11

Goodness-of-Fit and Contingency Tables

ISBN 0-558-58875-1

CHAPT

Is the nurse a serial killer?

Three alert nurses at the Veteran's Affairs Medical Center in Northampton, Massachusetts noticed an unusually high number of deaths at times when another nurse, Kristen Gilbert, was working. Those same nurses later noticed missing supplies of the drug epinephrine, which is a synthetic adrenaline that stimulates the heart. They reported their growing concerns, and an investigation followed. Kristen Gilbert was arrested and charged with four counts of murder and two counts of attempted murder. When seeking a grand jury indictment, prosecutors provided a key piece of evidence consisting of a two-way table showing the numbers of shifts with deaths when Gilbert was working. See Table 11-1.

Table 11-1 Two-Way Table with Deaths When Gilbert Was Working

	Shifts with a death	Shifts without a death
Gilbert was working	40	217
Gilbert was not working	34	1350

The numbers in Table 11-1 might be better understood with a graph, such as Figure 11-1, which shows the death rates during shifts when Gilbert was working and when she was not working. Figure 11-1 seems to make it clear that shifts when Gilbert was working had a much higher death rate than shifts when she was not working, but we need to determine whether those results are statistically significant.

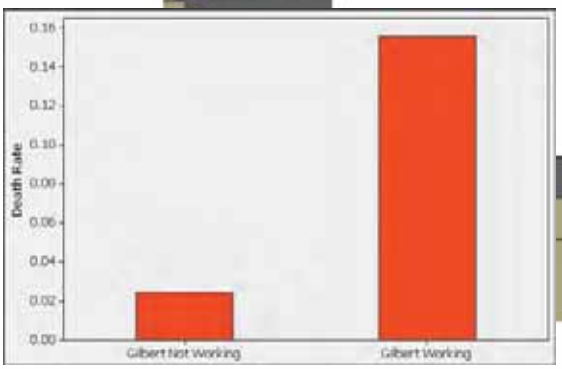


Figure 11-1 Bar Graph of Death Rates with Gilbert Working and Not Working

George Cobb, a leading statistician and statistics educator, became involved in the Gilbert case at the request of an attorney for the defense. Cobb wrote a report stating that the data in Table 11-1 should have been presented to the grand jury (as it was) for purposes of indictment, but that it should not be presented at the actual trial. He noted that the data in Table 11-1 are based on observations and do not show that Gilbert actually *caused* deaths. Also, Table 11-1 includes information about many other deaths that were not relevant to the trial. The judge ruled that the data in Table 11-1 could not be used at the trial. Kristen Gilbert was convicted on other evidence and is now serving a sentence of life in prison, without the possibility of parole.

This chapter will include methods for analyzing data in tables, such as Table 11-1. We will analyze Table 11-1 to see what conclusions could be presented to the grand jury that provided the indictment.

ISBN 0-558-58875-1

11-1 Review and Preview

We began a study of inferential statistics in Chapter 7 when we presented methods for estimating a parameter for a single population and in Chapter 8 when we presented methods of testing claims about a single population. In Chapter 9 we extended those methods to situations involving two populations. In Chapter 10 we considered methods of correlation and regression using paired sample data. In this chapter we use statistical methods for analyzing categorical (or qualitative, or attribute) data that can be separated into different cells. We consider hypothesis tests of a claim that the observed frequency counts agree with some claimed distribution. We also consider contingency tables (or two-way frequency tables), which consist of frequency counts arranged in a table with at least two rows and two columns. We conclude this chapter by considering two-way tables involving data consisting of matched pairs.

The methods of this chapter use the same χ^2 (chi-square) distribution that was first introduced in Section 7-5. See Section 7-5 for a quick review of properties of the χ^2 distribution.

11-2 Goodness-of-Fit

Key Concept In this section we consider sample data consisting of observed frequency counts arranged in a single row or column (called a one-way frequency table). We will use a hypothesis test for the claim that the observed frequency counts agree with some claimed distribution, so that there is a *good fit* of the observed data with the claimed distribution.

Because we test for how well an observed frequency distribution fits some specified theoretical distribution, the method of this section is called a *goodness-of-fit test.*

DEFINITION

A **goodness-of-fit test** is used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution.

Objective

Conduct a goodness-of-fit test.

Notation

O	represents the *observed frequency* of an outcome, found by tabulating the sample data.	k	represents the *number of different categories* or outcomes.
E	represents the *expected frequency* of an outcome, found by assuming that the distribution is as claimed.	n	represents the total *number of trials* (or observed sample values).

Requirements

1. The data have been randomly selected.

2. The sample data consist of frequency counts for each of the different categories.

ISBN 0-558-58875-1

3. For each category, the *expected* frequency is at least 5. (The expected frequency for a category is the frequency that would occur if the data actually have the distribution that is being claimed. There is no requirement that the *observed* frequency for each category must be at least 5.)

Test Statistic for Goodness-of-Fit Tests

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Critical Values

1. Critical values are found in Table A-4 by using $k - 1$ degrees of freedom, where k is the number of categories.
2. Goodness-of-fit hypothesis tests are always *right-tailed*.

P-Values

P-values are typically provided by computer software, or a range of *P*-values can be found from Table A-4.

Finding Expected Frequencies

Conducting a goodness-of-fit test requires that we identify the observed frequencies, then determine the frequencies expected with the claimed distribution. Table 11-2 on the next page includes observed frequencies with a sum of 80, so $n = 80$. If we assume that the 80 digits were obtained from a population in which all digits are equally likely, then we *expect* that each digit should occur in 1/10 of the 80 trials, so each of the 10 expected frequencies is given by $E = 8$. In general, if we are assuming that all of the expected frequencies are equal, each expected frequency is $E = n/k$, where n is the total number of observations and k is the number of categories. In other cases in which the expected frequencies are not all equal, we can often find the expected frequency for each category by multiplying the sum of all observed frequencies and the probability p for the category, so $E = np$. We summarize these two procedures here.

- **Expected frequencies are equal: $E = n/k$.**
- **Expected frequencies are not all equal: $E = np$ for each individual category.**

As good as these two preceding formulas for E might be, it is better to use an informal approach. Just ask, "How can the observed frequencies be split up among the different categories so that there is perfect agreement with the claimed distribution?" Also, note that the *observed* frequencies must all be whole numbers because they represent actual counts, but the *expected* frequencies need not be whole numbers. For example, when rolling a single die 33 times, the expected frequency for each possible outcome is $33/6 = 5.5$. The expected frequency for rolling a 3 is 5.5, even though it is impossible to have the outcome of 3 occur exactly 5.5 times.

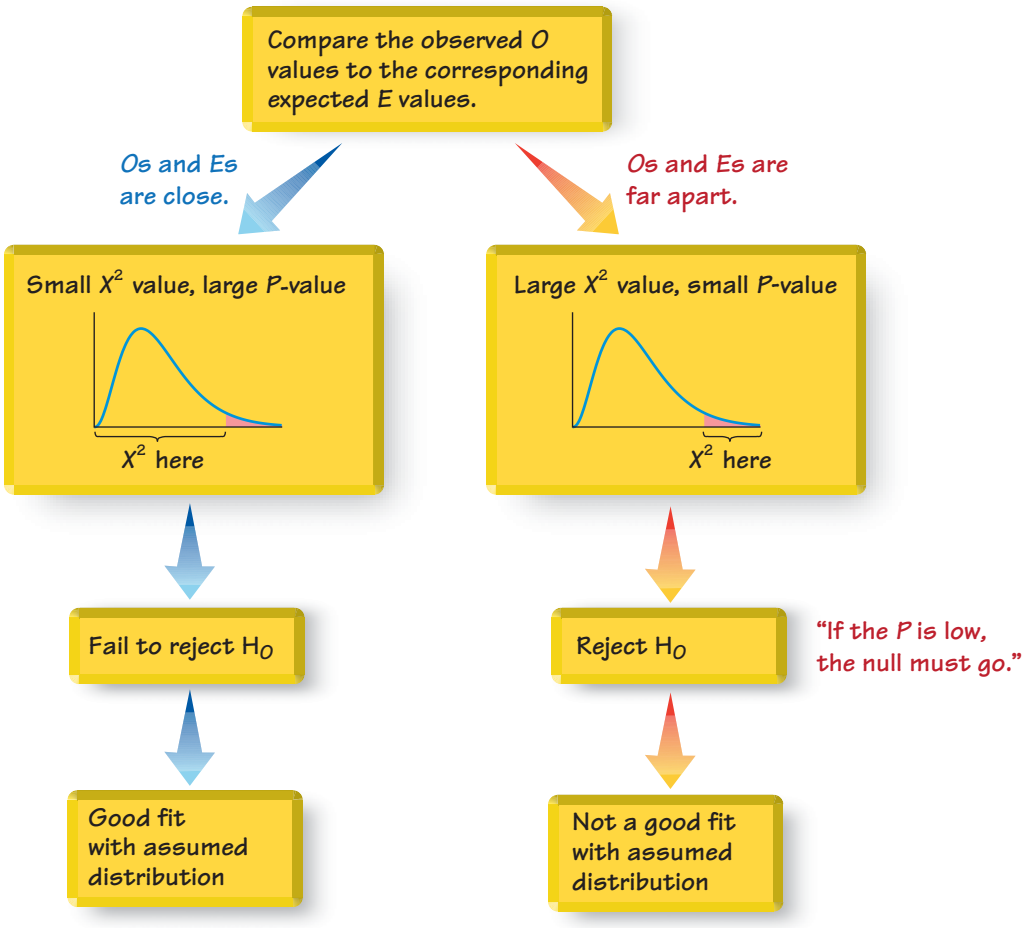
We know that sample frequencies typically deviate somewhat from the values we theoretically expect, so we now present the key question: Are the differences between the actual *observed* values O and the theoretically *expected* values E statistically significant? We need a measure of the discrepancy between the O and E values, so we use the test statistic given with the requirements and critical values. (Later, we will explain how this test statistic was developed, but you can see that it has differences of $O - E$ as a key component.)

The χ^2 test statistic is based on differences between the observed and expected values. If the observed and expected values are *close,* the χ^2 test statistic will be small and the *P*-value will be large. If the observed and expected frequencies are *not close,*

ISBN 0-558-58875-1

Figure 11-2

Relationships Among the χ^2 Test Statistic, *P*-Value, and Goodness-of-Fit



the χ^2 test statistic will be large and the *P*-value will be small. Figure 11-2 summarizes this relationship. The hypothesis tests of this section are always right-tailed, because the critical value and critical region are located at the extreme right of the distribution. If confused, just remember this:

"If the *P* is low, the null must go."
(If the *P*-value is small, reject the null hypothesis that the distribution is as claimed.)

Once we know how to find the value of the test statistic and the critical value, we can test hypotheses by using the same general procedures introduced in Chapter 8.

Table 11-2 Last Digits of Weights

Last Digit	Frequency
0	7
1	14
2	6
3	10
4	8
5	4
6	5
7	6
8	12
9	8

EXAMPLE 1 **Last Digits of Weights** Data Set 1 in Appendix B includes weights from 40 randomly selected adult males and 40 randomly selected adult females. Those weights were obtained as part of the National Health Examination Survey. When obtaining weights of subjects, it is extremely important to actually weigh individuals instead of asking them to report their weights. By analyzing the *last digits* of weights, researchers can verify that weights were obtained through actual measurements instead of being reported. When people report weights, they typically round to a whole number, so reported weights tend to have many last digits consisting of 0. In contrast, if people are actually weighed with a scale having precision to the nearest 0.1 pound, the weights tend to have last digits that are uniformly distributed, with 0, 1, 2, . . . , 9 all occurring with roughly the same frequencies. Table 11-2 shows the frequency distribution of the last digits from the

ISBN 0-558-58875-1

80 weights listed in Data Set 1 in Appendix B. (For example, the weight of 201.5 lb has a last digit of 5, and this is one of the data values included in Table 11-2.)

Test the claim that the sample is from a population of weights in which the last digits do *not* occur with the same frequency. Based on the results, what can we conclude about the procedure used to obtain the weights?

SOLUTION **REQUIREMENT CHECK** (1) The data come from randomly selected subjects. (2) The data do consist of frequency counts, as shown in Table 11-2. (3) With 80 sample values and 10 categories that are claimed to be equally likely, each expected frequency is 8, so each expected frequency does satisfy the requirement of being a value of at least 5. All of the requirements are satisfied.

The claim that the digits do not occur with the same frequency is equivalent to the claim that the relative frequencies or probabilities of the 10 cells ($p_0, p_1, \ldots, p_9$) are not all equal. We will use the traditional method for testing hypotheses (see Figure 8-9).

Step 1: The original claim is that the digits do not occur with the same frequency. That is, at least one of the probabilities $p_0, p_1, \ldots, p_9$ is different from the others.

Step 2: If the original claim is false, then all of the probabilities are the same. That is, $p_0 = p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9$.

Step 3: The null hypothesis must contain the condition of equality, so we have

$$H_0: p_0 = p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9$$

H_1: At least one of the probabilities is different from the others.

Step 4: No significance level was specified, so we select $\alpha = 0.05$.

Step 5: Because we are testing a claim about the distribution of the last digits being a uniform distribution, we use the goodness-of-fit test described in this section. The χ^2 distribution is used with the test statistic given earlier.

Step 6: The observed frequencies O are listed in Table 11-2. Each corresponding expected frequency E is equal to 8 (because the 80 digits would be uniformly distributed among the 10 categories). Table 11-3 on the next page shows the computation of the χ^2 test statistic. The test statistic is $\chi^2 = 11.250$. The critical value is $\chi^2 = 16.919$ (found in Table A-4 with $\alpha = 0.05$ in the right tail and degrees of freedom equal to $k - 1 = 9$). The test statistic and critical value are shown in Figure 11-3 on the next page.

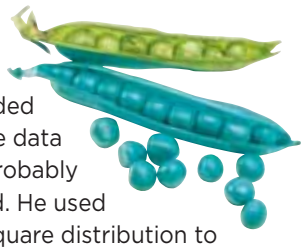
Step 7: Because the test statistic does not fall in the critical region, there is not sufficient evidence to reject the null hypothesis.

Step 8: There is not sufficient evidence to support the claim that the last digits do not occur with the same relative frequency.

INTERPRETATION This goodness-of-fit test suggests that the last digits provide a reasonably good fit with the claimed distribution of equally likely frequencies. Instead of asking the subjects how much they weigh, it appears that their weights were actually measured as they should have been.

Example 1 involves a situation in which the claimed frequencies for the different categories are all equal. The methods of this section can also be used when the hypothesized probabilities (or frequencies) are different, as shown in Example 2.

Mendel's Data Falsified?



Because some of Mendel's data from his famous genetics experiments seemed too perfect to be true, statistician R. A. Fisher concluded that the data were probably falsified. He used a chi-square distribution to show that when a test statistic is extremely far to the left and results in a *P*-value very close to 1, the sample data fit the claimed distribution almost perfectly, and this is evidence that the sample data have not been randomly selected. It has been suggested that Mendel's gardener knew what results Mendel's theory predicted, and subsequently adjusted results to fit that theory.

Ira Pilgrim wrote in *The Journal of Heredity* that this use of the chi-square distribution is not appropriate. He notes that the question is not about goodness-of-fit with a particular distribution, but whether the data are from a sample that is truly random. Pilgrim used the binomial probability formula to find the probabilities of the results obtained in Mendel's experiments. Based on his results, Pilgrim concludes that "there is no reason whatever to question Mendel's honesty." It appears that Mendel's results are not too good to be true, and they could have been obtained from a truly random process.

ISBN 0-558-58875-1

Which Car Seats Are Safest?

Many people believe that the back seat of a car is the safest place to sit, but is it?

University of Buffalo researchers analyzed more than 60,000 fatal car crashes and found that the middle back seat is the safest place to sit in a car. They found that sitting in that seat makes a passenger 86% more likely to survive than those who sit in the front seats, and they are 25% more likely to survive than those sitting in either of the back seats nearest the windows. An analysis of seat belt use showed that when not wearing a seat belt in the back seat, passengers are three times more likely to die in a crash than those wearing seat belts in that same seat. Passengers concerned with safety should sit in the middle back seat wearing a seat belt.

Table 11-3 Calculating the χ^2 Test Statistic for the Last Digits of Weights

Last Digit	Observed Frequency O	Expected Frequency E	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
0	7	8	−1	1	0.125
1	14	8	6	36	4.500
2	6	8	−2	4	0.500
3	10	8	2	4	0.500
4	8	8	0	0	0.000
5	4	8	−4	16	2.000
6	5	8	−3	9	1.125
7	6	8	−2	4	0.500
8	12	8	4	16	2.000
9	8	8	0	0	0.000

$$\chi^2 = \sum \frac{(O - E)^2}{E} = 11.250$$

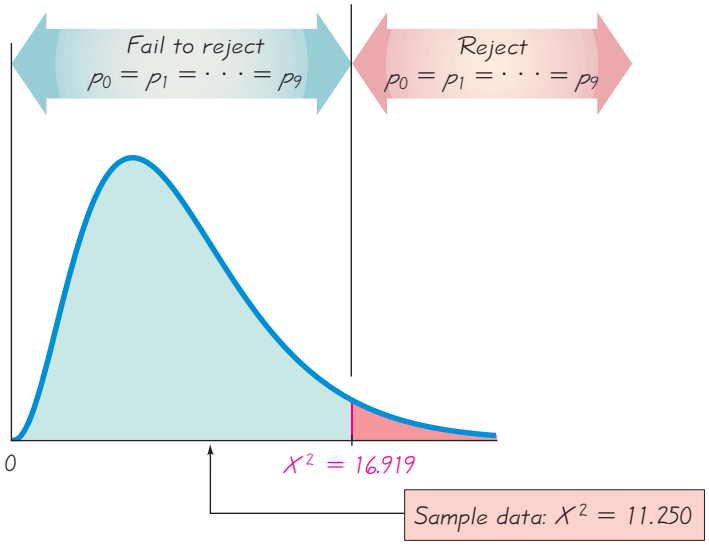


Figure 11-3 Test of $p_0 = p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9$

EXAMPLE 2 **World Series Games** Table 11-4 lists the numbers of games played in the baseball World Series, as of this writing. That table also includes the expected proportions for the numbers of games in a World Series, assuming that in each series, both teams have about the same chance of winning. Use a 0.05 significance level to test the claim that the actual numbers of games fit the distribution indicated by the probabilities.

Table 11-4 Numbers of Games in World Series Contests

Games played	4	5	6	7
Actual World Series contests	19	21	22	37
Expected proportion	2/16	4/16	5/16	5/16

ISBN 0-558-58875-1

SOLUTION

REQUIREMENT CHECK (1) We begin by noting that the observed numbers of games are not randomly selected from a larger population. However, we treat them as a random sample for the purpose of determining whether they are typical results that might be obtained from such a random sample. (2) The data do consist of frequency counts. (3) Each expected frequency is at least 5, as will be shown later in this solution. All of the requirements are satisfied.

Step 1: The original claim is that the actual numbers of games fit the distribution indicated by the expected proportions. Using subscripts corresponding to the number of games, we can express this claim as $p_4 = 2/16$ and $p_5 = 4/16$ and $p_6 = 5/16$ and $p_7 = 5/16$.

Step 2: If the original claim is false, then at least one of the proportions does not have the value as claimed.

Step 3: The null hypothesis must contain the condition of equality, so we have

H_0: $p_4 = 2/16$ and $p_5 = 4/16$ and $p_6 = 5/16$ and $p_7 = 5/16$.

H_1: At least one of the proportions is not equal to the given claimed value.

Step 4: The significance level is $\alpha = 0.05$.

Step 5: Because we are testing a claim that the distribution of numbers of games in World Series contests is as claimed, we use the goodness-of-fit test described in this section. The χ^2 distribution is used with the test statistic given earlier.

Step 6: Table 11-5 shows the calculations resulting in the test statistic of $\chi^2 = 7.885$. The critical value is $\chi^2 = 7.815$ (found in Table A-4 with $\alpha = 0.05$ in the right tail and degrees of freedom equal to $k - 1 = 3$). The Minitab display shows the value of the test statistic as well as the *P*-value of 0.048.

MINITAB

Category	Observed	Test Proportion	Expected	Contribution to Chi-Sq
1	19	0.1250	12.3750	3.54672
2	21	0.2500	24.7500	0.56818
3	22	0.3125	30.9375	2.58194
4	37	0.3125	30.9375	1.18801

N	DF	Chi-Sq	P-Value
99	3	7.88485	0.048

Table 11-5 Calculating the χ^2 Test Statistic for the Numbers of World Series Games

Number of Games	Observed Frequency *O*	Expected Frequency $E = np$	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
4	19	$99 \cdot \frac{2}{16} = 12.3750$	6.6250	43.8906	3.5467
5	21	$99 \cdot \frac{4}{16} = 24.7500$	−3.7500	14.0625	0.5682
6	22	$99 \cdot \frac{5}{16} = 30.9375$	−8.9375	79.8789	2.5819
7	37	$99 \cdot \frac{5}{16} = 30.9375$	6.0625	36.7539	1.1880

$$\chi^2 = \sum \frac{(O - E)^2}{E} = 7.885$$

STATISTICS IN THE NEWS

Which Airplane Seats Are Safest?

Because most crashes occur during takeoff or landing, passengers can improve their safety by flying non-stop. Also, larger planes are safer.

Many people believe that the rear seats are safest in an airplane crash. Todd Curtis is an aviation safety expert who maintains a database of airline incidents, and he says that it is not possible to conclude that some seats are safer than others. He says that each crash is unique, and there are far too many variables to consider. Also, Matt McCormick, a survival expert for the National Transportation Safety Board, told *Travel* magazine that "there is no one safe place to sit."

Goodness-of-fit tests can be used with a null hypothesis that all sections of an airplane are equally safe. Crashed airplanes could be divided into the front, middle, and rear sections. The observed frequencies of fatalities could then be compared to the frequencies that would be expected with a uniform distribution of fatalities. The χ^2 test statistic reflects the size of the discrepancies between observed and expected frequencies, and it would reveal whether some sections are safer than others.

ISBN 0-558-58875-1

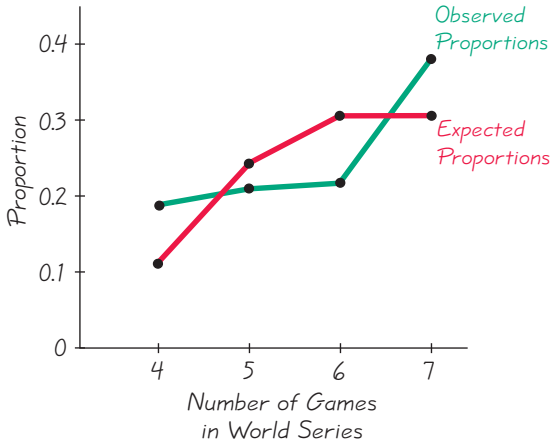
Step 7: The *P*-value of 0.048 is less than the significance level of 0.05, so there is sufficient evidence to reject the null hypothesis. (Also, the test statistic of $\chi^2 = 7.885$ is in the critical region bounded by the critical value of 7.815, so there is sufficient evidence to reject the null hypothesis.)

Step 8: There is sufficient evidence to warrant rejection of the claim that actual numbers of games in World Series contests fit the distribution indicated by the expected proportions given in Table 11-4.

INTERPRETATION This goodness-of-fit test suggests that the numbers of games in World Series contests do not fit the distribution expected from probability calculations. Different media reports have noted that seven-game series occur much more than expected. The results in Table 11-4 show that seven-game series occurred 37% of the time, but they were expected to occur only 31% of the time. (A *USA Today* headline stated that "Seven-game series defy odds.") So far, no reasonable explanations have been provided for the discrepancy.

In Figure 11-4 we graph the expected proportions of 2/16, 4/16, 5/16, and 5/16 along with the observed proportions of 19/99, 21/99, 22/99, and 37/99, so that we can visualize the discrepancy between the distribution that was claimed and the frequencies that were observed. The points along the red line represent the expected proportions, and the points along the green line represent the observed proportions. Figure 11-4 shows disagreement between the expected proportions (red line) and the observed proportions (green line), and the hypothesis test in Example 2 shows that the discrepancy is statistically significant.

Figure 11-4
Observed and Expected Proportions in the Numbers of World Series Games



P-Values

Computer software automatically provides *P*-values when conducting goodness-of-fit tests. If computer software is unavailable, a range of *P*-values can be found from Table A-4. Example 2 resulted in a test statistic of $\chi^2 = 7.885$, and if we refer to Table A-4 with 3 degrees of freedom, we find that the test statistic of 7.885 lies between the table values of 7.815 and 9.348. So, the *P*-value is between 0.025 and 0.05. In this case, we might state that "*P*-value < 0.05." The Minitab display shows that the *P*-value is 0.048. Because the *P*-value is less than the significance level of 0.05, we reject the null hypothesis. Remember, "if the *P* (value) is low, the null must go."

Rationale for the Test Statistic: Examples 1 and 2 show that the χ^2 test statistic is a measure of the discrepancy between observed and expected frequencies. Simply summing the differences between observed and expected values does not result in an

ISBN 0-558-58875-1

effective measure because that sum is always 0. Squaring the $O - E$ values provides a better statistic. (The reasons for squaring the $O - E$ values are essentially the same as the reasons for squaring the $x - \bar{x}$ values in the formula for standard deviation.) The value of $\Sigma(O - E)^2$ measures only the magnitude of the differences, but we need to find the magnitude of the differences relative to what was expected. This relative magnitude is found through division by the expected frequencies, as in the test statistic.

The theoretical distribution of $\Sigma(O - E)^2/E$ is a discrete distribution because the number of possible values is finite. The distribution can be approximated by a chi-square distribution, which is continuous. This approximation is generally considered acceptable, provided that all expected values E are at least 5. (There are ways of circumventing the problem of an expected frequency that is less than 5, such as combining categories so that all expected frequencies are at least 5. Also, there are other methods that can be used when not all expected frequencies are at least 5.)

The number of degrees of freedom reflects the fact that we can freely assign frequencies to $k - 1$ categories before the frequency for every category is determined. (Although we say that we can "freely" assign frequencies to $k - 1$ categories, we cannot have negative frequencies nor can we have frequencies so large that their sum exceeds the total of the observed frequencies for all categories combined.)

USING TECHNOLOGY

STATDISK First enter the observed frequencies in the first column of the Data Window. If the expected frequencies are not all equal, enter a second column that includes either expected proportions or actual expected frequencies. Select **Analysis** from the main menu bar, then select the option **Goodness-of-Fit.** Choose between "equal expected frequencies" and "unequal expected frequencies" and enter the data in the dialog box, then click on **Evaluate.**

MINITAB Enter observed frequencies in column C1. If the expected frequencies are not all equal, enter them as proportions in column C2. Select **Stat, Tables,** and **Chi-Square Goodness-of-Fit Test.** Make the entries in the window and click on **OK.**

EXCEL First enter the category names in one column, enter the observed frequencies in a second column, and use a third column to enter the expected *proportions* in decimal form (such as 0.20, 0.25, 0.25, and 0.30). If using Excel 2007, click on **Add-Ins,** then click on **DDXL;** if using Excel 2003, click on **DDXL.** Select the menu item of **Tables.** In the menu labeled **Function Type,** select **Goodness-of-Fit.** Click on the pencil icon for Category Names and enter the range of cells containing the category names, such as A1:A5. Click on the pencil icon for Observed Counts and enter the range of cells containing the observed frequencies, such as B1:B5. Click on the pencil icon for Test Distribution and enter the range of cells containing the expected proportions in decimal form, such as C1:C5. Click **OK** to get the chi-square test statistic and the *P*-value.

TI-83/84 PLUS Enter the observed frequencies in list L1, then identify the expected frequencies and enter them in list L2. With a TI-84 Plus calculator, press STAT, select **TESTS,** select χ^2 **GOF-Test,** then enter L1 and L2 and the number of degrees of freedom when prompted. (The number of degrees of freedom is 1 less than the number of categories.) With a TI-83 Plus calculator, use the program **X2GOF.** Press PRGM, select **X2GOF,** then enter L1 and L2 when prompted. Results will include the test statistic and *P*-value.

11-2 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

1. Goodness-of-Fit A *New York Times*/CBS News Poll typically involves the selection of random digits to be used for telephone numbers. The *New York Times* states that "within each (telephone) exchange, random digits were added to form a complete telephone number, thus permitting access to listed and unlisted numbers." When such digits are randomly generated, what is the distribution of those digits? Given such randomly generated digits, what is a test for "goodness-of-fit"?

ISBN 0-558-58875-1

2. Interpreting Values of χ^2 When generating random digits as in Exercise 1, we can test the generated digits for goodness-of-fit with the distribution in which all of the digits are equally likely. What does an exceptionally large value of the χ^2 test statistic suggest about the goodness-of-fit? What does an exceptionally small value of the χ^2 test statistic (such as 0.002) suggest about the goodness-of-fit?

3. Observed/Expected Frequencies A wedding caterer randomly selects clients from the past few years and records the months in which the wedding receptions were held. The results are listed below (based on data from *The Amazing Almanac*). Assume that you want to test the claim that weddings occur in different months with the same frequency. Briefly describe what O and E represent, then find the values of O and E.

Month	Jan.	Feb.	March	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Number	5	8	7	9	13	17	11	10	10	12	8	10

4. *P*-Value When using the data from Exercise 3 to conduct a hypothesis test of the claim that weddings occur in the 12 months with equal frequency, we obtain the P-value of 0.477. What does that P-value tell us about the sample data? What conclusion should be made?

In Exercises 5–20, conduct the hypothesis test and provide the test statistic, critical value and/or P-value, and state the conclusion.

5. Testing a Slot Machine The author purchased a slot machine (Bally Model 809), and tested it by playing it 1197 times. There are 10 different categories of outcome, including no win, win jackpot, win with three bells, and so on. When testing the claim that the observed outcomes agree with the expected frequencies, the author obtained a test statistic of $\chi^2 = 8.185$. Use a 0.05 significance level to test the claim that the actual outcomes agree with the expected frequencies. Does the slot machine appear to be functioning as expected?

6. Grade and Seating Location Do "A" students tend to sit in a particular part of the classroom? The author recorded the locations of the students who received grades of A, with these results: 17 sat in the front, 9 sat in the middle, and 5 sat in the back of the classroom. When testing the assumption that the "A" students are distributed evenly throughout the room, the author obtained the test statistic of $\chi^2 = 7.226$. If using a 0.05 significance level, is there sufficient evidence to support the claim that the "A" students are not evenly distributed throughout the classroom? If so, does that mean you can increase your likelihood of getting an A by sitting in the front of the room?

7. Pennies from Checks When considering effects from eliminating the penny as a unit of currency in the United States, the author randomly selected 100 checks and recorded the cents portions of those checks. The table below lists those cents portions categorized according to the indicated values. Use a 0.05 significance level to test the claim that the four categories are equally likely. The author expected that many checks for whole dollar amounts would result in a disproportionately high frequency for the first category, but do the results support that expectation?

Cents portion of check	0–24	25–49	50–74	75–99
Number	61	17	10	12

8. Flat Tire and Missed Class A classic tale involves four carpooling students who missed a test and gave as an excuse a flat tire. On the makeup test, the instructor asked the students to identify the particular tire that went flat. If they really didn't have a flat tire, would they be able to identify the same tire? The author asked 41 other students to identify the tire they would select. The results are listed in the following table (except for one student who selected the spare). Use a 0.05 significance level to test the author's claim that the results fit a uniform distribution. What does the result suggest about the ability of the four students to select the same tire when they really didn't have a flat?

Tire	Left front	Right front	Left rear	Right rear
Number selected	11	15	8	6

ISBN 0-558-58875-1

9. Pennies from Credit Card Purchases When considering effects from eliminating the penny as a unit of currency in the United States, the author randomly selected the amounts from 100 credit card purchases and recorded the cents portions of those amounts. The table below lists those cents portions categorized according to the indicated values. Use a 0.05 significance level to test the claim that the four categories are equally likely. The author expected that many credit card purchases for whole dollar amounts would result in a disproportionately high frequency for the first category, but do the results support that expectation?

Cents portion	0–24	25–49	50–74	75–99
Number	33	16	23	28

10. Occupational Injuries Randomly selected nonfatal occupational injuries and illnesses are categorized according to the day of the week that they first occurred, and the results are listed below (based on data from the Bureau of Labor Statistics). Use a 0.05 significance level to test the claim that such injuries and illnesses occur with equal frequency on the different days of the week.

Day	Mon	Tues	Wed	Thurs	Fri
Number	23	23	21	21	19

11. Loaded Die The author drilled a hole in a die and filled it with a lead weight, then proceeded to roll it 200 times. Here are the observed frequencies for the outcomes of 1, 2, 3, 4, 5, and 6, respectively: 27, 31, 42, 40, 28, 32. Use a 0.05 significance level to test the claim that the outcomes are not equally likely. Does it appear that the loaded die behaves differently than a fair die?

12. Births Records of randomly selected births were obtained and categorized according to the day of the week that they occurred (based on data from the National Center for Health Statistics). Because babies are unfamiliar with our schedule of weekdays, a reasonable claim is that births occur on the different days with equal frequency. Use a 0.01 significance level to test that claim. Can you provide an explanation for the result?

Day	Sun	Mon	Tues	Wed	Thurs	Fri	Sat
Number of births	77	110	124	122	120	123	97

13. Kentucky Derby The table below lists the frequency of wins for different post positions in the Kentucky Derby horse race. A post position of 1 is closest to the inside rail, so that horse has the shortest distance to run. (Because the number of horses varies from year to year, only the first ten post positions are included.) Use a 0.05 significance level to test the claim that the likelihood of winning is the same for the different post positions. Based on the result, should bettors consider the post position of a horse racing in the Kentucky Derby?

Post Position	1	2	3	4	5	6	7	8	9	10
Wins	19	14	11	14	14	7	8	11	5	11

14. Measuring Weights Example 1 in this section is based on the principle that when certain quantities are measured, the last digits tend to be uniformly distributed, but if they are estimated or reported, the last digits tend to have disproportionately more 0s or 5s. The last digits of the September weights in Data Set 3 in Appendix B are summarized in the table below. Use a 0.05 significance level to test the claim that the last digits of 0, 1, 2, . . . , 9 occur with the same frequency. Based on the observed digits, what can be inferred about the procedure used to obtain the weights?

Last digit	0	1	2	3	4	5	6	7	8	9
Number	7	5	6	7	14	5	5	8	6	4

15. UFO Sightings Cases of UFO sightings are randomly selected and categorized according to month, with the results listed in the table below (based on data from Larry Hatch). Use a 0.05 significance level to test the claim that UFO sightings occur in the different months with

ISBN 0-558-58875-1

equal frequency. Is there any reasonable explanation for the two months that have the highest frequencies?

Month	Jan.	Feb.	March	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Number	1239	1111	1428	1276	1102	1225	2233	2012	1680	1994	1648	1125

16. Violent Crimes Cases of violent crimes are randomly selected and categorized by month, with the results shown in the table below (based on data from the FBI). Use a 0.01 significance level to test the claim that the rate of violent crime is the same for each month. Can you explain the result?

Month	Jan.	Feb.	March	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Number	786	704	835	826	900	868	920	901	856	862	783	797

17. Genetics The Advanced Placement Biology class at Mount Pearl Senior High School conducted genetics experiments with fruit flies, and the results in the following table are based on the results that they obtained. Use a 0.05 significance level to test the claim that the observed frequencies agree with the proportions that were expected according to principles of genetics.

Characteristic	Red eye/ normal wing	Sepia eye/ normal wing	Red eye/ vestigial wing	Sepia eye/ vestigial wing
Frequency	59	15	2	4
Expected proportion	9/16	3/16	3/16	1/16

18. Do World War II Bomb Hits Fit a Poisson Distribution? In analyzing hits by V-1 buzz bombs in World War II, South London was subdivided into regions, each with an area of 0.25 km^2. Shown below is a table of actual frequencies of hits and the frequencies expected with the Poisson distribution. (The Poisson distribution is described in Section 5-5.) Use the values listed and a 0.05 significance level to test the claim that the actual frequencies fit a Poisson distribution.

Number of bomb hits	0	1	2	3	4 or more
Actual number of regions	229	211	93	35	8
Expected number of regions (from Poisson distribution)	227.5	211.4	97.9	30.5	8.7

19. M&M Candies Mars, Inc. claims that its M&M plain candies are distributed with the following color percentages: 16% green, 20% orange, 14% yellow, 24% blue, 13% red, and 13% brown. Refer to Data Set 18 in Appendix B and use the sample data to test the claim that the color distribution is as claimed by Mars, Inc. Use a 0.05 significance level.

20. Bias in Clinical Trials? Researchers investigated the issue of race and equality of access to clinical trials. The table below shows the population distribution and the numbers of participants in clinical trials involving lung cancer (based on data from "Participation in Cancer Clinical Trials," by Murthy, Krumholz, and Gross, *Journal of the American Medical Association,* Vol. 291, No. 22). Use a 0.01 significance level to test the claim that the distribution of clinical trial participants fits well with the population distribution. Is there a race/ethnic group that appears to be very underrepresented?

Race/ethnicity	White non-Hispanic	Hispanic	Black	Asian/Pacific Islander	American Indian/ Alaskan Native
Distribution of Population	75.6%	9.1%	10.8%	3.8%	0.7%
Number in Lung Cancer Clinical Trials	3855	60	316	54	12

ISBN 0-558-58875-1

Benford's Law. ***According to Benford's law, a variety of different data sets include numbers with leading (first) digits that follow the distribution shown in the table below. In Exercises 21–24, test for goodness-of-fit with Benford's law.***

Leading Digit	1	2	3	4	5	6	7	8	9
Benford's law: distribution of leading digits	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%

21. Detecting Fraud When working for the Brooklyn District Attorney, investigator Robert Burton analyzed the leading digits of the amounts from 784 checks issued by seven suspect companies. The frequencies were found to be 0, 15, 0, 76, 479, 183, 8, 23, and 0, and those digits correspond to the leading digits of 1, 2, 3, 4, 5, 6, 7, 8, and 9, respectively. If the observed frequencies are substantially different from the frequencies expected with Benford's law, the check amounts appear to result from fraud. Use a 0.01 significance level to test for goodness-of-fit with Benford's law. Does it appear that the checks are the result of fraud?

22. Author's Check Amounts Exercise 21 lists the observed frequencies of leading digits from amounts on checks from seven suspect companies. Here are the observed frequencies of the leading digits from the amounts on checks written by the author: 68, 40, 18, 19, 8, 20, 6, 9, 12. (Those observed frequencies correspond to the leading digits of 1, 2, 3, 4, 5, 6, 7, 8, and 9, respectively.) Using a 0.05 significance level, test the claim that these leading digits are from a population of leading digits that conform to Benford's law. Do the author's check amounts appear to be legitimate?

23. Political Contributions Amounts of recent political contributions are randomly selected, and the leading digits are found to have frequencies of 52, 40, 23, 20, 21, 9, 8, 9, and 30. (Those observed frequencies correspond to the leading digits of 1, 2, 3, 4, 5, 6, 7, 8, and 9, respectively, and they are based on data from "Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance," by Cho and Gaines, *American Statistician,* Vol. 61, No. 3.) Using a 0.01 significance level, test the observed frequencies for goodness-of-fit with Benford's law. Does it appear that the political campaign contributions are legitimate?

24. Check Amounts In the trial of *State of Arizona vs. Wayne James Nelson,* the defendant was accused of issuing checks to a vendor that did not really exist. The amounts of the checks are listed below in order by row. When testing for goodness-of-fit with the proportions expected with Benford's law, it is necessary to combine categories because not all expected values are at least 5. Use one category with leading digits of 1, a second category with leading digits of 2, 3, 4, 5, and a third category with leading digits of 6, 7, 8, 9. Using a 0.01 significance level, is there sufficient evidence to conclude that the leading digits on the checks do not conform to Benford's law?

$ 1,927.48	$27,902.31	$86,241.90	$72,117.46	$81,321.75	$97,473.96
$93,249.11	$89,658.16	$87,776.89	$92,105.83	$79,949.16	$87,602.93
$96,879.27	$91,806.47	$84,991.67	$90,831.83	$93,766.67	$88,336.72
$94,639.49	$83,709.26	$96,412.21	$88,432.86	$71,552.16	

11-2 Beyond the Basics

25. Testing Effects of Outliers In conducting a test for the goodness-of-fit as described in this section, does an outlier have much of an effect on the value of the χ^2 test statistic? Test for the effect of an outlier in Example 1 after changing the first frequency in Table 11-2 from 7 to 70. Describe the general effect of an outlier.

26. Testing Goodness-of-Fit with a Normal Distribution Refer to Data Set 21 in Appendix B for the axial loads (in pounds) of the aluminum cans that are 0.0109 in. thick.

ISBN 0-558-58875-1

An Eight-Year False Positive

The Associated Press recently released a report about Jim Malone, who had received a positive test result for an HIV infection. For eight years, he attended group support meetings, fought depression, and lost weight while fearing a death from AIDS. Finally, he was informed that the original test was wrong. He did not have an HIV infection. A follow-up test was given after the first positive test result, and the confirmation test showed that he did not have an HIV infection, but nobody told Mr. Malone about the new result. Jim Malone agonized for eight years because of a test result that was actually a false positive.

Axial load	Less than 239.5	239.5–259.5	259.5–279.5	More than 279.5
Frequency				

a. Enter the observed frequencies in the above table.

b. Assuming a normal distribution with mean and standard deviation given by the sample mean and standard deviation, use the methods of Chapter 6 to find the probability of a randomly selected axial load belonging to each class.

c. Using the probabilities found in part (b), find the expected frequency for each category.

d. Use a 0.01 significance level to test the claim that the axial loads were randomly selected from a normally distributed population. Does the goodness-of-fit test suggest that the data are from a normally distributed population?

11-3 Contingency Tables

Key Concept In this section we consider *contingency tables* (or *two-way frequency tables*), which include frequency counts for categorical data arranged in a table with at least two rows and at least two columns. In Part 1 of this section, we present a method for conducting a hypothesis test of the null hypothesis that the row and column variables are independent of each other. This test of independence is used in real applications quite often. In Part 2, we will use the same method for a test of homogeneity, whereby we test the claim that different populations have the same proportion of some characteristics.

Part 1: Basic Concepts of Testing for Independence

In this section we use standard statistical methods to analyze frequency counts in a contingency table (or two-way frequency table). We begin with the definition of a contingency table.

DEFINITION

A **contingency table** (or **two-way frequency table**) is a table in which frequencies correspond to two variables. (One variable is used to categorize rows, and a second variable is used to categorize columns.)

EXAMPLE 1 **Contingency Table from Echinacea Experiment** Table 11-6 is a contingency table with two rows and three columns. The cells of the table contain frequencies. The row variable identifies whether the subjects became infected, and the column variable identifies the treatment group (placebo, 20% extract group, or 60% extract group).

Table 11-6 Results from Experiment with Echinacea

	Treatment Group		
	Placebo	Echinacea: 20% extract	Echinacea: 60% extract
Infected	88	48	42
Not infected	15	4	10

ISBN 0-558-58875-1

We will now consider a hypothesis test of independence between the row and column variables in a contingency table. We first define a *test of independence*.

DEFINITION

A **test of independence** tests the null hypothesis that in a contingency table, the row and column variables are independent.

Objective

Conduct a hypothesis test for independence between the row variable and column variable in a contingency table.

Notation

O represents the *observed frequency* in a cell of a contingency table.

E represents the *expected frequency* in a cell, found by assuming that the row and column variables are independent.

r represents the number of rows in a contingency table (not including labels).

c represents the number of columns in a contingency table (not including labels).

Requirements

1. The sample data are randomly selected.
2. The sample data are represented as frequency counts in a two-way table.
3. For every cell in the contingency table, the *expected* frequency E is at least 5. (There is no requirement that every *observed* frequency must be at least 5. Also, there is no requirement that the population must have a normal distribution or any other specific distribution.)

Null and Alternative Hypotheses

The null and alternative hypotheses are as follows:

H_0: The row and column variables are *independent*.

H_1: The row and column variables are dependent.

Test Statistic for a Test of Independence

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed frequency in a cell and E is the expected frequency found by evaluating

$$E = \frac{(\text{row total})\,(\text{column total})}{(\text{grand total})}$$

Critical Values

1. The critical values are found in Table A-4 using

$$\textbf{degrees of freedom} = (r - 1)(c - 1)$$

where r is the number of rows and c is the number of columns.

2. Tests of independence with a contingency table are always *right-tailed*.

P-Values

P-values are typically provided by computer software, or a range of *P*-values can be found from Table A-4.

ISBN 0-558-58875-1

The test statistic allows us to measure the amount of disagreement between the frequencies actually observed and those that we would theoretically expect when the two variables are independent. Large values of the χ^2 test statistic are in the rightmost region of the chi-square distribution, and they reflect significant differences between observed and expected frequencies. The distribution of the test statistic χ^2 can be approximated by the chi-square distribution, provided that all expected frequencies are at least 5. The number of degrees of freedom $(r - 1)(c - 1)$ reflects the fact that because we know the total of all frequencies in a contingency table, we can freely assign frequencies to only $r - 1$ rows and $c - 1$ columns before the frequency for every cell is determined. (However, we cannot have negative frequencies or frequencies so large that any row (or column) sum exceeds the total of the observed frequencies for that row (or column).)

Finding Expected Values *E*

The test statistic χ^2 is found by using the values of O (observed frequencies) and the values of E (expected frequencies). The expected frequency E can be found for a cell by simply multiplying the total of the row frequencies by the total of the column frequencies, then dividing by the grand total of all frequencies, as shown in Example 2.

EXAMPLE 2 **Finding Expected Frequency** Refer to Table 11-6 and find the expected frequency for the first cell, where the observed frequency is 88.

SOLUTION The first cell lies in the first row (with a total frequency of 178) and the first column (with total frequency of 103). The "grand total" is the sum of all frequencies in the table, which is 207. The expected frequency of the first cell is

$$E = \frac{(\text{row total})\,(\text{column total})}{(\text{grand total})} = \frac{(178)\,(103)}{207} = 88.570$$

INTERPRETATION We know that the first cell has an observed frequency of $O = 88$ and an expected frequency of $E = 88.570$. We can interpret the expected value by stating that if we assume that getting an infection is independent of the treatment, then we expect to find that 88.570 of the subjects would be given a placebo and would get an infection. There is a discrepancy between $O = 88$ and $E = 88.570$, and such discrepancies are key components of the test statistic.

To better understand expected frequencies, pretend that we know only the row and column totals, as in Table 11-7, and that we must fill in the cell expected frequencies by assuming independence (or no relationship) between the row and column variables. In the first row, 178 of the 207 subjects got infections, so $P(\text{infection}) = 178/207$. In the first column, 103 of the 207 subjects were given a placebo, so $P(\text{placebo}) = 103/207$. Because we are assuming independence between getting an infection and the treatment group, the multiplication rule for independent events $[P(A \text{ and } B) = P(A) \cdot P(B)]$ is expressed as

$$P(\text{infection and placebo}) = P(\text{infection}) \cdot P(\text{placebo})$$
$$= \frac{178}{207} \cdot \frac{103}{207}$$

ISBN 0-558-58875-1

Table 11-7 Results from Experiment with Echinacea

	Treatment Group			Row totals:
	Placebo	Echinacea: 20% extract	Echinacea: 60% extract	
Infected				178
Not infected				29
Column totals:	**103**	**52**	**52**	**Grand total: 207**

We can now find the *expected value* for the first cell by multiplying the probability for that cell by the total number of subjects, as shown here:

$$E = n \cdot p = 207\left[\frac{178}{207} \cdot \frac{103}{207}\right] = 88.570$$

The form of this product suggests a general way to obtain the expected frequency of a cell:

$$\text{Expected frequency } E = (\text{grand total}) \cdot \frac{(\text{row total})}{(\text{grand total})} \cdot \frac{(\text{column total})}{(\text{grand total})}$$

This expression can be simplified to

$$E = \frac{(\text{row total}) \cdot (\text{column total})}{(\text{grand total})}$$

We can now proceed to conduct a hypothesis test of independence, as in Example 3.

EXAMPLE 3 **Does Echinacea Have an Effect on Colds?** Common colds are typically caused by a rhinovirus. In a test of the effectiveness of echinacea, some test subjects were treated with echinacea extracted with 20% ethanol, some were treated with echinacea extracted with 60% ethanol, and others were given a placebo. All of the test subjects were then exposed to rhinovirus. Results are summarized in Table 11-6 (based on data from "An Evaluation of *Echinacea angustifolia* in Experimental Rhinovirus Infections," by Turner, et al., *New England Journal of Medicine,* Vol. 353, No. 4). Use a 0.05 significance level to test the claim that getting an infection (cold) is independent of the treatment group. What does the result indicate about the effectiveness of echinacea as a treatment for colds?

SOLUTION **REQUIREMENT CHECK** (1) The subjects were recruited and were randomly assigned to the different treatment groups. (2) The results are expressed as frequency counts in Table 11-6. (3) The expected frequencies are all at least 5. (The expected frequencies are 88.570, 44.715, 44.715, 14.430, 7.285, and 7.285.) The requirements are satisfied.

The null hypothesis and alternative hypothesis are as follows:

H_0: Getting an infection is independent of the treatment.

H_1: Getting an infection and the treatment are dependent.

The significance level is $\alpha = 0.05$.

Because the data are in the form of a contingency table, we use the χ^2 distribution with this test statistic:

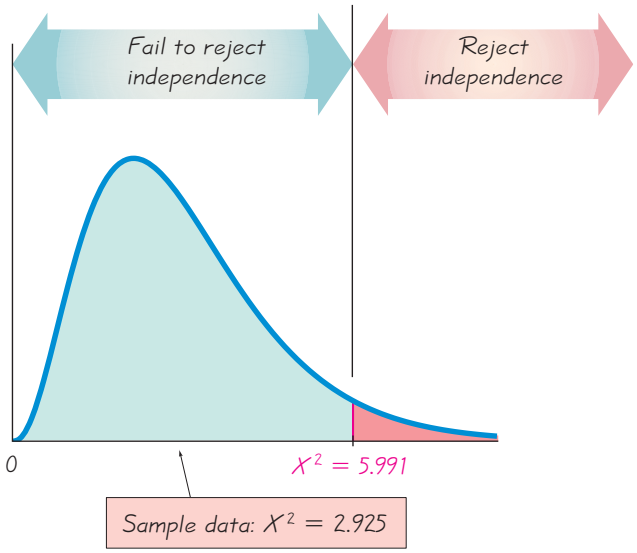
$$\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(88 - 88.570)^2}{88.570} + \cdots + \frac{(10 - 7.285)^2}{7.285}$$
$$= 2.925$$

ISBN 0-558-58875-1

The critical value of $\chi^2 = 5.991$ is found from Table A-4 with $\alpha = 0.05$ in the right tail and the number of degrees of freedom given by $(r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$. The test statistic and critical value are shown in Figure 11-5. Because the test statistic does not fall within the critical region, we fail to reject the null hypothesis of independence between getting an infection and treatment.

INTERPRETATION It appears that getting an infection is independent of the treatment group. This suggests that echinacea is not an effective treatment for colds.

Figure 11-5
Test of Independence for the Echinacea Data



P-Values

The preceding example used the traditional approach to hypothesis testing, but we can easily use the *P*-value approach. STATDISK, Minitab, Excel, and the TI-83/84 Plus calculator all provide *P*-values for tests of independence in contingency tables. (See Example 4.) If you don't have a suitable calculator or statistical software, estimate *P*-values from Table A-4 by finding where the test statistic falls in the row corresponding to the appropriate number of degrees of freedom.

EXAMPLE 4 **Is the Nurse a Serial Killer?** Table 11-1 provided with the Chapter Problem consists of a contingency table with a row variable (whether Kristen Gilbert was on duty) and a column variable (whether the shift included a death). Test the claim that whether Gilbert was on duty for a shift is independent of whether a patient died during the shift. Because this is such a serious analysis, use a significance level of 0.01. What does the result suggest about the charge that Gilbert killed patients?

SOLUTION **REQUIREMENT CHECK** (1) The data in Table 11-1 can be treated as random data for the purpose of determining whether such random data could easily occur by chance. (2) The sample data are represented as frequency counts in a two-way table. (3) Each expected frequency is at least 5. (The expected frequencies are 11.589, 245.411, 62.411, and 1321.589.) The requirements are satisfied.

ISBN 0-558-58875-1

The null hypothesis and alternative hypothesis are as follows:

H_0: Whether Gilbert was working is independent of whether there was a death during the shift.

H_1: Whether Gilbert was working and whether there was a death during the shift are dependent.

Minitab shows that the test statistic is $\chi^2 = 86.481$ and the P-value is 0.000. Because the P-value is less than the significance level of 0.01, we reject the null hypothesis of independence. There is sufficient evidence to warrant rejection of independence between the row and column variables.

MINITAB

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

        Death  No Death  Total
    1      40       217    257
        11.59    245.41
       69.648     3.289

    2      34      1350   1384
        62.41   1321.59
      12.933     0.611

Total      74      1567   1641

Chi-Sq = 86.481, DF = 1, P-Value = 0.000
```

INTERPRETATION We reject independence between whether Gilbert was working and whether a patient died during a shift. It appears that there is an association between Gilbert working and patients dying. (Note that this does not show that Gilbert *caused* the deaths, so this is not evidence that could be used at her trial, but it was evidence that led investigators to pursue other evidence that eventually led to her conviction for murder.)

As in Section 11-2, if observed and expected frequencies are close, the χ^2 test statistic will be small and the P-value will be large. If observed and expected frequencies are not close, the χ^2 test statistic will be large and the P-value will be small. These relationships are summarized and illustrated in Figure 11-6 on the next page.

Part 2: Test of Homogeneity and the Fisher Exact Test

Test of Homogeneity

In Part 1 of this section, we focused on the test of independence between the row and column variables in a contingency table. In Part 1, the sample data are from one population, and individual sample results are categorized with the row and column variables. However, we sometimes obtain samples drawn from *different* populations, and we want to determine whether those populations have the same proportions of the characteristics being considered. The *test of homogeneity* can be used in such cases. (The word *homogeneous* means "having the same quality," and in this context, we are testing to determine whether the proportions are the same.)

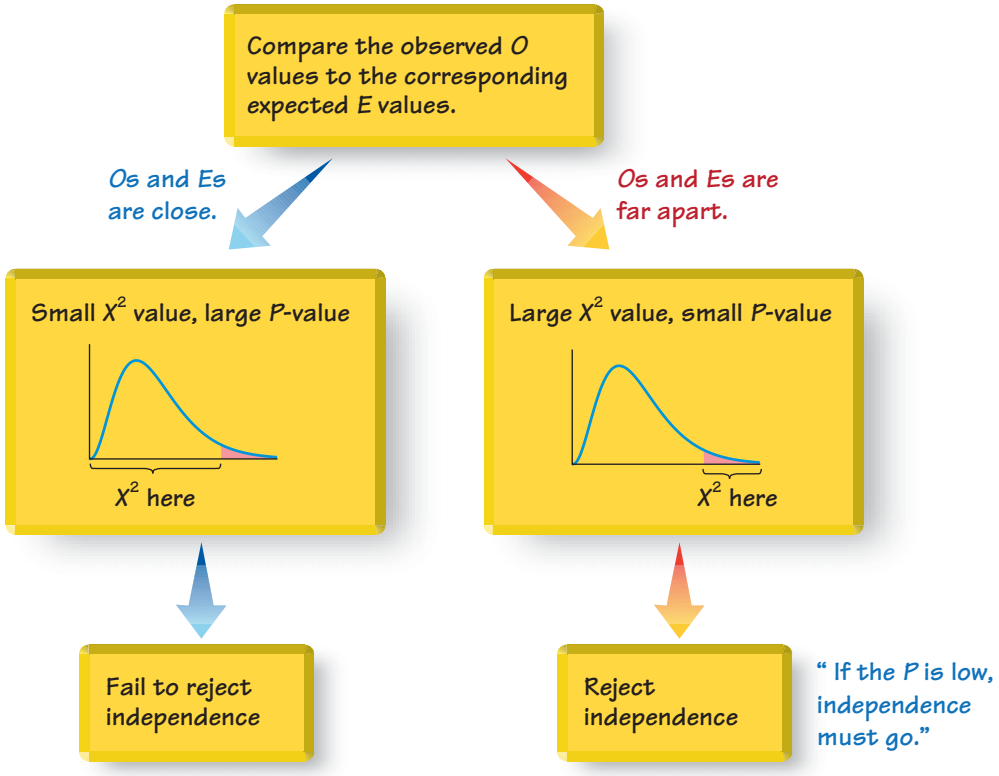
DEFINITION

In a **test of homogeneity,** we test the claim that *different populations* have the same proportions of some characteristics.

ISBN 0-558-58875-1

Figure 11-6

Relationships Among Key Components in Test of Independence



In conducting a test of homogeneity, we can use the same notation, requirements, test statistic, critical value, and procedures presented in Part 1 of this section, with one exception: Instead of testing the null hypothesis of *independence* between the row and column variables, we test the null hypothesis that *the different populations have the same proportions of some characteristics.*

EXAMPLE 5 **Influence of Gender** Does a pollster's gender have an effect on poll responses by men? A *U.S. News & World Report* article about polls stated: "On sensitive issues, people tend to give 'acceptable' rather than honest responses; their answers may depend on the gender or race of the interviewer." To support that claim, data were provided for an Eagleton Institute poll in which surveyed men were asked if they agreed with this statement: "Abortion is a private matter that should be left to the woman to decide without government intervention." We will analyze the effect of gender on male survey subjects only. Table 11-8 is based on the responses of surveyed men. Assume that the survey was designed so that male interviewers were instructed to obtain 800 responses from male subjects, and female interviewers were instructed to obtain 400 responses from male subjects. Using a 0.05 significance level, test the claim that the proportions of agree/disagree responses are the same for the subjects interviewed by men and the subjects interviewed by women.

Table 11-8 Gender and Survey Responses

	Gender of Interviewer	
	Man	Woman
Men who agree	560	308
Men who disagree	240	92

ISBN 0-558-58875-1

SOLUTION **REQUIREMENT CHECK** (1) The data are random. (2) The sample data are represented as frequency counts in a two-way table. (3) The expected frequencies (shown in the accompanying Minitab display as 578.67, 289.33, 221.33, and 110.67) are all at least 5. All of the requirements are satisfied.

Because this is a test of homogeneity, we test the claim that the proportions of agree/disagree responses are the same for the subjects interviewed by males and the subjects interviewed by females. We have two separate populations (subjects interviewed by men and subjects interviewed by women), and we test for homogeneity with these hypotheses:

H_0: The proportions of agree/disagree responses are the same for the subjects interviewed by men and the subjects interviewed by women.

H_1: The proportions are different.

The significance level is $\alpha = 0.05$. We use the same χ^2 test statistic described earlier, and it is calculated using the same procedure. Instead of listing the details of that calculation, we provide the Minitab display for the data in Table 11-8.

MINITAB

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

          C1       C2  Total
1        560      308    868
      578.67   289.33
       0.602    1.204

2        240       92    332
      221.33   110.67
       1.574    3.149

Total    800      400   1200

Chi-Sq = 6.529, DF = 1, P-Value = 0.011
```

The Minitab display shows the expected frequencies of 578.67, 289.33, 221.33, and 110.67. It also includes the test statistic of $\chi^2 = 6.529$ and the P-value of 0.011. Using the P-value approach to hypothesis testing, we reject the null hypothesis of equal (homogeneous) proportions (because the P-value of 0.011 is less than 0.05). There is sufficient evidence to warrant rejection of the claim that the proportions are the same.

INTERPRETATION It appears that response and the gender of the interviewer are dependent. Although this statistical analysis cannot be used to justify any statement about causality, it does appear that men are influenced by the gender of the interviewer.

Fisher Exact Test

The procedures for testing hypotheses with contingency tables with two rows and two columns (2×2) have the requirement that every cell must have an expected frequency of at least 5. This requirement is necessary for the χ^2 distribution to be a suitable approximation to the exact distribution of the χ^2 test statistic. The *Fisher exact test* is often used for a 2×2 contingency table with one or more expected frequencies that are below 5. The Fisher exact test provides an *exact* P-value and does not require an approximation technique. Because the calculations are quite complex, it's a good idea to use computer software when using the Fisher exact test. STATDISK and Minitab both have the ability to perform the Fisher exact test.

ISBN 0-558-58875-1

USING TECHNOLOGY

STATDISK Enter the observed frequencies in the Data Window as they appear in the contingency table. Select **Analysis** from the main menu, then select **Contingency Tables.** Enter a significance level and proceed to identify the columns containing the frequencies. Click on **Evaluate.** The STATDISK results include the test statistic, critical value, *P*-value, and conclusion, as shown in the display resulting from Table 11-1.

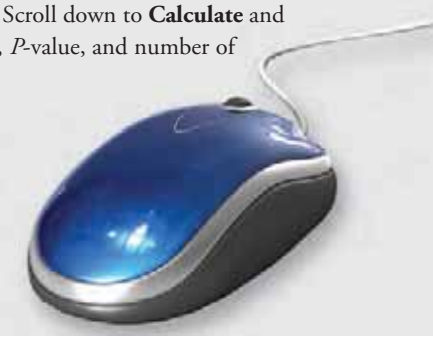
STATDISK

```
Degrees of freedom: 1

Test Statistic, X^2: 86.4809
Critical X^2:        6.634903
P-Value:             0.0000

Reject the Null Hypothesis.
Data provides evidence that the
rows and columns are related.
```

MINITAB First enter the observed frequencies in columns, then select **Stat** from the main menu bar. Next select the option **Tables,** then select **Chi Square Test (Two-Way Table in Worksheet)** and enter the names of the columns containing the observed frequencies, such as C1 C2 C3 C4. Minitab provides the test statistic and *P*-value, the expected frequencies, and the individual terms of the χ^2 test statistic. See the Minitab displays that accompany Examples 4 and 5.

EXCEL You must enter the observed frequencies, and you must also determine and enter the expected frequencies. When finished, click on the ***fx*** icon in the menu bar, select the function category **Statistical,** and then select the function name **CHITEST.** You must enter the range of values for the observed frequencies and the range of values for the expected frequencies. Only the *P*-value is provided. (DDXL can also be used by selecting **Tables,** then **Indep. Test for Summ Data.**)

TI-83/84 PLUS First enter the contingency table as a *matrix* by pressing **2nd** x^{-1} to get the **MATRIX** menu (or the **MATRIX** key on the TI-83). Select **EDIT,** and press **ENTER.** Enter the dimensions of the matrix (rows by columns) and proceed to enter the individual frequencies. When finished, press **STAT,** select **TESTS,** and then select the option χ^2**-Test.** Be sure that the observed matrix is the one you entered, such as matrix A. The expected frequencies will be automatically calculated and stored in the separate matrix identified as "Expected." Scroll down to **Calculate** and press **ENTER** to get the test statistic, *P*-value, and number of degrees of freedom.



11-3 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

1. Polio Vaccine Results of a test of the Salk vaccine against polio are summarized in the table below. If we test the claim that getting paralytic polio is independent of whether the child was treated with the Salk vaccine or was given a placebo, the TI-83/84 Plus calculator provides a *P*-value of 1.732517E−11, which is in scientific notation. Write the *P*-value in a standard form that is not in scientific notation. Based on the *P*-value, what conclusion should we make? Does the vaccine appear to be effective?

	Paralytic polio	No paralytic polio
Salk vaccine	33	200,712
Placebo	115	201,114

2. Cause and Effect Based on the data in the table provided with Exercise 1, can we conclude that the Salk vaccine causes a decrease in the rate of paralytic polio? Why or why not?

3. Interpreting *P*-Value Refer to the *P*-value given in Exercise 1. Interpret that *P*-value by completing this statement: The *P*-value is the probability of .

4. Right-Tailed Test Why are the hypothesis tests described in this section always right-tailed, as in Example 1?

In Exercises 5 and 6, test the given claim using the displayed software results.

5. Home Field Advantage Winning team data were collected for teams in different sports, with the results given in the accompanying table (based on data from "Predicting Professional

ISBN 0-558-58875-1

Sports Game Outcomes from Intermediate Game Scores," by Copper, DeNeve, and Mosteller, *Chance*, Vol. 5, No. 3–4). The TI-83/84 Plus results are also displayed. Use a 0.05 level of significance to test the claim that home/visitor wins are independent of the sport.

	Basketball	Baseball	Hockey	Football
Home team wins	127	53	50	57
Visiting team wins	71	47	43	42

TI-83/84 PLUS

```
χ²-Test
 χ²=4.737208763
 p=.1920828463
 df=3
```

6. Crime and Strangers The Minitab display results from the table below, which lists data obtained from randomly selected crime victims (based on data from the U.S. Department of Justice). What can we conclude?

	Homicide	Robbery	Assault
Criminal was a stranger	12	379	727
Criminal was acquaintance or relative	39	106	642

MINITAB

Chi-Sq = 119.330, DF = 2, P-Value = 0.000

In Exercises 7–22, test the given claim.

7. Instant Replay in Tennis The table below summarizes challenges made by tennis players in the first U.S. Open that used the Hawk-Eye electronic instant replay system. Use a 0.05 significance level to test the claim that success in challenges is independent of the gender of the player. Does either gender appear to be more successful?

	Was the challenge to the call successful?	
	Yes	No
Men	201	288
Women	126	224

8. Open Roof or Closed Roof? In a recent baseball World Series, the Houston Astros wanted to close the roof on their domed stadium so that fans could make noise and give the team a better advantage at home. However, the Astros were ordered to keep the roof open, unless weather conditions justified closing it. But does the closed roof really help the Astros? The table below shows the results from home games during the season leading up to the World Series. Use a 0.05 significance level to test for independence between wins and whether the roof is open or closed. Does it appear that a closed roof really gives the Astros an advantage?

	Win	Loss
Closed roof	36	17
Open roof	15	11

9. Testing a Lie Detector The table below includes results from polygraph (lie detector) experiments conducted by researchers Charles R. Honts (Boise State University) and Gordon H. Barland (Department of Defense Polygraph Institute). In each case, it was known if the subject lied or did not lie, so the table indicates when the polygraph test was correct. Use a 0.05 significance level to test the claim that whether a subject lies is independent of the polygraph test indication. Do the results suggest that polygraphs are effective in distinguishing between truths and lies?

	Did the Subject Actually Lie?	
	No (Did Not Lie)	Yes (Lied)
Polygraph test indicated that the subject *lied.*	15	42
Polygraph test indicated that the subject did *not lie.*	32	9

ISBN 0-558-58875-1

10. Clinical Trial of Chantix Chantix is a drug used as an aid for those who want to stop smoking. The adverse reaction of nausea has been studied in clinical trials, and the table below summarizes results (based on data from Pfizer). Use a 0.01 significance level to test the claim that nausea is independent of whether the subject took a placebo or Chantix. Does nausea appear to be a concern for those using Chantix?

	Placebo	Chantix
Nausea	10	30
No nausea	795	791

11. Amalgam Tooth Fillings The table below shows results from a study in which some patients were treated with amalgam restorations and others were treated with composite restorations that do not contain mercury (based on data from "Neuropsychological and Renal Effects of Dental Amalgam in Children," by Bellinger, et al., *Journal of the American Medical Association,* Vol. 295, No. 15). Use a 0.05 significance level to test for independence between the type of restoration and the presence of any adverse health conditions. Do amalgam restorations appear to affect health conditions?

	Amalgam	Composite
Adverse health condition reported	135	145
No adverse health condition reported	132	122

12. Amalgam Tooth Fillings In recent years, concerns have been expressed about adverse health effects from amalgam dental restorations, which include mercury. The table below shows results from a study in which some patients were treated with amalgam restorations and others were treated with composite restorations that do not contain mercury (based on data from "Neuropsychological and Renal Effects of Dental Amalgam in Children," by Bellinger, et al., *Journal of the American Medical Association,* Vol. 295, No. 15). Use a 0.05 significance level to test for independence between the type of restoration and sensory disorders. Do amalgam restorations appear to affect sensory disorders?

	Amalgam	Composite
Sensory disorder	36	28
No sensory disorder	231	239

13. Is Sentence Independent of Plea? Many people believe that criminals who plead guilty tend to get lighter sentences than those who are convicted in trials. The accompanying table summarizes randomly selected sample data for San Francisco defendants in burglary cases (based on data from "Does It Pay to Plead Guilty? Differential Sentencing and the Functioning of the Criminal Courts," by Brereton and Casper, *Law and Society Review,* Vol. 16, No. 1). All of the subjects had prior prison sentences. Use a 0.05 significance level to test the claim that the sentence (sent to prison or not sent to prison) is independent of the plea. If you were an attorney defending a guilty defendant, would these results suggest that you should encourage a guilty plea?

	Guilty Plea	Not Guilty Plea
Sent to prison	392	58
Not sent to prison	564	14

14. Is the Vaccine Effective? In a *USA Today* article about an experimental vaccine for children, the following statement was presented: "In a trial involving 1602 children, only 14 (1%) of the 1070 who received the vaccine developed the flu, compared with 95 (18%) of the 532 who got a placebo." The data are shown in the table below. Use a 0.05 significance level to test for independence between the variable of treatment (vaccine or placebo) and the variable representing flu (developed flu, did not develop flu). Does the vaccine appear to be effective?

ISBN 0-558-58875-1

	Developed Flu?	
	Yes	No
Vaccine treatment	14	1056
Placebo	95	437

15. Which Treatment Is Better? A randomized controlled trial was designed to compare the effectiveness of splinting versus surgery in the treatment of carpal tunnel syndrome. Results are given in the table below (based on data from "Splinting vs. Surgery in the Treatment of Carpal Tunnel Syndrome," by Gerritsen, et al., *Journal of the American Medical Association,* Vol. 288, No. 10). The results are based on evaluations made one year after the treatment. Using a 0.01 significance level, test the claim that success is independent of the type of treatment. What do the results suggest about treating carpal tunnel syndrome?

	Successful Treatment	Unsuccessful Treatment
Splint treatment	60	23
Surgery treatment	67	6

16. Norovirus on Cruise Ships The *Queen Elizabeth II* cruise ship and Royal Caribbean's *Freedom of the Seas* cruise ship both experienced outbreaks of norovirus within two months of each other. Results are shown in the table below. Use a 0.05 significance level to test the claim that getting norovirus is independent of the ship. Based on these results, does it appear that an outbreak of norovirus has the same effect on different ships?

	Norovirus	No norovirus
Queen Elizabeth II	276	1376
Freedom of the Seas	338	3485

17. Global Warming Survey A Pew Research poll was conducted to investigate opinions about global warming. The respondents who answered yes when asked if there is solid evidence that the earth is getting warmer were then asked to select a cause of global warming. The results are given in the table below. Use a 0.05 significance level to test the claim that the sex of the respondent is independent of the choice for the cause of global warming. Do men and women appear to agree, or is there a substantial difference?

	Human activity	Natural patterns	Don't know or refused to answer
Male	314	146	44
Female	308	162	46

18. Global Warming Survey A Pew Research poll was conducted to investigate opinions about global warming. The respondents who answered yes when asked if there is solid evidence that the earth is getting warmer were then asked to select a cause of global warming. The results for two age brackets are given in the table below. Use a 0.01 significance level to test the claim that the age bracket is independent of the choice for the cause of global warming. Do respondents from both age brackets appear to agree, or is there a substantial difference?

	Human activity	Natural patterns	Don't know or refused to answer
Under 30	108	41	7
65 and over	121	71	43

19. Clinical Trial of Campral Campral is a drug used to help patients continue their abstinence from the use of alcohol. Adverse reactions of Campral have been studied in clinical trials, and the table below summarizes results for digestive system effects among patients from different treatment groups (based on data from Forest Pharmaceuticals, Inc.). Use a 0.01 significance level to test the claim that experiencing an adverse reaction in the digestive system is

ISBN 0-558-58875-1

independent of the treatment group. Does Campral treatment appear to have an effect on the digestive system?

	Placebo	Campral 1332 mg	Campral 1998 mg
Adverse effect on digestive system	344	89	8
No effect on digestive system	1362	774	71

20. Is Seat Belt Use Independent of Cigarette Smoking? A study of seat belt users and nonusers yielded the randomly selected sample data summarized in the given table (based on data from "What Kinds of People Do Not Use Seat Belts?" by Helsing and Comstock, *American Journal of Public Health,* Vol. 67, No. 11). Test the claim that the amount of smoking is independent of seat belt use. A plausible theory is that people who smoke more are less concerned about their health and safety and are therefore less inclined to wear seat belts. Is this theory supported by the sample data?

	Number of Cigarettes Smoked per Day			
	0	1–14	15–34	35 and over
Wear seat belts	175	20	42	6
Don't wear seat belts	149	17	41	9

21. Clinical Trial of Lipitor Lipitor is the trade name of the drug atorvastatin, which is used to reduce cholesterol in patients. (This is the largest-selling drug in the world, with $13 billion in sales for a recent year.) Adverse reactions have been studied in clinical trials, and the table below summarizes results for infections in patients from different treatment groups (based on data from Parke-Davis). Use a 0.05 significance level to test the claim that getting an infection is independent of the treatment. Does the atorvastatin treatment appear to have an effect on infections?

	Placebo	Atorvastatin 10 mg	Atorvastatin 40 mg	Atorvastatin 80 mg
Infection	27	89	8	7
No infection	243	774	71	87

22. Injuries and Motorcycle Helmet Color A case-control (or retrospective) study was conducted to investigate a relationship between the colors of helmets worn by motorcycle drivers and whether they are injured or killed in a crash. Results are given in the table below (based on data from "Motorcycle Rider Conspicuity and Crash Related Injury: Case-Control Study," by Wells, et al., *BMJ USA,* Vol. 4). Test the claim that injuries are independent of helmet color. Should motorcycle drivers choose helmets with a particular color? If so, which color appears best?

	Color of Helmet				
	Black	White	Yellow/Orange	Red	Blue
Controls (not injured)	491	377	31	170	55
Cases (injured or killed)	213	112	8	70	26

11-3 Beyond the Basics

23. Test of Homogeneity Table 11-8 summarizes data for male survey subjects, but the table on the next page summarizes data for a sample of women (based on data from an Eagleton Institute poll). Using a 0.01 significance level, and assuming that the sample sizes of 800 men and 400 women are predetermined, test the claim that the proportions of agree/disagree responses are the same for the subjects interviewed by men and the subjects interviewed by women. Does it appear that the gender of the interviewer affected the responses of women?

ISBN 0-558-58875-1

	Gender of Interviewer	
	Man	Woman
Women who agree	512	336
Women who disagree	288	64

24. Using Yates' Correction for Continuity The chi-square distribution is continuous, whereas the test statistic used in this section is discrete. Some statisticians use *Yates' correction for continuity* in cells with an expected frequency of less than 10 or in all cells of a contingency table with two rows and two columns. With Yates' correction, we replace

$$\sum \frac{(O - E)^2}{E} \quad \text{with} \quad \sum \frac{(|O - E| - 0.5)^2}{E}$$

Given the contingency table in Exercise 7, find the value of the χ^2 test statistic with and without Yates' correction. What effect does Yates' correction have?

25. Equivalent Tests A χ^2 test involving a 2×2 table is equivalent to the test for the difference between two proportions, as described in Section 9-2. Using the table in Exercise 7, verify that the χ^2 test statistic and the z test statistic (found from the test of equality of two proportions) are related as follows: $z^2 = \chi^2$. Also show that the critical values have that same relationship.

11-4 McNemar's Test for Matched Pairs

Key Concept The methods in Section 11-3 for analyzing two-way tables are based on *independent* data. For 2×2 tables consisting of frequency counts that result from *matched pairs,* the frequency counts within each matched pair are not independent and, for such cases, we can use McNemar's test for matched pairs. In this section we present the method of using McNemar's test for testing the null hypothesis that the frequencies from the discordant (different) categories occur in the same proportion.

Table 11-9 shows a general format for summarizing results from data consisting of frequency counts from matched pairs. Table 11-9 refers to two different treatments (such as two different eye drop solutions) applied to two different parts of each subject (such as left eye and right eye). It's a bit difficult to correctly read a table such as Table 11-9. The total number of subjects is $a + b + c + d$, and each of those subjects yields results from each of two parts of a matched pair. If $a = 100$, then 100 subjects were cured with both treatments. If $b = 50$ in Table 11-9, then each of 50 subjects had no cure with treatment X but they were each cured with treatment Y. Remember, the entries in Table 11-9 are frequency counts of *subjects,* not the total number of individual components in the matched pairs. If 500 people have each eye treated with two different ointments, the value of $a + b + c + d$ is 500 (the number of subjects), not 1000 (the number of treated eyes).

Table 11-9 2×2 Table with Frequency Counts from Matched Pairs

		Treatment X	
		Cured	Not Cured
Treatment Y	Cured	a	b
	Not cured	c	d

ISBN 0-558-58875-1

Because the frequency counts in Table 11-9 result from *matched* pairs, the data are not independent and we cannot use the methods from Section 11-3. Instead, we use McNemar's test.

DEFINITION

McNemar's test uses frequency counts from *matched pairs* of nominal data from two categories to test the null hypothesis that for a 2×2 table such as Table 11-9, the frequencies b and c occur in the same proportion.

Objective

Test for a difference in proportions by using McNemar's test for matched pairs.

Notation

a, b, c, and d represent the frequency counts from a 2×2 table consisting of frequency counts from *matched pairs*. (The total number of subjects is $a + b + c + d$.)

Requirements

1. The sample data have been randomly selected.
2. The sample data consist of *matched pairs* of frequency counts.
3. The data are at the nominal level of measurement, and each observation can be classified two ways: (1) According to the category distinguishing values with each matched pair (such as left eye and right eye), and (2) according to another category with two possible values (such as cured/not cured).
4. For tables such as Table 11-9, the frequencies are such that $b + c \geq 10$.

Null and Alternative Hypotheses

H_0: The proportions of the frequencies b and c (as in Table 11-9) are the same.

H_1: The proportions of the frequencies b and c (as in Table 11-9) are different.

Test Statistic (for testing the null hypothesis that for tables such as Table 11-9, the frequencies b and c occur in the same proportion):

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}$$

where the frequencies of b and c are obtained from the 2×2 table with a format similar to Table 11-9. (The frequencies b and c must come from "discordant" (or different) pairs, as described later in this section.)

Critical Values

1. The critical region is located in the *right tail only*.
2. The critical values are found in Table A-4 by using **degrees of freedom = 1.**

P-Values

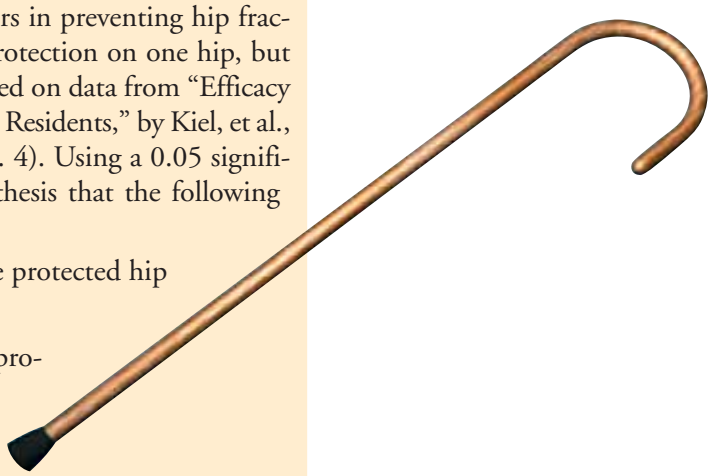
P-values are typically provided by computer software, or a range of *P*-values can be found from Table A-4.

ISBN 0-558-58875-1

EXAMPLE 1 **Are Hip Protectors Effective?** A randomized controlled trial was designed to test the effectiveness of hip protectors in preventing hip fractures in the elderly. Nursing home residents each wore protection on one hip, but not the other. Results are summarized in Table 11-10 (based on data from "Efficacy of Hip Protector to Prevent Hip Fracture in Nursing Home Residents," by Kiel, et al., *Journal of the American Medical Association,* Vol. 298, No. 4). Using a 0.05 significance level, apply McNemar's test to test the null hypothesis that the following two proportions are the same:

- The proportion of subjects with no hip fracture on the protected hip and a hip fracture on the unprotected hip.
- The proportion of subjects with a hip fracture on the protected hip and no hip fracture on the unprotected hip.

Based on the results, do the hip protectors appear to be effective in preventing hip fractures?



SOLUTION

REQUIREMENT CHECK (1) The data are from randomly selected subjects. (2) The data consist of matched pairs of frequency counts. (3) The data are at the nominal level of measurement and each observation can be categorized according to two variables. (One variable has values of "hip protection was worn" and "hip protection was not worn," and the other variable has values of "hip was fractured" and "hip was not fractured.") (4) For Table 11-10, $b = 10$ and $c = 15$, so that $b + c = 25$, which is at least 10. All of the requirements are satisfied.

Although Table 11-10 might appear to be a 2×2 contingency table, we cannot use the procedures of Section 11-3 because the data come from *matched pairs* (instead of being independent). Instead, we use McNemar's test.

After comparing the frequency counts in Table 11-9 to those given in Table 11-10, we see that $b = 10$ and $c = 15$, so the test statistic can be calculated as follows:

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} = \frac{(|10 - 15| - 1)^2}{10 + 15} = 0.640$$

With a 0.05 significance level and degrees of freedom given by df $= 1$, we refer to Table A-4 to find the critical value of $\chi^2 = 3.841$ for this right-tailed test. The test statistic of $\chi^2 = 0.640$ does not exceed the critical value of $\chi^2 = 3.841$, so we fail to reject the null hypothesis. (Also, the *P*-value is 0.424, which is greater than 0.05, indicating that the null hypothesis should be rejected.)

INTERPRETATION The proportion of hip fractures with the protectors worn is not significantly different from the proportion of hip fractures without the protectors worn. The hip protectors do not appear to be effective in preventing hip fractures.

Table 11-10 Randomized Controlled Trial of Hip Protectors

		No Hip Protector Worn	
		No Hip Fracture	Hip Fracture
Hip Protector Worn	No Hip Fracture	309	10
	Hip Fracture	15	2

ISBN 0-558-58875-1

Note that in the calculation of the test statistic in Example 1, we did not use the 309 subjects with no fractured hips, nor did we use the frequency of 2 representing subjects with both hips fractured. We used only those subjects with a fracture in one hip but not in the other. That is, we are using only the results from the categories that are *different*. Such pairs of different categories are referred to as *discordant pairs*.

DEFINITION

Discordant pairs of results come from matched pairs of results in which the two categories are different (as in the frequencies *b* and *c* in Table 11-9).

When trying to determine whether hip protectors are effective, we are not helped by any subjects with no fractures, and we are not helped by any subjects with both hips fractured. The differences are reflected in the discordant results from the subjects with one hip fractured while the other hip is not fractured. Consequently, the test statistic includes only the two frequencies that result from the two discordant (or different) pairs of categories.

CAUTION

When applying McNemar's test, be careful to use only the frequencies from the pairs of categories that are *different*. Do not blindly use the frequencies in the upper right and lower left corners, because they do not necessarily represent the discordant pairs. If Table 11-10 were reconfigured as shown below, it would be inconsistent in its format, but it would be technically correct in summarizing the same results as Table 11-10; however, blind use of the frequencies of 2 and 309 would result in the *wrong* test statistic.

		No Hip Protector Worn	
		No Hip Fracture	Hip Fracture
Hip Protector Worn	Hip Fracture	15	2
	No Hip Fracture	309	10

In this reconfigured table, the discordant pairs of frequencies are these:

Hip fracture/No hip fracture: 15

No hip fracture/Hip fracture: 10

With this reconfigured table, we should again use the frequencies of 15 and 10 (as in Example 1), not 2 and 309. In a more perfect world, all such 2×2 tables would be configured with a consistent format, and we would be much less likely to use the wrong frequencies.

In addition to comparing treatments given to matched pairs (as in Example 1), McNemar's test is often used to test a null hypothesis of no change in before/after types of experiments. (See Exercises 5–12.)

ISBN 0-558-58875-1

USING TECHNOLOGY

STATDISK Select **Analysis,** then select **McNemar's Test.** Enter the frequencies in the table that appears, then enter the significance level, then click on **Evaluate.** The STATDISK results include the test statistic, critical value, *P*-value, and conclusion.

MINITAB, EXCEL, and TI-83/84 Plus McNemar's test is not available.

11-4 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

1. McNemar's Test The table below summarizes results from a study in which 186 students in an introductory statistics course were each given algebra problems in two different formats: a symbolic format and a verbal format (based on data from "Changing Student's Perspectives of McNemar's Test of Change," by Levin and Serlin, *Journal of Statistics Education,* Vol. 8, No. 2). Assume that the data are randomly selected. Using only an examination of the table entries, does either format appear to be better? If so, which one? Why?

		Verbal Format	
		Mastery	Nonmastery
Symbolic Format	Mastery	74	31
	Nonmastery	33	48

2. Discordant Pairs Refer to the table in Exercise 1. Identify the discordant pairs of results.

3. Discordant Pairs Refer to the data in Exercise 1. Explain why McNemar's test ignores the frequencies of 74 and 48.

4. Requirement Check Refer to the data in Exercise 1. Identify which requirements are satisfied for McNemar's test.

In Exercises 5–12, refer to the following table. The table summarizes results from an experiment in which subjects were first classified as smokers or nonsmokers, then they were given a treatment, then later they were again classified as smokers or nonsmokers (based on data from Pfizer Pharmaceuticals in clinical trials of Chantix).

		Before Treatment	
		Smoke	Don't Smoke
After treatment	Smoke	460	4
	Don't smoke	361	192

5. Sample Size How many subjects are included in the experiment?

6. Treatment Effectiveness How many subjects changed their smoking status after the treatment?

ISBN 0-558-58875-1

7. Treatment Ineffectiveness How many subjects appear to be unaffected by the treatment one way or the other?

8. Why Not *t* Test? Section 9-4 presented procedures for data consisting of matched pairs. Why can't we use the procedures of Section 9-4 for the analysis of the results summarized in the table?

9. Discordant Pairs Which of the following pairs of before/after results are *discordant?*

a. smoke/smoke

b. smoke/don't smoke

c. don't smoke/smoke

d. don't smoke/don't smoke

10. Test Statistic Using the appropriate frequencies, find the value of the test statistic.

11. Critical Value Using a 0.01 significance level, find the critical value.

12. Conclusion Based on the preceding results, what do you conclude? How does the conclusion make sense in terms of the original sample results?

13. Testing Hip Protectors Example 1 in this section used results from subjects who used hip protection at least 80% of the time. Results from a larger data set were obtained from the same study, and the results are shown in the table below (based on data from "Efficacy of Hip Protector to Prevent Hip Fracture in Nursing Home Residents," by Kiel, et al., *Journal of the American Medical Association,* Vol. 298, No. 4). Use a 0.05 significance level to test the effectiveness of the hip protectors.

		No Hip Protector Worn	
		No Hip Fracture	Hip Fracture
Hip Protector Worn	No Hip Fracture	1004	17
	Hip Fracture	21	0

14. Predicting Measles Immunity Pregnant women were tested for immunity to the rubella virus, and they were also tested for immunity to measles, with results given in the following table (based on data from "Does Rubella Predict Measles Immunity? A Serosurvey of Pregnant Women," by Kennedy, et al., *Infectious Diseases in Obstetrics and Gynecology,* Vol. 2006). Use a 0.05 significance level to apply McNemar's test. What does the result tell us? If a woman is likely to become pregnant and she is found to have rubella immunity, should she also be tested for measles immunity?

		Measles	
		Immune	Not Immune
Rubella	Immune	780	62
	Not Immune	10	7

15. Treating Athlete's Foot Randomly selected subjects are inflicted with tinea pedis (athlete's foot) on each of their feet. One foot is treated with a fungicide solution while the other foot is given a placebo. The results are given in the accompanying table. Using a 0.05 significance level, test the effectiveness of the treatment.

		Fungicide Treatment	
		Cure	No Cure
Placebo	Cure	5	12
	No cure	22	55

ISBN 0-558-58875-1

16. Treating Athlete's Foot Repeat Exercise 15 after changing the frequency of 22 to 66.

17. PET/CT Compared to MRI In the article "Whole-Body Dual-Modality PET/CT and Whole Body MRI for Tumor Staging in Oncology" (Antoch, et al., *Journal of the American Medical Association,* Vol. 290, No. 24), the authors cite the importance of accurately identifying the stage of a tumor. Accurate staging is critical for determining appropriate therapy. The article discusses a study involving the accuracy of positron emission tomography (PET) and computed tomography (CT) compared to magnetic resonance imaging (MRI). Using the data in the given table for 50 tumors analyzed with both technologies, does there appear to be a difference in accuracy? Does either technology appear to be better?

		PET/CT	
		Correct	Incorrect
MRI	Correct	36	1
	Incorrect	11	2

18. Testing a Treatment In the article "Eradication of Small Intestinal Bacterial Overgrowth Reduces Symptoms of Irritable Bowel Syndrome" (Pimentel, Chow, and Lin, *American Journal of Gastroenterology,* Vol. 95, No. 12), the authors include a discussion of whether antibiotic treatment of bacteria overgrowth reduces intestinal complaints. McNemar's test was used to analyze results for those subjects with eradication of bacterial overgrowth. Using the data in the given table, does the treatment appear to be effective against abdominal pain?

		Abdominal Pain Before Treatment?	
		Yes	No
Abdominal pain after treatment?	Yes	11	1
	No	14	3

11-4 Beyond the Basics

19. Correction for Continuity The test statistic given in this section includes a correction for continuity. The test statistic given below does not include the correction for continuity, and it is sometimes used as the test statistic for McNemar's test. Refer to Exercise 18 and find the value of the test statistic using the expression given below, and compare the result to the one found in the exercise.

$$\chi^2 = \frac{(b - c)^2}{b + c}$$

20. Using Common Sense Consider the table given in Exercise 17. The frequencies of 36 and 2 are not included in the computations, but how are your conclusions modified if those two frequencies are changed to 8000 and 7000 respectively?

21. Small Sample Case The requirements for McNemar's test include the condition that $b + c \geq 10$ so that the distribution of the test statistic can be approximated by the chi-square distribution. Refer to the table on the next page. McNemar's test should not be used because the condition of $b + c \geq 10$ is not satisfied since $b = 2$ and $c = 6$. Instead, use the binomial distribution to find the probability that among 8 equally likely outcomes, the results consist of 6 items in one category and 2 in the other category, or the results are more extreme. That is, use a probability of 0.5 to find the probability that among $n = 8$ trials, the number of successes x is 6 or 7 or 8. Double that probability to find the P-value for this test. Compare the result to the P-value of 0.289 that results from using the chi-square approximation, even though the condition of $b + c \geq 10$ is violated. What do you conclude about the two treatments?

ISBN 0-558-58875-1

		Treatment with Pedacream	
		Cured	Not Cured
Treatment with Fungacream	Cured	12	2
	Not cured	6	20

Review

The three sections of this chapter all involve applications of the χ^2 distribution to categorical data consisting of frequency counts. In Section 11-2 we described methods for using frequency counts from different categories for testing goodness-of-fit with some claimed distribution. The test statistic given below is used in a right-tailed test in which the χ^2 distribution has $k - 1$ degrees of freedom, where k is the number of categories. This test requires that each of the expected frequencies must be at least 5.

$$\text{Test statistic is } \chi^2 = \sum \frac{(O - E)^2}{E}$$

In Section 11-3 we described methods for testing claims involving contingency tables (or two-way frequency tables), which have at least two rows and two columns. Contingency tables incorporate two variables: One variable is used for determining the row that describes a sample value, and the second variable is used for determining the column that describes a sample value. We conduct a test of independence between the row and column variables by using the test statistic given below. This test statistic is used in a right-tailed test in which the χ^2 distribution has the number of degrees of freedom given by $(r - 1)(c - 1)$, where r is the number of rows and c is the number of columns. This test requires that each of the expected frequencies must be at least 5.

$$\text{Test statistic is } \chi^2 = \sum \frac{(O - E)^2}{E}$$

In Section 11-4 we introduced McNemar's test for testing the null hypothesis that a sample of matched pairs of data comes from a population in which the discordant (different) pairs occur in the same proportion. The test statistic is given below. The frequencies of b and c must come from "discordant" pairs. This test statistic is used in a right-tailed test in which the χ^2 distribution has 1 degree of freedom.

$$\text{Test statistic is } \chi^2 = \frac{(|b - c| - 1)^2}{b + c}$$

Statistical Literacy and Critical Thinking

1. Categorical Data In what sense are the data in the table below *categorical* data? (The data are from Pfizer, Inc.)

	Celebrex	Ibuprofen	Placebo
Nausea	145	23	78
No Nausea	4001	322	1786

2. Terminology Refer to the table given in Exercise 1. Why is that table referred to as a *two-way* table?

3. Cause/Effect Refer to the table given in Exercise 1. After analysis of the data in such a table, can we ever conclude that a treatment of Celebrex and/or Ibuprofen *causes* nausea? Why or why not?

ISBN 0-558-58875-1

4. Observed and Expected Frequencies Refer to the table given in Exercise 1. The cell with the observed frequency of 145 has an expected frequency of 160.490. Describe what that expected frequency represents.

Chapter Quick Quiz

Questions 1–4 refer to the sample data in the following table (based on data from the Dutchess County STOP-DWI Program). The table summarizes results from randomly selected fatal car crashes in which the driver had a blood-alcohol level greater than 0.10.

Day	Sun	Mon	Tues	Wed	Thurs	Fri	Sat
Number	40	24	25	28	29	32	38

1. What are the null and alternative hypotheses corresponding to a test of the claim that fatal DWI crashes occur equally on the different days of the week?

2. When testing the claim in Question 1, what are the observed and expected frequencies for Sunday?

3. If using a 0.05 significance level for a test of the claim that the proportions of DWI fatalities are the same for the different days of the week, what is the critical value?

4. Given that the *P*-value for the hypothesis test is 0.2840, what do you conclude?

5. When testing the null hypothesis of independence between the row and column variables in a contingency table, is the test two-tailed, left-tailed, or right-tailed?

6. What distribution is used for testing the null hypothesis that the row and column variables in a contingency table are independent? (normal, *t*, *F*, chi-square, uniform)

Questions 7–10 refer to the sample data in the following table (based on data from a Gallup poll). The table summarizes results from a survey of workers and senior-level bosses who were asked if it was seriously unethical to monitor employee e-mail.

	Yes	No
Workers	192	244
Bosses	40	81

7. If using the given sample data for a hypothesis test, what are the appropriate null and alternative hypotheses?

8. If testing the null hypothesis with a 0.05 significance level, find the critical value.

9. Given that the *P*-value for the hypothesis test is 0.0302, what do you conclude when using a 0.05 significance level?

10. Given that the *P*-value for the hypothesis test is 0.0302, what do you conclude when using a 0.01 significance level?

Review Exercises

1. Testing for Adverse Reactions The table on the next page summarizes results from a clinical trial (based on data from Pfizer, Inc). Use a 0.05 significance level to test the claim that experiencing nausea is independent of whether a subject is treated with Celebrex, Ibuprofen, or a placebo. Does the adverse reaction of nausea appear to be about the same for the different treatments?

ISBN 0-558-58875-1

	Celebrex	Ibuprofen	Placebo
Nausea	145	23	78
No Nausea	4001	322	1786

2. Lightning Deaths Listed below are the numbers of deaths from lightning on the different days of the week. The deaths were recorded for a recent period of 35 years (based on data from the National Oceanic and Atmospheric Administration). Use a 0.01 significance level to test the claim that deaths from lightning occur on the different days with the same frequency. Can you provide an explanation for the result?

Day	Sun	Mon	Tues	Wed	Thurs	Fri	Sat
Number of deaths	574	445	429	473	428	422	467

3. Participation in Clinical Trials by Race Researchers conducted a study to investigate racial disparity in clinical trials of cancer. Among the randomly selected participants, 644 were white, 23 were Hispanic, 69 were black, 14 were Asian/Pacific Islander, and 2 were American Indian/Alaskan Native. The proportions of the U.S. population of the same groups are 0.757, 0.091, 0.108, 0.038, and 0.007, respectively. (Based on data from "Participation in Clinical Trials," by Murthy, Krumholz, and Gross, *Journal of the American Medical Association,* Vol. 291, No. 22.) Use a 0.05 significance level to test the claim that the participants fit the same distribution as the U.S. population. Why is it important to have proportionate representation in such clinical trials?

4. Effectiveness of Treatment A clinical trial tested the effectiveness of bupropion hydrochloride in helping people who want to stop smoking. Results of abstinence from smoking 52 weeks after the treatment are summarized in the table below (based on data from "A Double-Blind, Placebo-Controlled, Randomized Trial of Bupropion for Smoking Cessation in Primary Care," by Fossatti, et al., *Archives of Internal Medicine,* Vol. 167, No. 16). Use a 0.05 significance level to test the claim that whether a subject smokes is independent of whether the subject was treated with bupropion hydrochloride or a placebo. Does the bupropion hydrochloride treatment appear to be better than a placebo? Is the bupropion hydrochloride treatment highly effective?

	Bupropion Hydrochloride	Placebo
Smoking	299	167
Not Smoking	101	26

5. McNemar's Test Parents and their children were surveyed in a study of children's respiratory systems. They were asked if the children coughed early in the morning, and results are shown in the table below (based on data from "Cigarette Smoking and Children's Respiratory Symptoms: Validity of Questionnaire Method," by Bland, et al., *Revue d'Epidemiologie et Sante Publique,* Vol. 27). Use a 0.05 significance level to test the claim that the following proportions are the same: (1) the proportion of cases in which the child indicated no cough while the parent indicated coughing; (2) the proportion of cases in which the child indicated coughing while the parent indicated no coughing. What do the results tell us?

		Child Response	
		Cough	No Cough
Parent Response	Cough	29	104
	No Cough	172	5097

ISBN 0-558-58875-1

Cumulative Review Exercises

1. Cleanliness The American Society for Microbiology and the Soap and Detergent Association released survey results indicating that among 3065 men observed in public restrooms, 2023 of them washed their hands, and among 3011 women observed, 2650 washed their hands (based on data from *USA Today*).

a. Is the study an experiment or an observational study?

b. Are the given numbers discrete or continuous?

c. Are the given numbers statistics or parameters?

d. Is there anything about the study that might make the results questionable?

2. Cleanliness Refer to the results given in Exercise 1 and use a 0.05 significance level to test the claim that the proportion of men who wash their hands is equal to the proportion of women who wash their hands. Is there a significant difference?

3. Cleanliness Refer to the results given in Exercise 1. Construct a two-way frequency table and use a 0.05 significance level to test the claim that hand washing is independent of gender.

4. Golf Scores Listed below are first round and fourth round golf scores of randomly selected golfers in a Professional Golf Association Championship (based on data from the *New York Times*). Find the mean, median, range, and standard deviation for the first round scores, then find those same statistics for the fourth round scores. Compare the results.

First round	71	68	75	72	74	67
Fourth round	69	69	69	72	70	73

5. Golf Scores Refer to the sample data given in Exercise 4. Use a 0.05 significance level to test for a linear correlation between the first round scores and the fourth round scores.

6. Golf Scores Using only the first round golf scores given in Exercise 4, construct a 95% confidence interval estimate of the mean first round golf score for all golfers. Interpret the result.

7. Wise Action for Job Applicants In an Accountemps survey of 150 randomly selected senior executives, 88% said that sending a thank-you note after a job interview increases the applicant's chances of being hired (based on data from *USA Today*). Construct a 95% confidence interval estimate of the percentage of all senior executives who believe that a thank-you note is helpful. What very practical advice can be gained from these results?

8. Testing a Claim Refer to the sample results given in Exercise 7 and use a 0.01 significance level to test the claim that more than 75% of all senior executives believe that a thank-you note after a job interview increases the applicant's chances of being hired.

9. Ergonomics When designing the cockpit of a single-engine aircraft, engineers must consider the upper leg lengths of men. Those lengths are normally distributed with a mean of 42.6 cm and a standard deviation of 2.9 cm (based on Data Set 1 in Appendix B).

a. If one man is randomly selected, find the probability that his upper leg length is greater than 45 cm.

b. If 16 men are randomly selected, find the probability that their mean upper leg length is greater than 45 cm.

c. When designing the aircraft cockpit, which result is more meaningful: the result from part (a) or the result from part (b)? Why?

10. Tall Women The probability of randomly selecting a woman who is more than 5 feet tall is 0.925 (based on data from the National Health and Nutrition Examination Survey). Find the probability of randomly selecting five women and finding that all of them are more than 5 feet tall. Is it unusual to randomly select five women and find that all of them are more than 5 feet tall? Why or why not?

ISBN 0-558-58875-1

Technology Project

Use STATDISK, Minitab, Excel, or a TI-83/84 Plus calculator, or any other software package or calculator capable of generating equally likely random digits between 0 and 9 inclusive. Generate 5000 digits and record the results in the accompanying table. Use a 0.05 significance level to test the claim that the sample digits come from a population with a uniform distribution (so that all digits are equally likely). Does the random number generator appear to be working as it should?

Digit	0	1	2	3	4	5	6	7	8	9
Frequency										

INTERNET PROJECT

Contingency Tables

Go to: **http://www.aw.com/triola**

An important characteristic of tests of independence with contingency tables is that the data collected need not be quantitative in nature. A contingency table summarizes observations by the categories or labels of the rows and columns. As a result, characteristics such as gender, race, and political party all become fair game for formal hypothesis testing procedures. In the Internet Project for this chapter you will find links to a variety of demographic data. With these data sets, you will conduct tests in areas as diverse as academics, politics, and the entertainment industry. In each test, you will draw conclusions related to the independence of interesting characteristics.

APPLET PROJECT

Open the Applets folder on the CD and double-click on **Start.** Select the menu item of **Random numbers.** Randomly generate 100 whole numbers between 0 and 9 inclusive. Construct a frequency distribution of the results, then use the methods of this chapter to test the claim that the whole numbers between 0 and 9 are equally likely.

ISBN 0-558-58875-1

FROM DATA TO DECISION

Critical Thinking: Was the law of "women and children first" followed in the sinking of the *Titanic*?

One of the most notable sea disasters occurred with the sinking of the *Titanic* on Monday, April 15, 1912. The table below summarizes the fate of the passengers and crew. A common rule of the sea is that when a ship is threatened with sinking, women and children are the first to be saved.

Fate of Passengers and Crew on the *Titanic*

	Men	Women	Boys	Girls
Survived	332	318	29	27
Died	1360	104	35	18

Analyzing the Results

If we examine the data, we see that 19.6% of the men (332 out of 1692) survived, 75.4% of the women (318 out of 422) survived, 45.3% of the boys (29 out of 64) survived, and 60% of the girls (27 out of 45) survived. There do appear to be differences, but are the differences really *significant?*

First construct a bar graph showing the percentage of survivors in each of the four categories (men, women, boys, girls). What does the graph suggest?

Next, treat the 2223 people aboard the *Titanic* as a *sample*. We could take the position that the *Titanic* data in the above table constitute a *population* and therefore should not be treated as a sample, so that methods of inferential statistics do not apply. But let's stipulate that the data in the table are sample data randomly selected from the population of all theoretical people who would find themselves in the same conditions. Realistically, no other people will actually find themselves in the same conditions, but we will make that assumption for the purposes of this discussion and analysis. We can then determine whether the observed differences have statistical significance. Use one or more formal hypothesis tests to investigate the claim that although some men survived while some women and children died, the rule of "women and children first" was essentially followed. Identify the hypothesis test(s) used and interpret the results by addressing the claim that when the *Titanic* sank on its maiden voyage, the rule of "women and children first" was essentially followed.

Cooperative Group Activities

1. Out-of-class activity Divide into groups of four or five students. The instructions for Exercises 21–24 in Section 11-2 noted that according to Benford's law, a variety of different data sets include numbers with leading (first) digits that follow the distribution shown in the table below. Collect original data and use the methods of Section 11-2 to support or refute the claim that the data conform reasonably well to Benford's law. Here are some possibilities that might be considered: (1) amounts on the checks that you wrote; (2) prices of stocks; (3) populations of counties in the United States; (4) numbers on street addresses; (5) lengths of rivers in the world.

Leading Digit	1	2	3	4	5	6	7	8	9
Benford's law:	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%

2. Out-of-class activity Divide into groups of four or five students and collect past results from a state lottery. Such results are often available on Web sites for individual state lotteries. Use the methods of Section 11-2 to test that the numbers are selected in such a way that all possible outcomes are equally likely.

ISBN 0-558-58875-1

3. Out-of-class activity Divide into groups of four or five students. Each group member should survey at least 15 male students and 15 female students at the same college by asking two questions: (1) Which political party does the subject favor most? (2) If the subject were to make up an absence excuse of a flat tire, which tire would he or she say went flat if the instructor asked? (See Exercise 8 in Section 11-2.) Ask the subject to write the two responses on an index card, and also record the gender of the subject and whether the subject wrote with the right or left hand. Use the methods of this chapter to analyze the data collected. Include these tests:

- The four possible choices for a flat tire are selected with equal frequency.
- The tire identified as being flat is independent of the gender of the subject.
- Political party choice is independent of the gender of the subject.
- Political party choice is independent of whether the subject is right- or left-handed.
- The tire identified as being flat is independent of whether the subject is right- or left-handed.
- Gender is independent of whether the subject is right- or left-handed.
- Political party choice is independent of the tire identified as being flat.

4. Out-of-class activity Divide into groups of four or five students. Each group member should select about 15 other students and first ask them to "randomly" select four digits each. After the four digits have been recorded, ask each subject to write the last four digits of his or her social security number. Take the "random" sample results and mix them into one big sample, then mix the social security digits into a second big sample. Using the "random" sample set, test the claim that students select digits randomly. Then use the social security digits to test the claim that they come from a population of random digits. Compare the results. Does it appear that students can randomly select digits? Are they likely to select any digits more often than others? Are they likely to select any digits less often than others? Do the last digits of social security numbers appear to be randomly selected?

5. In-class activity Divide into groups of three or four students. Each group should be given a die along with the instruction that it should be tested for "fairness." Is the die fair or is it biased? Describe the analysis and results.

6. Out-of-class activity Divide into groups of two or three students. The analysis of last digits of data can sometimes reveal whether values are the results of actual measurements or whether they are reported estimates. Refer to an almanac and find the lengths of rivers in the world, then analyze the last digits to determine whether those lengths appear to be actual measurements or whether they appear to be reported estimates. (Instead of lengths of rivers, you could use heights of mountains, heights of the tallest buildings, lengths of bridges, and so on.)

ISBN 0-558-58875-1

NAME:	Jackie Macmullan
JOB:	Associate Editor, Sports Columnist
COMPANY:	*Boston Globe*

A previous edition of this book included an interview with Bill James, a recognized baseball expert who specializes in the analysis of baseball statistics and identifying the best game strategies based on past results. The Boston Red Sox hired Bill James as an advisor, and he is credited with some of the changes that led to the first World Series victory by the Red Sox in 86 years. Another Bostonian who makes extensive use of statistics in the world of sports is Jackie Macmullan, who is Associate Editor and sports columnist for the Boston Globe.

Q: How do you use statistics when writing about sports?

A: Sports is all about statistics, really. Especially baseball. Yesterday I wrote about why C.C. Sabathia won the Cy Young award over Josh Beckett. Beckett had more wins but lost to Sabathia because of "deeper numbers," which are the underlying statistics. The big one in this case was quality starts—a pretty good indicator of how proficient you were on the mound (Sabathia's 25 vs. Beckett's 20). In many of my stories on sports, and baseball in particular, the use of statistics gives readers a deeper understanding of the game and its players.

Q: Please describe a specific example illustrating the use of statistics.

A: For many years, as I covered the NBA as a beat writer for the Boston Celtics, I kept a binder full of information on every aspect of the game. After each game, I would record everything—points per game, turnovers, assists, etc., for each player on the team. I would also keep track of the opposing team and what they did—any player who scored 10 or more points, for example. I would break down all of the action by quarters, noting where the Celtics turned the game around. These meticulous records kept me on top of every player and their tendencies. I relied on that book for a tremendous amount of obscure little things, too, which I think are what makes a strong sportswriter.

Q: Is your use of probability and statistics increasing, decreasing, or remaining stable?

A: Now, everyone is so sophisticated in their use of statistics. Every college and professional team has their own statistical team that records all their data as they play. Every single team in every single sport has its own scouting system that breaks every statistic down. Here's an example: There was a company hired by the Chicago Bulls that would break down everything Michael Jordan did. Things like, "when Jordan gets ball on left block, shoots 36% of time, on right, 31% of time" so you'd know where he'd prefer to shoot from and his rate of success in every location. Data told you players may like to do a certain thing more, but weren't necessarily successful at it. This really influences other teams and strategies, because they know everything these players do and how good they are at doing it.

Q: How critical do you find your knowledge of statistics for performing your responsibilities?

A: Using statistics really enhances my writing — people don't keep their own stats now, but if you do keep them yourself it takes time, effort, and concentration, and the data become more valuable to you. Collecting data like this provides information that you didn't necessarily know you were looking for. For example, from my NBA binder, I would often find that three quarters way through season, I had learned about player tendencies that I didn't see on my own, or even realize I was recording.

Q: In terms of statistics, what would you recommend for prospective employees?

A: An introductory course would be fine. You really have to know how to crunch numbers to do my job effectively. Some of the best articles are written by people who make these unique statistical observations.

ISBN 0-558-58875-1