CHAPTER 12

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 $E_1, E_2, E_3, \ldots, E_k$ (see Table 12.1) are observed to occur with frequencies $o_1, o_2, o_3, \ldots, o_k$, called *observed frequencies*, and that according to probability rules they are expected to occur with frequencies $e_1, e_2, e_3, \ldots, e_k$, 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	$\cdots$	E_k
Observed frequency	o_1	o_2	o_3	$\cdots$	o_k
Expected frequency	e_1	e_2	e_3	$\cdots$	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

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

where if the total frequency is N,

$$\sum o_j = \sum e_j = N \tag{2}$$

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

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N \tag{3}$$

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

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

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 jth 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

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

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

$$\chi^2 = \frac{N(a_1 b_2 - a_2 b_1)^2}{(a_1 + b_1)(a_2 + b_2)(a_1 + a_2)(b_1 + b_2)} = \frac{N\Delta^2}{N_1 N_2 N_A N_B} \tag{7}$$

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

$$\chi^2 \text{ (corrected)} = \frac{N(|a_1b_2 - a_2b_1| - \frac{1}{2}N)^2}{(a_1 + b_1)(a_2 + b_2)(a_1 + a_2)(b_1 + b_2)} = \frac{N(|\Delta| - \frac{1}{2}N)^2}{N_1N_2N_AN_B} \tag{8}$$

2 × 3 Tables

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

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

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N \tag{10}$$

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

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \tag{11}$$

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

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

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 $\chi_1^2, \chi_2^2, \chi_3^2, \ldots$ with $\nu_1, \nu_2, \nu_3, \ldots$ degrees of freedom, respectively. Then the result of all these experiments can be considered equivalent to a χ^2 value given by $\chi_1^2 + \chi_2^2 + \chi_3^2 + \cdots$ with $\nu_1 + \nu_2 + \nu_3 + \cdots$ 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

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

$$= \frac{(14.5)^2}{100} + \frac{(14.5)^2}{100} = 4.205$$

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.