

The Chi-Square Test

OBSERVED AND THEORETICAL FREQUENCIES

As we have already seen many times, the results obtained in samples do not always agree exactly with the theoretical results expected according to the rules of probability. For example, although theoretical considerations lead us to expect 50 heads and 50 tails when we toss a fair coin 100 times, it is rare that these results are obtained exactly.

Suppose that in a particular sample a set of possible events $E_1, E_2, E_3, \dots, E_k$ (see Table 12.1) are observed to occur with frequencies $o_1, o_2, o_3, \dots, o_k$, called *observed frequencies*, and that according to probability rules they are expected to occur with frequencies $e_1, e_2, e_3, \dots, e_k$, called *expected, or theoretical, frequencies*. Often we wish to know whether the observed frequencies differ significantly from the expected frequencies.

Table 12.1

Event	E_1	E_2	E_3	\dots	E_k
Observed frequency	o_1	o_2	o_3	\dots	o_k
Expected frequency	e_1	e_2	e_3	\dots	e_k

DEFINITION OF χ^2

A measure of the discrepancy existing between the observed and expected frequencies is supplied by the statistic χ^2 (read chi-square) given by

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \dots + \frac{(o_k - e_k)^2}{e_k} = \sum_{j=1}^k \frac{(o_j - e_j)^2}{e_j} \quad (1)$$

where if the total frequency is N ,

$$\sum o_j = \sum e_j = N \quad (2)$$

An expression equivalent to formula (1) is (see Problem 12.11)

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N \quad (3)$$

If $\chi^2 = 0$, the observed and theoretical frequencies agree exactly; while if $\chi^2 > 0$, they do not agree exactly. The larger the value of χ^2 , the greater is the discrepancy between the observed and expected frequencies.

The sampling distribution of χ^2 is approximated very closely by the chi-square distribution

$$Y = Y_0(\chi^2)^{1/2(\nu-2)}e^{-1/2\chi^2} = Y_0\chi^{\nu-2}e^{-1/2\chi^2} \tag{4}$$

(already considered in Chapter 11) if the expected frequencies are at least equal to 5. The approximation improves for larger values.

The number of degrees of freedom, ν , is given by

- (1) $\nu = k - 1$ if the expected frequencies can be computed without having to estimate the population parameters from sample statistics. Note that we subtract 1 from k because of constraint condition (2), which states that if we know $k - 1$ of the expected frequencies, the remaining frequency can be determined.
- (2) $\nu = k - 1 - m$ if the expected frequencies can be computed only by estimating m population parameters from sample statistics.

SIGNIFICANCE TESTS

In practice, expected frequencies are computed on the basis of a hypothesis H_0 . If under this hypothesis the computed value of χ^2 given by equation (1) or (3) is greater than some critical value (such as $\chi^2_{.95}$ or $\chi^2_{.99}$, which are the critical values of the 0.05 and 0.01 significance levels, respectively), we would conclude that the observed frequencies differ *significantly* from the expected frequencies and would reject H_0 at the corresponding level of significance; otherwise, we would accept it (or at least not reject it). This procedure is called *the chi-square test* of hypothesis or significance.

It should be noted that we must look with suspicion upon circumstances where χ^2 is *too close to zero*, since it is rare that observed frequencies agree *too well* with expected frequencies. To examine such situations, we can determine whether the computed value of χ^2 is less than $\chi^2_{.05}$ or $\chi^2_{.01}$, in which cases we would decide that the agreement is *too good* at the 0.05 or 0.01 significance levels, respectively.

THE CHI-SQUARE TEST FOR GOODNESS OF FIT

The chi-square test can be used to determine how well theoretical distributions (such as the normal and binomial distributions) fit empirical distributions (i.e., those obtained from sample data). See Problems 12.12 and 12.13.

EXAMPLE 1. A pair of dice is rolled 500 times with the sums in Table 12.2 showing on the dice:

Table 12.2

Sum	2	3	4	5	6	7	8	9	10	11	12
Observed	15	35	49	58	65	76	72	60	35	29	6

The expected number, if the dice are fair, are determined from the distribution of x as in Table 12.3.

Table 12.3

x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

We have the observed and expected frequencies in Table 12.4.

Table 12.4

Observed	15	35	49	58	65	76	72	60	35	29	6
Expected	13.9	27.8	41.7	55.6	69.5	83.4	69.5	55.6	41.7	27.8	13.9

If the observed and expected are entered into B1:L2 in the EXCEL worksheet, the expression $=(B1-B2)^2/B2$ is entered into B4, a click-and-drag is executed from B4 to L4, and then the quantities in B4:L4 are summed we obtain 10.34 for $\chi^2 = \sum_j ((o_j - e_j)^2/e_j)$.

The p -value corresponding to 10.34 is given by the EXCEL expression $=CHIDIST(10.34,10)$. The p -value is 0.411. Because of this large p -value, we have no reason to doubt the fairness of the dice.

CONTINGENCY TABLES

Table 12.1, in which the observed frequencies occupy a single row, is called a *one-way classification table*. Since the number of columns is k , this is also called a $1 \times k$ (read “1 by k ”) *table*. By extending these ideas, we can arrive at *two-way classification tables*, or $h \times k$ *tables*, in which the observed frequencies occupy h rows and k columns. Such tables are often called *contingency tables*.

Corresponding to each observed frequency in an $h \times k$ contingency table, there is an *expected* (or *theoretical*) *frequency* that is computed subject to some hypothesis according to rules of probability. These frequencies, which occupy the *cells* of a contingency table, are called *cell frequencies*. The total frequency in each row or each column is called the *marginal frequency*.

To investigate agreement between the observed and expected frequencies, we compute the statistic

$$\chi^2 = \sum_j \frac{(o_j - e_j)^2}{e_j} \quad (5)$$

where the sum is taken over all cells in the contingency table and where the symbols o_j and e_j represent, respectively, the observed and expected frequencies in the j th cell. This sum, which is analogous to equation (1), contains hk terms. The sum of all observed frequencies is denoted by N and is equal to the sum of all expected frequencies [compare with equation (2)].

As before, statistic (5) has a sampling distribution given very closely by (4), provided the expected frequencies are not too small. The number of degrees of freedom, ν , of this chi-square distribution is given for $h > 1$ and $k > 1$ by

1. $\nu = (h - 1)(k - 1)$ if the expected frequencies can be computed without having to estimate population parameters from sample statistics. For a proof of this, see Problem 12.18.
2. $\nu = (h - 1)(k - 1) - m$ if the expected frequencies can be computed only by estimating m population parameters from sample statistics.

Significance tests for $h \times k$ tables are similar to those for $1 \times k$ tables. The expected frequencies are found subject to a particular hypothesis H_0 . A hypothesis commonly assumed is that the two classifications are independent of each other.

Contingency tables can be extended to higher dimensions. Thus, for example, we can have $h \times k \times l$ tables, where three classifications are present.

EXAMPLE 2. The data in Table 12.5 were collected on how individuals prepared their taxes and their education level. The null hypothesis is that the way people prepare their taxes (computer software or pen and paper) is independent of their education level. Table 12.5 is a contingency table.

Table 12.5

	Education Level		
Tax prep.	High school	Bachelors	Masters
computer	23	35	42
Pen and paper	45	30	25

If MINITAB is used to analyze this data, the following results are obtained.

Chi-Square Test: highschool, bachelors, masters

Expected counts are printed below observed counts

Chi-Square contributions are printed below expected counts

	highschool	bachelors	masters	Total
1	23	35	42	100
	34.00	32.50	33.50	
	3.559	0.192	2.157	
2	45	30	25	100
	34.00	32.50	33.50	
	3.559	0.192	2.157	
Total	68	65	67	200

Chi-Sq = 11.816, DF = 2, P-Value = 0.003

Because of the small p -value, the hypothesis of independence would be rejected and we would conclude that tax preparation would be contingent upon education level.

YATES' CORRECTION FOR CONTINUITY

When results for continuous distributions are applied to discrete data, certain corrections for continuity can be made, as we have seen in previous chapters. A similar correction is available when the chi-square distribution is used. The correction consists in rewriting equation (1) as

$$\chi^2(\text{corrected}) = \frac{(|o_1 - e_1| - 0.5)^2}{e_1} + \frac{(|o_2 - e_2| - 0.5)^2}{e_2} + \dots + \frac{(|o_k - e_k| - 0.5)^2}{e_k} \quad (6)$$

and is often referred to as *Yates' correction*. An analogous modification of equation (5) also exists.

In general, the correction is made only when the number of degrees of freedom is $\nu = 1$. For large samples, this yields practically the same results as the uncorrected χ^2 , but difficulties can arise near critical values (see Problem 12.8). For small samples where each expected frequency is between 5 and 10, it is perhaps best to compare both the corrected and uncorrected values of χ^2 . If both values lead to the same conclusion regarding a hypothesis, such as rejection at the 0.05 level, difficulties are rarely encountered. If they lead to different conclusions, one can resort to increasing the sample sizes or, if this proves impractical, one can employ methods of probability involving the *multinomial distribution* of Chapter 6.

SIMPLE FORMULAS FOR COMPUTING χ^2

Simple formulas for computing χ^2 that involve only the observed frequencies can be derived. The following gives the results for 2×2 and 2×3 contingency tables (see Tables 12.6 and 12.7, respectively).

2 × 2 Tables

$$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{(a_1 + b_1)(a_2 + b_2)(a_1 + a_2)(b_1 + b_2)} = \frac{N\Delta^2}{N_1N_2N_A N_B} \quad (7)$$

Table 12.6

	I	II	Total
A	a_1	a_2	N_A
B	b_1	b_2	N_B
Total	N_1	N_2	N

Table 12.7

	I	II	III	Total
A	a_1	a_2	a_3	N_A
B	b_1	b_2	b_3	N_B
Total	N_1	N_2	N_3	N

where $\Delta = a_1b_2 - a_2b_1$, $N = a_1 + a_2 + b_1 + b_2$, $N_1 = a_1 + b_1$, $N_2 = a_2 + b_2$, $N_A = a_1 + a_2$, and $N_B = b_1 + b_2$ (see Problem 12.19). With Yates' correction, this becomes

$$\chi^2 (\text{corrected}) = \frac{N(|a_1b_2 - a_2b_1| - \frac{1}{2}N)^2}{(a_1 + b_1)(a_2 + b_2)(a_1 + a_2)(b_1 + b_2)} = \frac{N(|\Delta| - \frac{1}{2}N)^2}{N_1N_2N_A N_B} \quad (8)$$

2 × 3 Tables

$$\chi^2 = \frac{N}{N_A} \left[\frac{a_1^2}{N_1} + \frac{a_2^2}{N_2} + \frac{a_3^2}{N_3} \right] + \frac{N}{N_B} \left[\frac{b_1^2}{N_1} + \frac{b_2^2}{N_2} + \frac{b_3^2}{N_3} \right] - N \quad (9)$$

where we have used the general result valid for all contingency tables (see Problem 12.43):

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N \quad (10)$$

Result (9) for $2 \times k$ tables where $k > 3$ can be generalized (see Problem 12.46).

COEFFICIENT OF CONTINGENCY

A measure of the degree of relationship, association, or dependence of the classifications in a contingency table is given by

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \quad (11)$$

which is called the *coefficient of contingency*. The larger the value of C , the greater is the degree of association. The number of rows and columns in the contingency table determines the maximum value of C , which is never greater than 1. If the number of rows and columns of a contingency table is equal to k , the maximum value of C is given by $\sqrt{(k-1)/k}$ (see Problems 12.22, 12.52, and 12.53).

EXAMPLE 3. Find the coefficient of contingency for Example 2.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{11.816}{11.816 + 200}} = 0.236$$

CORRELATION OF ATTRIBUTES

Because classifications in a contingency table often describe characteristics of individuals or objects, they are often referred to as *attributes*, and the degree of dependence, association, or relationship is called the *correlation* of attributes. For $k \times k$ tables, we define

$$r = \sqrt{\frac{\chi^2}{N(k-1)}} \quad (12)$$

as the correlation coefficient between attributes (or classifications). This coefficient lies between 0 and 1 (see Problem 12.24). For 2×2 tables in which $k = 2$, the correlation is often called *tetrachoric correlation*.

The general problem of correlation of numerical variables is considered in Chapter 14.

ADDITIVE PROPERTY OF χ^2

Suppose that the results of repeated experiments yield sample values of χ^2 given by $\chi_1^2, \chi_2^2, \chi_3^2, \dots$ with $\nu_1, \nu_2, \nu_3, \dots$ degrees of freedom, respectively. Then the result of all these experiments can be considered equivalent to a χ^2 value given by $\chi_1^2 + \chi_2^2 + \chi_3^2 + \dots$ with $\nu_1 + \nu_2 + \nu_3 + \dots$ degrees of freedom (see Problem 12.25).

Solved Problems

THE CHI-SQUARE TEST

- 12.1** In 200 tosses of a coin, 115 heads and 85 tails were observed. Test the hypothesis that the coin is fair, using Appendix IV and significance levels of (a) 0.05 and (b) 0.01. Test the hypothesis by computing the p -value and (c) comparing it to levels 0.05 and 0.01.

SOLUTION

The observed frequencies of heads and tails are $o_1 = 115$ and $o_2 = 85$, respectively, and the expected frequencies of heads and tails (if the coin is fair) are $e_1 = 100$ and $e_2 = 100$, respectively. Thus

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} = \frac{(115 - 100)^2}{100} + \frac{(85 - 100)^2}{100} = 4.50$$

Since the number of categories, or classes (heads, tails), is $k = 2$, $\nu = k - 1 = 2 - 1 = 1$.

- (a) The critical value $\chi_{.95}^2$ for 1 degree of freedom is 3.84. Thus, since $4.50 > 3.84$, we reject the hypothesis that the coin is fair at the 0.05 significance level.
- (b) The critical value $\chi_{.99}^2$ for 1 degree of freedom is 6.63. Thus, since $4.50 < 6.63$, we cannot reject the hypothesis that the coin is fair at the 0.02 significance level.

We conclude that the observed results are *probably significant* and that the coin is *probably not fair*. For a comparison of this method with previous methods used, see Problem 12.3.

Using EXCEL, the p -value is given by =CHIDIST(4.5,1), which equals 0.0339. And we see, using the p -value approach that the results are significant at 0.05 but not at 0.01. Either of these methods of testing may be used.

- 12.2** Work Problem 12.1 by using Yates' correction.

SOLUTION

$$\begin{aligned} \chi^2 (\text{corrected}) &= \frac{(|o_1 - e_1| - 0.5)^2}{e_1} + \frac{(|o_2 - e_2| - 0.5)^2}{e_2} = \frac{(|115 - 100| - 0.5)^2}{100} + \frac{(|85 - 100| - 0.5)^2}{100} \\ &= \frac{(14.5)^2}{100} + \frac{(14.5)^2}{100} = 4.205 \end{aligned}$$

Since $4.205 > 3.84$ and $4.205 < 6.63$, the conclusions reached in Problem 12.1 are valid. For a comparison with previous methods, see Problem 12.3.

12.3 Work Problem 12.1 by using the normal approximation to the binomial distribution.

SOLUTION

Under the hypothesis that the coin is fair, the mean and standard deviation of the number of heads expected in 200 tosses of a coin are $\mu = Np = (200)(0.5) = 100$ and $\sigma = \sqrt{Npq} = \sqrt{(200)(0.5)(0.5)} = 7.07$, respectively.

First method

$$115 \text{ heads in standard units} = \frac{115 - 100}{7.07} = 2.12$$

Using the 0.05 significance level and a two-tailed test, we would reject the hypothesis that the coin is fair if the z score were outside the interval -1.96 to 1.96 . With the 0.01 level, the corresponding interval would be -2.58 to 2.58 . It follows (as in Problem 12.1) that we can reject the hypothesis at the 0.05 level but cannot reject it at the 0.01 level.

Note that the square of the above standard score, $(2.12)^2 = 4.50$, is the same as the value of χ^2 obtained in Problem 12.1. This is always the case for a chi-square test involving two categories (see Problem 12.10).

Second method

Using the correction for continuity, 115 or more heads is equivalent to 114.5 or more heads. Then 114.5 in standard units $= (114.5 - 100)/7.07 = 2.05$. This leads to the same conclusions as in the first method.

Note that the square of this standard score is $(2.05)^2 = 4.20$, agreeing with the value of χ^2 corrected for continuity by using Yates' correction of Problem 12.2. This is always the case for a chi-square test involving two categories in which Yates' correction is applied.

12.4 Table 12.8 shows the observed and expected frequencies in tossing a die 120 times.

- Test the hypothesis that the die is fair using a 0.05 significance level by calculating χ^2 and giving the 0.05 critical value and comparing the computed test statistic with the critical value.
- Compute the p -value and compare it with 0.05 to test the hypothesis.

Table 12.8

Die face	1	2	3	4	5	6
Observed frequency	25	17	15	23	24	16
Expected frequency	20	20	20	20	20	20

SOLUTION

$$\begin{aligned} \chi^2 &= \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \frac{(o_3 - e_3)^2}{e_3} + \frac{(o_4 - e_4)^2}{e_4} + \frac{(o_5 - e_5)^2}{e_5} + \frac{(o_6 - e_6)^2}{e_6} \\ &= \frac{(25 - 20)^2}{20} + \frac{(17 - 20)^2}{20} + \frac{(15 - 20)^2}{20} + \frac{(23 - 20)^2}{20} + \frac{(24 - 20)^2}{20} + \frac{(16 - 20)^2}{20} = 5.00 \end{aligned}$$

- The 0.05 critical value is given by the EXCEL expression =CHIINV(0.05, 5) or 11.0705. The computed value of the test statistic is 5.00. Since the computed test statistic is not in the 0.05 critical region, do not reject the null that the die is fair.
- The p -value is given by the EXCEL expression =CHIDIST(5.00, 5) or 0.4159. Since the p -value is not less than 0.05, do not reject the null that the die is fair.

- 12.5** Table 12.9 shows the distribution of the digits 0, 1, 2, ..., 9 in a random-number table of 250 digits. (a) Find the value of the test statistic χ^2 , (b) find the 0.01 critical value and give your conclusion for $\alpha=0.01$, and (c) find the p -value for the value you found in (a) and give your conclusion for $\alpha=0.01$.

Table 12.9

Digit	0	1	2	3	4	5	6	7	8	9
Observed frequency	17	31	29	18	14	20	35	30	20	36
Expected frequency	25	25	25	25	25	25	25	25	25	25

SOLUTION

$$(a) \quad \chi^2 = \frac{(17-25)^2}{25} + \frac{(31-25)^2}{25} + \frac{(29-25)^2}{25} + \frac{(18-25)^2}{25} + \cdots + \frac{(36-25)^2}{25} = 23.3$$

(b) The 0.01 critical value is given by =CHIINV(0.01, 9) which equals 21.6660. Since the computed value of χ^2 exceeds this value, we reject the hypothesis that the numbers are random.

(c) The p -value is given by the EXCEL expression =CHIDIST(23.3, 9) which equals 0.0056, which is less than 0.01. By the p -value technique, we reject the null.

- 12.6** In his experiments with peas, Gregor Mendel observed that 315 were round and yellow, 108 were round and green, 101 were wrinkled and yellow, and 32 were wrinkled and green. According to his theory of heredity, the numbers should be in the proportion 9 : 3 : 3 : 1. Is there any evidence to doubt his theory at the (a) 0.01 and (b) 0.05 significance levels?

SOLUTION

The total number of peas is $315 + 108 + 101 + 32 = 556$. Since the expected numbers are in the proportion 9 : 3 : 3 : 1 (and $9 + 3 + 3 + 1 = 16$), we would expect

$$\begin{aligned} \frac{9}{16}(556) &= 312.75 \text{ round and yellow} & \frac{3}{16}(556) &= 104.25 \text{ wrinkled and yellow} \\ \frac{3}{16}(556) &= 104.25 \text{ round and green} & \frac{1}{16}(556) &= 34.75 \text{ wrinkled and green} \end{aligned}$$

$$\text{Thus } \chi^2 = \frac{(315 - 312.75)^2}{312.75} + \frac{(108 - 104.25)^2}{104.25} + \frac{(101 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} = 0.470$$

Since there are four categories, $k = 4$ and the number of degrees of freedom is $\nu = 4 - 1 = 3$.

(a) For $\nu = 3$, $\chi_{.99}^2 = 11.3$, and thus we cannot reject the theory at the 0.01 level.

(b) For $\nu = 3$, $\chi_{.95}^2 = 7.81$, and thus we cannot reject the theory at the 0.05 level.

We conclude that the theory and experiment are in agreement.

Note that for 3 degrees of freedom, $\chi_{.05}^2 = 7.81$ and $\chi^2 = 0.470 < 7.81$. Thus, although the agreement is good, the results obtained are subject to a reasonable amount of sampling error.

- 12.7** An urn contains a very large number of marbles of four different colors: red, orange, yellow, and green. A sample of 12 marbles drawn at random from the urn revealed 2 red, 5 orange, 4 yellow, and 1 green marble. Test the hypothesis that the urn contains equal proportions of the differently colored marbles.

SOLUTION

Under the hypothesis that the urn contains equal proportions of the differently colored marbles, we would expect 3 of each kind in a sample of 12 marbles. Since these expected numbers are less than 5,

the chi-square approximation will be in error. To avoid this, we combine categories so that the expected number in each category is at least 5.

If we wish to reject the hypothesis, we should combine categories in such a way that the evidence against the hypothesis shows up best. This is achieved in our case by considering the categories “red or green” and “orange or yellow,” for which the sample revealed 3 and 9 marbles, respectively. Since the expected number in each category under the hypothesis of equal proportions is 6, we have

$$\chi^2 = \frac{(3-6)^2}{6} + \frac{(9-6)^2}{6} = 3$$

For $\nu = 2 - 1 = 1$, $\chi_{.95}^2 = 3.84$. Thus we cannot reject the hypothesis at the 0.05 significance level (although we can at the 0.10 level). Conceivably the observed results could arise on the basis of chance even when equal proportions of the colors are present.

Another method

Using Yates' correction, we find

$$\chi^2 = \frac{(|3-6|-0.5)^2}{6} + \frac{(|9-6|-0.5)^2}{6} = \frac{(2.5)^2}{6} + \frac{(2.5)^2}{6} = 2.1$$

which leads to the same conclusion given above. This is to be expected, of course, since Yates' correction always *reduces* the value of χ^2 .

It should be noted that if the χ^2 approximation is used despite the fact that the frequencies are too small, we would obtain

$$\chi^2 = \frac{(2-3)^2}{3} + \frac{(5-3)^2}{3} + \frac{(4-3)^2}{3} + \frac{(1-3)^2}{3} = 3.33$$

Since for $\nu = 4 - 1 = 3$, $\chi_{.95}^2 = 7.81$, we would arrive at the same conclusions as above. Unfortunately, the χ^2 approximation for small frequencies is poor; hence, when it is not advisable to combine frequencies, we must resort to the exact probability methods of Chapter 6.

- 12.8** In 360 tosses of a pair of dice, 74 sevens and 24 elevens are observed. Using the 0.05 significance level, test the hypothesis that the dice are fair.

SOLUTION

A pair of dice can fall 36 ways. A seven can occur in 6 ways, an eleven in 2 ways. Then $\Pr\{\text{seven}\} = \frac{6}{36} = \frac{1}{6}$ and $\Pr\{\text{eleven}\} = \frac{2}{36} = \frac{1}{18}$. Thus in 360 tosses we would expect $\frac{1}{6}(360) = 60$ sevens and $\frac{1}{18}(360) = 20$ elevens, so that

$$\chi^2 = \frac{(74-60)^2}{60} + \frac{(24-20)^2}{20} = 4.07$$

For $\nu = 2 - 1 = 1$, $\chi_{.95}^2 = 3.84$. Thus, since $4.07 > 3.84$, we would be inclined to reject the hypothesis that the dice are fair. Using Yates' correction, however, we find

$$\chi^2 \text{ (corrected)} = \frac{(|74-60|-0.5)^2}{60} + \frac{(|24-20|-0.5)^2}{20} = \frac{(13.5)^2}{60} + \frac{(3.5)^2}{20} = 3.65$$

Thus on the basis of the corrected χ^2 we could not reject the hypothesis at the 0.05 level.

In general, for large samples such as we have here, results using Yates' correction prove to be more reliable than uncorrected results. However, since even the corrected value of χ^2 lies so close to the critical value, we are hesitant about making decisions one way or the other. In such cases it is perhaps best to increase the sample size by taking more observations if we are interested especially in the 0.05 level for some reason; otherwise, we could reject the hypothesis at some other level (such as 0.10) if this is satisfactory.

- 12.9** A survey of 320 families with 5 children revealed the distribution shown in Table 12.10. Is the result consistent with the hypothesis that male and female births are equally probable?

Table 12.10

Number of boys and girls	5 boys 0 girls	4 boys 1 girl	3 boys 2 girls	2 boys 3 girls	1 boy 4 girls	0 boys 5 girls	Total
Number of families	18	56	110	88	40	8	320

SOLUTION

Let p = probability of a male birth, and let $q = 1 - p$ = probability of a female birth. Then the probabilities of (5 boys), (4 boys and 1 girl), . . . , (5 girls) are given by the terms in the binomial expansion

$$(p + q)^5 = p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5$$

If $p = q = \frac{1}{2}$, we have

$$\begin{aligned} \Pr\{5 \text{ boys and } 0 \text{ girls}\} &= \left(\frac{1}{2}\right)^5 = \frac{1}{32} & \Pr\{2 \text{ boys and } 3 \text{ girls}\} &= 10\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 = \frac{10}{32} \\ \Pr\{4 \text{ boys and } 1 \text{ girl}\} &= 5\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right) = \frac{5}{32} & \Pr\{1 \text{ boy and } 4 \text{ girls}\} &= 5\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^4 = \frac{5}{32} \\ \Pr\{3 \text{ boys and } 2 \text{ girls}\} &= 10\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 = \frac{10}{32} & \Pr\{0 \text{ boys and } 5 \text{ girls}\} &= \left(\frac{1}{2}\right)^5 = \frac{1}{32} \end{aligned}$$

Then the expected number of families with 5, 4, 3, 2, 1, and 0 boys are obtained by multiplying the above probabilities by 320, and the results are 10, 50, 100, 100, 50, and 10, respectively. Hence

$$\chi^2 = \frac{(18 - 10)^2}{10} + \frac{(56 - 50)^2}{50} + \frac{(110 - 100)^2}{100} + \frac{(88 - 100)^2}{100} + \frac{(40 - 50)^2}{50} + \frac{(8 - 10)^2}{10} = 12.0$$

Since $\chi_{.95}^2 = 11.1$ and $\chi_{.99}^2 = 15.1$ for $\nu = 6 - 1 = 5$ degrees of freedom, we can reject the hypothesis at the 0.05 but not at the 0.01 significance level. Thus we conclude that the results are probably significant and male and female births are not equally probable.

12.10 In a survey of 500 individuals, it was found that 155 of the 500 rented at least one video from a video rental store during the past week. Test the hypothesis that 25% of the population rented at least one video during the past week using a two-tailed alternative and $\alpha = 0.05$. Perform the test using both the standard normal distribution and the chi-square distribution. Show that the chi-square test involving only two categories is equivalent to the significance test for proportions given in Chapter 10.

SOLUTION

If the null hypothesis is true, then $\mu = Np = 500(0.25) = 125$ and $\sigma = \sqrt{Npq} = \sqrt{500(0.25)(0.75)} = 9.68$. The computed test statistic is $Z = (155 - 125)/9.68 = 3.10$. The critical values are ± 1.96 , and the null hypothesis is rejected.

The solution using the chi-square distribution is found by using the results as displayed in Table 12.11.

Table 12.11

Frequency	Rented Video	Did Not Rent Video	Total
Observed	155	345	500
Expected	125	375	500

The computed chi-square statistic is determined as follows:

$$\chi^2 = \frac{(155 - 125)^2}{125} + \frac{(345 - 375)^2}{375} = 9.6$$

The critical value for one degree of freedom is 3.84, and the null hypothesis is rejected. Note that $(3.10)^2 = 9.6$ and $(\pm 1.96)^2 = 3.84$ or $Z^2 = \chi^2$. The two procedures are equivalent.

12.11 (a) Prove that formula (1) of this chapter can be written

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N$$

(b) Use the result of part (a) to verify the value of χ^2 computed in Problem 12.6.

SOLUTION

(a) By definition,

$$\begin{aligned} \chi^2 &= \sum \frac{(o_j - e_j)^2}{e_j} = \sum \left(\frac{o_j^2 - 2o_j e_j + e_j^2}{e_j} \right) \\ &= \sum \frac{o_j^2}{e_j} - 2 \sum o_j + \sum e_j = \sum \frac{o_j^2}{e_j} - 2N + N = \sum \frac{o_j^2}{e_j} - N \end{aligned}$$

where formula (2) of this chapter has been used.

$$(b) \quad \chi^2 = \sum \frac{o_j^2}{e_j} - N = \frac{(315)^2}{312.75} + \frac{(108)^2}{104.25} + \frac{(101)^2}{104.25} + \frac{(32)^2}{34.75} - 556 = 0.470$$

GOODNESS OF FIT

12.12 A racquetball player plays 3 game sets for exercise several times over the years and keeps records of how he does in the three game sets. For 250 days, his records show that he wins 0 games on 25 days, 1 game on 75 days, 2 games on 125 days, and wins all 3 games on 25 days. Test that X = the number of wins in a 3 game series is binomial distributed at $\alpha = 0.05$.

SOLUTION

The mean number of wins in 3 game sets is $(0 \times 25 + 1 \times 75 + 2 \times 125 + 3 \times 25)/250 = 1.6$. If X is binomial, the mean is $np = 3p$ which is set equal to the statistic 1.6 and solving for p , we find that $p = 0.53$. We wish to test that X is binomial with $n = 3$ and $p = 0.53$. If X is binomial with $p = 0.53$, the distribution of X and the expected number of wins is shown in the following EXCEL output. Note that the binomial probabilities, $p(x)$ are found by entering =BINOMDIST(A2, 3, 0.53, 0) and performing a click-and-drag from B2 to B5. This gives the values shown under $p(x)$.

x	$p(x)$	expected wins	observed wins
0	0.103823	25.95575	25
1	0.351231	87.80775	75
2	0.396069	99.01725	125
3	0.148877	37.21925	25

The expected wins are found by multiplying the $p(x)$ values by 250.

$$\chi^2 = \frac{(25 - 30.0)^2}{30.0} + \frac{(75 - 87.8)^2}{87.8} + \frac{(125 - 99.0)^2}{99.0} + \frac{(25 - 37.2)^2}{37.2} = 12.73.$$

Since the number of parameters used in estimating the expected frequencies is $m = 1$ (namely, the parameter p of the binomial distribution), $v = k - 1 - m = 4 - 1 - 1 = 2$. The p -value is given by the EXCEL expression =CHIDIST(12.73, 2) = 0.0017 and the hypothesis that the variable X is binomial distributed is rejected.

12.13 The number of hours per week that 200 college students spend on the Internet is grouped into the classes 0 to 3, 4 to 7, 8 to 11, 12 to 15, 16 to 19, 20 to 23, and 24 to 27 with the observed frequencies 12, 25, 36, 45, 34, 31, and 17. The grouped mean and the grouped standard deviation

are found from the data. The null hypothesis is that the data are normally distributed. Using the mean and the standard deviation that are found from the grouped data and assuming a normal distribution, the expected frequencies are found, after rounding off, to be the following: 10, 30, 40, 50, 36, 28, and 6.

- (a) Find χ^2 .
- (b) How many degrees of freedom does χ^2 have?
- (c) Use EXCEL to find the 5% critical value and give your conclusion at 5%.
- (d) Use EXCEL to find the p -value for your result.

SOLUTION

- (a) A portion of the EXCEL worksheet is shown in Fig. 12-1. $=(A2-B2)^2/B2$ is entered into C2 and a click-and-drag is executed from C2 to C8. $=SUM(C2:C8)$ is entered into C9. We see that $\chi^2 = 22.7325$.

	A	B	C	D
1	observed	expected	$(O - E)^2/E$	
2	12	10	0.4	
3	25	30	0.8333333	
4	36	40	0.4	
5	45	50	0.5	
6	34	36	0.1111111	
7	31	28	0.3214286	
8	17	6	20.166667	
9			22.73254	
10				

Fig. 12-1 Portion of EXCEL worksheet for Problem 12.13.

- (b) Since the number of parameters used in estimating the expected frequencies is $m = 2$ (namely, the mean μ and the standard deviation σ of the normal distribution), $v = k - 1 - m = 7 - 1 - 2 = 4$. Note that no classes needed to be combined, since the expected frequencies all exceeded 5.
- (c) The 5% critical value is given by $=CHIINV(0.05, 4)$ or 9.4877. Reject the null hypothesis that the data came from a normal distribution since 22.73 exceeds the critical value.
- (d) The p -value is given by $=CHIDIST(22.7325, 4)$ or we have p -value = 0.000143.

CONTINGENCY TABLES

12.14 Work Problem 10.20 by using the chi-square test. Also work using MINITAB and compare the two solutions.

SOLUTION

The conditions of the problem are presented in Table 12.12(a). Under the null hypothesis H_0 that the serum has no effect, we would expect 70 people in each of the groups to recover and 30 in each group not to recover, as shown in Table 12.12(b). Note that H_0 is equivalent to the statement that recovery is *independent* of the use of the serum (i.e., the classifications are independent).

Table 12.12(a) Frequencies Observed

	Recover	Do Not Recover	Total
Group A (using serum)	75	25	100
Group B (not using serum)	65	35	100
Total	140	60	200

Table 12.12(b) Frequencies Expected under H_0

	Recover	Do Not Recover	Total
Group A (using serum)	70	30	100
Group B (not using serum)	70	30	100
Total	140	60	200

$$\chi^2 = \frac{(75 - 70)^2}{70} + \frac{(65 - 70)^2}{70} + \frac{(25 - 30)^2}{30} + \frac{(35 - 30)^2}{30} = 2.38$$

To determine the number of degrees of freedom, consider Table 12.13, which is the same as Table 12.12 except that only the totals are shown. It is clear that we have the freedom of placing only one number in any of the four empty cells, since once this is done the numbers in the remaining cells are uniquely determined from the indicated totals. Thus there is 1 degree of freedom.

Table 12.13

	Recover	Do Not Recover	Total
Group A			100
Group B			100
Total	140	60	200

Another method

By formula (see Problem 12.18), $\nu = (h - 1)(k - 1) = (2 - 1)(2 - 1) = 1$. Since $\chi_{.95}^2 = 3.84$ for 1 degree of freedom and since $\chi^2 = 2.38 < 3.84$, we conclude that the results are *not significant* at the 0.05 level. We are thus unable to reject H_0 at this level, and we either conclude that the serum is not effective or withhold decision, pending further tests.

Note that $\chi^2 = 2.38$ is the square of the z score, $z = 1.54$, obtained in Problem 10.20. In general the chi-square test involving sample proportions in a 2×2 contingency table is equivalent to a test of significance of differences in proportions using the normal approximation.

Note also that a one-tailed test using χ^2 is equivalent to a two-tailed test using χ since, for example, $\chi^2 > \chi_{.95}^2$ corresponds to $\chi > \chi_{.95}$ or $\chi < -\chi_{.95}$. Since for 2×2 tables χ^2 is the square of the z score, it follows that χ is the same as z for this case. Thus a rejection of a hypothesis at the 0.05 level using χ^2 is equivalent to a rejection in a two-tailed test at the 0.10 level using z .

Chi-Square Test: Recover, Not-recover

Expected counts are printed below observed counts

Chi-Square contributions are printed below expected counts

	Recover	Not-recover	Total
1	75	25	100
	70.00	30.00	
	0.357	0.833	
2	65	35	100
	70.00	30.00	
	0.357	0.833	
Total	140	60	200

Chi-Sq=2.381, DF=1, P-Value=0.123

12.15 Work Problem 12.14 by using Yates' correction.

SOLUTION

$$\chi^2 \text{ (corrected)} = \frac{(|75 - 70| - 0.5)^2}{70} + \frac{(|65 - 70| - 0.5)^2}{70} + \frac{(|25 - 30| - 0.5)^2}{30} + \frac{(|35 - 30| - 0.5)^2}{30} = 1.93$$

Thus the conclusions reached in Problem 12.14 are valid. This could have been realized at once by noting that Yates' correction always decreases the value of χ^2 .

12.16 A cellular phone company conducts a survey to determine the ownership of cellular phones in different age groups. The results for 1000 households are shown in Table 12.14. Test the hypothesis that the proportions owning cellular phones are the same for the different age groups.

Table 12.14

Cellular phone	18–24	25–54	55–64	≥ 65	Total
Yes	50	80	70	50	250
No	200	170	180	200	750
Total	250	250	250	250	1000

SOLUTION

Under the hypothesis H_0 that the proportions owning cellular phones are the same for the different age groups, $250/1000 = 25\%$ is an estimate of the percentage owning a cellular phone in each age group, and 75% is an estimate of the percent not owning a cellular phone in each age group. The frequencies expected under H_0 are shown in Table 12.15.

The computed value of the chi-square statistic can be found as illustrated in Table 12.16.

The degrees of freedom for the chi-square distribution is $\nu = (h - 1)(k - 1) = (2 - 1)(4 - 1) = 3$. Since $\chi^2_{.95} = 7.81$, and 14.3 exceeds 7.81, we reject the null hypothesis and conclude that the percentages are not the same for the four age groups.

Table 12.15

Cellular phone	18–24	25–54	55–64	≥ 65	Total
Yes	25% of 250 = 62.5	25% of 250 = 62.5	25% of 250 = 62.5	25% of 250 = 62.5	250
No	75% of 250 = 187.5	75% of 250 = 187.5	75% of 250 = 187.5	75% of 250 = 187.5	750
Total	250	250	250	250	1000

Table 12.16

Row, column	o	e	$(o - e)$	$(o - e)^2$	$(o - e)^2/e$
1, 1	50	62.5	-12.5	156.25	2.5
1, 2	80	62.5	17.5	306.25	4.9
1, 3	70	62.5	7.5	56.25	0.9
1, 4	50	62.5	-12.5	156.25	2.5
2, 1	200	187.5	12.5	156.25	0.8
2, 2	170	187.5	-17.5	306.25	1.6
2, 3	180	187.5	-7.5	56.25	0.3
2, 4	200	187.5	12.5	156.25	0.8
Sum	1000	1000	0		14.3

12.17 Use MINITAB to solve Problem 12.16.**SOLUTION**

The MINITAB solution to Problem 12.16 is shown below. The observed and the expected counts are shown along with the computation of the test statistic. Note that the null hypothesis would be rejected for any level of significance exceeding 0.002.

Data Display

Row	18-24	25-54	55-64	65 or more
1	50	80	70	50
2	200	170	180	200

MTB > chisquare c1-c4

Chi-Square Test

Expected counts are printed below observed counts

	18-24	25-54	55-64	65 or mo	Total
1	50	80	70	50	250
	62.50	62.50	62.50	62.50	
2	200	170	180	200	750
	187.50	187.50	187.50	187.50	
Total	250	250	250	250	1000
Chi-Sq =	2.500 +	4.900 +	0.900 +	2.500 +	
	0.833 +	1.633 +	0.300 +	0.833 =	14.400

DF = 3, P-Value = 0.002

12.18 Show that for an $h \times k$ contingency table the number of degrees of freedom is $(h - 1) \times (k - 1)$, where $h > 1$ and $k > 1$.**SOLUTION**

In a table with h rows and k columns, we can leave out a single number in each row and column, since such numbers can easily be restored from a knowledge of the totals of each column and row. It follows that we have the freedom of placing only $(h - 1)(k - 1)$ numbers into the table, the others being then automatically determined uniquely. Thus the number of degrees of freedom is $(h - 1)(k - 1)$. Note that this result holds if the population parameters needed in obtaining the expected frequencies are known.

12.19 (a) Prove that for the 2×2 contingency table shown in Table 12.17(a),

$$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{N_1N_2N_A N_B}$$

(b) Illustrate the result in part (a) with reference to the data of Problem 12.14.

Table 12.17(a) Results Observed

	I	II	Total
A	a_1	a_2	N_A
B	b_1	b_2	N_B
Total	N_1	N_2	N

Table 12.17(b) Results Expected

	I	II	Total
A	N_1N_A/N	N_2N_A/N	N_A
B	N_1N_B/N	N_2N_B/N	N_B
Total	N_1	N_2	N

SOLUTION

(a) As in Problem 12.14, the results expected under a null hypothesis are shown in Table 12.17(b). Then

$$\chi^2 = \frac{(a_1 - N_1N_A/N)^2}{N_1N_A/N} + \frac{(a_2 - N_2N_A/N)^2}{N_2N_A/N} + \frac{(b_1 - N_1N_B/N)^2}{N_1N_B/N} + \frac{(b_2 - N_2N_B/N)^2}{N_2N_B/N}$$

But
$$a_1 - \frac{N_1N_A}{N} = a_1 - \frac{(a_1 + b_1)(a_1 + a_2)}{a_1 + b_1 + a_2 + b_2} = \frac{a_1b_2 - a_2b_1}{N}$$

Similarly
$$a_2 - \frac{N_2N_A}{N} \quad \text{and} \quad b_1 - \frac{N_1N_B}{N} \quad \text{and} \quad b_2 - \frac{N_2N_B}{N}$$

are also equal to
$$\frac{a_1b_2 - a_2b_1}{N}$$

Thus we can write

$$\begin{aligned} \chi^2 &= \frac{N}{N_1N_A} \left(\frac{a_1b_2 - a_2b_1}{N} \right)^2 + \frac{N}{N_2N_A} \left(\frac{a_1b_2 - a_2b_1}{N} \right)^2 \\ &\quad + \frac{N}{N_1N_B} \left(\frac{a_1b_2 - a_2b_1}{N} \right)^2 + \frac{N}{N_2N_B} \left(\frac{a_1b_2 - a_2b_1}{N} \right)^2 \end{aligned}$$

which simplifies to
$$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{N_1N_2N_AN_B}$$

(b) In Problem 12.14, $a_1 = 75$, $a_2 = 25$, $b_1 = 65$, $b_2 = 35$, $N_1 = 140$, $N_2 = 60$, $N_A = 100$, $N_B = 100$, and $N = 200$; then, as obtained before,

$$\chi^2 = \frac{200[(75)(35) - (25)(65)]^2}{(140)(60)(100)(100)} = 2.38$$

Using Yates' correction, the result is the same as in Problem 12.15:

$$\chi^2 \text{ (corrected)} = \frac{N(|a_1b_2 - a_2b_1| - \frac{1}{2}N)^2}{N_1N_2N_AN_B} = \frac{200[|(75)(35) - (25)(65)| - 100]^2}{(140)(60)(100)(100)} = 1.93$$

12.20 Nine hundred males and 900 females were asked whether they would prefer more federal programs to assist with childcare. Forty percent of the females and 36 percent of the males responded yes. Test the null hypothesis of equal percentages versus the alternative hypothesis of unequal percentages at $\alpha = 0.05$. Show that a chi-square test involving two sample proportions is equivalent to a significance test of differences using the normal approximation of Chapter 10.

SOLUTION

Under hypothesis H_0 ,

$$\mu_{P_1 - P_2} = 0 \text{ and } \sigma_{P_1 - P_2} = \sqrt{pq \left(\frac{1}{N_1} + \frac{1}{N_2} \right)} = \sqrt{(0.38)(0.62) \left(\frac{1}{900} + \frac{1}{900} \right)} = 0.0229$$

where p is estimated by pooling the proportions in the two samples. That is

$$p = \frac{360 + 324}{900 + 900} = 0.38 \quad \text{and} \quad q = 1 - 0.38 = 0.62$$

The normal approximation test statistic is as follows:

$$Z = \frac{P_1 - P_2}{\sigma_{P_1 - P_2}} = \frac{0.40 - 0.36}{0.0229} = 1.7467$$

The MINITAB solution for the chi-square analysis is as follows.

Chi-Square Test

Expected counts are printed below observed counts

	males	females	Total
1	324 342.00	360 342.00	684
2	576 558.00	549 558.00	1116
Total	900	900	1800

Chi-Sq = 0.947 + 0.947 +
0.581 + 0.581 = 3.056
DF = 1, P-Value = 0.080

The square of the normal test statistic is $(1.7467)^2 = 3.056$, the value for the chi-square statistic. The two tests are equivalent. The p -values are always the same for the two tests.

COEFFICIENT OF CONTINGENCY

12.21 Find the coefficient of contingency for the data in the contingency table of Problem 12.14.

SOLUTION

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{2.38}{2.38 + 200}} = \sqrt{0.01176} = 0.1084$$

12.22 Find the maximum value of C for the 2×2 table of Problem 12.14.

SOLUTION

The maximum value of C occurs when the two classifications are perfectly dependent or associated. In such case all those who take the serum will recover, and all those who do not take the serum will not recover. The contingency table then appears as in Table 12.18.

Table 12.18

	Recover	Do Not Recover	Total
Group A (using serum)	100	0	100
Group B (not using serum)	0	100	100
Total	100	100	200

Since the expected cell frequencies, assuming complete independence, are all equal to 50,

$$\chi^2 = \frac{(100 - 50)^2}{50} + \frac{(0 - 50)^2}{50} + \frac{(0 - 50)^2}{50} + \frac{(100 - 50)^2}{50} = 200$$

Thus the maximum value of C is $\sqrt{\chi^2/(\chi^2 + N)} = \sqrt{200/(200 + 200)} = 0.7071$.