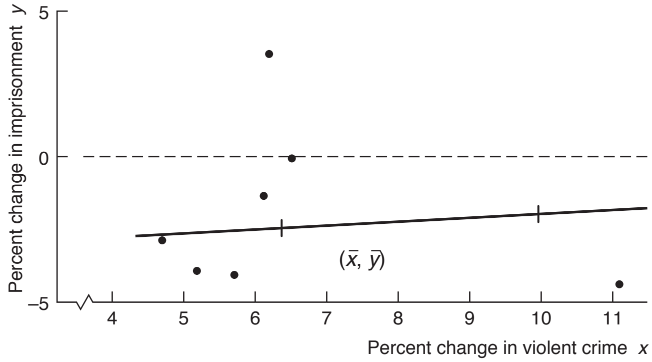
10. (a) Fouls and Basketball Wins



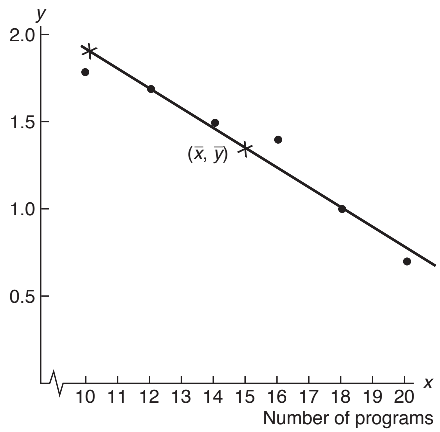
12. (a) Age and Percentage of Fatal Accidents Due to Failure to Yield



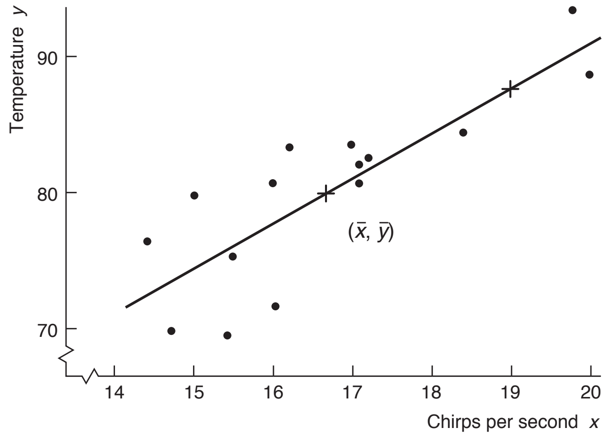
14. (a) Percent Change in Rate of Violent Crime and Percent Change in Rate of Imprisonment in U.S. Population



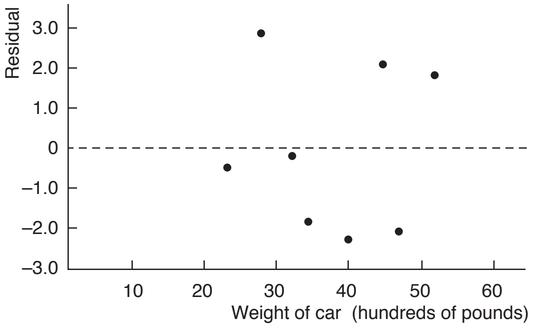
16. (a) Number of Research Programs and Mean Number of Patents per Program



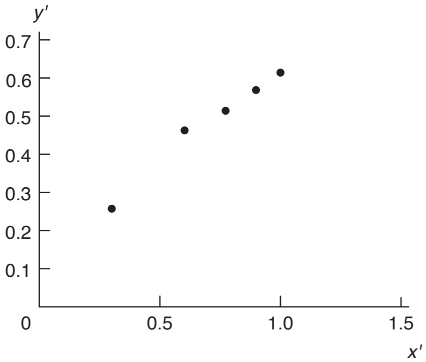
18. (a) Chirps per Second and Temperature (°F)



20. (a) Residuals: 2.9; 2.1; −0.1; −2.1; −0.5; −2.3; −1.9; 1.9.
Residual Plot

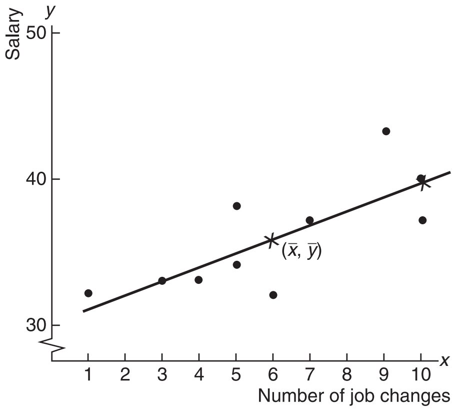


24. (a) Model with $(x'y')$ Data Pairs

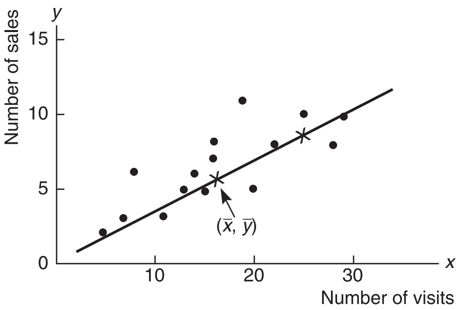


Chapter 9 Review

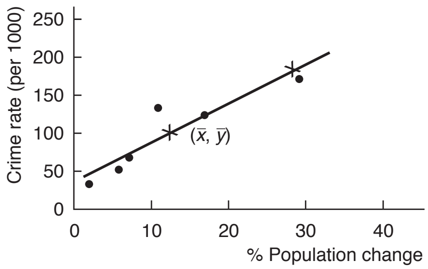
6. (a) Annual Salary (thousands) and Number of Job Changes



8. (a) Number of Insurance Sales and Number of Visits



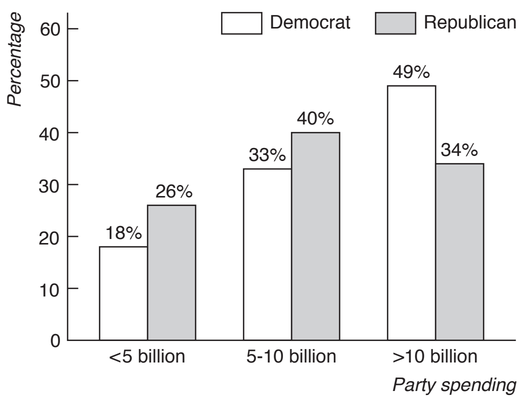
10. (a) Percent Population Change and Crime Rate



CHAPTER 10

Section 10.1

14. (i) Percentage of Each Party Spending Designated Amount



Section 10.5

2. (a) $\alpha = 0.05$; H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$; H_1: Not all the means are equal.
(b–f)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F* Ratio	*P*-value	Test Decision
Between groups	421.033	3	140.344	1.573	> 0.100	Do not reject H_0
Within groups	1516.967	17	89.233			
Total	1938.000	20				

From TI-84, P-value ≈ 0.2327.

4. (a) $\alpha = 0.01$; H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: Not all the means are equal.
(b–f)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F* Ratio	*P*-value	Test Decision
Between groups	215.680	2	107.840	0.816	> 0.100	Do not reject H_0
Within groups	1981.725	15	132.115			
Total	2197.405	17				

From TI-84, P-value ≈ 0.4608.

6. (a) $\alpha = 0.05$; H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: Not all the means are equal.
(b–f)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F* Ratio	*P*-value	Test Decision
Between groups	2.441	2	1.2207	2.95	between	Do not reject H_0
Within groups	7.448	18	0.4138		0.050 and 0.100	
Total	9.890	20				

From TI-84, P-value ≈ 0.0779.

8. (a) $\alpha = 0.05$; H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$; H_1: Not all the means are equal.
(b–f)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F* Ratio	*P*-value	Test Decision
Between groups	18.965	3	6.322	14.910	< 0.001	Reject H_0
Within groups	5.517	13	0.424			
Total	24.482	16				

From TI-84, P-value ≈ 0.0002.

Chapter 10 Review

8. One-way ANOVA. H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: Not all the means are equal.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F* Ratio	*P*-value	Test Decision
Between groups	1.002	2	0.501	0.443	> 0.100	Fail to reject H_0
Within groups	10.165	9	1.129			
Total	11.167	11				

TI-84 gives P-value ≈ 0.6651.

Index

FREQUENTLY USED FORMULAS

n = sample size $\quad N$ = population size $\quad f$ = frequency

Chapter 2

Class width $= \dfrac{\text{high} - \text{low}}{\text{number of classes}}$ (increase to next integer)

Class midpoint $= \dfrac{\text{upper limit} + \text{lower limit}}{2}$

Lower boundary = lower boundary of previous class + class width

Chapter 3

Sample mean $\bar{x} = \dfrac{\Sigma x}{n}$

Population mean $\mu = \dfrac{\Sigma x}{N}$

Weighted average $= \dfrac{\Sigma xw}{\Sigma w}$

Range = largest data value − smallest data value

Sample standard deviation $s = \sqrt{\dfrac{\Sigma(x - \bar{x})^2}{n - 1}}$

Computation formula $s = \sqrt{\dfrac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1}}$

Population standard deviation $\sigma = \sqrt{\dfrac{\Sigma(x - \mu)^2}{N}}$

Sample variance s^2

Population variance σ^2

Sample coefficient of variation $CV = \dfrac{s}{\bar{x}} \cdot 100$

Sample mean for grouped data $\bar{x} = \dfrac{\Sigma xf}{n}$

Sample standard deviation for grouped data

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2 f}{n - 1}} = \sqrt{\frac{\Sigma x^2 f - (\Sigma xf)^2/n}{n - 1}}$$

Chapter 4

Probability of the complement of event A
$P(A^c) = 1 - P(A)$

Multiplication rule for independent events
$P(A \text{ and } B) = P(A) \cdot P(B)$

General multiplication rules
$P(A \text{ and } B) = P(A) \cdot P(B|A)$
$P(A \text{ and } B) = P(B) \cdot P(A|B)$

Addition rule for mutually exclusive events
$P(A \text{ or } B) = P(A) + P(B)$

General addition rule
$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Permutation rule $P_{n,r} = \dfrac{n!}{(n - r)!}$

Combination rule $C_{n,r} = \dfrac{n!}{r!(n - r)!}$

Chapter 5

Mean of a discrete probability distribution $\mu = \Sigma x P(x)$

Standard deviation of a discrete probability distribution

$\sigma = \sqrt{\Sigma(x - \mu)^2 P(x)}$

Given $L = a + bx$

$\mu_L = a + b\mu$

$\sigma_L = |b|\sigma$

Given $W = ax_1 + bx_2$ (x_1 and x_2 independent)

$\mu_W = a\mu_1 + b\mu_2$

$\sigma_W = \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}$

For Binomial Distributions

r = number of successes; p = probability of success;
$q = 1 - p$

Binomial probability distribution $P(r) = C_{n,r}\, p^r q^{n-r}$

Mean $\mu = np$

Standard deviation $\sigma = \sqrt{npq}$

Geometric Probability Distribution

n = number of trial on which first success occurs

$P(n) = p(1 - p)^{n-1}$

Poisson Probability Distribution

r = number of successes

λ = mean number of successes over given interval

$P(r) = \dfrac{e^{-\lambda}\lambda^r}{r!}$

Chapter 6

Raw score $x = z\sigma + \mu$ $\qquad$ Standard score $z = \dfrac{x - \mu}{\sigma}$

Mean of $\bar{x}$ distribution $\mu_{\bar{x}} = \mu$

Standard deviation of $\bar{x}$ distribution $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

Standard score for $\bar{x}$ $\quad z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Mean of $\hat{p}$ distribution $\mu_{\hat{p}} = p$

Standard deviation of $\hat{p}$ distribution $\sigma_{\hat{p}} = \sqrt{\dfrac{pq}{n}}$; $q = 1 - p$

Chapter 7

Confidence Interval

for μ

$\bar{x} - E < \mu < \bar{x} + E$

where $E = z_c \dfrac{\sigma}{\sqrt{n}}$ when σ is known

$E = t_c \dfrac{s}{\sqrt{n}}$ when σ is unknown

with $d.f. = n - 1$

for p ($np > 5$ and $n(1 - p) > 5$)

$\hat{p} - E < p < \hat{p} + E$

where $E = z_c \sqrt{\dfrac{\hat{p}(1 - \hat{p})}{n}}$

$\hat{p} = \dfrac{r}{n}$

for $\mu_1 - \mu_2$ (independent samples)

$(\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$

where $E = z_c \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$ when σ_1 and σ_2 are known

$E = t_c \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$ when σ_1 or σ_2 is unknown

with $d.f.$ = smaller of $n_1 - 1$ and $n_2 - 1$

(*Note*: Software uses Satterthwaite's approximation for degrees of freedom $d.f.$)

for difference of proportions $p_1 - p_2$

$(\hat{p}_1 - \hat{p}_2) - E < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + E$

where $E = z_c \sqrt{\dfrac{\hat{p}_1\hat{q}_1}{n_1} + \dfrac{\hat{p}_2\hat{q}_2}{n_2}}$

$\hat{p}_1 = r_1/n_1; \ \hat{p}_2 = r_2/n_2$

$\hat{q}_1 = 1 - \hat{p}_1; \ \hat{q}_2 = 1 - \hat{p}_2$

Sample Size for Estimating

means $n = \left(\dfrac{z_c\sigma}{E}\right)^2$

proportions

$n = p(1 - p)\left(\dfrac{z_c}{E}\right)^2$ with preliminary estimate for p

$n = \dfrac{1}{4}\left(\dfrac{z_c}{E}\right)^2$ without preliminary estimate for p

Chapter 8

Sample Test Statistics for Tests of Hypotheses

for μ (σ known) $\quad z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

for μ (σ unknown) $\quad t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$; $d.f. = n - 1$

for p ($np > 5$ and $nq > 5$) $\quad z = \dfrac{\hat{p} - p}{\sqrt{pq/n}}$

where $q = 1 - p; \ \hat{p} = r/n$

for paired differences $d \quad t = \dfrac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$; $d.f. = n - 1$

for difference of means, σ_1 and σ_2 known

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

for difference of means, σ_1 or σ_2 unknown

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

$d.f.$ = smaller of $n_1 - 1$ and $n_2 - 1$

(*Note*: Software uses Satterthwaite's approximation for degrees of freedom $d.f.$)

for difference of proportions

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\bar{p}\,\bar{q}}{n_1} + \dfrac{\bar{p}\,\bar{q}}{n_2}}}$$

where $\bar{p} = \dfrac{r_1 + r_2}{n_1 + n_2}$ and $\bar{q} = 1 - \bar{p}$

$\hat{p}_1 = r_1/n_1; \ \hat{p}_2 = r_2/n_2$

Chapter 9

Regression and Correlation

Pearson product-moment correlation coefficient

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

Least-squares line $\hat{y} = a + bx$

where $b = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$

$a = \bar{y} - b\bar{x}$

Coefficient of determination $= r^2$

Sample test statistic for r

$t = \dfrac{r\sqrt{n - 2}}{\sqrt{1 - r^2}}$ with $d.f. = n - 2$

Standard error of estimate $S_e = \sqrt{\dfrac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n - 2}}$

Confidence interval for y

$\hat{y} - E < y < \hat{y} + E$

where $E = t_c S_e \sqrt{1 + \dfrac{1}{n} + \dfrac{n(x - \bar{x})^2}{n\Sigma x^2 - (\Sigma x)^2}}$

with $d.f. = n - 2$

Sample test statistic for slope b

$$t = \frac{b}{S_e}\sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2} \text{ with } d.f. = n - 2$$

Confidence interval for β

$$b - E < \beta < b + E$$

$$\text{where } E = \frac{t_c S_e}{\sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}} \text{ with } d.f. = n - 2$$

Chapter 10

$$\chi^2 = \Sigma \frac{(O - E)^2}{E} \text{ where}$$

O = observed frequency and

E = expected frequency

For tests of independence and tests of homogeneity

$$E = \frac{(\text{row total})(\text{column total})}{\text{sample size}}$$

For goodness of fit test E = (given percent)(sample size)

Tests of independence $d.f. = (R - 1)(C - 1)$

Test of homogeneity $d.f. = (R - 1)(C - 1)$

Goodness of fit $d.f.$ = (number of categories) − 1

Confidence interval for σ^2; $d.f. = n - 1$

$$\frac{(n-1)s^2}{\chi^2_U} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}$$

Sample test statistic for σ^2

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \text{ with } d.f. = n - 1$$

Testing Two Variances

$$\text{Sample test statistic } F = \frac{s_1^2}{s_2^2}$$

where $s_1^2 \geq s_2^2$

$d.f._N = n_1 - 1$; $d.f._D = n_2 - 1$

ANOVA

k = number of groups; N = total sample size

$$SS_{TOT} = \Sigma x^2_{TOT} - \frac{(\Sigma x_{TOT})^2}{N}$$

$$SS_{BET} = \sum_{\text{all groups}} \left(\frac{(\Sigma x_i)^2}{n_i}\right) - \frac{(\Sigma x_{TOT})^2}{N}$$

$$SS_W = \sum_{\text{all groups}} \left(\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n_i}\right)$$

$$SS_{TOT} = SS_{BET} + SS_W$$

$$MS_{BET} = \frac{SS_{BET}}{d.f._{BET}} \text{ where } d.f._{BET} = k - 1$$

$$MS_W = \frac{SS_W}{d.f._W} \text{ where } d.f._W = N - k$$

$$F = \frac{MS_{BET}}{MS_W} \text{ where } d.f. \text{ numerator} = d.f._{BET} = k - 1;$$

$$d.f. \text{ denominator} = d.f._W = N - k$$

Two-Way ANOVA

r = number of rows; c = number of columns

$$\text{Row factor } F\text{: } \frac{MS \text{ row factor}}{MS \text{ error}}$$

$$\text{Column factor } F\text{: } \frac{MS \text{ column factor}}{MS \text{ error}}$$

$$\text{Interaction } F\text{: } \frac{MS \text{ interaction}}{MS \text{ error}}$$

with degrees of freedom for

row factor $= r - 1$ interaction $= (r - 1)(c - 1)$

column factor $= c - 1$ error $= rc(n - 1)$

Chapter 11

Sample test statistic for x = proportion of plus signs to all signs ($n \geq 12$)

$$z = \frac{x - 0.5}{\sqrt{0.25/n}}$$

Sample test statistic for R = sum of ranks

$$z = \frac{R - \mu_R}{\sigma_R} \text{ where } \mu_R = \frac{n_1(n_1 + n_2 + 1)}{2} \text{ and}$$

$$\sigma_R = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$$

Spearman rank correlation coefficient

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} \text{ where } d = x - y$$

Sample test statistic for runs test

R = number of runs in sequence

Procedure for Hypothesis Testing

Use appropriate experimental design and obtain random samples of data (see Sections 1.2 and 1.3).

In the context of the application:

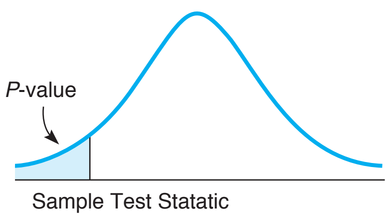
1. State the null hypothesis H_0 and the alternate hypothesis H_1. Set the level of significance α for the test.
2. Determine the appropriate sampling distribution and compute the sample test statistic.
3. Use the type of test (one-tailed or two-tailed) and the sampling distribution to compute the P-value of the corresponding sample test statistic.
4. Conclude the test. If P-value $\leq \alpha$ then reject H_0. If P-value $> \alpha$ then do not reject H_0.
5. Interpret the conclusion in the context of the application.

Finding the *P*-Value Corresponding to a Sample Test Statistic

Use the appropriate sampling distribution as described in procedure displays for each of the various tests.

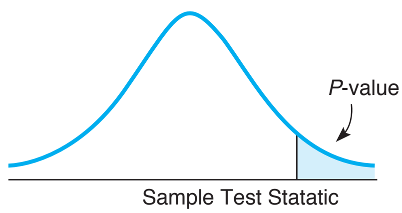
Left-Tailed Test

P-value = area to the left of the sample test statistic



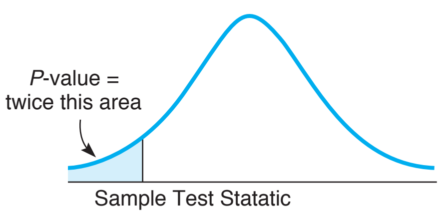
Right-Tailed Test

P-value = area to the right of the sample test statistic

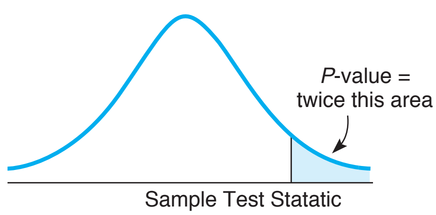


Two-Tailed Test

Sample test statistic lies to *left* of center
P-value = twice area to the left of sample test statistic



Sample test statistic lies to *right* of center
P-value = twice area to the right of sample test statistic



Sampling Distributions for Inferences Regarding μ or p

Parameter	Condition	Sampling Distribution
μ	σ is known and x has a normal distribution or $n \geq 30$	Normal distribution
μ	σ is not known and x has a normal or mound-shaped, symmetric distribution or $n \geq 30$	Student's t distribution with $d.f. = n - 1$
p	$np > 5$ and $n(1 - p) > 5$	Normal distribution