# Analysis of Categorical Data

LEARNING OBJECTIVES

The overall objective of this chapter is to give you an understanding of two statistical techniques used to analyze categorical data, thereby enabling you to:

1. Use the chi-square goodness-of-fit test to analyze probabilities of multinomial distribution trials along a single dimension and test a population proportion
2. Use the chi-square test of independence to perform contingency analysis
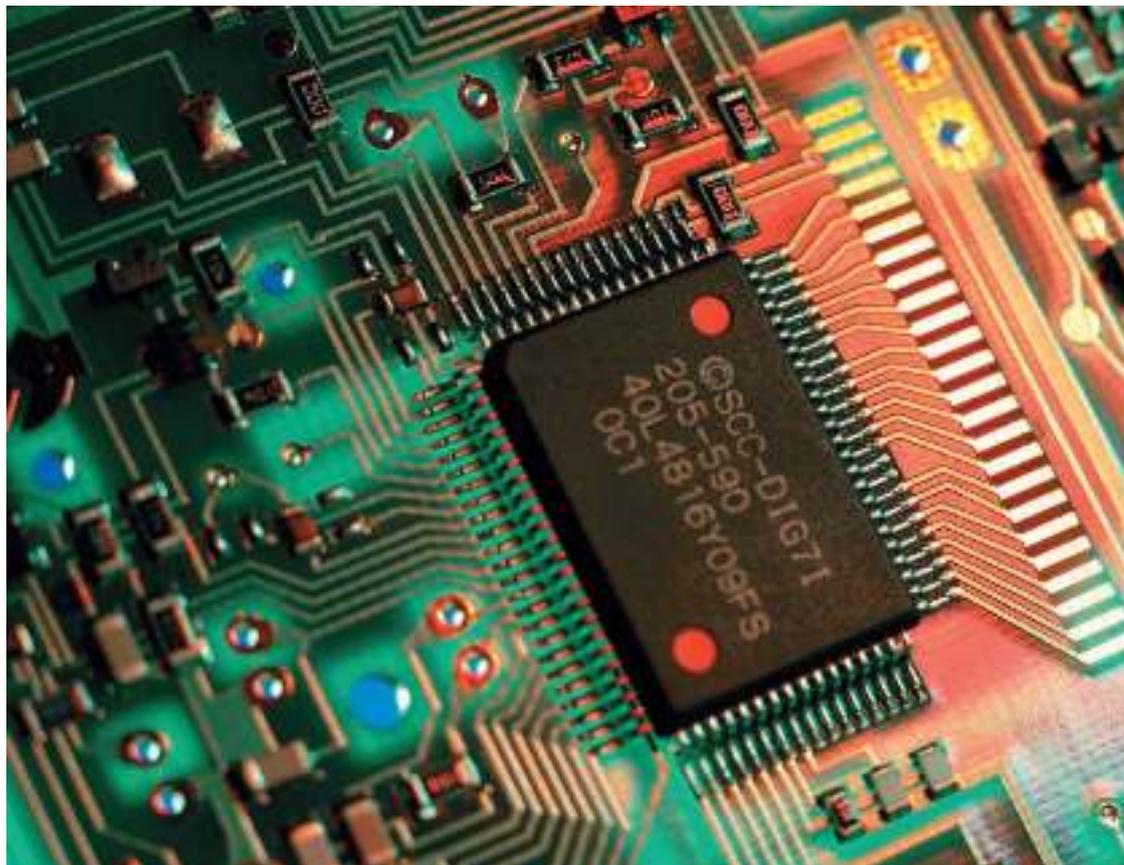
Selecting Suppliers in the Electronics Industry

What criteria are used in the electronics industry to select a supplier? In years past, price has been the dominant criterion of suppliers in many industries, and the supplier with the low bid often got the job. In more recent years, companies have been forced by global competition and a marked increase in quality to examine other aspects of potential suppliers.

Pearson and Ellram investigated the techniques used by firms in the electronics industry to select suppliers, and they wanted to determine if there is a difference between small and large firms in supplier selection. They sent out a survey instrument with questions about criteria used to select and evaluate suppliers, the participation of various functional areas in the selection process, and the formality of methods used in the selection. Of the 210 survey responses received, 87 were from small companies and 123 from large companies. The average sales were $33 million for the small companies and $583 million for the large companies.

Survey questions were stated in such a way as to generate frequencies. The respondents were given a series of supplier selection and evaluation criteria such as quality, cost, current technology, design capabilities, speed to market, manufacturing process, location, and so on. They were asked to check off the criteria used in supplier selection and evaluation and to rank the criteria that they checked. As part of the analysis, the researchers recorded how many of each of the small and large company respondents ranked a criterion first, how many ranked it second, and how many ranked it third. The results are shown in the following table of raw numbers for the criteria of quality, cost, and current technology.

In addition, companies in the study sometimes involved departments such as research/development or engineering in the supplier search process. For example, 41.38% of the small companies in the study included research/development in the supplier search process versus 48.78% of the large companies.

Quality	Company Size	
	Small	Large
1	48	70
2	17	27
3	7	6

Cost	Company Size	
	Small	Large
1	8	14
2	29	36
3	26	37

Current Technology	Company Size	
	Small	Large
1	5	13
2	8	11
3	5	12

Managerial and Statistical Questions

1. Is there a difference between small and large companies in the ranking of criteria for the evaluation and selection of suppliers in the electronics industry?

2. The authors of the study used frequencies to measure the relative rankings of criteria. What is the appropriate statistical technique to analyze these data?

3. In comparing the participation of company employees in the process of selection and evaluation of suppliers by function and by company size, the researchers reported percentages. Are the differences in percentages merely chance differences from samples or is there a significant difference between small and large companies in the extent to which they involve people from various functional areas in the supplier selection and evaluation process? What statistical technique is appropriate for analyzing these data?

Source: Adapted from John N. Pearson and Lisa M. Ellram. "Supplier Selection and Evaluation in Small Versus Large Electronics Firms," *Journal of Small Business Management,* vol. 33, no. 4 (October 1995), pp. 53–65.
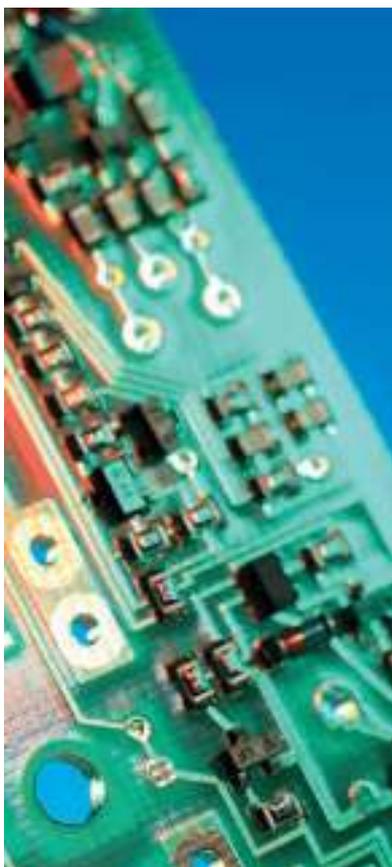
In this chapter, we explore techniques for analyzing categorical data. **Categorical data** are *nonnumerical data that are frequency counts of categories from one or more variables.* For example, it is determined that of the 790 people attending a convention, 240 are engineers, 160 are managers, 310 are sales reps, and 80 are information technologists. The variable is "position in company" with four categories: engineers, managers, sales reps, and information technologists. The data are not ratings or sales figures but rather

frequency counts of how many of each position attended. Research questions producing this type of data are often analyzed using chi-square techniques. The chi-square distribution was introduced in Chapters 8 and 9. The techniques presented here for analyzing categorical data, the *chi-square goodness-of-fit test* and *the chi-square test of independence,* are an outgrowth of the binomial distribution and the inferential techniques for analyzing population proportions.

16.1 CHI-SQUARE GOODNESS-OF-FIT TEST

In Chapter 5, we studied the binomial distribution in which only two possible outcomes could occur on a single trial in an experiment. An extension of the binomial distribution is a multinomial distribution in which more than two possible outcomes can occur in a single trial. **The chi-square goodness-of-fit test** is *used to analyze probabilities of multinomial distribution trials along a single dimension.* For example, if the variable being studied is economic class with three possible outcomes of lower income class, middle income class, and upper income class, the single dimension is economic class and the three possible outcomes are the three classes. On each trial, one and only one of the outcomes can occur. In other words, a family unit must be classified either as lower income class, middle income class, or upper income class and cannot be in more than one class.

The chi-square goodness-of-fit test compares the *expected,* or theoretical, *frequencies* of categories from a population distribution to the *observed,* or actual, *frequencies* from a distribution to determine whether there is a difference between what was expected and what was observed. For example, airline industry officials might theorize that the ages of airline ticket purchasers are distributed in a particular way. To validate or reject this expected distribution, an actual sample of ticket purchaser ages can be gathered randomly, and the observed results can be compared to the expected results with the chi-square goodness-of-fit test. This test also can be used to determine whether the observed arrivals at teller windows at a bank are Poisson distributed, as might be expected. In the paper industry, manufacturers can use the chi-square goodness-of-fit test to determine whether the demand for paper follows a uniform distribution throughout the year.

Formula 16.1 is used to compute a chi-square goodness-of-fit test.

CHI-SQUARE GOODNESS-OF-FIT TEST (16.1)

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\text{df} = k - 1 - c$$

where

f_o = frequency of observed values
f_e = frequency of expected values
k = number of categories
c = number of parameters being estimated from the sample data

---

This formula compares the frequency of observed values to the frequency of the expected values across the distribution. The test loses one degree of freedom because the total number of expected frequencies must equal the number of observed frequencies; that is, the observed total taken from the sample is used as the total for the expected frequencies. In addition, in some instances a population parameter, such as λ, μ, or σ, is estimated from the sample data to determine the frequency distribution of expected values. Each time this estimation occurs, an additional degree of freedom is lost. As a rule, if a uniform distribution is being used as the expected distribution or if an expected distribution of values is given, $k - 1$ degrees of freedom are used in the test. In testing to determine whether an

observed distribution is Poisson, the degrees of freedom are $k - 2$ because an additional degree of freedom is lost in estimating λ. In testing to determine whether an observed distribution is normal, the degrees of freedom are $k - 3$ because two additional degrees of freedom are lost in estimating both μ and σ from the observed sample data.

Karl Pearson introduced the chi-square test in 1900. The **chi-square distribution** is *the sum of the squares of k independent random variables* and therefore can never be less than zero; it extends indefinitely in the positive direction. Actually the chi-square distributions constitute a family, with each distribution defined by the degrees of freedom (df) associated with it. For small df values the chi-square distribution is skewed considerably to the right (positive values). As the df increase, the chi-square distribution begins to approach the normal curve. Table values for the chi-square distribution are given in Appendix A. Because of space limitations, chi-square values are listed only for certain probabilities.

How can the chi-square goodness-of-fit test be applied to business situations? One survey of U.S. consumers conducted by *The Wall Street Journal* and NBC News asked the question: "In general, how would you rate the level of service that American businesses provide?" The distribution of responses to this question was as follows:

Excellent	8%
Pretty good	47%
Only fair	34%
Poor	11%

Suppose a store manager wants to find out whether the results of this consumer survey apply to customers of supermarkets in her city. To do so, she interviews 207 randomly selected consumers as they leave supermarkets in various parts of the city. She asks the customers how they would rate the level of service at the supermarket from which they had just exited. The response categories are excellent, pretty good, only fair, and poor. The observed responses from this study are given in Table 16.1. Now the manager can use a chi-square goodness-of-fit test to determine whether the observed frequencies of responses from this survey are the same as the frequencies that would be expected on the basis of the national survey.

TABLE 16.1

Results of a Local Survey of Consumer Satisfaction

Response	Frequency (f_o)
Excellent	21
Pretty good	109
Only fair	62
Poor	15

HYPOTHESIZE:

STEP 1. The hypotheses for this example follows.

H_o: The observed distribution is the same as the expected distribution.

H_a: The observed distribution is not the same as the expected distribution.

TEST:

STEP 2. The statistical test being used is

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

STEP 3. Let $\alpha = .05$.

STEP 4. Chi-square goodness-of-fit tests are one tailed because a chi-square of zero indicates perfect agreement between distributions. Any deviation from zero difference occurs in the positive direction only because chi-square is determined by a sum of squared values and can never be negative. With four categories in this example (excellent, pretty good, only fair, and poor), $k = 4$. The degrees of freedom are $k - 1$ because the expected distribution is given: $k - 1 = 4 - 1 = 3$. For $\alpha = .05$ and df $= 3$, the critical chi-square value is

$$\chi^2_{.05,3} = 7.8147$$

After the data are analyzed, an observed chi-square greater than 7.8147 must be computed in order to reject the null hypothesis.

STEP 5. The observed values gathered in the sample data from Table 16.1 sum to 207. Thus $n = 207$. The expected proportions are given, but the expected frequencies must be calculated by multiplying the expected proportions by the sample total of the observed frequencies, as shown in Table 16.2.

Response	Expected Proportion	Expected Frequency (f_e) (proportion × sample total)
Excellent	.08	(.08)(207) = 16.56
Pretty good	.47	(.47)(207) = 97.29
Only fair	.34	(.34)(207) = 70.38
Poor	.11	(.11)(207) = 22.77
		207.00

Response	f_o	f_e	$\dfrac{(f_o - f_e)^2}{f_e}$
Excellent	21	16.56	1.19
Pretty good	109	97.29	1.41
Only fair	62	70.38	1.00
Poor	15	22.77	2.65
	207	207.00	6.25

STEP 6. The chi-square goodness-of-fit can then be calculated, as shown in Table 16.3.

ACTION:

STEP 7. Because the observed value of chi-square of 6.25 is not greater than the critical table value of 7.8147, the store manager will not reject the null hypothesis.

BUSINESS IMPLICATIONS:

STEP 8. Thus the data gathered in the sample of 207 supermarket shoppers indicate that the distribution of responses of supermarket shoppers in the manager's city is not significantly different from the distribution of responses to the national survey.

The store manager may conclude that her customers do not appear to have attitudes different from those people who took the survey. Figure 16.1 depicts the chi-square distribution produced by using Minitab for this example, along with the observed and critical values.

FIGURE 16.1

Minitab Graph of Chi-Square Distribution for Service Satisfaction Example

Nonrejection region

Observed $\chi^2 = 6.25$

$\alpha = .05$

$\chi^2_{.05, 3} = 7.8147$

χ^2
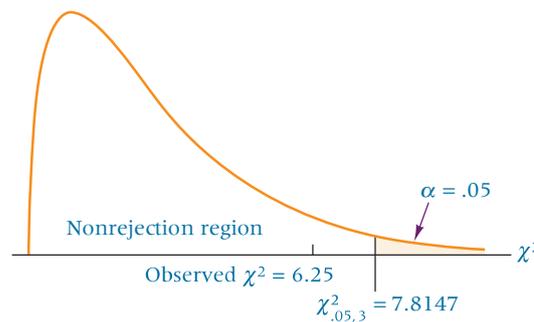
DEMONSTRATION PROBLEM 16.1

Dairies would like to know whether the sales of milk are distributed uniformly over a year so they can plan for milk production and storage. A uniform distribution means that the frequencies are the same in all categories. In this situation, the producers are attempting to determine whether the amounts of milk sold are the same for each month of the year. They ascertain the number of gallons of milk sold by sampling one

large supermarket each month during a year, obtaining the following data. Use $\alpha = .01$ to test whether the data fit a uniform distribution.

Month	Gallons	Month	Gallons
January	1610	August	1350
February	1585	September	1495
March	1649	October	1564
April	1590	November	1602
May	1540	December	1655
June	1397	Total	18,447
July	1410		

Solution

HYPOTHESIZE:

STEP 1. The hypotheses follow.

H_0: The monthly figures for milk sales are uniformly distributed.

H_a: The monthly figures for milk sales are not uniformly distributed.

TEST:

STEP 2. The statistical test used is

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

STEP 3. Alpha is .01.

STEP 4. There are 12 categories and a uniform distribution is the expected distribution, so the degrees of freedom are $k - 1 = 12 - 1 = 11$. For $\alpha = .01$, the critical value is $\chi^2_{.01,11} = 24.725$. An observed chi-square value of more than 24.725 must be obtained to reject the null hypothesis.

STEP 5. The data are given in the preceding table.

STEP 6. The first step in calculating the test statistic is to determine the expected frequencies. The total for the expected frequencies must equal the total for the observed frequencies (18,447). If the frequencies are uniformly distributed, the same number of gallons of milk is expected to be sold each month. The expected monthly figure is

$$\frac{18,447}{12} = 1537.25 \text{ gallons}$$

The following table shows the observed frequencies, the expected frequencies, and the chi-square calculations for this problem.

| Month	f_o	f_e	$\frac{(f_o - f_e)^2}{f_e}$
January	1610	1537.25	3.44
February	1585	1537.25	1.48
March	1649	1537.25	8.12
April	1590	1537.25	1.81
May	1540	1537.25	0.00
June	1397	1537.25	12.80
July	1410	1537.25	10.53
August	1350	1537.25	22.81
September	1495	1537.25	1.16
October	1564	1537.25	0.47
November	1602	1537.25	2.73
December	1655	1537.25	9.02
Total	18,447	18,447.00	$\chi^2 = 74.37$

ACTION:

STEP 7. The observed χ^2 value of 74.37 is greater than the critical table value of $\chi^2_{.01,11} = 24.725$, so the decision is to reject the null hypothesis. This problem provides enough evidence to indicate that the distribution of milk sales is not uniform.

BUSINESS IMPLICATIONS:

STEP 8. Because retail milk demand is not uniformly distributed, sales and production managers need to generate a production plan to cope with uneven demand. In times of heavy demand, more milk will need to be processed or on reserve; in times of less demand, provision for milk storage or for a reduction in the purchase of milk from dairy farmers will be necessary.

The following Minitab graph depicts the chi-square distribution, critical chi-square value, and observed chi-square value.
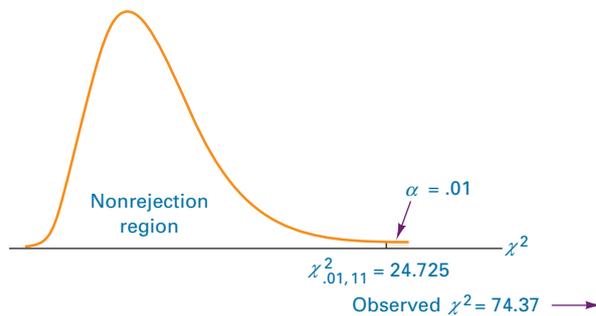
DEMONSTRATION PROBLEM 16.2

Chapter 5 indicated that, quite often in the business world, random arrivals are Poisson distributed. This distribution is characterized by an average arrival rate, λ, per some interval. Suppose a teller supervisor believes the distribution of random arrivals at a local bank is Poisson and sets out to test this hypothesis by gathering information. The following data represent a distribution of frequency of arrivals during 1-minute intervals at the bank. Use $\alpha = .05$ to test these data in an effort to determine whether they are Poisson distributed.

Number of Arrivals	Observed Frequencies
0	7
1	18
2	25
3	17
4	12
≥ 5	5

Solution

HYPOTHESIZE:

STEP 1. The hypotheses follow.

H_0: The frequency distribution is Poisson.
H_a: The frequency distribution is not Poisson.

TEST:

STEP 2. The appropriate statistical test for this problem is

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

STEP 3. Alpha is .05.

STEP 4. The degrees of freedom are $k - 2 = 6 - 1 - 1 = 4$ because the expected distribution is Poisson. An extra degree of freedom is lost, because the value of lambda must be calculated by using the observed sample data. For $\alpha = .05$, the critical table value is $\chi^2_{.05,4} = 9.4877$. The decision rule is to reject the null hypothesis if the observed chi-square is greater than $\chi^2_{.05,4} = 9.4877$.

STEP 5. To determine the expected frequencies, the supervisor must obtain the probability of each category of arrivals and then multiply each by the total of the observed frequencies. These probabilities are obtained by determining lambda and then using the Poisson table. As it is the mean of a Poisson distribution, lambda can be determined from the observed data by computing the mean of the data. In this case, the supervisor computes a weighted average by summing the product of number of arrivals and frequency of those arrivals and dividing that sum by the total number of observed frequencies.

Number of Arrivals	Observed Frequencies	Arrival × Observed
0	7	0
1	18	18
2	25	50
3	17	51
4	12	48
≥ 5	5	25
	84	192

$$\lambda = \frac{192}{84} = 2.3$$

With this value of lambda and the Poisson distribution table in Appendix A, the supervisor can determine the probabilities of the number of arrivals in each category. The expected probabilities are determined from Table A.3 by looking up the values of $x = 0, 1, 2, 3,$ and 4 in the column under $\lambda = 2.3$, shown in the following table as expected probabilities. The probability for $x \geq 5$ is determined by summing the probabilities for the values of $x = 5, 6, 7, 8,$ and so on. Using these probabilities and the total of 84 from the observed data, the supervisor computes the expected frequencies by multiplying each expected probability by the total (84).

Arrivals	Expected Probabilities	Expected Frequencies
0	.1003	8.42
1	.2306	19.37
2	.2652	22.28
3	.2033	17.08
4	.1169	9.82
≥ 5	.0837	7.03
		84.00

STEP 6. The supervisor uses these expected frequencies and the observed frequencies to compute the observed value of chi-square.

| Arrivals	Observed Frequencies	Expected Frequencies	$\frac{(f_o - f_e)^2}{f_e}$
0	7	8.42	.24
1	18	19.37	.10
2	25	22.28	.33
3	17	17.08	.00
4	12	9.82	.48
≥ 5	5	7.03	.59
	84	84.00	$\chi^2 = 1.74$

ACTION:

STEP 7. The observed value of 1.74 is not greater than the critical chi-square value of 9.4877, so the supervisor's decision is to not reject the null hypothesis. In other words, he fails to reject the hypothesis that the distribution of bank arrivals is Poisson.

BUSINESS IMPLICATIONS:

STEP 8. The supervisor can use the Poisson distribution as the basis for other types of analysis, such as queuing modeling.

The following Minitab graph depicts the chi-square distribution, critical value, and computed value.
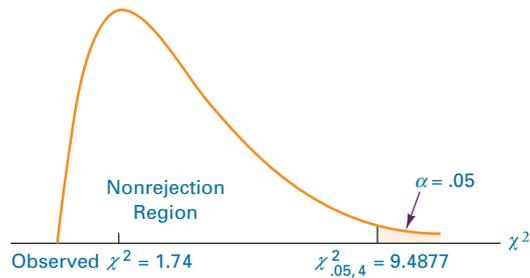
Caution: *When the expected value of a category is small, a large chi-square value can be obtained erroneously, leading to a Type I error. To control for this potential error, the chi-square goodness-of-fit test should not be used when any of the expected frequencies is less than 5. If the observed data produce expected values of less than 5, combining adjacent categories (when meaningful) to create larger frequencies may be possible.*

Testing a Population Proportion by Using the Chi-Square Goodness-of-Fit Test as an Alternative Technique to the *z* Test

In Chapter 9 we discussed a technique for testing the value of a population proportion. When sample size is large enough ($n \cdot p \geq 5$ and $n \cdot q \geq 5$), sample proportions are normally distributed and the following formula can be used to test hypotheses about p.

$$z = \frac{\hat{p} - p}{\sqrt{\dfrac{p \cdot q}{n}}}$$

The chi-square goodness-of-fit test can also be used to conduct tests about p; this situation can be viewed as a special case of the chi-square goodness-of-fit test where the number of classifications equals two (binomial distribution situation). The observed chi-square is computed in the same way as in any other chi-square goodness-of-fit test, but because

the test contains only two classifications (success or failure), $k = 2$ and the degrees of freedom are $k - 1 = 2 - 1 = 1$.

As an example, we will work two problems from Section 9.4 by using chi-square methodology. The first example in Section 9.4 tests the hypothesis that exactly 8% of a manufacturer's products are defective. The following hypotheses are being tested.

$$H_o: p = .08$$
$$H_a: p \neq .08$$

The value of alpha was given to be .10. To test these hypotheses, a business researcher randomly selected a sample of 200 items and determined that 33 of the items had at least one flaw.

Working this problem by the chi-square goodness-of-fit test, we view it as a two-category expected distribution in which we expect .08 defects and .92 nondefects. The observed categories are 33 defects and $200 - 33 = 167$ nondefects. Using the total observed items (200), we can determine an expected distribution as $.08(200) = 16$ and $.92(200) = 184$. Shown here are the observed and expected frequencies.

	f_o	f_e
Defects	33	16
Nondefects	167	184

Alpha is .10 and this test is two-tailed, so $\alpha/2 = .05$. The degrees of freedom are 1. The critical table chi-square value is

$$\chi^2_{.05,1} = 3.8415$$

An observed chi-square value greater than this value must be obtained to reject the null hypothesis. The chi-square for this problem is calculated as follows.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(33 - 16)^2}{16} + \frac{(167 - 184)^2}{184} = 18.06 + 1.57 = 19.63$$

Notice that this observed value of chi-square, 19.63, is greater than the critical table value, 3.8415. The decision is to reject the null hypotheses. The manufacturer does not produce 8% defects according to this analysis. Observing the actual sample result, in which .165 of the sample was defective, indicates that the proportion of the population that is defective might be greater than 8%.

The results obtained are approximately the same as those computed in Chapter 9, in which an observed z value of 4.43 was determined and compared to a critical z value of 1.645, causing us to reject the null hypothesis. This result is not surprising to researchers who understand that when the degrees of freedom equal 1, the value of χ^2 equals z^2.

DEMONSTRATION PROBLEM 16.3

Rework Demonstration Problem 9.3 using the chi-square goodness-of-fit technique.

Solution

In this problem, we tested to determine whether the residents of Wisconsin consume a significantly higher proportion of milk as their primary breakfast beverage than the .17 figure for the United States. The hypotheses were

$$H_o: p = .17$$
$$H_a: p > .17$$

The value of alpha was .05, and it is a one-tailed test. The degrees of freedom are $k - 1 = 2 - 1 = 1$, as there are $k = 2$ categories (milk or not milk). The critical table value for chi-square is

$$\chi^2_{.05,1} = 3.8415$$

To test these hypotheses, a sample of 550 people were contacted. Of these, 115 declared that milk was their primary breakfast beverage. The observed categories are 115 and 550 − 115 = 435. The expected categories are determined by multiplying .17 and .83 by the observed total number (550). Thus, the expected categories are .17(550) = 93.5 and .83(550) = 456.5. These frequencies follow.

	f_o	f_e
Milk	115	93.5
Not milk	435	456.5

The observed chi-square is determined by

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(115 - 93.5)^2}{93.5} + \frac{(435 - 456.5)^2}{456.5} = 4.94 + 1.01 = 5.95$$

This observed chi-square, 5.95, is greater than the critical chi-square value of 3.8415. The decision is to reject the null hypothesis. The proportion of residents who drink milk as their primary breakfast beverage is higher in Wisconsin than in the United States as a whole. In Demonstration Problem 9.3, an observed z value of 2.44 was obtained, which was greater than the critical value of 1.645, allowing us to reject the null hypothesis. The results obtained by the two different methods (χ^2 and z) are essentially the same, with the observed value of χ^2 approximately equal to z^2 ($z = 2.44$, $z^2 = 5.95 = \chi^2$).

16.1 PROBLEMS

16.1 Use a chi-square goodness-of-fit test to determine whether the observed frequencies are distributed the same as the expected frequencies ($\alpha = .05$).

Category	f_o	f_e
1	53	68
2	37	42
3	32	33
4	28	22
5	18	10
6	15	8

16.2 Use the following data and $\alpha = .01$ to determine whether the observed frequencies represent a uniform distribution.

Category	f_o
1	19
2	17
3	14
4	18
5	19
6	21
7	18
8	18

16.3 Are the following data Poisson distributed? Use $\alpha = .05$ and the chi-square goodness-of-fit test to answer this question. What is your estimated lambda?

Number of Arrivals	f_o
0	28
1	17
2	11
3	5

16.4 Use the chi-square goodness-of-fit to test to determine if the following observed data are normally distributed. Let $\alpha = .05$. What are your estimated mean and standard deviation?

Category	Observed
10–under 20	6
20–under 30	14
30–under 40	29
40–under 50	38
50–under 60	25
60–under 70	10
70–under 80	7

16.5 In one survey, successful female entrepreneurs were asked to state their personal definition of success in terms of several categories from which they could select. Thirty-nine percent responded that happiness was their definition of success, 12% said that sales/profit was their definition, 18% responded that helping others was their definition, and 31% responded that achievements/challenge was their definition. Suppose you wanted to determine whether male entrepreneurs felt the same way and took a random sample of men, resulting in the following data. Use the chi-square goodness-of-fit test to determine whether the observed frequency distribution of data for men is the same as the distribution for women. Let $\alpha = .05$.

Definition	f_o
Happiness	42
Sales/profit	95
Helping others	27
Achievements/challenge	63

16.6 The following percentages come from a national survey of the ages of prerecorded-music shoppers. A local survey produced the observed values. Does the evidence in the observed data indicate that we should reject the national survey distribution for local prerecorded-music shoppers? Use $\alpha = .01$.

Age	Percent from Survey	f_o
10–14	9	22
15–19	23	50
20–24	22	43
25–29	14	29
30–34	10	19
≥ 35	22	49

16.7 The general manager of a Major League Baseball team believes the ages of purchasers of game tickets are normally distributed. The following data represent the distribution of ages for a sample of observed purchasers of Major League Baseball game tickets. Use the chi-square goodness-of-fit test to determine whether this distribution is significantly different from the normal distribution. Assume that $\alpha = .05$.

Age of Purchaser	Frequency
10–under 20	16
20–under 30	44
30–under 40	61
40–under 50	56
50–under 60	35
60–under 70	19

16.8 The Springfield Emergency Medical Service keeps records of emergency telephone calls. A study of 150 five-minute time intervals resulted in the distribution of number of calls as follows. For example, during 18 of the 5-minute intervals, no calls occurred. Use the chi-square goodness-of-fit test and $\alpha = .01$ to determine whether this distribution is Poisson.

Number of Calls (per 5-minute interval)	Frequency
0	18
1	28
2	47
3	21
4	16
5	11
6 or more	9

16.9 According to an extensive survey conducted for *Business Marketing* by Leo J. Shapiro & Associates, 66% of all computer companies are going to spend more on marketing this year than in previous years. Only 33% of other information technology companies and 28% of non–information technology companies are going to spend more. Suppose a researcher wanted to conduct a survey of her own to test the claim that 28% of all non–information technology companies are spending more on marketing next year than this year. She randomly selects 270 companies and determines that 62 of the companies do plan to spend more on marketing next year. Use $\alpha = .05$, the chi-square goodness-of-fit test, and the sample data to test to determine whether the 28% figure holds for all non–information technology companies.

16.10 Cross-cultural training is rapidly becoming a popular way to prepare executives for foreign management positions within their company. This training includes such aspects as foreign language, previsit orientations, meetings with former expatriates, and cultural background information on the country. According to Runzheimer International, 30% of all major companies provide formal cross-cultural programs to their executives being relocated in foreign countries. Suppose a researcher wants to test this figure for companies in the communications industry to determine whether the figure is too high for that industry. In a random sample, 180 communications firms are contacted; 42 provide such a program. Let $\alpha = .05$ and use the chi-square goodness-of-fit test to determine whether the .30 proportion for all major companies is too high for this industry.

16.2 CONTINGENCY ANALYSIS: CHI-SQUARE TEST OF INDEPENDENCE

Interactive Applet

The chi-square goodness-of-fit test is used to analyze the distribution of frequencies for categories of *one* variable, such as age or number of bank arrivals, to determine whether the distribution of these frequencies is the same as some hypothesized or expected distribution. However, the goodness-of-fit test cannot be used to analyze *two* variables simultaneously. A different chi-square test, the **chi-square test of independence**, can be *used to analyze the frequencies of two variables with multiple categories to determine whether the two variables are independent.* Many times this type of analysis is desirable. For example, a market researcher might want to determine whether the type of soft drink preferred by a consumer is independent of the consumer's age. An organizational behaviorist might want to know whether absenteeism

is independent of job classification. Financial investors might want to determine whether type of preferred stock investment is independent of the region where the investor resides.

The chi-square test of independence can be used to analyze any level of data measurement, but it is particularly useful in analyzing nominal data. Suppose a business researcher is interested in determining whether geographic region is independent of type of financial investment. On a questionnaire, the following two questions might be used to measure geographic region and type of financial investment.

In which region of the country do you reside?

A. Northeast   B. Midwest   C. South   D. West

Which type of financial investment are you most likely to make today?

E. Stocks   F. Bonds   G. Treasury Bills

The business researcher would *tally the frequencies of responses* to these two questions into a two-way table called a **contingency table**. Because the chi-square test of independence uses a contingency table, this test is sometimes referred to as **contingency analysis**.

Depicted in Table 16.4 is a contingency table for these two variables. Variable 1, geographic region, uses four categories: A, B, C, and D. Variable 2, type of financial investment, uses three categories: E, F, and G. The observed frequency for each cell is denoted as o_{ij}, where i is the row and j is the column. Thus, o_{13} is the observed frequency for the cell in the first row and third column. The expected frequencies are denoted in a similar manner.

If the two variables are independent, they are not related. In a sense, the chi-square test of independence is a test of whether the variables are related. The null hypothesis for a chi-square test of independence is that the two variables are independent (not related). If the null hypothesis is rejected, the conclusion is that the two variables are not independent and are related.

Assume at the beginning that variable 1 and variable 2 are independent. The probability of the intersection of two of their respective categories, A and F, can be found by using the multiplicative law for independent events presented in Chapter 4:

$$P(A \cap F) = P(A) \cdot P(F)$$

If A and F are independent, then

$$P(A) = \frac{n_A}{N}, P(F) = \frac{n_F}{N}, \text{ and } P(A \cap F) = \frac{n_A}{N} \cdot \frac{n_F}{N}$$

<table>
<tr><td>

TABLE 16.4

Contingency Table for the Investment Example

</td><td>

Type of Financial Investment

		E	F	G	
	A			o_{13}	n_A
Geographic	B				n_B
Region	C				n_C
	D				n_D
		n_E	n_F	n_G	N

</td></tr>
</table>

If $P(A \cap F)$ is multiplied by the total number of frequencies, N, the expected frequency for the cell of A and F can be determined.

$$e_{AF} = \frac{n_A}{N} \cdot \frac{n_F}{N}(N) = \frac{n_A \cdot n_F}{N}$$

In general, if the two variables are independent, the expected frequency values of each cell can be determined by

$$e_{ij} = \frac{n_i \cdot n_j}{N}$$

where:

i = the row
j = the column
n_i = the total of row i
n_j = the total of column j
N = the total of all frequencies

Using these expected frequency values and the observed frequency values, we can compute a chi-square test of independence to determine whether the variables are independent. Formula 16.2 is the formula for accomplishing this test.

CHI-SQUARE TEST OF INDEPENDENCE (16.2)

$$\chi^2 = \sum\sum \frac{(f_o - f_e)^2}{f_e}$$

where:

$df = (r-1)(c-1)$
r = number of
c = number of columns

The null hypothesis for a chi-square test of independence is that the two variables are independent. The alternative hypothesis is that the variables are not independent. This test is one-tailed. The degrees of freedom are $(r-1)(c-1)$. Note that formula 16.2 is similar to formula 16.1, with the exception that the values are summed across both rows and columns and the degrees of freedom are different.

Suppose a business researcher wants to determine whether type of gasoline preferred is independent of a person's income. She takes a random survey of gasoline purchasers, asking them one question about gasoline preference and a second question about income. The respondent is to check whether he or she prefers (1) regular gasoline, (2) premium gasoline, or (3) extra premium gasoline. The respondent also is to check his or her income brackets as being (1) less than \$30,000, (2) \$30,000 to \$49,999, (3) \$50,000 to \$99,999, or (4) more than \$100,000. The business researcher tallies the responses and obtains the results in Table 16.5. Using $\alpha = .01$, she can use the chi-square test of independence to determine whether type of gasoline preferred is independent of income level.

HYPOTHESIZE:

STEP 1. The hypotheses follow.

H$_o$: Type of gasoline is independent of income.
H$_a$: Type of gasoline is not independent of income.

		Type of Gasoline			
		Regular	Premium	Extra Premium	
	Less than $30,000	85	16	6	107
	$30,000 to $49,999	102	27	13	142
Income	$50,000 to $99,999	36	22	15	73
	More than $100,000	15	23	25	63
		238	88	59	385

TEST:

STEP 2. The appropriate statistical test is

$$\chi^2 = \sum \sum \frac{(f_o - f_e)^2}{f_e}$$

STEP 3. Alpha is .01.

STEP 4. Here, there are four rows ($r = 4$) and three columns ($c = 3$). The degrees of freedom are $(4 - 1)(3 - 1) = 6$. The critical value of chi-square for $\alpha = .01$ is $\chi^2_{.01,6} = 16.8119$. The decision rule is to reject the null hypothesis if the observed chi-square is greater than 16.8119.

STEP 5. The observed data appear in Table 16.5.

STEP 6. To determine the observed value of chi-square, the researcher must compute the expected frequencies. The expected values for this example are calculated as follows, with the first term in the subscript (and numerator) representing the row and the second term in the subscript (and numerator) representing the column.

$$e_{11} = \frac{(n_{1\cdot})(n_{\cdot 1})}{N} = \frac{(107)(238)}{385} = 66.15 \qquad e_{21} = \frac{(n_{2\cdot})(n_{\cdot 1})}{N} = \frac{(142)(238)}{385} = 87.78$$

$$e_{12} = \frac{(n_{1\cdot})(n_{\cdot 2})}{N} = \frac{(107)(88)}{385} = 24.46 \qquad e_{22} = \frac{(n_{2\cdot})(n_{\cdot 2})}{N} = \frac{(142)(88)}{385} = 32.46$$

$$e_{13} = \frac{(n_{1\cdot})(n_{\cdot 3})}{N} = \frac{(107)(59)}{385} = 16.40 \qquad e_{23} = \frac{(n_{2\cdot})(n_{\cdot 3})}{N} = \frac{(142)(59)}{385} = 21.76$$

$$e_{31} = \frac{(n_{3\cdot})(n_{\cdot 1})}{N} = \frac{(73)(238)}{385} = 45.13 \qquad e_{41} = \frac{(n_{4\cdot})(n_{\cdot 1})}{N} = \frac{(63)(238)}{385} = 38.95$$

$$e_{32} = \frac{(n_{3\cdot})(n_{\cdot 2})}{N} = \frac{(73)(88)}{385} = 16.69 \qquad e_{42} = \frac{(n_{4\cdot})(n_{\cdot 2})}{N} = \frac{(63)(88)}{385} = 14.40$$

$$e_{33} = \frac{(n_{3\cdot})(n_{\cdot 3})}{N} = \frac{(73)(59)}{385} = 11.19 \qquad e_{43} = \frac{(n_{4\cdot})(n_{\cdot 3})}{N} = \frac{(63)(59)}{385} = 9.65$$

The researcher then lists the expected frequencies in the cells of the contingency tables along with observed frequencies. In this text, expected frequencies are enclosed in parentheses. Table 16.6 provides the contingency table for this example.

Contingency Table of Observed and Expected Frequencies for Gasoline Consumer Example

		Type of Gasoline			
		Regular	Premium	Extra Premium	
Income	Less than $30,000	(66.15) 85	(24.46) 16	(16.40) 6	107
	$30,000 to $49,999	(87.78) 102	(32.46) 27	(21.76) 13	142
	$50,000 to $99,999	(45.13) 36	(16.69) 22	(11.19) 15	73
	More than $100,000	(38.95) 15	(14.40) 23	(9.65) 25	63
		238	88	59	385

Next, the researcher computes the chi-square value by summing $(f_o - f_e)^2 / f_e$ for all cells.

$$\chi^2 = \frac{(85 - 66.15)^2}{66.15} + \frac{(16 - 24.46)^2}{24.46} + \frac{(6 - 16.40)^2}{16.40} + \frac{(102 - 87.78)^2}{87.78} + \frac{(27 - 32.46)^2}{32.46}$$

$$+ \frac{(13 - 21.76)^2}{21.76} + \frac{(36 - 45.13)^2}{45.13} + \frac{(22 - 16.69)^2}{16.69} + \frac{(15 - 11.19)^2}{11.19}$$

$$+ \frac{(15 - 38.95)^2}{38.95} + \frac{(23 - 14.40)^2}{14.40} + \frac{(25 - 9.65)^2}{9.65} = 5.37 + 2.93 + 6.60 + 2.30$$

$$+ 0.92 + 3.53 + 1.85 + 1.69 + 1.30 + 14.73 + 5.14 + 24.42 = 70.78$$

ACTION:

STEP 7. The observed value of chi-square, 70.78, is greater than the critical value of chi-square, 16.8119, obtained from Table A.8. The business researcher's decision is to reject the null hypothesis; that is, type of gasoline preferred is not independent of income.

Minitab Output for Gasoline Consumer Example

```
Chi-Square Test: Regular, Premium, Extra Premium
Expected counts are printed below observed counts
Chi-square contributions are printed below expected counts

                             Extra
         Regular  Premium   Premium    Total
    1         85       16         6      107
           66.15    24.46     16.40
           5.374    2.924     6.593

    2        102       27        13      142
           87.78    32.46     21.76
           2.303    0.918     3.527

    3         36       22        15       73
           45.13    16.69     11.19
           1.846    1.693     1.300

    4         15       23        25       63
           38.95    14.40      9.65
          14.723    5.136    24.391

Total       238       88        59      385
Chi-Sq = 70.727, DF = 6, P-Value = 0.000
```

FIGURE 16.3

Minitab Graph of Chi-Square
Distribution for Gasoline
Consumer Example

Nonrejection
Region

$\alpha = .01$

$\chi^2_{.01,6} = 16.8119$

χ^2

Observed $\chi^2 = 70.78 \longrightarrow$
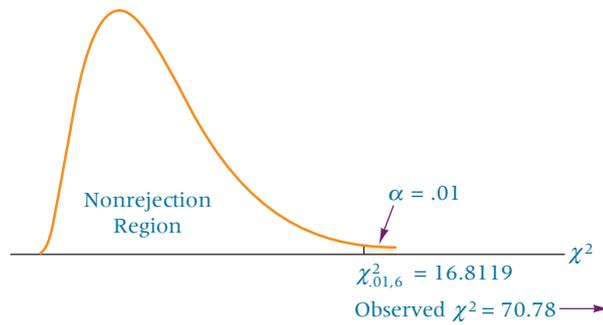
BUSINESS IMPLICATION:

STEP 8. Having established that conclusion, the business researcher can then examine the outcome to determine which people, by income brackets, tend to purchase which type of gasoline and use this information in market decisions.

Figure 16.2 is the Minitab output for calculating the chi-square value. Figure 16.3 is the Minitab chi-square graph with the critical value, the rejection region, and the observed χ^2.

**DEMONSTRATION
PROBLEM 16.4**

Is the type of beverage ordered with lunch at a restaurant independent of the age of the consumer? A random poll of 309 lunch customers is taken, resulting in the following contingency table of observed values. Use $\alpha = .01$ to determine whether the two variables are independent.

Preferred Beverage

		Coffee/Tea	Soft Drink	Other (Milk, etc.)	
	21–34	26	95	18	139
Age	35–55	41	40	20	101
	>55	24	13	32	69
		91	148	70	309

Solution

HYPOTHESIZE:

STEP 1. The hypotheses follow.

H$_0$: Type of beverage preferred is independent of age.
H$_a$: Type of beverage preferred is not independent of age.

TEST:

STEP 2. The appropriate statistical test is

$$\chi^2 = \sum\sum\frac{(f_o - f_e)^2}{f_e}$$

STEP 3. Alpha is .01.

STEP 4. The degrees of freedom are $(3 - 1)(3 - 1) = 4$, and the critical value is $\chi^2_{.01,4} = 13.2767$. The decision rule is to reject the null hypothesis if the observed value of chi-square is greater than 13.2767.

STEP 5. The sample data were shown previously.

STATISTICS IN BUSINESS TODAY

Risk Taking by Ad Agencies

A study was conducted by Douglas C. West to examine under what conditions, if any, advertising agencies take more risk. The primary statistical technique used to analyze the data in the study was the chi-square test of independence. Although several studies previously examined risk on the part of advertisers, little research addresses the willingness of advertising agencies to take risks on behalf of their clients. West theorized that people are more apt to take risks when the outcome affects someone else's livelihood and income rather than their own; consequently, advertising agencies might be more willing to take risks than advertisers. In addition, he theorized that advertising agencies might be more willing to take risks with smaller clients rather than large clients and that newer agencies might tend to be more willing to take risks than older agencies.

The study involved 64 account directors and/or heads of creative departments of advertising agencies selected from a standard directory. Respondents were presented with two advertising options under a plan to launch a new product. Plan A was a standard one with an average rate of return (risk averse), and Plan B was an uncertain one in which there is a 50% chance of getting a lower rate of return than the client's worst forecast and a 50% chance of getting a better rate of return than the client's highest forecast (risk seeking). Using a chi-square test of independence, the percentages of respondents selecting each plan were compared to percentages produced in a similar study with advertisers. The result was that the proportions of agency respondents that were risk averse/risk seeking were not significantly different from the proportions of advertisers that were risk averse/risk seeking ($\chi^2 = 3.165$, $p = .076$). Agencies and advertisers were also compared on four degrees of risk in light of the risk taken with their most recent client. The result showed no significant difference between agencies and advertisers on the amount of risk taken with their most recent client ($\chi^2 = 3.165$, $p = .076$, $\alpha = .05$). Thus, on two questions, there was no difference in the risk taking between agencies and advertisers.

Are there circumstances under which an advertising agency might be more willing to take risks? Respondents were asked to what degree they were willing to take risks if the client is their smallest client versus if the client is their largest client. A 4×2 contingency table was constructed with four degrees of risk and the two client sizes. Analysis of these data produced a significant chi-square of 9.819 ($p = .021$) showing that agencies tended to be more risk taking with smaller clients than with large clients.

The effect of agency age on participant selection of Plan A versus Plan B was analyzed using a 2×2 contingency table. Agencies were separated into two age categories (3–19 years versus 20–135 years) with Plan A and Plan B as the risk options. Using a chi-square test of independence, it was determined that a significantly higher proportion of the younger agencies were more risk seeking than older agencies ($\chi^2 = 6.75$, $p = .01$).

In this study, the chi-square test of independence allowed for categorical comparisons between agencies and advertisers, between small clients and large clients, and between young and old agencies. In many other studies, chi-square categorical statistics can be used to study business phenomena.

Source: Adapted from Douglas C. West, "360° of Creative Risk," *Journal of Advertising Research*, vol. 39, no. 1 (January/February 1999), pp. 39–50.

STEP 6.

The expected frequencies are the product of the row and column totals divided by the grand total. The contingency table, with expected frequencies, follows.

<table>
<tr><td colspan="5" align="center">Preferred Beverage</td></tr>
<tr><td></td><td></td><td>Coffee/Tea</td><td>Soft Drink</td><td>Other (Milk, etc.)</td><td></td></tr>
<tr><td></td><td>21–34</td><td>(40.94)
26</td><td>(66.58)
95</td><td>(31.49)
18</td><td>139</td></tr>
<tr><td>Age</td><td>35–55</td><td>(29.74)
41</td><td>(48.38)
40</td><td>(22.88)
20</td><td>101</td></tr>
<tr><td></td><td>>55</td><td>(20.32)
24</td><td>(33.05)
13</td><td>(15.63)
32</td><td>69</td></tr>
<tr><td></td><td></td><td>91</td><td>148</td><td>70</td><td>309</td></tr>
</table>

For these values, the observed χ^2 is

$$\chi^2 = \frac{(26 - 40.94)^2}{40.94} + \frac{(95 - 66.58)^2}{66.58} + \frac{(18 - 31.49)^2}{31.49} + \frac{(41 - 29.74)^2}{29.74}$$
$$+ \frac{(40 - 48.38)^2}{48.38} + \frac{(20 - 22.88)^2}{22.88} + \frac{(24 - 20.32)^2}{20.32} + \frac{(13 - 33.05)^2}{33.05}$$
$$+ \frac{(32 - 15.63)^2}{15.63} = 5.45 + 12.13 + 5.78 + 4.26 + 1.45 + 0.36 + 0.67$$
$$+ 12.16 + 17.15 = 59.41$$

ACTION:

STEP 7. The observed value of chi-square, 59.41, is greater than the critical value, 13.2767, so the null hypothesis is rejected.

BUSINESS IMPLICATIONS:

STEP 8. The two variables—preferred beverage and age—are not independent. The type of beverage that a customer orders with lunch is related to or dependent on age. Examination of the categories reveals that younger people tend to prefer soft drinks and older people prefer other types of beverages. Managers of eating establishments and marketers of beverage products can utilize such information in targeting their market and in providing appropriate products.

Caution: *As with the chi-square goodness-of-fit test, small expected frequencies can lead to inordinately large chi-square values with the chi-square test of independence. Hence contingency tables should not be used with expected cell values of less than 5. One way to avoid small expected values is to collapse (combine) columns or rows whenever possible and whenever doing so makes sense.*

16.2 PROBLEMS

16.11 Use the following contingency table to test whether variable 1 is independent of variable 2. Let $\alpha = .01$.

		Variable 2	
Variable 1		203	326
		68	110

16.12 Use the following contingency table to determine whether variable 1 is independent of variable 2. Let $\alpha = .01$.

	Variable 2			
Variable 1	24	13	47	58
	93	59	187	244

16.13 Use the following contingency table and the chi-square test of independence to determine whether social class is independent of number of children in a family. Let $\alpha = .05$.

		Social Class		
		Lower	*Middle*	*Upper*
Number of Children	*0*	7	18	6
	1	9	38	23
	2 or 3	34	97	58
	More than 3	47	31	30

16.14 A group of 30-year-olds is interviewed to determine whether the type of music most listened to by people in their age category is independent of the geographic location of their residence. Use the chi-square test of independence, $\alpha = .01$, and the following contingency table to determine whether music preference is independent of geographic location.

		Type of Music Preferred			
		Rock	*R & B*	*Country*	*Classical*
	Northeast	140	32	5	18
Geographic Region	*South*	134	41	52	8
	West	154	27	8	13

16.15 Is the transportation mode used to ship goods independent of type of industry? Suppose the following contingency table represents frequency counts of types of transportation used by the publishing and the computer hardware industries. Analyze the data by using the chi-square test of independence to determine whether type of industry is independent of transportation mode. Let $\alpha = .05$.

		Transportation Mode		
		Air	*Train*	*Truck*
	Publishing	32	12	41
Industry	*Computer Hardware*	5	6	24

16.16 According to data released by the U.S. Department of Housing and Urban Development about new homes built in the United States, there is an almost 50–50 split between one-story and two-story homes. In addition, more than half of all new homes have three bedrooms. Suppose a study is done to determine whether the number of bedrooms in a new home is independent of the number of stories. Use $\alpha = .10$ and the following contingency table to conduct a chi-square test of independence to determine whether, in fact, the number of bedrooms is independent of the number of stories.

		Number of Bedrooms		
		≤2	*3*	*≥4*
Number of Stories	*1*	116	101	57
	2	90	325	160

16.17 A study was conducted to determine the impact of a major Mexican peso devaluation on U.S. border retailers. As a part of the study, data were gathered on the magnitude of business that U.S. border retailers were doing with Mexican citizens. Forty-one shoppers of border city department stores were interviewed; 24 were Mexican citizens, and the rest were U.S. citizens. Thirty-five discount store shoppers were interviewed, as were 30 hardware store and 60 shoe store customers. In these three groups, 20, 11, and 32 were Mexican citizens, and the remaining shoppers were U.S. citizens. Use a chi-square contingency analysis to determine whether the shoppers' citizenship (Mexican versus American) is independent of type of border city retailer (department, discount, hardware, shoe) for these data. Let $\alpha = .05$.

Selecting Suppliers

Pearson and Ellram examined the comparative rankings of selection and evaluation criteria for suppliers of small and large electronics firms. These researchers chose to analyze the relative rankings by using frequency counts for various categories of rankings. The three tables of data displayed in the Decision Dilemma contain frequencies for small and large company respondents on three of these criteria: quality, cost, and current technology. Because each of these tables contains categorical data with two variables, company size and rank category, a contingency analysis (chi-square test of independence) is an appropriate statistical technique for analyzing the data. The null hypothesis is that company size is independent of category of rank for each criterion. The alternative hypothesis is that category of rank for each criterion is not independent of company size (company size makes a difference in how the criterion is ranked). Minitab chi-square analysis produced the following output. For quality:

$$\text{Chi-Sq} = 0.991, \text{DF} = 2, P\text{-Value} = 0.609$$

The p-value of .609 indicates that we fail to reject the null hypothesis. On quality as a criterion, rankings of small company respondents are independent of rankings of large company respondents. There appears to be no difference between small and large company respondents on the importance of quality as a selection and evaluation criterion. Observe in looking at the table that more than half of the respondents for both small and large companies ranked quality as the number 1 criterion.

For cost:

$$\text{Chi-Sq} = 0.483, \text{DF} = 2, P\text{-Value} = 0.785$$

The p-value of .785 indicates failure to reject the null hypothesis. Company size is independent of the ranking of this criterion as a supplier selection and evaluation tool. In perusing the raw data, it is evident that about one-third of the respondents in both large and small companies ranked cost either as the second or third most important criterion.

For current technology:

$$\text{Chi-Sq} = 1.026, \text{DF} = 2, P\text{-Value} = 0.599$$

The p-value of .599 indicates failure to reject the null hypothesis. Company size is independent of the ranking of this criterion as a supplier selection and evaluation tool.

Pearson and Ellram also found that of small companies' responses to the question of what functional areas are involved in the selection and evaluation of suppliers, 41.38% included research and development versus 48.78% for large companies. Because sample statistics are used, is enough evidence provided to declare that there is a significant difference in proportions between small and large companies? Techniques presented in [Chapter 10] can be used to statistically test this hypothesis. However, the chi-square test of independence can also be used. The two variables are company size (small, large) and whether or not the functional area is included (Yes, No) producing a 2 × 2 table:

	Company Size	
	Small	Large
Yes	36	60
No	51	63

Recall that 87 small companies and 123 large companies took part in the study. The raw numbers are obtained by multiplying the samples sizes by the percentages [e.g., sample size for small companies = 87 and 87(.4138) = 36]. Minitab gives a chi-square value of 1.125 with a p-value of 0.289. Based on this result, there is no significant difference between small and large companies on the inclusion of research and development in the supplier selection and evaluation process.

ETHICAL CONSIDERATIONS

The usage of chi-square goodness-of-fit tests and chi-square tests of independence becomes an issue when the expected frequencies are too small. Considerable debate surrounds the discussion of how small is too small. In this chapter, we used an expected frequency of less than 5 as too small. As an example, suppose an expected frequency is 2. If the observed value is 6, then the calculation of $(f_o - f_e)^2/f_e$ results in $(6 - 2)^2/2 = 8$ just for this pair of observed and expected frequencies. Such a contribution to the overall computed chi-square can inordinately affect the total chi-square value and skew the analysis. Researchers should exercise caution in using small expected frequencies with chi-square tests lest they arrive at an incorrect statistical outcome.

SUMMARY

The chapter presented two techniques for analyzing categorical data. Categorical data are *nonnumerical data that are frequency counts of categories from one or more variables*. Categorical data producing this type of data are often analyzed using chi-square techniques. The two techniques presented for analyzing categorical data are the chi-square goodness-of-fit test and the chi-square test of independence. These techniques are an outgrowth of the binomial distribution and the inferential techniques for analyzing population proportions.

The chi-square goodness-of-fit test is used to compare a theoretical or expected distribution of measurements for several categories of a variable with the actual or observed distribution of measurements. It can be used to determine whether a distribution of values fits a given distribution, such as the Poisson or normal distribution. If only two categories are used, the test offers the equivalent of a z test for a single proportion.

The chi-square test of independence is used to analyze frequencies for categories of two variables to determine whether the two variables are independent. The data used in analysis by a chi-square test of independence are arranged in a two-dimensional table called a contingency table. For this reason, the test is sometimes referred to as contingency analysis. A chi-square test of independence is computed in a manner similar to that used with the chi-square goodness-of-fit test. Expected values are computed for each cell of the contingency table and then compared to observed values with the chi-square statistic. Both the chi-square test of independence and the chi-square goodness-of-fit test require that expected values be greater than or equal to 5.

KEY TERMS

Flash Cards

categorical data
chi-square
  distribution

chi-square goodness-of-fit test
chi-square test of
  independence

contingency analysis
contingency table

FORMULAS

χ^2 goodness-of-fit test

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$
$$\mathrm{df} = k - 1 - c$$

χ^2 test of independence

$$\chi^2 = \sum \sum \frac{(f_o - f_e)^2}{f_e}$$
$$\mathrm{df} = (r - 1)(c - 1)$$

SUPPLEMENTARY PROBLEMS

CALCULATING THE STATISTICS

16.18 Use a chi-square goodness-of-fit test to determine whether the following observed frequencies are distributed the same as the expected frequencies. Let $\alpha = .01$.

Category	f_o	f_e
1	214	206
2	235	232
3	279	268
4	281	284
5	264	268
6	254	232
7	211	206

16.19 Use the chi-square contingency analysis to test to determine whether variable 1 is independent of variable 2. Use 5% level of significance.

	Variable 2		
Variable 1	12	23	21
	8	17	20
	7	11	18

TESTING YOUR UNDERSTANDING

16.20 Is a manufacturer's geographic location independent of type of customer? Use the following data for companies with primarily industrial customers and companies with primarily retail customers to test this question. Let $\alpha = .10$.

		Geographic Location		
		Northeast	*West*	*South*
Customer Type	*Industrial Customer*	230	115	68
	Retail Customer	185	143	89

16.21 A national youth organization sells six different kinds of cookies during its annual cookie campaign. A local leader is curious about whether national sales of the six kinds of cookies are uniformly distributed. He randomly selects the amounts of each kind of cookies sold from five youths and combines them into the observed data that follow. Use $\alpha = .05$ to determine whether the data indicate that sales for these six kinds of cookies are uniformly distributed.

Kind of Cookie	Observed Frequency
Chocolate chip	189
Peanut butter	168
Cheese cracker	155
Lemon flavored	161
Chocolate mint	216
Vanilla filled	165

16.22 A researcher interviewed 2067 people and asked whether they were the primary decision makers in the household when buying a new car last year. Two hundred seven were men and had bought a new car last year. Sixty-five were women and had bought a new car last year. Eight hundred eleven of the responses were from men who did not buy a car last year. Nine hundred eighty-four were from women who did not buy a car last year. Use these data to determine whether gender is independent of being a major decision maker in purchasing a car last year. Let $\alpha = .05$.

16.23 Are random arrivals at a shoe store at the local mall Poisson distributed? Suppose a mall employee researches this question by gathering data for arrivals during one-minute intervals on a weekday between 6:30 P.M. and 8:00 P.M. The data obtained follow. Use $\alpha = .05$ to determine whether the observed data seem to be from a Poisson distribution.

Arrivals per Minute	Observed Frequency
0	26
1	40
2	57
3	32
4	17
5	12
6	8

16.24 According to *Beverage Digest*/Maxwell Report, the distribution of market share for the top seven soft drinks in the United States was: Coca-Cola 17.9%, Pepsi-Cola 11.5%, Diet Coke 9.7%, Mountain Dew 6.3%, Diet Pepsi 6.1%, Sprite 5.7%, and Dr. Pepper 5.6%. Suppose a marketing analyst wants to determine whether this distribution fits that of her geographic region. She randomly surveys 1726 people and asks them to name their favorite soft drink. The responses are: Coca-Cola 314, Pepsi 219, Diet Coke 212, Mountain Dew 121, Diet Pepsi 98, Sprite 93, Dr. Pepper 88, and others 581. She then tests to determine whether the local distribution of soft drink preferences is the same or different from the national figures, using $\alpha = .05$. What does she find?

16.25 Are the types of professional jobs held in the computing industry independent of the number of years a person has worked in the industry? Suppose 246 workers are interviewed. Use the results obtained to determine whether type of professional job held in the computer industry is independent of years worked in the industry. Let $\alpha = .01$.

		Professional Position			
		Manager	*Programmer*	*Operator*	*Systems Analyst*
Years	*0–3*	6	37	11	13
	4–8	28	16	23	24
	More than 8	47	10	12	19

16.26 A study by Market Facts/TeleNation for Personnel Decisions International (PDI) found that the average work-week is getting longer for U.S. full-time workers. Forty-three percent of the responding workers in the survey cited "more work, more business" as the number one reason for this increase in workweek. Suppose you want to test this figure in California to determine whether California workers feel the same way. A random sample of 315 California full-time workers whose workweek has been getting longer is chosen. They are offered a selection of possible reasons for this increase and 120 pick "more work, more business." Use techniques presented in this chapter and an alpha of .05 to test to determine whether the 43% U.S. figure for this reason holds true in California.

16.27 Is the number of children that a college student currently has independent of the type of college or university being attended? Suppose students were randomly selected from three types of colleges and universities and the data shown represent the results of a survey of those students. Use a chi-square test of independence of answer the question. Let $\alpha = .05$.

		Type of College or University		
		Community College	*Large University*	*Small College*
Number of Children	*0*	25	178	31
	1	49	141	12
	2	31	54	8
	3 or more	22	14	6

INTERPRETING THE OUTPUT

16.28 A survey by Ipsos-Reid reported in *American Demographics* showed that if a person was given a $1,000 wind-fall, 36% would spend the money on home improvement, 24% on leisure travel/vacation, 15% on clothing, 15% on home entertainment or electronic products, and 10% on local entertainment including restaurants and movies. Suppose a researcher believes that these results would not be the same if posed to adults between 21 and 30 years of age. The researcher conducts a new survey interviewing 200 adults between 21 and 30 years of age asking these same questions. A chi-square goodness-of-fit test is conducted to compare the results of the new survey to the one taken by Ipsos-Reid. The Excel results follow. The observed and expected values are for the categories as already listed and appear in the same order. Discuss the findings. How did the distribution of results from the new survey compare to the old? Discuss the business implications of this outcome.

21–30 Years of Age	General Population
Observed	Expected
36	72
64	48
42	30
38	30
20	20

The *p*-value for the chi-square goodness-of-fit test is: **0.0000043**. The observed chi-square for the goodness-of-fit test is: **30.18**.

16.29 Do men and women prefer the same colors of cars? That is, is sex independent of color preference for cars? Suppose a study is undertaken to address this question. A random sample of men and women are asked which of five colors (silver, white, black, green, blue) they prefer in a car. The results as analyzed using Minitab are shown here. Discuss the test used, the hypotheses, the findings, and the business implications.

Chi-Square Test: Men, Women

```
Expected counts are printed below observed counts.
Chi-square contributions are printed below expected
counts.
        Men      Women     Total
1        90         52       142
       85.20      56.80
        0.270      0.406

2        75         58       133
       79.80      53.20
        0.289      0.433

3        63         30        93
       55.80      37.20
        0.929      1.394

4        39         33        72
       43.20      28.80
        0.408      0.612

5        33         27        60
       36.00      24.00
        0.250      0.375

Total   300        200       500
Chi-Sq = 5.366, DF = 4, P-Value = 0.252
```

ANALYZING THE DATABASE

1. The Financial database contains seven different types of companies. These seven are denoted by the variable Type. Use a chi-square goodness-of-fit test to determine whether the seven types of companies are uniformly distributed in this database.

2. In the Manufacturing database, is the Value of Industrial Shipments (a four-category variable) uniformly distributed across the database?

3. Use a chi-square test of independence to determine whether Control is independent of Service in the Hospital database. Comment on the results of this test.

4. In the Consumer database, is Location independent of Region? Use a chi-square test of independence to answer the question.

CASE

FOOT LOCKER IN THE SHOE MIX

Foot Locker, Inc., is the world's number one retailer of athletic footwear and apparel. The company has approximately 3,800 retail stores in 21 countries across North America, Europe, and Australia under various brand names. Foot Locker estimates that it controls about 18% of the U.S. $15 billion athletic footwear market. The company intends to increase its share of the worldwide market by adding additional stores and by growing its Internet and catalog business.

In recent years, Foot Locker officials have been rethinking the company's retail mix. Determining the shoe mix that will maximize profits is an important decision for Foot Locker. By the year 2002, in an effort to stock more lower-priced footwear,

the company had reduced its inventory of sneakers priced at $120 or more by 50%.

Discussion

Suppose the data presented below represented the number of unit sales (million $) for athletic footwear in the years 2000 and 2009. Use techniques presented in this chapter to analyze these data and discuss the business implications for Foot Locker.

Price Category	2000	2009
Less than $30	115	126
$30–less than $40	38	40
$40–less than $50	37	35
$50–less than $60	30	27
$60–less than $70	22	20
$70–less than $85	21	20
$85–less than $100	11	11
$100 or more	17	18

Suppose Foot Locker strongly encourages its employees to make formal suggestions to improve the store, the product, and the working environment. Suppose a quality auditor keeps records of the suggestions, the persons who submitted them, and the geographic region from which they come. A possible breakdown of the number of suggestions over a 3-year period by employee sex and geographic location follows. Is there any relationship between the sex of the employee and the geographic location in terms of number of suggestions? If they are related, what does this relationship mean to the company? What business implications might there be for such an analysis?

		Sex	
		Male	Female
	U.S. West	29	43
	U.S. South	48	20
Location	U.S. East	52	61
	U.S. North	28	25
	Europe	78	32
	Australia	47	29

According to Wells Fargo Securities and the NPDFashionworld Consumer Panel, Foot Locker holds 19.4% of the sneaker market. Suppose that due to its relationship with Nike and Nike's presence in the U.S. West region, Foot Locker believes that its share of the market in that region is higher than it is in the rest of the country. Foot Locker hires a market research company to test this notion. The research company randomly samples 1000 people in the U.S. West who have just purchased a pair of sneakers and 230 of the sampled shoppers purchased their sneakers at Foot Locker. Is this result enough evidence to declare that Foot Locker's share of the market in the U.S. West is significantly higher than it is otherwise? Use techniques from this chapter to test this hypothesis. What business implications might there be for Foot Locker if this market share information is true?

Source: Adapted from Christopher Lawton and Maureen Tkacik. "Foot Locker Changes Mix of Sneakers," *The Wall Street Journal* (July 22, 2002), p. B3; Foot Locker, Inc., available at: http://www.footlocker-inc.com; "Venator Group, Inc. Announces Name Change to Foot Locker, Inc.," *PR NEWSWIRE*, (November 1, 2001), p. 1. http://www.footlocker-inc.com/ (2008)

USING THE COMPUTER

EXCEL

■ Excel can compute a chi-square goodness-of-fit test but not a chi-square test of independence.

■ To compute a chi-square goodness-of-fit test using Excel, begin with the **Insert Function** (f_x). To access the **Insert Function**, go to the **Formulas** tab on an Excel worksheet (top center tab). The **Insert Function** is on the far left of the menu bar. In the **Insert Function** dialog box at the top, there is a pulldown menu where it says **Or select a category**. From the pulldown menu associated with this command, select **Statistical**. Select **CHITEST** from the **Insert Function**'s **Statistical** menu. In the **CHITEST** dialog box, place the location of the observed values in **Actual_range**. Place the location of the expected values in **Expected_range**. The output will consist of a *p*-value. To determine the observed chi-square from this *p*-value, go back to the **Insert Function** (f_x) and select **Statistical** and then **CHIINV**. In the **CHIINV** dialog box, place the *p*-value in **Probability** and the degrees of freedom in **Deg_freedom**. The output is the chi-square value.

MINITAB

■ Minitab can compute the chi-square test of independence but does not directly compute the chi-square goodness-of-fit test. To compute the chi-square test of independence, from the **Stat** pulldown menu, select **Tables**. From the **Tables** pulldown menu, select **Chi-Square Test (table in work-sheet)**. Enter the columns containing the tabular data in the space beside **Columns containing the data**.