CHAPTER 12

ANALYSIS OF FREQUENCY DATA: AN INTRODUCTION TO THE CHI-SQUARE DISTRIBUTION

CHAPTER OVERVIEW

This chapter explores techniques that are commonly used in the analysis of count or frequency data. Uses of the chi-square distribution, which was mentioned briefly in Chapter 6, are discussed and illustrated in greater detail. Additionally, statistical techniques often used in epidemiological studies are introduced and demonstrated by means of examples.

TOPICS

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand the mathematical properties of the chi-square distribution.
2. be able to use the chi-square distribution for goodness-of-fit tests.
3. be able to construct and use contingency tables to test independence and homogeneity.

4. be able to apply Fisher's exact test for 2 × 2 tables.
5. understand how to calculate and interpret the epidemiological concepts of relative risk, odds ratios, and the Mantel-Haenszel statistic.

12.1 INTRODUCTION

In the chapters on estimation and hypothesis testing, brief mention is made of the chi-square distribution in the construction of confidence intervals for, and the testing of, hypotheses concerning a population variance. This distribution, which is one of the most widely used distributions in statistical applications, has many other uses. Some of the more common ones are presented in this chapter along with a more complete description of the distribution itself, which follows in the next section.

The chi-square distribution is the most frequently employed statistical technique for the analysis of count or frequency data. For example, we may know for a sample of hospitalized patients how many are male and how many are female. For the same sample we may also know how many have private insurance coverage, how many have Medicare insurance, and how many are on Medicaid assistance. We may wish to know, for the population from which the sample was drawn, if the type of insurance coverage differs according to gender. For another sample of patients, we may have frequencies for each diagnostic category represented and for each geographic area represented. We might want to know if, in the population from which the same was drawn, there is a relationship between area of residence and diagnosis. We will learn how to use chi-square analysis to answer these types of questions.

There are other statistical techniques that may be used to analyze frequency data in an effort to answer other types of questions. In this chapter we will also learn about these techniques.

12.2 THE MATHEMATICAL PROPERTIES OF THE CHI-SQUARE DISTRIBUTION

The chi-square distribution may be derived from normal distributions. Suppose that from a normally distributed random variable Y with mean μ and variance σ^2 we randomly and independently select samples of size $n = 1$. Each value selected may be transformed to the standard normal variable z by the familiar formula

$$z_i = \frac{y_i - \mu}{\sigma} \qquad (12.2.1)$$

Each value of z may be squared to obtain z^2. When we investigate the sampling distribution of z^2, we find that it follows a chi-square distribution with 1 degree of freedom. That is,

$$\chi^2_{(1)} = \left(\frac{y - \mu}{\sigma}\right)^2 = z^2$$

Now suppose that we randomly and independently select samples of size $n = 2$ from the normally distributed population of Y values. Within each sample we may transform each value of y to the standard normal variable z and square as before. If the resulting values of z^2 for each sample are added, we may designate this sum by

$$\chi^2_{(2)} = \left(\frac{y_1 - \mu}{\sigma}\right)^2 + \left(\frac{y_2 - \mu}{\sigma}\right)^2 = z_1^2 + z_2^2$$

since it follows the chi-square distribution with 2 degrees of freedom, the number of independent squared terms that are added together.

The procedure may be repeated for any sample size n. The sum of the resulting z^2 values in each case will be distributed as chi-square with n degrees of freedom. In general, then,

$$\chi^2_{(n)} = z_1^2 + z_2^2 + \cdots + z_n^2 \tag{12.2.2}$$

follows the chi-square distribution with n degrees of freedom. The mathematical form of the chi-square distribution is as follows:

$$f(u) = \frac{1}{\left(\frac{k}{2} - 1\right)!} \frac{1}{2^{k/2}} u^{(k/2)-1} e^{-(u/2)}, \quad u > 0 \tag{12.2.3}$$

where e is the irrational number 2.71828 $\cdots$ and k is the number of degrees of freedom. The variate u is usually designated by the Greek letter chi (χ) and, hence, the distribution is called the chi-square distribution. As we pointed out in Chapter 6, the chi-square distribution has been tabulated in Appendix Table F. Further use of the table is demonstrated as the need arises in succeeding sections.

The mean and variance of the chi-square distribution are k and $2k$, respectively. The modal value of the distribution is $k - 2$ for values of k greater than or equal to 2 and is zero for $k = 1$.

The shapes of the chi-square distributions for several values of k are shown in Figure 6.9.1. We observe in this figure that the shapes for $k = 1$ and $k = 2$ are quite different from the general shape of the distribution for $k > 2$. We also see from this figure that chi-square assumes values between 0 and infinity. It cannot take on negative values, since it is the sum of values that have been squared. A final characteristic of the chi-square distribution worth noting is that the sum of two or more independent chi-square variables also follows a chi-square distribution.

Types of Chi-Square Tests As already noted, we make use of the chi-square distribution in this chapter in testing hypotheses where the data available for analysis are in the form of frequencies. These hypothesis testing procedures are discussed under the topics of *tests of goodness-of-fit*, *tests of independence*, and *tests of homogeneity*. We will discover that, in a sense, all of the chi-square tests that we employ may be thought of as goodness-of-fit tests, in that they test the goodness-of-fit of observed frequencies to frequencies that one would expect if the data were generated under some particular theory or hypothesis. We, however, reserve the phrase "goodness-of-fit" for use in a more restricted sense. We use it to refer to a comparison of a sample distribution to some theoretical distribution that it is assumed describes the population from which the sample came. The justification of our use of the distribution in these situations is due to Karl Pearson (1), who showed that the chi-square distribution may be used as a test of the agreement between observation and hypothesis whenever the data are in the form of frequencies. An extensive treatment of the chi-square distribution is to be found in the book by Lancaster (2). Nikulin and Greenwood (3) offer practical advice for conducting chi-square tests.

Observed Versus Expected Frequencies The chi-square statistic is most appropriate for use with categorical variables, such as marital status, whose values are the categories married, single, widowed, and divorced. The quantitative data used in the computation of the test statistic are the frequencies associated with each category of the one or more variables under study. There are two sets of frequencies with which we are concerned, *observed frequencies* and *expected frequencies.* The observed frequencies are the number of subjects or objects in our sample that fall into the various categories of the variable of interest. For example, if we have a sample of 100 hospital patients, we may observe that 50 are married, 30 are single, 15 are widowed, and 5 are divorced. Expected frequencies are the number of subjects or objects in our sample that we would expect to observe if some null hypothesis about the variable is true. For example, our null hypothesis might be that the four categories of marital status are equally represented in the population from which we drew our sample. In that case we would expect our sample to contain 25 married, 25 single, 25 widowed, and 25 divorced patients.

The Chi-Square Test Statistic The test statistic for the chi-square tests we discuss in this chapter is

$$X^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] \tag{12.2.4}$$

When the null hypothesis is true, X^2 is distributed approximately as χ^2 with $k - r$ degrees of freedom. In determining the degrees of freedom, k is equal to the number of groups for which observed and expected frequencies are available, and r is the number of restrictions or constraints imposed on the given comparison. A restriction is imposed when we force the sum of the expected frequencies to equal the sum of the observed frequencies, and an additional restriction is imposed for each parameter that is estimated from the sample.

In Equation 12.2.4, O_i is the observed frequency for the ith category of the variable of interest, and E_i is the expected frequency (given that H_0 is true) for the ith category.

The quantity X^2 is a measure of the extent to which, in a given situation, pairs of observed and expected frequencies agree. As we will see, the nature of X^2 is such that when there is close agreement between observed and expected frequencies it is small, and when the agreement is poor it is large. Consequently, only a sufficiently large value of X^2 will cause rejection of the null hypothesis.

If there is perfect agreement between the observed frequencies and the frequencies that one would expect, given that H_0 is true, the term $O_i - E_i$ in Equation 12.2.4 will be equal to zero for each pair of observed and expected frequencies. Such a result would yield a value of X^2 equal to zero, and we would be unable to reject H_0.

When there is disagreement between observed frequencies and the frequencies one would expect given that H_0 is true, at least one of the $O_i - E_i$ terms in Equation 12.2.4 will be a nonzero number. In general, the poorer the agreement between the O_i and the E_i, the greater or the more frequent will be these nonzero values. As noted previously, if the agreement between the O_i and the E_i is sufficiently poor (resulting in a sufficiently large X^2 value,) we will be able to reject H_0.

When there is disagreement between a pair of observed and expected frequencies, the difference may be either positive or negative, depending on which of the two frequencies is the larger. Since the measure of agreement, X^2, is a sum of component quantities whose magnitudes depend on the difference $O_i - E_i$, positive and negative differences must be given equal weight. This is achieved by squaring each $O_i - E_i$ difference. Dividing the squared differences by the appropriate expected frequency converts the quantity to a term that is measured in original units. Adding these individual $(O_i - E_i)^2/E_i$ terms yields X^2, a summary statistic that reflects the extent of the overall agreement between observed and expected frequencies.

The Decision Rule The quantity $\sum[(O_i - E_i)^2/E_i]$ will be small if the observed and expected frequencies are close together and will be large if the differences are large.

The computed value of X^2 is compared with the tabulated value of χ^2 with $k - r$ degrees of freedom. The decision rule, then, is: Reject H_0 if X^2 is greater than or equal to the tabulated χ^2 for the chosen value of α.

Small Expected Frequencies Frequently in applications of the chi-square test the expected frequency for one or more categories will be small, perhaps much less than 1. In the literature the point is frequently made that the approximation of X^2 to χ^2 is not strictly valid when some of the expected frequencies are small. There is disagreement among writers, however, over what size expected frequencies are allowable before making some adjustment or abandoning χ^2 in favor of some alternative test. Some writers, especially the earlier ones, suggest lower limits of 10, whereas others suggest that all expected frequencies should be no less than 5. Cochran (4, 5), suggests that for goodness-of-fit tests of unimodal distributions (such as the normal), the minimum expected frequency can be as low as 1. If, in practice, one encounters one or more expected frequencies less than 1, adjacent categories may be combined to achieve the suggested minimum. Combining reduces the number of categories and, therefore, the number of degrees of freedom. Cochran's suggestions appear to have been followed extensively by practitioners in recent years.

12.3 TESTS OF GOODNESS-OF-FIT

As we have pointed out, a goodness-of-fit test is appropriate when one wishes to decide if an observed distribution of frequencies is incompatible with some preconceived or hypothesized distribution.

We may, for example, wish to determine whether or not a sample of observed values of some random variable is compatible with the hypothesis that it was drawn from a population of values that is normally distributed. The procedure for reaching a decision consists of placing the values into mutually exclusive categories or class intervals and noting the frequency of occurrence of values in each category. We then make use of our knowledge of normal distributions to determine the frequencies for each category that one could expect if the sample had come from a normal distribution. If the discrepancy is of such magnitude that it could have come about due to chance, we conclude that the sample may have come from a normal distribution. In a similar manner, tests of goodness-of-fit may be carried out in cases where the hypothesized distribution is the binomial, the Poisson, or any other distribution. Let us illustrate in more detail with some examples of tests of hypotheses of goodness-of-fit.

EXAMPLE 12.3.1 *The Normal Distribution*

Cranor and Christensen (A-1) conducted a study to assess short-term clinical, economic, and humanistic outcomes of pharmaceutical care services for patients with diabetes in community pharmacies. For 47 of the subjects in the study, cholesterol levels are summarized in Table 12.3.1.

We wish to know whether these data provide sufficient evidence to indicate that the sample did not come from a normally distributed population. Let $\alpha = .05$

Solution:

1. **Data.** See Table 12.3.1.
2. **Assumptions.** We assume that the sample available for analysis is a simple random sample.

TABLE 12.3.1 Cholesterol Levels as Described in Example 12.3.1

Cholesterol Level (mg/dl)	Number of Subjects
100.0–124.9	1
125.0–149.9	3
150.0–174.9	8
175.0–199.9	18
200.0–224.9	6
225.0–249.9	4
250.0–274.9	4
275.0–299.9	3

Source: Carole W. Cranor, and Dale B. Christensen, "The Asheville Project: Short-Term Outcomes of a Community Pharmacy Diabetes Care Program," *Journal of the American Pharmaceutical Association*, *43* (2003), 149–159. Used with permission.

3. Hypotheses.

H_0: In the population from which the sample was drawn, cholesterol levels are normally distributed.

H_A: The sampled population is not normally distributed.

4. Test statistic. The test statistic is

$$X^2 = \sum_{i=1}^{k}\left[\frac{(O_i - E_i)^2}{E_i}\right]$$

5. Distribution of test statistic. If H_0 is true, the test statistic is distributed approximately as chi-square with $k - r$ degrees of freedom. The values of k and r will be determined later.

6. Decision rule. We will reject H_0 if the computed value of X^2 is equal to or greater than the critical value of chi-square.

7. Calculation of test statistic. Since the mean and variance of the hypothesized distribution are not specified, the sample data must be used to estimate them. These parameters, or their estimates, will be needed to compute the frequency that would be expected in each class interval when the null hypothesis is true. The mean and standard deviation computed from the grouped data of Table 12.3.1 are

$$\bar{x} = 198.67$$
$$s = 41.31$$

As the next step in the analysis, we must obtain for each class interval the frequency of occurrence of values that we would expect when the null hypothesis is true, that is, if the sample were, in fact, drawn from a normally distributed population of values. To do this, we first determine the expected relative frequency of occurrence of values for each class interval and then multiply these expected relative frequencies by the total number of values to obtain the expected number of values for each interval.

The Expected Relative Frequencies

It will be recalled from our study of the normal distribution that the relative frequency of occurrence of values equal to or less than some specified value, say, x_0, of the normally distributed random variable X is equivalent to the area under the curve and to the left of x_0 as represented by the shaded area in Figure 12.3.1. We obtain the numerical value of this area by converting x_0 to a standard normal deviation by the formula $z_0 = (x_0 - \mu)/\sigma$ and finding the appropriate value in Appendix Table D. We use this procedure to obtain the expected relative frequencies corresponding to each of the class intervals in Table 12.3.1. We estimate μ and σ with $\bar{x}$ and s as computed from the grouped sample data. The first step consists of obtaining z values corresponding to the lower limit of each class interval. The area between two successive z values will give the expected relative frequency of occurrence of values for the corresponding class interval.

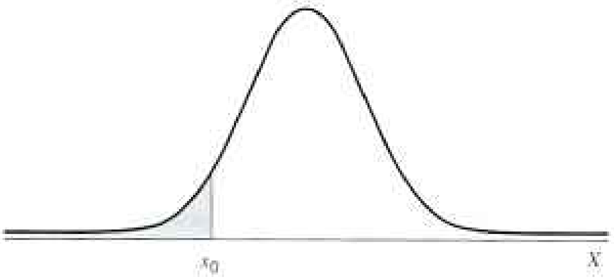


FIGURE 12.3.1 A normal distribution showing the relative frequency of occurrence of values less than or equal to x_0. The shaded area represents the relative frequency of occurrence of values equal to or less than x_0.

For example, to obtain the expected relative frequency of occurrence of values in the interval 100.0 to 124.9 we proceed as follows:

The z value corresponding to $X = 100.0$ is $z = \dfrac{100.0 - 198.67}{41.31} = -2.39$

The z value corresponding to $X = 125.0$ is $z = \dfrac{125.0 - 198.67}{41.31} = -1.78$

In Appendix Table D we find that the area to the left of -2.39 is .0084, and the area to the left of -1.78 is .0375. The area between -1.78 and -2.39 is equal to $.0375 - .0084 = .0291$, which is equal to the expected relative frequency of occurrence of cholesterol levels within the interval 100.0 to 124.9. This tells us that if the null hypothesis is true, that is, if the cholesterol levels are normally distributed, we should expect 2.91 percent of the values in our sample to be between 100.0 and 124.9. When we multiply our total sample size, 47, by .0291 we find the expected frequency for the interval to be 1.4. Similar calculations will give the expected frequencies for the other intervals as shown in Table 12.3.2.

TABLE 12.3.2 Class Intervals and Expected Frequencies for Example 12.3.1

Class Interval	$z = (x_i - \bar{x})/s$ At Lower Limit of Interval	Expected Relative Frequency	Expected Frequency	
< 100		.0084	.4	1.8
100.0–124.9	−2.39	.0291	1.4	
125.0–149.9	−1.78	.0815	3.8	
150.0–174.9	−1.18	.1653	7.8	
175.0–199.9	−.57	.2277	10.7	
200.0–224.9	.03	.2269	10.7	
225.0–249.9	.64	.1536	7.2	
250.0–274.9	1.24	.0753	3.5	
275.0–299.9	1.85	.0251	1.2	1.5
300.0 and greater	2.45	.0071	.3	

Comparing Observed and Expected Frequencies

We are now interested in examining the magnitudes of the discrepancies between the observed frequencies and the expected frequencies, since we note that the two sets of frequencies do not agree. We know that even if our sample were drawn from a normal distribution of values, sampling variability alone would make it highly unlikely that the observed and expected frequencies would agree perfectly. We wonder, then, if the discrepancies between the observed and expected frequencies are small enough that we feel it reasonable that they could have occurred by chance alone, when the null hypothesis is true. If they are of this magnitude, we will be unwilling to reject the null hypothesis that the sample came from a normally distributed population.

If the discrepancies are so large that it does not seem reasonable that they could have occurred by chance alone when the null hypothesis is true, we will want to reject the null hypothesis. The criterion against which we judge whether the discrepancies are "large" or "small" is provided by the chi-square distribution.

The observed and expected frequencies along with each value of $(O_i - E_i)^2/E_i$ are shown in Table 12.3.3. The first entry in the last column, for example, is computed from $(1 - 1.8)^2/1.8 = .356$. The other values of $(O_i - E_i)^2/E_i$ are computed in a similar manner.

From Table 12.3.3 we see that $X^2 = \sum[(O_i - E_i)^2/E_i] = 10.566$. The appropriate degrees of freedom are 8 (the number of groups or class intervals) − 3 (for the three restrictions: making $\sum E_i = \sum O_i$, and estimating μ and σ from the sample data) = 5.

8. **Statistical decision.** When we compare $X^2 = 10.566$ with values of χ^2 in Appendix Table F, we see that it is less than $\chi^2_{.95} = 11.070$, so that, at the .05 level of significance, we cannot reject the null hypothesis that the sample came from a normally distributed population.

TABLE 12.3.3 Observed and Expected Frequencies and $(O_i - E_i)^2/E_i$ for Example 12.3.1

Class Interval	Observed Frequency (O_i)	Expected Frequency (E_i)	$(O_i - E_i)^2/E_i$
< 100	0	.4 } 1.8	.356
100.0–124.9	1	1.4 }	
125.0–149.9	3	3.8	.168
150.0–174.9	8	7.8	.005
175.0–199.9	18	10.7	4.980
200.0–224.9	6	10.7	2.064
225.0–249.9	4	7.2	1.422
250.0–274.9	4	3.5	.071
275.0–299.9	3	1.2 } 1.5	1.500
300.0 and greater	0	.3 }	
Total	47	47	10.566

9. **Conclusion.** We conclude that in the sampled population, cholesterol levels may follow a normal distribution.

10. **p value.** Since $11.070 > 10.566 > 9.236$, $.05 < p < .10$. In other words, the probability of obtaining a value of X^2 as large as 10.566, when the null hypothesis is true, is between .05 and .10. Thus we conclude that such an event is not sufficiently rare to reject the null hypothesis that the data come from a normal distribution. ∎

Sometimes the parameters are specified in the null hypothesis. It should be noted that had the mean and variance of the population been specified as part of the null hypothesis in Example 12.3.1, we would not have had to estimate them from the sample and our degrees of freedom would have been $8 - 1 = 7$.

Alternatives Although one frequently encounters in the literature the use of chi-square to test for normality, it is not the most appropriate test to use when the hypothesized distribution is continuous. The Kolmogorov–Smirnov test, described in Chapter 13, was especially designed for goodness-of-fit tests involving continuous distributions.

EXAMPLE 12.3.2 *The Binomial Distribution*

In a study designed to determine patient acceptance of a new pain reliever, 100 physicians each selected a sample of 25 patients to participate in the study. Each patient, after trying the new pain reliever for a specified period of time, was asked whether it was preferable to the pain reliever used regularly in the past.

The results of the study are shown in Table 12.3.4.

TABLE 12.3.4 Results of Study Described in Example 12.3.2

Number of Patients Out of 25 Preferring New Pain Reliever	Number of Doctors Reporting this Number	Total Number of Patients Preferring New Pain Reliever by Doctor
0	5	0
1	6	6
2	8	16
3	10	30
4	10	40
5	15	75
6	17	102
7	10	70
8	10	80
9	9	81
10 or more	0	0
Total	100	500

We are interested in determining whether or not these data are compatible with the hypothesis that they were drawn from a population that follows a binomial distribution. Again, we employ a chi-square goodness-of-fit test.

Solution: Since the binomial parameter, p, is not specified, it must be estimated from the sample data. A total of 500 patients out of the 2500 patients participating in the study said they preferred the new pain reliever, so that our point estimate of p is $\hat{p} = 500/2500 = .20$. The expected relative frequencies can be obtained by evaluating the binomial function

$$f(x) = \binom{25}{x} .2^x .8^{25-x}$$

for $x = 0, 1, \ldots, 25$. For example, to find the probability that out of a sample of 25 patients none would prefer the new pain reliever, when in the total population the true proportion preferring the new pain reliever is .2, we would evaluate

$$f(0) = \binom{25}{0} .2^0 .8^{25-0}$$

This can be done most easily by consulting Appendix Table B, where we see that $P(X = 0) = .0038$. The relative frequency of occurrence of samples of size 25 in which no patients prefer the new pain reliever is .0038. To obtain the corresponding expected frequency, we multiply .0038 by 100 to get .38. Similar calculations yield the remaining expected frequencies, which, along with the observed frequencies, are shown in Table 12.3.5. We

TABLE 12.3.5 Calculations for Example 12.3.2

Number of Patients Out of 25 Preferring New Pain Reliever	Number of Doctors Reporting This Number (Observed Frequency, O_i)	Expected Relative Frequency	Expected Frequency E_i
0	5 } 11	.0038	.38 } 2.74
1	6	.0236	2.36
2	8	.0708	7.08
3	10	.1358	13.58
4	10	.1867	18.67
5	15	.1960	19.60
6	17	.1633	16.33
7	10	.1109	11.09
8	10	.0623	6.23
9	9	.0295	2.95
10 or more	0	.0173	1.73
Total	100	1.0000	100.00

see in this table that the first expected frequency is less than 1, so that we follow Cochran's suggestion and combine this group with the second group. When we do this, all the expected frequencies are greater than 1.

From the data, we compute

$$X^2 = \frac{(11 - 2.74)^2}{2.74} + \frac{(8 - 7.08)^2}{7.08} + \cdots + \frac{(0 - 1.73)^2}{1.73} = 47.624$$

The appropriate degrees of freedom are 10 (the number of groups left after combining the first two) less 2, or 8. One degree of freedom is lost because we force the total of the expected frequencies to equal the total observed frequencies, and one degree of freedom is sacrificed because we estimated p from the sample data.

We compare our computed X^2 with the tabulated χ^2 with 8 degrees of freedom and find that it is significant at the .005 level of significance; that is, $p < .005$. We reject the null hypothesis that the data came from a binomial distribution. ∎

EXAMPLE 12.3.3 *The Poisson Distribution*

A hospital administrator wishes to test the null hypothesis that emergency admissions follow a Poisson distribution with $\lambda = 3$. Suppose that over a period of 90 days the numbers of emergency admissions were as shown in Table 12.3.6.

TABLE 12.3.6 Number of Emergency Admissions to a Hospital During a 90-Day Period

Day	Emergency Admissions	Day	Emergency Admissions	Day	Emergency Admissions	Day	Emergency Admissions
1	2	24	5	47	4	70	3
2	3	25	3	48	2	71	5
3	4	26	2	49	2	72	4
4	5	27	4	50	3	73	1
5	3	28	4	51	4	74	1
6	2	29	3	52	2	75	6
7	3	30	5	53	3	76	3
8	0	31	1	54	1	77	3
9	1	32	3	55	2	78	5
10	0	33	2	56	3	79	2
11	1	34	4	57	2	80	1
12	0	35	2	58	5	81	7
13	6	36	5	59	2	82	7
14	4	37	0	60	7	83	1
15	4	38	6	61	8	84	5
16	4	39	4	62	3	85	1

(*Continued*)

Day	Emergency Admissions	Day	Emergency Admissions	Day	Emergency Admissions	Day	Emergency Admissions
17	3	40	4	63	1	86	4
18	4	41	5	64	3	87	4
19	3	42	1	65	1	88	9
20	3	43	3	66	0	89	2
21	3	44	1	67	3	90	3
22	4	45	2	68	2		
23	3	46	3	69	1		

The data of Table 12.3.6 are summarized in Table 12.3.7.

Solution: To obtain the expected frequencies we first obtain the expected relative frequencies by evaluating the Poisson function given by Equation 4.4.1 for each entry in the left-hand column of Table 12.3.7. For example, the first expected relative frequency is obtained by evaluating

$$f(0) = \frac{e^{-3}3^0}{0!}$$

We may use Appendix Table C to find this and all the other expected relative frequencies that we need. Each of the expected relative frequencies

TABLE 12.3.7 Summary of Data Presented in Table 12.3.6

Number of Emergency Admissions in a Day	Number of Days This Number of Emergency Admissions Occurred
0	5
1	14
2	15
3	23
4	16
5	9
6	3
7	3
8	1
9	1
10 or more	0
Total	90

TABLE 12.3.8 Observed and Expected Frequencies and Components of X^2 for Example 12.3.3

Number of Emergency Admissions	Number of Days this Number Occurred, O_i	Expected Relative Frequency	Expected Frequency	$\frac{(O_i - E_i)^2}{E_i}$
0	5	.050	4.50	.056
1	14	.149	13.41	.026
2	15	.224	20.16	1.321
3	23	.224	20.16	.400
4	16	.168	15.12	.051
5	9	.101	9.09	.001
6	3	.050	4.50	.500
7	3	.022	1.98	.525
8	1 }	.008	.72 }	
9	1 } 2	.003	.27 } 1.08	.784
10 or more	0 }	.001	.09 }	
Total	90	1.000	90.00	3.664

is multiplied by 90 to obtain the corresponding expected frequencies. These values along with the observed and expected frequencies and the components of X^2, $(O_i - E_i)^2/E_i$, are displayed in Table 12.3.8, in which we see that

$$X^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] = \frac{(5 - 4.50)^2}{4.50} + \cdots + \frac{(2 - 1.08)^2}{1.08} = 3.664$$

We also note that the last three expected frequencies are less than 1, so that they must be combined to avoid having any expected frequencies less than 1. This means that we have only nine effective categories for computing degrees of freedom. Since the parameter, λ, was specified in the null hypothesis, we do not lose a degree of freedom for reasons of estimation, so that the appropriate degrees of freedom are $9 - 1 = 8$. By consulting Appendix Table F, we find that the critical value of χ^2 for 8 degrees of freedom and $\alpha = .05$ is 15.507, so that we cannot reject the null hypothesis at the .05 level, or for that matter any reasonable level, of significance ($p > .10$). We conclude, therefore, that emergency admissions at this hospital may follow a Poisson distribution with $\lambda = 3$. At least the observed data do not cast any doubt on that hypothesis.

If the parameter λ has to be estimated from sample data, the estimate is obtained by multiplying each value x by its frequency, summing these products, and dividing the total by the sum of the frequencies. ∎

EXAMPLE 12.3.4 *The Uniform Distribution*

The flu season in southern Nevada for 2005–2006 ran from December to April, the coldest months of the year. The Southern Nevada Health District reported the numbers of vaccine-preventable influenza cases shown in Table 12.3.9. We are interested in knowing whether the numbers of flu cases in the district are equally distributed among the five flu season months. That is, we wish to know if flu cases follow a uniform distribution.

Solution:

1. **Data.** See Table 12.3.9.
2. **Assumptions.** We assume that the reported cases of flu constitute a simple random sample of cases of flu that occurred in the district.
3. **Hypotheses.**

   H_0: Flu cases in southern Nevada are uniformly distributed over the five flu season months.

   H_A: Flu cases in southern Nevada are not uniformly distributed over the five flu season months.

   Let $\alpha = .01$.
4. **Test statistic.** The test statistic is

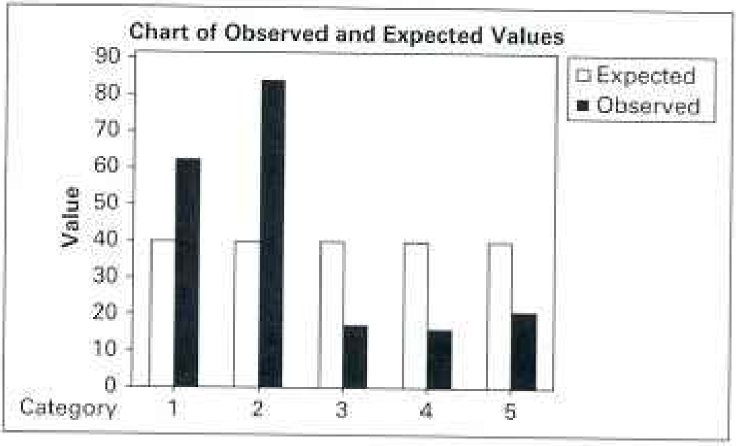
$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

5. **Distribution of test statistic.** If H_0 is true, X^2 is distributed approximately as χ^2 with $(5 - 1) = 4$ degrees of freedom.
6. **Decision rule.** Reject H_0 if the computed value of X^2 is equal to or greater than 13.277.

TABLE 12.3.9 Reported Vaccine-Preventable Influenza Cases from Southern Nevada, December 2005–April 2006

Month	Number of Reported Cases of Influenza
December 2005	62
January 2006	84
February 2006	17
March 2006	16
April 2006	21
Total	200

Source: http://www.southernnevadahealthdistrict.org/epidemiology/disease_statistics.htm.



Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: C1

Category	Observed	Test Proportion	Expected	Contribution to Chi-Sq
1	62	0.2	40	12.100
2	84	0.2	40	48.400
3	17	0.2	40	13.225
4	16	0.2	40	14.400
5	21	0.2	40	9.025

N	DF	Chi-Sq	P-Value
200	4	97.15	0.000

FIGURE 12.3.2 MINITAB output for Example 12.3.4.

7. **Calculation of test statistic.** If the null hypothesis is true, we would expect to observe $200/5 = 40$ cases per month. Figure 12.3.2 shows the computer printout obtained from MINITAB. The bar graph shows the observed and expected frequencies per month. The chi-square table provides the observed frequencies, the expected frequencies based on a uniform distribution, and the individual chi-square contribution for each test value.

8. **Statistical decision.** Since 97.15, the computed value of X^2, is greater than 13.277, we reject, based on these data, the null hypothesis of a

uniform distribution of flu cases during the flu season in southern Nevada.

9. **Conclusion.** We conclude that the occurrence of flu cases does not follow a uniform distribution.

10. **p value.** From the MINITAB output we see that $p = .000$ (i.e., $< .001$). ■

EXAMPLE 12.3.5

A certain human trait is thought to be inherited according to the ratio 1:2:1 for homozygous dominant, heterozygous, and homozygous recessive. An examination of a simple random sample of 200 individuals yielded the following distribution of the trait: dominant, 43; heterozygous, 125; and recessive, 32. We wish to know if these data provide sufficient evidence to cast doubt on the belief about the distribution of the trait.

Solution:

1. **Data.** See statement of the example.

2. **Assumptions.** *We assume that the data meet the requirements for the application of the chi-square goodness-of-fit test.*

3. **Hypotheses.**

   H_0: The trait is distributed according to the ratio 1:2:1 for homozygous dominant, heterozygous, and homozygous recessive.
   H_A: The trait is not distributed according to the ratio 1:2:1.

4. **Test statistic.** The test statistic is

$$X^2 = \sum\left[\frac{(O - E)^2}{E}\right]$$

5. **Distribution of test statistic.** If H_0 is true, X^2 is distributed as chi-square with 2 degrees of freedom.

6. **Decision rule.** Suppose we let the probability of committing a type I error be .05. Reject H_0 if the computed value of X^2 is equal to or greater than 5.991.

7. **Calculation of test statistic.** If H_0 is true, the expected frequencies for the three manifestations of the trait are 50, 100, and 50 for dominant, heterozygous, and recessive, respectively. Consequently,

$$X^2 = (43 - 50)^2/50 + (125 - 100)2/100 + (32 - 50)^2/50 = 13.71$$

8. **Statistical decision.** Since $13.71 > 5.991$, we reject H_0.

9. **Conclusion.** We conclude that the trait is not distributed according to the ratio 1:2:1.

10. **p value.** Since $13.71 > 10.597$, the p value for the test is $p < .005$. ■

EXERCISES

12.3.1 The following table shows the distribution of uric acid determinations taken on 250 patients. Test the goodness-of-fit of these data to a normal distribution with $\mu = 5.74$ and $\sigma = 2.01$. Let $\alpha = .01$.

Uric Acid Determination	Observed Frequency	Uric Acid Determination	Observed Frequency
< 1	1	6 to 6.99	45
1 to 1.99	5	7 to 7.99	30
2 to 2.99	15	8 to 8.99	22
3 to 3.99	24	9 to 9.99	10
4 to 4.99	43	10 or higher	5
5 to 5.99	50		
Total			250

12.3.2 The following data were collected on 300 eight-year-old girls. Test, at the .05 level of significance, the null hypothesis that the data are drawn from a normally distributed population. The sample mean and standard deviation computed from grouped data are 127.02 and 5.08.

Height in Centimeters	Observed Frequency	Height in Centimeters	Observed Frequency
114 to 115.9	5	128 to 129.9	43
116 to 117.9	10	130 to 131.9	42
118 to 119.9	14	132 to 133.9	30
120 to 121.9	21	134 to 135.9	11
122 to 123.9	30	136 to 137.9	5
124 to 125.9	40	138 to 139.9	4
126 to 127.9	45		
Total			300

12.3.3 The face sheet of patients' records maintained in a local health department contains 10 entries. A sample of 100 records revealed the following distribution of erroneous entries:

Number of Erroneous Entries Out of 10	Number of Records
0	8
1	25
2	32
3	24
4	10
5 or more	1
Total	100

Test the goodness-of-fit of these data to the binomial distribution with $p = .20$. Find the p value for this test.

12.3.4 In a study conducted by Byers et al. (A-2), researchers tested a Poisson model for the distribution of activities of daily living (ADL) scores after a 7-month prehabilitation program designed to prevent functional decline among physically frail, community-living older persons. ADL measured the ability of individuals to perform essential tasks, including walking inside the house, bathing, upper and lower body dressing, transferring from a chair, toileting, feeding, and grooming. The scoring method used in this study assigned a value of 0 for no (personal) help and no difficulty, 1 for difficulty but no help, and 2 for help regardless of difficulty. Scores were summed to produce an overall score ranging from 0 to 16 (for eight tasks). There were 181 subjects who completed the study. Suppose we use the authors' scoring method to assess the status of another group of 181 subjects relative to their activities of daily living. Let us assume that the following results were obtained.

X	Observed Frequency X	Expected Frequency	X	Observed Frequency X	Expected Frequency
0	74	11.01	7	4	2.95
1	27	30.82	8	3	1.03
2	14	43.15	9	2	0.32
3	14	40.27	10	3	0.09
4	11	28.19	11	4	0.02
5	7	15.79	12 or more	13	0.01
6	5	7.37			

Source: Hypothetical data based on procedure reported by Amy L. Byers, Heather Allore, Thomas M. Gill, and Peter N. Peduzzi, "Application of Negative Binomial Modeling for Discrete Outcomes: A Case Study in Aging Research," *Journal of Clinical Epidemiology*, 56 (2003), 559–564.

Test the null hypothesis that these data were drawn from a Poisson distribution with $\lambda = 2.8$. Let $\alpha = .01$.

12.3.5 The following are the numbers of a particular organism found in 100 samples of water from a pond:

Number of Organisms per Sample	Frequency	Number of Organisms per Sample	Frequency
0	15	4	5
1	30	5	4
2	25	6	1
3	20	7	0
Total			100

Test the null hypothesis that these data were drawn from a Poisson distribution. Determine the p value for this test.

12.3.6 A research team conducted a survey in which the subjects were adult smokers. Each subject in a sample of 200 was asked to indicate the extent to which he or she agreed with the statement: "I would like to quit smoking." The results were as follows:

Response:	Strongly agree	Agree	Disagree	Strongly Disagree
Number Responding:	102	30	60	8

Can one conclude on the basis of these data that, in the sampled population, opinions are not equally distributed over the four levels of agreement? Let the probability of committing a type I error be .05 and find the p value.

12.4 TESTS OF INDEPENDENCE

Another, and perhaps the most frequent, use of the chi-square distribution is to test the null hypothesis that two criteria of classification, when applied to the same set of entities, are independent. We say that two criteria of classification are independent if the distribution of one criterion is the same no matter what the distribution of the other criterion. For example, if socioeconomic status and area of residence of the inhabitants of a certain city are independent, we would expect to find the same proportion of families in the low, medium, and high socioeconomic groups in all areas of the city.

The Contingency Table The classification, according to two criteria, of a set of entities, say, people, can be shown by a table in which the r rows represent the various levels of one criterion of classification and the c columns represent the various levels of the second criterion. Such a table is generally called a *contingency table*. The classification according to two criteria of a finite population of entities is shown in Table 12.4.1.

We will be interested in testing the null hypothesis that in the population the two criteria of classification are independent. If the hypothesis is rejected, we will conclude

TABLE 12.4.1 Two-Way Classification of a Finite Population of Entities

Second Criterion of Classification Level	First Criterion of Classification Level					
	1	2	3	...	c	Total
1	N_{11}	N_{12}	N_{13}	...	N_{1c}	$N_{1.}$
2	N_{21}	N_{22}	N_{23}	...	N_{2c}	$N_{2.}$
3	N_{31}	N_{32}	N_{33}	...	N_{3c}	$N_{3.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	N_{r1}	N_{r2}	N_{r3}	...	N_{rc}	$N_{r.}$
Total	$N_{.1}$	$N_{.2}$	$N_{.3}$	...	$N_{.c}$	N

TABLE 12.4.2 Two-Way Classification of a Sample of Entities

Second Criterion of Classification Level	First Criterion of Classification Level					
	1	2	3	...	c	Total
1	n_{11}	n_{12}	n_{13}	...	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	...	n_{2c}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	...	n_{3c}	$n_{3.}$
⋮	⋮	⋮	⋮		⋮	⋮
r	n_{r1}	n_{r2}	n_{r3}	...	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	...	$n_{.c}$	n

that the two criteria of classification are not independent. A sample of size n will be drawn from the population of entities, and the frequency of occurrence of entities in the sample corresponding to the cells formed by the intersections of the rows and columns of Table 12.4.1 along with the marginal totals will be displayed in a table such as Table 12.4.2.

Calculating the Expected Frequencies The expected frequency, under the null hypothesis that the two criteria of classification are independent, is calculated for each cell.

We learned in Chapter 3 (see Equation 3.4.4) that if two events are independent, the probability of their joint occurrence is equal to the product of their individual probabilities. Under the assumption of independence, for example, we compute the probability that one of the n subjects represented in Table 12.4.2 will be counted in Row 1 and Column 1 of the table (that is, in Cell 11) by multiplying the probability that the subject will be counted in Row 1 by the probability that the subject will be counted in Column 1. In the notation of the table, the desired calculation is

$$\left(\frac{n_{1.}}{n}\right)\left(\frac{n_{.1}}{n}\right)$$

To obtain the expected frequency for Cell 11, we multiply this probability by the total number of subjects, n. That is, the expected frequency for Cell 11 is given by

$$\left(\frac{n_{1.}}{n}\right)\left(\frac{n_{.1}}{n}\right)(n)$$

Since the n in one of the denominators cancels into numerator n, this expression reduces to

$$\frac{(n_{1.})(n_{.1})}{n}$$

In general, then, we see that to obtain the expected frequency for a given cell, we multiply the total of the row in which the cell is located by the total of the column in which the cell is located and divide the product by the grand total.

Observed Versus Expected Frequencies The expected frequencies and observed frequencies are compared. If the discrepancy is sufficiently small, the null hypothesis is tenable. If the discrepancy is sufficiently large, the null hypothesis is rejected, and we conclude that the two criteria of classification are not independent. The decision as to whether the discrepancy between observed and expected frequencies is sufficiently large to cause rejection of H_0 will be made on the basis of the size of the quantity computed when we use Equation 12.2.4, where O_i and E_i refer, respectively, to the observed and expected frequencies in the cells of Table 12.4.2. It would be more logical to designate the observed and expected frequencies in these cells by O_{ij} and E_{ij}, but to keep the notation simple and to avoid the introduction of another formula, we have elected to use the simpler notation. It will be helpful to think of the cells as being numbered from 1 to k, where 1 refers to Cell 11 and k refers to Cell rc. It can be shown that X^2 as defined in this manner is distributed approximately as χ^2 with $(r-1)(c-1)$ degrees of freedom when the null hypothesis is true. If the computed value of X^2 is equal to or larger than the tabulated value of χ^2 for some α, the null hypothesis is rejected at the α level of significance. The hypothesis testing procedure is illustrated with the following example.

EXAMPLE 12.4.1

In 1992, the U.S. Public Health Service and the Centers for Disease Control and Prevention recommended that all women of childbearing age consume 400 mg of folic acid daily to reduce the risk of having a pregnancy that is affected by a neural tube defect such as spina bifida or anencephaly. In a study by Stepanuk et al. (A-3), 693 pregnant women called a teratology information service about their use of folic acid supplementation. The researchers wished to determine if preconceptional use of folic acid and race are independent. The data appear in Table 12.4.3.

Solution:

1. **Data.** See Table 12.4.3.
2. **Assumptions.** We assume that the sample available for analysis is equivalent to a simple random sample drawn from the population of interest.

TABLE 12.4.3 Race of Pregnant Caller and Use of Folic Acid

	Preconceptional Use of Folic Acid		
	Yes	No	Total
White	260	299	559
Black	15	41	56
Other	7	14	21
Total	282	354	636

Source: Kathleen M. Stepanuk, Jorge E. Tolosa, Dawneete Lewis, Victoria Meyers, Cynthia Royds, Juan Carlos Saogal, and Ron Librizzi, "Folic Acid Supplementation Use Among Women Who Contact a Teratology Information Service," *American Journal of Obstetrics and Gynecology, 187* (2002), 964–967.

3. **Hypotheses.**

   H_0: Race and preconceptional use of folic acid are independent.
   H_A: The two variables are not independent.

   Let $\alpha = .05$.

4. **Test statistic.** The test statistic is

$$X^2 = \sum_{i=1}^{k}\left[\frac{(O_i - E_i)^2}{E_i}\right]$$

5. **Distribution of test statistic.** When H_0 is true, X^2 is distributed approximately as χ^2 with $(r - 1)(c - 1) = (3 - 1)(2 - 1) = (2)(1) = 2$ degrees of freedom.

6. **Decision rule.** Reject H_0 if the computed value of X^2 is equal to or greater than 5.991.

7. **Calculation of test statistic.** The expected frequency for the first cell is $(559 \times 282)/636 = 247.86$. The other expected frequencies are calculated in a similar manner. Observed and expected frequencies are displayed in Table 12.4.4. From the observed and expected frequencies we may compute

$$\begin{aligned} X^2 &= \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] \\ &= \frac{(260 - 247.86)^2}{247.86} + \frac{(299 - 311.14)^2}{311.14} + \cdots + \frac{(14 - 11.69)^2}{11.69} \\ &= .59461 + .47368 + \cdots + .45647 = 9.08960 \end{aligned}$$

8. **Statistical decision.** We reject H_0 since $9.08960 > 5.991$.

9. **Conclusion.** We conclude that H_0 is false, and that there is a relationship between race and preconceptional use of folic acid.

10. **p value.** Since $7.378 < 9.08960 < 9.210$, $.01 < p < .025$.

TABLE 12.4.4 Observed and Expected Frequencies for Example 12.4.1

	Preconceptional Use of Folic Acid		
	Yes	No	Total
White	260 (247.86)	299 (311.14)	559
Black	15 (24.83)	41 (31.17)	56
Other	7 (9.31)	14 (11.69)	21
Total	282	354	636

■

Computer Analysis The computer may be used to advantage in calculating X^2 for tests of independence and tests of homogeneity. Figure 12.4.1 shows the procedure and printout for Example 12.4.1 when the MINITAB program for computing X^2 from contingency tables is used. The data were entered into MINITAB Columns 1 and 2, corresponding to the columns of Table 12.4.3.

We may use SAS® to obtain an analysis and printout of contingency table data by using the PROC FREQ statement. Figure 12.4.2 shows a partial SAS® printout reflecting the analysis of the data of Example 12.4.1.

Data:

```
C1: 260 15  7
C2: 299 41 14
```

Dialog Box:

Stat ➤ Tables ➤ Chi-square Test

Type *C1-C2* in Columns containing the table.
Click **OK.**

Session command:

```
MTB > CHISQUARE C1-C3
```

Output:

Chi-Square Test: C1, C2

```
Expected counts are printed below observed counts

            C1        C2    Total
    1      260       299      559
        247.86    311.14

    2       15        41       56
         24.83     31.17

    3        7        14       21
          9.31     11.69

Total      282       354      636

Chi-Sq =  0.595 + 0.474 +
          3.892 + 3.100 +
          0.574 + 0.457 = 9.091
DF = 2, P-Value = 0.011
```

FIGURE 12.4.1 MINITAB procedure and output for chi-square analysis of data in Table 12.4.3

The SAS System

The FREQ Procedure

Table of race by folic

race Frequency Percent Row Pct Col Pct	folic No	Yes	Total
Black	41 6.45 73.21 11.58	15 2.36 26.79 5.32	56 8.81
Other	14 2.20 66.67 3.95	7 1.10 33.33 2.48	21 3.30
White	299 47.01 53.49 84.46	260 40.88 46.51 92.20	559 87.89
Total	354 55.66	282 44.34	636 100.00

Statistics for Table of race by folic

Statistic	DF	Value	Prob
Chi-Square	2	9.0913	0.0106
Likelihood Ratio Chi-Square	2	9.4808	0.0087
Mantel–Haenszel Chi-Square	1	8.9923	0.0027
Phi Coefficient		0.1196	
Contingency Coefficient		0.1187	
Cramer's V		0.1196	

Sample Size = 636

FIGURE 12.4.2 Partial SAS® printout for the chi-square analysis of the data from Example 12.4.1.

Note that the SAS® printout shows, in each cell, the percentage that cell frequency is of its row total, its column total, and the grand total. Also shown, for each row and column total, is the percentage that the total is of the grand total. In addition to the X^2 statistic, SAS® gives the value of several other statistics that may be computed from contingency table data. One of these, the Mantel–Haenszel chi-square statistic, will be discussed in a later section of this chapter.

Small Expected Frequencies The problem of small expected frequencies discussed in the previous section may be encountered when analyzing the data of contingency tables. Although there is a lack of consensus on how to handle this problem, many authors currently follow the rule given by Cochran (5). He suggests that for contingency tables with more than 1 degree of freedom a minimum expectation of 1 is allowable if no more than 20 percent of the cells have expected frequencies of less than 5. To meet this rule, adjacent rows and/or adjacent columns may be combined when to do so is logical in light of other considerations. If X^2 is based on less than 30 degrees of freedom, expected frequencies as small as 2 can be tolerated. We did not experience the problem of small expected frequencies in Example 12.4.1, since they were all greater than 5.

The 2 × 2 Contingency Table Sometimes each of two criteria of classification may be broken down into only two categories, or levels. When data are cross-classified in this manner, the result is a contingency table consisting of two rows and two columns. Such a table is commonly referred to as a 2 × 2 table. The value of X^2 may be computed by first calculating the expected cell frequencies in the manner discussed above. In the case of a 2 × 2 contingency table, however, X^2 may be calculated by the following shortcut formula:

$$X^2 = \frac{n(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)} \qquad (12.4.1)$$

where a, b, c, and d are the observed cell frequencies as shown in Table 12.4.5. When we apply the $(r - 1)(c - 1)$ rule for finding degrees of freedom to a 2 × 2 table, the result is 1 degree of freedom. Let us illustrate this with an example.

TABLE 12.4.5 A 2 × 2 Contingency Table

Second Criterion of Classification	First Criterion of Classification		
	1	2	Total
1	a	b	$a + b$
2	c	d	$c + d$
Total	$a + c$	$b + d$	n

EXAMPLE 12.4.2

According to Silver and Aiello (A-4), falls are of major concern among polio survivors. Researchers wanted to determine the impact of a fall on lifestyle changes. Table 12.4.6 shows the results of a study of 233 polio survivors on whether fear of falling resulted in lifestyle changes.

Solution:

1. **Data.** From the information given we may construct the 2×2 contingency table displayed as Table 12.5.6.
2. **Assumptions.** We assume that the sample is equivalent to a simple random sample.
3. **Hypotheses.**

   H_0: Fall status and lifestyle change because of fear of falling are independent.

   H_1: The two variables are not independent.

   Let $\alpha = .05$.
4. **Test statistic.** The test statistic is

$$X^2 = \sum_{i=1}^{k}\left[\frac{(O_i - E_i)^2}{E_i}\right]$$

5. **Distribution of test statistic.** When H_0 is true, X^2 is distributed approximately as χ^2 with $(r - 1)(c - 1) = (2 - 1)(2 - 1) = (1)(1) = 1$ degree of freedom.
6. **Decision rule.** Reject H_0 if the computed value of X^2 is equal to or greater than 3.841.
7. **Calculation of test statistic.** By Equation 12.4.1 we compute

$$X^2 = \frac{233[(131)(36) - (52)(14)]^2}{(145)(88)(183)(50)} = 31.7391$$

8. **Statistical decision.** We reject H_0 since $31.7391 > 3.841$.

TABLE 12.4.6 Contingency Table for the Data of Example 12.4.2

	Made Lifestyle Changes Because of Fear of Falling		
	Yes	No	Total
Fallers	131	52	183
Nonfallers	14	36	50
Total	145	88	233

Source: J. K. Silver and D. D. Aiello, "Polio Survivors: Falls and Subsequent Injuries," *American Journal of Physical Medicine and Rehabilitation*, 81 (2002), 567–570.

9. **Conclusion.** We conclude that H_0 is false, and that there is a relationship between experiencing a fall and changing one's lifestyle because of fear of falling.

10. **p value.** Since $31.7391 > 7.879$, $p < .005$. ■

Small Expected Frequencies The problems of how to handle small expected frequencies and small total sample sizes may arise in the analysis of 2×2 contingency tables. Cochran (5) suggests that the χ^2 test should not be used if $n < 20$ or if $20 < n < 40$ and any expected frequency is less than 5. When $n = 40$, an expected cell frequency as small as 1 can be tolerated.

Yates's Correction The observed frequencies in a contingency table are discrete and thereby give rise to a discrete statistic, X^2, which is approximated by the χ^2 distribution, which is continuous. Yates (6) in 1934 proposed a procedure for correcting for this in the case of 2×2 tables. The correction, as shown in Equation 12.4.2, consists of subtracting half the total number of observations from the absolute value of the quantity $ad - bc$ before squaring. That is,

$$X^2_{\text{corrected}} = \frac{n(|ad - bc| - .5n)^2}{(a + c)(b + d)(a + b)(c + d)} \tag{12.4.2}$$

It is generally agreed that no correction is necessary for larger contingency tables. Although Yates's correction for 2×2 tables has been used extensively in the past, more recent investigators have questioned its use. As a result, some practitioners recommend against its use.

We may, as a matter of interest, apply the correction to our current example. Using Equation 12.4.2 and the data from Table 12.4.6, we may compute

$$X^2 = \frac{233[|(131)(36) - (52)(14)| - .5(233)]^2}{(145)(88)(183)(50)} = 29.9118$$

As might be expected, with a sample this large, the difference in the two results is not dramatic.

Tests of Independence: Characteristics The characteristics of a chi-square test of independence that distinguish it from other chi-square tests are as follows:

1. A single sample is selected from a population of interest, and the subjects or objects are cross-classified on the basis of the two variables of interest.
2. The rationale for calculating expected cell frequencies is based on the probability law, which states that if two events (here the two criteria of classification) are independent, the probability of their joint occurrence is equal to the product of their individual probabilities.
3. The hypotheses and conclusions are stated in terms of the independence (or lack of independence) of two variables.

EXERCISES

In the exercises that follow perform the test at the indicated level of significance and determine the p value.

12.4.1 In the study by Silver and Aiello (A-4) cited in Example 12.4.2, a secondary objective was to determine if the frequency of falls was independent of wheelchair use. The following table gives the data for falls and wheelchair use among the subjects of the study.

	Wheelchair Use	
	Yes	No
Fallers	62	121
Nonfallers	18	32

Source: J. K. Silver and D. D. Aiello, "Polio Survivors: Falls and Subsequent Injuries," *American Journal of Physical Medicine and Rehabilitation*, *81* (2002), 567–570.

Do these data provide sufficient evidence to warrant the conclusion that wheelchair use and falling are related? Let $\alpha = .05$.

12.4.2 Sternal surgical site infection (SSI) after coronary artery bypass graft surgery is a complication that increases patient morbidity and costs for patients, payers, and the health care system. Segal and Anderson (A-5) performed a study that examined two types of preoperative skin preparation before performing open heart surgery. These two preparations used aqueous iodine and insoluble iodine with the following results.

	Comparison of Aqueous and Insoluble Preps	
Prep Group	Infected	Not Infected
Aqueous iodine	14	94
Insoluble iodine	4	97

Source: Cynthia G. Segal and Jacqueline J. Anderson, "Preoperative Skin Preparation of Cardiac Patients," *AORN Journal*, *76* (2002), 8231–827.

Do these data provide sufficient evidence at the $\alpha = .05$ level to justify the conclusion that the type of skin preparation and infection are related?

12.4.3 The side effects of nonsteroidal antiinflammatory drugs (NSAIDs) include problems involving peptic ulceration, renal function, and liver disease. In 1996, the American College of Rheumatology issued and disseminated guidelines recommending baseline tests (CBC, hepatic panel, and renal tests) when prescribing NSAIDs. A study was conducted by Rothenberg and Holcomb (A-6) to determine if physicians taking part in a national database of computerized medical records performed the recommended baseline tests when prescribing NSAIDs. The researchers classified physicians in the study into four categories—those practicing in internal medicine, family practice, academic family practice, and multispeciality groups. The data appear in the following table.

	Performed Baseline Tests	
Practice Type	Yes	No
Internal medicine	294	921
Family practice	98	2862
Academic family practice	50	3064
Multispecialty groups	203	2652

Source: Ralph Rothenberg and John P. Holcomb, "Guidelines for Monitoring of NSAIDs: Who Listened?," *Journal of Clinical Rheumatology*, 6 (2000), 258–265.

Do the data above provide sufficient evidence for us to conclude that type of practice and performance of baseline tests are related? Use $\alpha = .01$.

12.4.4 Boles and Johnson (A-7) examined the beliefs held by adolescents regarding smoking and weight. Respondents characterized their weight into three categories: underweight, overweight, or appropriate. Smoking status was categorized according to the answer to the question, "Do you currently smoke, meaning one or more cigarettes per day?" The following table shows the results of a telephone study of adolescents in the age group 12–17.

	Smoking	
	Yes	No
Underweight	17	97
Overweight	25	142
Appropriate	96	816

Source: Sharon M. Boles and Patrick B. Johnson, "Gender, Weight Concerns, and Adolescent Smoking," *Journal of Addictive Diseases*, 20 (2001), 5–14.

Do the data provide sufficient evidence to suggest that weight perception and smoking status are related in adolescents? $\alpha = .05$.

12.4.5 A sample of 500 college students participated in a study designed to evaluate the level of college students' knowledge of a certain group of common diseases. The following table shows the students classified by major field of study and level of knowledge of the group of diseases:

	Knowledge of Diseases		
Major	Good	Poor	Total
Premedical	31	91	122
Other	19	359	378
Total	50	450	500

Do these data suggest that there is a relationship between knowledge of the group of diseases and major field of study of the college students from which the present sample was drawn? Let $\alpha = .05$.

12.4.6 The following table shows the results of a survey in which the subjects were a sample of 300 adults residing in a certain metropolitan area. Each subject was asked to indicate which of three policies they favored with respect to smoking in public places.

Highest Education Level	Policy Favored: No Restrictions on Smoking	Smoking Allowed in Designated Areas Only	No Smoking at All	No Opinion	Total
College graduate	5	44	23	3	75
High-school graduate	15	100	30	5	150
Grade-school graduate	15	40	10	10	75
Total	35	184	63	18	300

Can one conclude from these data that, in the sampled population, there is a relationship between level of education and attitude toward smoking in public places? Let $\alpha = .05$.

12.5 TESTS OF HOMOGENEITY

A characteristic of the examples and exercises presented in the last section is that, in each case, the total sample was assumed to have been drawn before the entities were classified according to the two criteria of classification. That is, the observed number of entities falling into each cell was determined after the sample was drawn. As a result, the row and column totals are chance quantities not under the control of the investigator. We think of the sample drawn under these conditions as a single sample drawn from a single population. On occasion, however, either row or column totals may be under the control of the investigator; that is, the investigator may specify that independent samples be drawn from each of several populations. In this case, one set of marginal totals is said to be *fixed*, while the other set, corresponding to the criterion of classification applied to the samples, is *random*. The former procedure, as we have seen, leads to a chi-square test of independence. The latter situation leads to a chi-square *test of homogeneity*. The two situations not only involve different sampling procedures; they lead to different questions and null hypotheses. The test of independence is concerned with the question: Are the two criteria of classification independent? The homogeneity test is concerned with the question: Are the samples drawn from populations that are homogeneous with respect to some criterion of classification? In the latter case the null hypothesis states that the samples are drawn from the same population. Despite these differences in concept and sampling procedure, the two tests are mathematically identical, as we see when we consider the following example.

Calculating Expected Frequencies Either the row categories or the column categories may represent the different populations from which the samples are drawn. If, for example, three populations are sampled, they may be designated as populations 1, 2, and 3, in which case these labels may serve as either row or column headings. If the variable of interest has three categories, say, *A*, *B*, and *C*, these labels may serve as headings for rows or columns, whichever is not used for the populations. If we use notation similar to that adopted for Table 12.4.2, the contingency table for this situation, with columns used to represent the populations, is shown as Table 12.5.1. Before computing our test statistic we

TABLE 12.5.1 A Contingency Table for Data for a Chi-Square Test of Homogeneity

	Population			
Variable Category	1	2	3	Total
A	n_{A1}	n_{A2}	n_{A3}	$n_{A.}$
B	n_{B1}	n_{B2}	n_{B3}	$n_{B.}$
C	n_{C1}	n_{C2}	n_{C3}	$n_{C.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

need expected frequencies for each of the cells in Table 12.5.1. If the populations are indeed homogeneous, or, equivalently, if the samples are all drawn from the same population, with respect to the categories A, B, and C, our best estimate of the proportion in the combined population who belong to category A is $n_{A.}/n$. By the same token, if the three populations are homogeneous, we interpret this probability as applying to each of the populations individually. For example, under the null hypothesis, $n_{A.}$ is our best estimate of the probability that a subject picked at random from the combined population will belong to category A. We would expect, then, to find $n_{.1}(n_{A.}/n)$ of those in the sample from population 1 to belong to category A, $n_{.2}(n_{A.}/n)$ of those in the sample from population 2 to belong to category A, and $n_{.3}(n_{A.}/n)$ of those in the sample from population 3 to belong to category A. These calculations yield the expected frequencies for the first row of Table 12.5.1. Similar reasoning and calculations yield the expected frequencies for the other two rows.

We see again that the shortcut procedure of multiplying appropriate marginal totals and dividing by the grand total yields the expected frequencies for the cells.

From the data in Table 12.5.1 we compute the following test statistic:

$$X^2 = \sum_{i=1}^{k}\left[\frac{(O_i - E_i)^2}{E_i}\right]$$

EXAMPLE 12.5.1

Narcolepsy is a disease involving disturbances of the sleep–wake cycle. Members of the German Migraine and Headache Society (A-8) studied the relationship between migraine headaches in 96 subjects diagnosed with narcolepsy and 96 healthy controls. The results are shown in Table 12.5.2. We wish to know if we may conclude, on the basis of these

TABLE 12.5.2 Frequency of Migraine Headaches by Narcolepsy Status

	Reported Migraine Headaches		
	Yes	No	Total
Narcoleptic subjects	21	75	96
Healthy controls	19	77	96
Total	40	152	192

Source: The DMG Study Group, "Migraine and Idiopathic Narcolepsy—A Case-Control Study," *Cephalagia, 23* (2003), 786–789.

data, that the narcolepsy population and healthy populations represented by the samples are not homogeneous with respect to migraine frequency.

Solution:

1. **Data.** See Table 12.5.2.
2. **Assumptions.** We assume that we have a simple random sample from each of the two populations of interest.
3. **Hypotheses.**

   H_0: The two populations are homogeneous with respect to migraine frequency.

   H_A: The two populations are not homogeneous with respect to migraine frequency.

   Let $\alpha = .05$.
4. **Test statistic.** The test statistic is

$$X^2 = \sum[(O_i - E_i)^2/E_i]$$

5. **Distribution of test statistic.** If H_0 is true, X^2 is distributed approximately as χ^2 with $(2-1)(2-1) = (1)(1) = 1$ degree of freedom.
6. **Decision rule.** Reject H_0 if the computed value of X^2 is equal to or greater than 3.841.
7. **Calculation of test statistic.** The MINITAB output is shown in Figure 12.5.1.

Chi-Square Test

```
Expected counts are printed below observed counts

Rows: Narcolepsy   Columns: Migraine

            No       Yes       All

No          77        19        96
         76.00     20.00     96.00

Yes         75        21        96
         76.00     20.00     96.00

All        152        40       192
        152.00     40.00    192.00

Chi-Square = 0.126, DF = 1, P-Value = 0.722
```

FIGURE 12.5.1 MINITAB output for Example 12.5.1.

8. **Statistical decision.** Since .126 is less than the critical value of 3.841, we are unable to reject the null hypothesis.
9. **Conclusion.** We conclude that the two populations may be homogeneous with respect to migraine frequency.
10. ***p* value.** From the MINITAB output we see that $p = .722$. ■

Small Expected Frequencies The rules for small expected frequencies given in the previous section are applicable when carrying out a test of homogeneity.

In summary, the chi-square test of homogeneity has the following characteristics:

1. Two or more populations are identified in advance, and an independent sample is drawn from each.
2. Sample subjects or objects are placed in appropriate categories of the variable of interest.
3. The calculation of expected cell frequencies is based on the rationale that if the populations are homogeneous as stated in the null hypothesis, the best estimate of the probability that a subject or object will fall into a particular category of the variable of interest can be obtained by pooling the sample data.
4. The hypotheses and conclusions are stated in terms of homogeneity (with respect to the variable of interest) of populations.

Test of Homogeneity and H_0: $p_1 = p_2$ The chi-square test of homogeneity for the two-sample case provides an alternative method for testing the null hypothesis that two population proportions are equal. In Section 7.6, it will be recalled, we learned to test H_0: $p_1 = p_2$ against H_A: $p_1 \neq p_2$ by means of the statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)_0}{\sqrt{\dfrac{\bar{p}(1-\bar{p})}{n_1} + \dfrac{\bar{p}(1-\bar{p})}{n_2}}}$$

where $\bar{p}$ is obtained by pooling the data of the two independent samples available for analysis.

Suppose, for example, that in a test of H_0: $p_1 = p_2$ against H_A: $p_1 \neq p_2$, the sample data were as follows: $n_1 = 100$, $\hat{p}_1 = .60$ $n_2 = 120$, $\hat{p}_2 = .40$. When we pool the sample data we have

$$\bar{p} = \frac{.60(100) + .40(120)}{100 + 120} = \frac{108}{220} = .4909$$

and

$$z = \frac{.60 - .40}{\sqrt{\dfrac{(.4909)(.5091)}{100} + \dfrac{(.4909)(.5091)}{120}}} = 2.95469$$

which is significant at the .05 level since it is greater than the critical value of 1.96.

If we wish to test the same hypothesis using the chi-square approach, our contingency table will be

	Characteristic Present		
Sample	**Yes**	**No**	**Total**
1	60	40	100
2	48	72	120
Total	108	112	220

By Equation 12.4.1 we compute

$$X^2 = \frac{220[(60)(72) - (40)(48)]^2}{(108)(112)(100)(120)} = 8.7302$$

which is significant at the .05 level because it is greater than the critical value of 3.841. We see, therefore, that we reach the same conclusion by both methods. This is not surprising because, as explained in Section 12.2, $X^2_{(1)} = z^2$. We note that $8.7302 = (2.95469)^2$ and that $3.841 = (1.96)^2$.

EXERCISES

In the exercises that follow perform the test at the indicated level of significance and determine the p value.

12.5.1 Refer to the study by Carter et al. (A-9), who investigated the effect of age at onset of bipolar disorder on the course of the illness. One of the variables studied was subjects' family history. Table 3.4.1 shows the frequency of a family history of mood disorders in the two groups of interest: early age at onset (18 years or younger) and later age at onset (later than 18 years).

Family History of Mood Disorders	Early $\leq$ 18(E)	Later > 18(L)	Total
Negative (A)	28	35	63
Bipolar disorder (B)	19	38	57
Unipolar (C)	41	44	85
Unipolar and bipolar (D)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research*, 37 (2003), 297–303.

Can we conclude on the basis of these data that subjects 18 or younger differ from subjects older than 18 with respect to family histories of mood disorders? Let $\alpha = .05$.

12.5.2 Coughlin et al. (A-10) examined breast and cervical screening practices of Hispanic and non-Hispanic women in counties that approximate the U.S. southern border region. The study used data from the Behavioral Risk Factor Surveillance System surveys of adults ages 18 years or older conducted in 1999 and 2000. The following table shows the number of observations of Hispanic and non-Hispanic women who had received a mammogram in the past 2 years cross-classified by marital status.

Marital Status	Hispanic	Non-Hispanic	Total
Currently married	319	738	1057
Divorced or separated	130	329	459
Widowed	88	402	490
Never married or living as an unmarried couple	41	95	136
Total	578	1564	2142

Source: Steven S. Coughlin, Robert J. Uhler, Thomas Richards, and Katherine M. Wilson, "Breast and Cervical Cancer Screening Practices Among Hispanic and Non-Hispanic Women Residing Near the United States–Mexico Border, 1999–2000," *Family and Community Health*, 26, (2003), 130–139.

We wish to know if we may conclude on the basis of these data that marital status and ethnicity (Hispanic and non-Hispanic) in border counties of the southern United States are not homogeneous. Let $\alpha = .05$.

12.5.3 Swor et al. (A-11) examined the effectiveness of cardiopulmonary resuscitation (CPR) training in people over 55 years of age. They compared the skill retention rates of subjects in this age group who completed a course in traditional CPR instruction with those who received chest-compression-only cardiopulmonary resuscitation (CC-CPR). Independent groups were tested 3 months after training. Among the 27 subjects receiving traditional CPR, 12 were rated as competent. In the CC-CPR group, 15 out of 29 were rated competent. Do these data provide sufficient evidence for us to conclude that the two populations are not homogeneous with respect to competency rating 3 months after training? Let $\alpha = .05$.

12.5.4 In an air pollution study, a random sample of 200 households was selected from each of two communities. A respondent in each household was asked whether or not anyone in the household was bothered by air pollution. The responses were as follows:

	Any Member of Household Bothered by Air Pollution?		
Community	Yes	No	Total
I	43	157	200
II	81	119	200
Total	124	276	400

Can the researchers conclude that the two communities differ with respect to the variable of interest? Let $\alpha = .05$.