Hypothesis Testing: Categorical Data

Solved Problems

10.1 Introduction

If the variable under study is not continuous but is instead classified into categories, which may or may not be ordered, then different methods of inference should be used. Therefore, in this chapter (chapter 10), we study the most widely used techniques for analyzing qualitative (or categorical) data as follows:

- First, the problem of how to compare binomial proportions from two independent samples was studied. For the large-sample case, this problem was solved in two different (but equivalent) ways:
  (a) Using the two-sample test for binomial proportions.
  (b) Using the chi-square test for 2 × 2 contingency tables.

- Second, a chi-square test for R × C contingency tables was developed, which is a direct generalization of the 2 × 2 contingency-table test to investigate the relationship between two qualitative variables, in which one or both variables have more than two possible categories of response.

- Finally, we studied how to assess the goodness-of-fit of probability models using the chi-square goodness-of-fit test.

10.2 Two-Sample Test for Binomial Proportions

In this section, we discuss the problem of testing for some constant p (*population proportion*) for a two-sample (*independent samples*) problem comparing two binomial proportions p_1 and p_2 the following hypothesis:

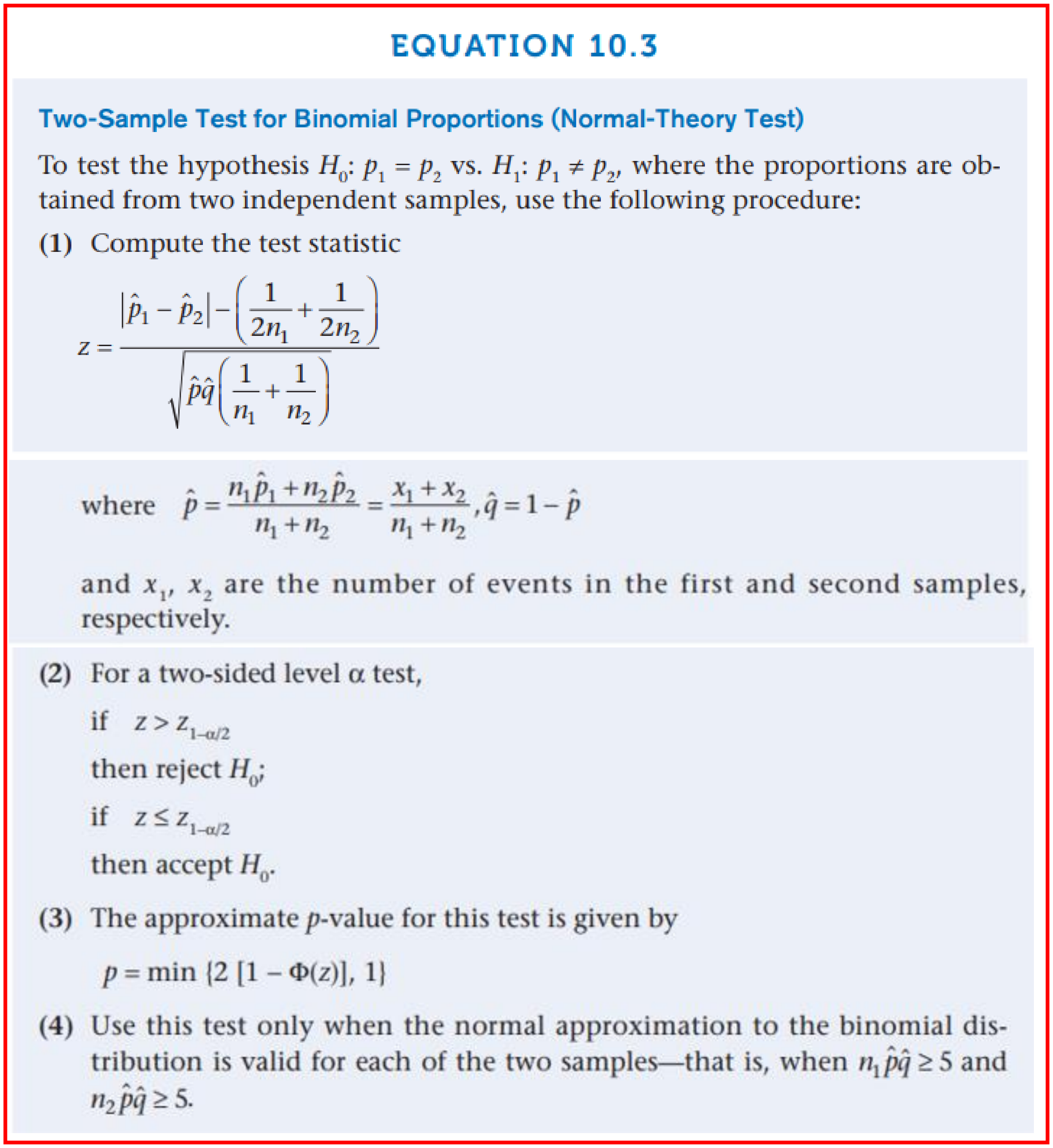
$$\boxed{H_0\text{: } p_1 = p_2 = p \text{ vs. } H_1\text{: } p_1 \neq p_2}$$

Two *equivalent* approaches for testing the hypothesis are presented as follows:

- ➢ The first approach uses normal-theory methods similar to those developed in Chapter 8.
- ➢ The second approach uses contingency-table methods.

10.2.1 Normal-Theory Method

In this section, the following test procedure for a two-sample problem comparing two binomial proportions is suggested:



EQUATION 10.3

Two-Sample Test for Binomial Proportions (Normal-Theory Test)

To test the hypothesis H_0: $p_1 = p_2$ vs. H_1: $p_1 \neq p_2$, where the proportions are obtained from two independent samples, use the following procedure:

(1) Compute the test statistic

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}, \hat{q} = 1 - \hat{p}$

and x_1, x_2 are the number of events in the first and second samples, respectively.

(2) For a two-sided level α test,

if $z > z_{1-\alpha/2}$

then reject H_0;

if $z \leq z_{1-\alpha/2}$

then accept H_0.

(3) The approximate p-value for this test is given by

$$p = \min\{2[1 - \Phi(z)], 1\}$$

(4) Use this test only when the normal approximation to the binomial distribution is valid for each of the two samples—that is, when $n_1\hat{p}\hat{q} \geq 5$ and $n_2\hat{p}\hat{q} \geq 5$.

Example

A sample of 50 randomly selected men with high triglyceride levels consumed 2 tablespoons of oat bran daily for six weeks. After six weeks, 60% of the men had lowered their triglyceride level. A sample of 80 men consumed 2 tablespoons of wheat bran for six weeks. After six weeks, 25% had lower triglyceride levels. Is there a significance difference in the two proportions at $\alpha = 0.01$?

Solution

Since the statistics are given in percentages (proportions) then we have to use a two-sample problem comparing two binomial proportions as follows:

Let

p₁ = proportion of men consumed 2 tablespoons daily of oat bran who had lowered their triglyceride level after six weeks.

p₂ = proportion of men consumed 2 tablespoons daily of wheat bran who had lowered their triglyceride level after six weeks.

Step (1): Sample Proportions

- Sample proportion of men consumed 2 tablespoons daily of oat bran is:

$$\hat{p}_1 = 60\% = 0.60$$

- Sample proportion of men consumed 2 tablespoons daily of wheat bran is:

$$\hat{p}_2 = 25\% = 0.25$$

Step (2): In order to compute $\hat{p}$, we must find x_1 and x_2 as follows:

$$x_1 = n_1 * \hat{p}_1 = (50)(0.60) = 30$$
$$x_2 = n_2 * \hat{p}_2 = (80)(0.25) = 20$$

Step (3): Estimated common proportions $\hat{p}$ and $\hat{q}$ are obtained as follows:

$\hat{p} = (30 + 20) / (50 + 80) = 50/130 = 0.385$

$\hat{q} = 1 - \hat{p} = 1 - 0.385 = 0.615$

Step (4): Hypotheses to be tested are:

$$H_0: p_1 = p_2 \text{ vs. } H_1: p_1 \neq p_2$$

Step (5): Compute the Test Statistic (Z)

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\mathrm{Z} = \frac{|0.60 - 0.25| - \left(\frac{1}{2(50)} + \frac{1}{2(80)}\right)}{\sqrt{(0.385)(0.615)(\frac{1}{50} + \frac{1}{80})}} = \frac{0.33375}{0.08772} = 3.80$$

Step (6): Critical Value

$$Z_{1-(\alpha/2)} = Z_{1-(0.01/2)} = Z_{0.995} = 2.575 \approx 2.58$$

Step (7): Decision

Now by using the critical value method, we get Z = 3.80 > $Z_{0.995} = 2.58$, then the decision will be reject H$_0$ and accept H$_1$ at level of significance $\alpha = 0.01$.

Conclusion

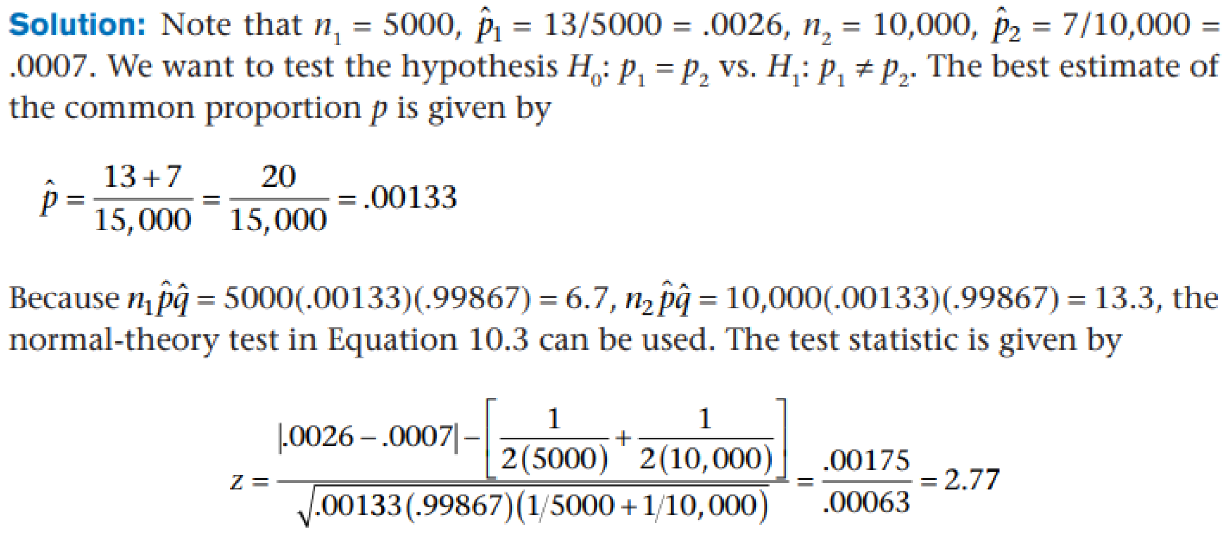
The results are highly significant. Therefore, we can conclude that there is enough evidence to support the claim that there is a difference in proportions.

Notations

- $n_1\hat{p}\hat{q} =$ (50)(0.385)(0.615) = 11.839 > 5
- $n_2\hat{p}\hat{q} =$ (80)(0.385)(0.615) = 18.942 > 5
- The p-value = 2 × [1 − Φ (3.80)]= 2 × [1 – 0.9999] = 0.0001 < 0.05

EXAMPLE 10.6

Cardiovascular Disease A study looked at the effects of OC use on heart disease in women 40 to 44 years of age. The researchers found that among 5000 current OC users at baseline, 13 women developed a myocardial infarction (MI) over a 3-year period, whereas among 10,000 never-OC users, 7 developed an MI over a 3-year period. Assess the statistical significance of the results? Use $\alpha = 0.05$?



Solution: Note that $n_1 = 5000$, $\hat{p}_1 = 13/5000 = .0026$, $n_2 = 10{,}000$, $\hat{p}_2 = 7/10{,}000 = .0007$. We want to test the hypothesis H_0: $p_1 = p_2$ vs. H_1: $p_1 \neq p_2$. The best estimate of the common proportion p is given by

$$\hat{p} = \frac{13+7}{15{,}000} = \frac{20}{15{,}000} = .00133$$

Because $n_1\hat{p}\hat{q} = 5000(.00133)(.99867) = 6.7$, $n_2\hat{p}\hat{q} = 10{,}000(.00133)(.99867) = 13.3$, the normal-theory test in Equation 10.3 can be used. The test statistic is given by

$$z = \frac{|.0026 - .0007| - \left[\frac{1}{2(5000)} + \frac{1}{2(10{,}000)}\right]}{\sqrt{.00133(.99867)(1/5000 + 1/10{,}000)}} = \frac{.00175}{.00063} = 2.77$$

The p-value is given by 2 × [1 − Φ(2.77)] =2 × [1 – 0.9972] = 0.0056 ≈ 0.006.

Decision

Now by using the p-value method, we get p-value = 0.006 < $\alpha = 0.05$, then the decision will be reject H_0 and accept H_1 at level of significance $\alpha = 0.05$.

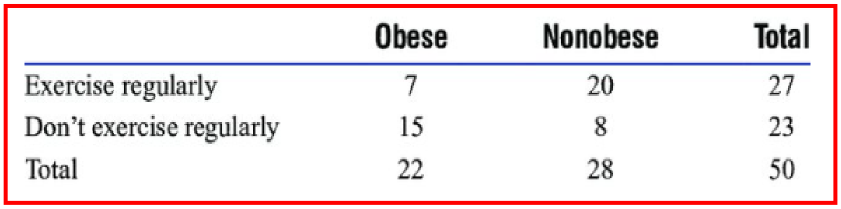
Conclusion

There is a highly significant difference between MI incidence rates for current OC users versus never-OC users. In other words, OC use is significantly associated with higher MI incidence over a 3-year period.

10.2.2 Contingency-Table Method

DEFINITION 10.1

A **2 × 2 contingency table** is a table composed of two rows cross-classified by two columns and it is an appropriate way to display data that can be classified by two different variables, each of which has only two possible outcomes. One variable is arbitrarily assigned to the rows and the other to the columns. Each of the four cells represents the number of units (*frequencies*), with a specific value for each of the two variables.



	Obese	Nonobese	Total
Exercise regularly	7	20	27
Don't exercise regularly	15	8	23
Total	22	28	50

Notations

(1) The cells are sometimes referred to by number as follows:

- The (1, 1) cell being the cell in the first row and first column.
- The (1, 2) cell being the cell in the first row and second column.
- The (2, 1) cell being the cell in the second row and first column.
- The (2, 2) cell being the cell in the second row and second column.

(2) The observed number of units in the four cells are likewise referred to as O_{11}, O_{12}, O_{21}, and O_{22}, respectively.

(3) Furthermore, it is customary to total

- The number of units in each row and display them in the right margins, which are called **row marginal totals** or **row margins**.
- The number of units in each column and display them in the bottom margins, which are called **column marginal totals** or **column margins**.

- The total number of units in the four cells, which is displayed in the lower right hand corner of the table and is called the **grand total**.

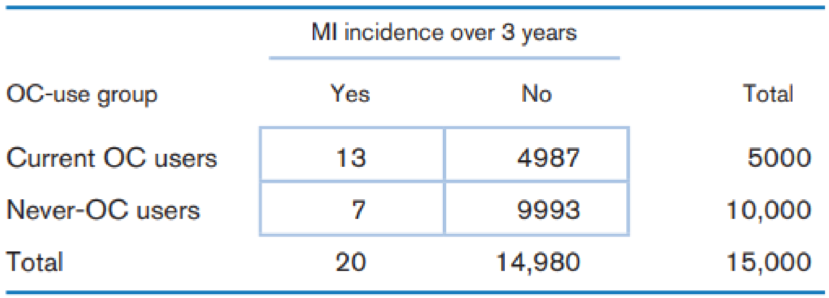
EXAMPLE 10.8

Cardiovascular Disease A study looked at the effects of OC use on heart disease in women 40 to 44 years of age. The researchers found that among 5000 current OC users at baseline, 13 women developed a myocardial infarction (MI) over a 3-year period, whereas among 10,000 never-OC users, 7 developed an MI over a 3-year period. Display the MI data in this example in the form of a 2 × 2 contingency table?

Solution

We studied 5000 current OC users, of whom 13 developed MI and 4987 did not. We studied 10,000 never-OC users, of whom 7 developed MI and 9993 did not. Thus, the contingency table should look like Table 10.2 given as follows:

TABLE 10.2 2 × 2 contingency table for the OC–MI data in Example 10.6



	MI incidence over 3 years		
OC-use group	Yes	No	Total
Current OC users	13	4987	5000
Never-OC users	7	9993	10,000
Total	20	14,980	15,000

Note that in the OC–MI data in Example 10.8 there are two independent samples of women with different contraceptive-use patterns, and we want to compare the proportion of women in each group who develop an MI. In both instances, we want to test whether the proportions are the same in the two independent samples. This test is called **a test for homogeneity of binomial proportions**. In this situation, one set of margins is fixed (e.g., the rows) and the number of successes in each row is a random variable.

Notation

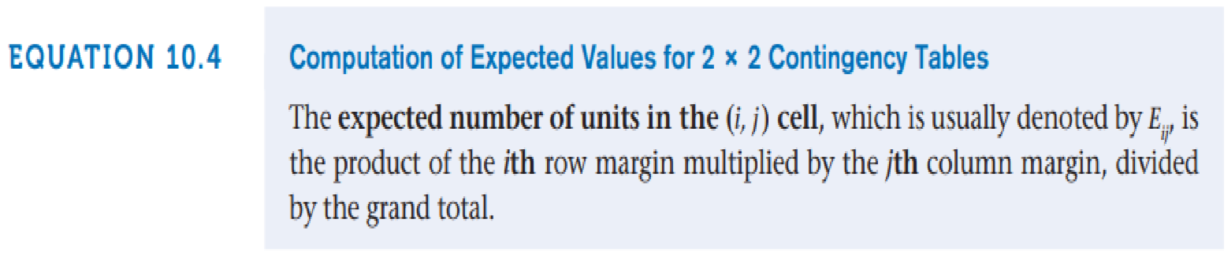
Another possible design from which contingency tables arise is in testing for the independence of two characteristics in the same sample when neither characteristic is particularly appropriate as a denominator. In this setting, both sets of margins are assumed to be fixed. The number of units in one particular cell of the table [e.g., the (1, 1) cell] is a random variable, and all other cells can be determined from the fixed margins and the (1, 1) cell. A test used in this case is called **a test of independence** or **a test of association** between the two characteristics.

Notation

The same test procedure is used whether **a test of homogeneity** or **a test of independence** is performed.

Expected Table

For a contingency table or an observed table, to determine statistical significance, we need to develop an expected table, which is the contingency table that would be expected if there were no relationship between the two variables, for example, between breast cancer and age at first birth. In general, the following rule can be applied to find the expected value:



EQUATION 10.4

Computation of Expected Values for 2 × 2 Contingency Tables

The **expected number of units in the** (i, j) **cell**, which is usually denoted by E_{ij}, is the product of the ith row margin multiplied by the jth column margin, divided by the grand total.

Example

Compute the expected table for the data given in Table 10.1 shown below:

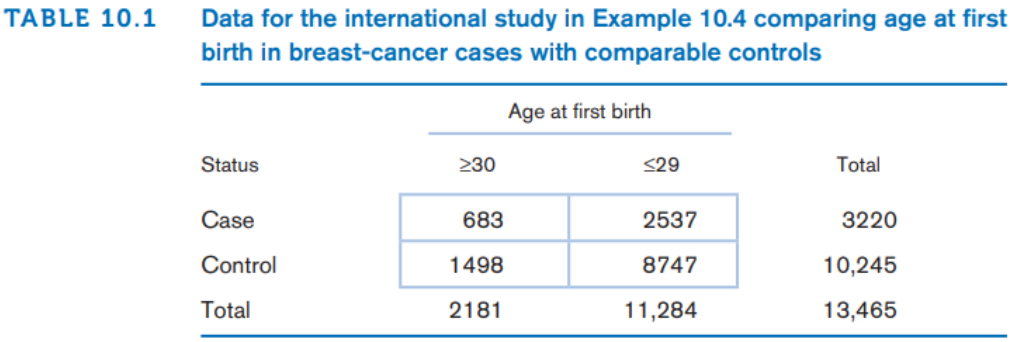


TABLE 10.1 Data for the international study in Example 10.4 comparing age at first birth in breast-cancer cases with comparable controls

	Age at first birth		
Status	≥30	≤29	Total
Case	683	2537	3220
Control	1498	8747	10,245
Total	2181	11,284	13,465

Solution

- The **row totals** are 3220 and 10245.
- The **column totals** are 2181 and 11,284.
- The **grand total** is 13465.

Thus, the four expected values can be calculated as follows:

E_{11} = expected number of units in the (1, 1) cell

= [(3220)(2181)]/[13,465]

= 521.6

E_{12} = expected number of units in the (1, 2) cell

= [(3220)(11,284)]/[13,465]

= 2698.4

E_{21} = expected number of units in the (2, 1) cell

= [(10,245)(2181)]/[13,465]

= 1659.4

E_{22} = expected number of units in the (2, 2) cell

= [(10,245(11,284)]/[13,465]

= 8585.6

Note that

$E_{11} + E_{12} + E_{21} + E_{22}$ = 521.6 + 2698.4 + 1659.4 + 8585.6 = 13,465 = **Grand Total**

These expected values are shown in Table 10.5 given below:

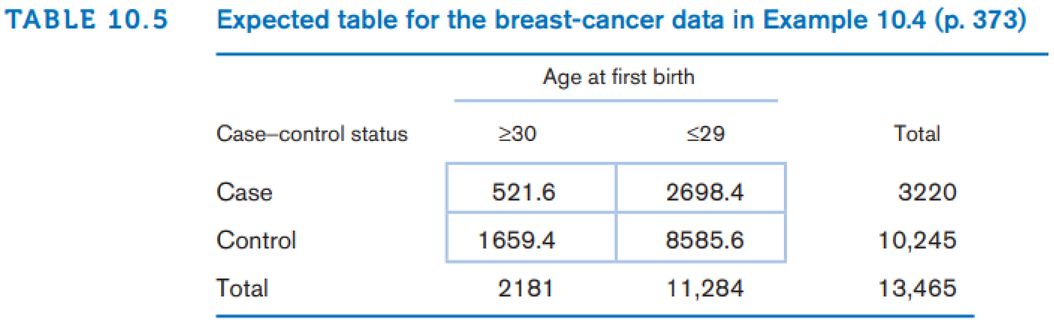


TABLE 10.5 Expected table for the breast-cancer data in Example 10.4 (p. 373)

	Age at first birth		
Case–control status	≥30	≤29	Total
Case	521.6	2698.4	3220
Control	1659.4	8585.6	10,245
Total	2181	11,284	13,465

Notation

We can show from Equation 10.4 that the total of the expected number of units in any row or column should be the same as the corresponding observed row or column total. This relationship provides a useful check that the expected values are computed correctly.

Example

Check that the expected values in Table 10.5 are computed correctly?

Solution

(1) The total of the expected values in the first row

= $E_{11} + E_{12}$ = 521.6 + 2698.4 = 3220 = First row total in the observed table.

(2) The total of the expected values in the second row

= $E_{21} + E_{22}$ = 1659.4 + 8585.6 = 10,245 = Second row total in the observed table.

(3) The total of the expected values in the first column

= $E_{11} + E_{21}$ = 521.6 + 1659.4 = 2181 = First column total in the observed table.

(4) The total of the expected values in the second column

= $E_{12} + E_{22}$ = 2698.4 + 8585.6 = 11,284 = Second column total in observed table.

Objective: We now want to compare the observed table in Table 10.1 with the expected table in Table 10.5. Then:

- If the corresponding cells in these two tables are close, then H_0 will be accepted.
- If the corresponding cells in these two tables are differ enough, then H_0 will be rejected.

Question: How should we decide how different the cells should be for us in order to reject H_0?

Answer: It can be shown that the best way of comparing the cells in the two tables is to use the statistic $(O - E)^2/E$, where O and E are the observed and expected number of units, respectively, in a particular cell. This is usually referred to as the Pearson Chi-Square Statistic.

Notations

- In particular, under H_0 it can be shown that the sum of $(O - E)^2/E$ over the 4 cells in the contingency table approximately follows a chi-square distribution with 1 degree of freedom ($df = 1$).

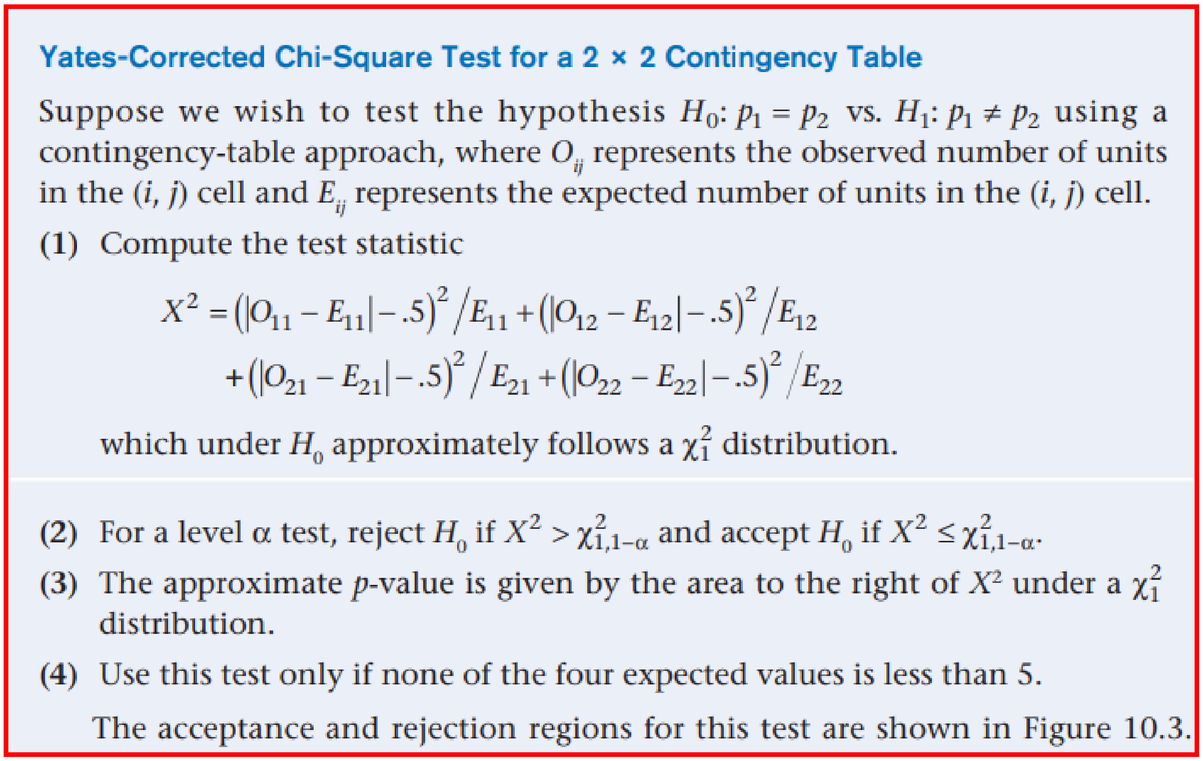
- H_0 is rejected only if this sum is large and is accepted otherwise because small values of this sum correspond to good agreement between the two tables, whereas large values correspond to poor agreement.

- This test procedure will be used only when the normal approximation to the binomial distribution is valid. In this setting the normal approximation can be shown to be **approximately true if no expected value in the table is less than 5** (sometimes known as "the rule of five").

- Under certain circumstances a version of this test statistic with a continuity correction yields more accurate p-values than does the uncorrected version when approximated by a chi-square distribution.
- For the continuity-corrected version, the statistic $\frac{\left(|O - E| - \frac{1}{2}\right)^2}{E}$ rather than $(O - E)^2/E$ is computed for each cell and the preceding expression is summed over the four cells. This test procedure is called the Yates-Corrected Chi-Square.

Yates-Corrected Chi-Square Test

EQUATION 10.5



Yates-Corrected Chi-Square Test for a 2 × 2 Contingency Table

Suppose we wish to test the hypothesis H_0: $p_1 = p_2$ vs. H_1: $p_1 \neq p_2$ using a contingency-table approach, where O_{ij} represents the observed number of units in the (i, j) cell and E_{ij} represents the expected number of units in the (i, j) cell.

(1) Compute the test statistic

$$X^2 = \left(|O_{11} - E_{11}| - .5\right)^2 / E_{11} + \left(|O_{12} - E_{12}| - .5\right)^2 / E_{12}$$
$$+ \left(|O_{21} - E_{21}| - .5\right)^2 / E_{21} + \left(|O_{22} - E_{22}| - .5\right)^2 / E_{22}$$

which under H_0 approximately follows a χ_1^2 distribution.

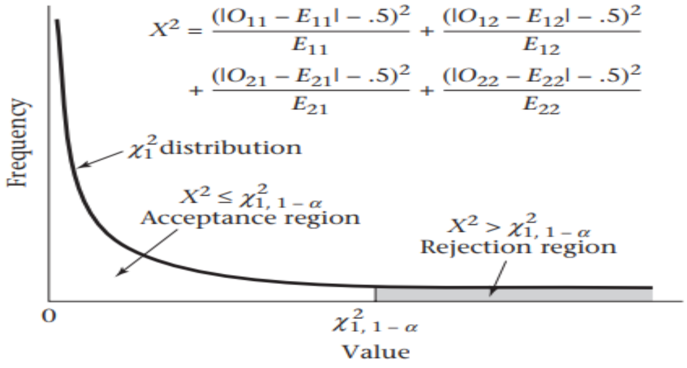
(2) For a level α test, reject H_0 if $X^2 > \chi^2_{1,1-\alpha}$ and accept H_0 if $X^2 \le \chi^2_{1,1-\alpha}$.

(3) The approximate p-value is given by the area to the right of X^2 under a χ_1^2 distribution.

(4) Use this test only if none of the four expected values is less than 5.

The acceptance and rejection regions for this test are shown in Figure 10.3.

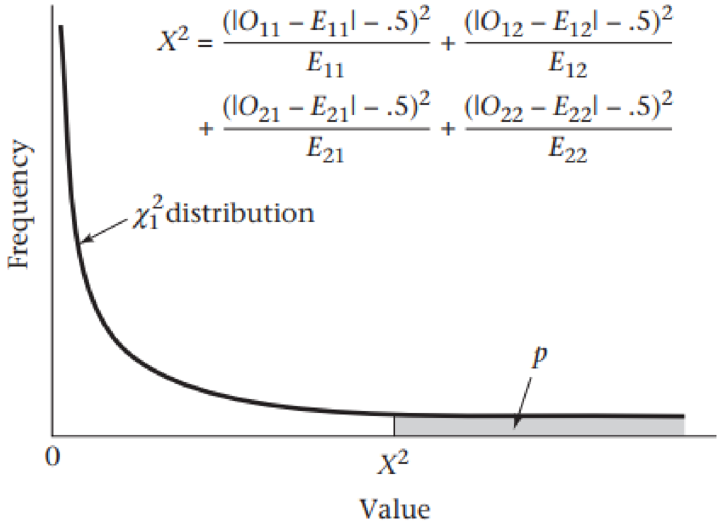
FIGURE 10.3 Acceptance and rejection regions for the Yates-corrected chi-square test for a 2 × 2 contingency table



p-value

The computation of the *p*-value is illustrated in Figure 10.4 shown below:

FIGURE 10.4 Computation of the p-value for the Yates-corrected chi-square test for a 2 × 2 contingency table



Example

A sample of 50 randomly selected men with high triglyceride levels consumed 2 tablespoons of oat bran daily for six weeks. After six weeks, 60% of the men had lowered their triglyceride level. A sample of 80 men consumed 2 tablespoons of wheat bran for six weeks. After six weeks, 25% had lower triglyceride levels. By using a 2 x 2 contingency-table approach can we conclude that there is a significance difference in the two proportions at $\alpha = 0.01$?

Solution

Step (1): First compute the observed and expected tables as given below respectively:

Observed Table

Triglyceride level	Type of consumed food for six weeks		Total
	Oat bran	Wheat bran	
Lowered	30	20	50
Non-Lowered	20	60	80
Total	50	80	130

Expected Table

Triglyceride level	Type of consumed food for six weeks		Total
	Oat bran	Wheat bran	
Lowered	19.231	30.769	50
Non-Lowered	30.769	49.231	80
Total	50	80	130

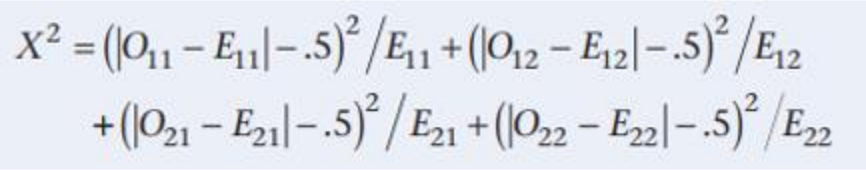
Note that the minimum expected value is 19.231, which is > 5.

Step (2): Use Table 6 (*Percentage points of the chi-square distribution*) page 880 in the Appendix to find the critical value $\chi^2_{(1,\ \ 1-\alpha)}$ as follows:

$$\chi^2_{(1,\ \ 1-\alpha)} = \chi^2_{(1,\ \ 1-0.01)} = \chi^2_{(1,\ \ 0.99)} = \ 6.63$$

Step (3): Thus, Equation 10.5, can be applied as follows:

$$X^2 = \frac{\left(|\mathrm{O}_{ij}\ -\ \mathrm{E}_{ij}| - \frac{1}{2}\right)^2}{\mathrm{E}_{ij}}$$



$$X^2 = (|O_{11} - E_{11}| - .5)^2 / E_{11} + (|O_{12} - E_{12}| - .5)^2 / E_{12} + (|O_{21} - E_{21}| - .5)^2 / E_{21} + (|O_{22} - E_{22}| - .5)^2 / E_{22}$$

$$X^2 = \frac{(|30 - 19.231| - 0.5)^2}{19.231} + \frac{(|20 - 30.769| - 0.5)^2}{30.796} + \frac{(|20 - 30.769| - 0.5)^2}{30.769} + \frac{(|60 - 49.231| - 0.5)^2}{49.231}$$
$$= 5.483 + 3.427 + 3.427 + 2.142 = 14.299 \approx 14.30$$

Step (4): Decision and Conclusion

Because we get:

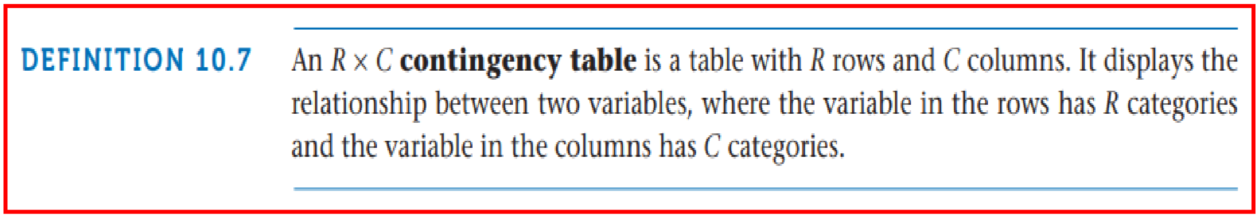
$X^2 = 14.30 > \chi^2_{(1,\ \ 0.99)} = \ 6.63$ and → p < 1 – 0.99 → p < 0.01

therefore the results are highly significant. Thus there is a significant difference between the two proportions at $\alpha = 0.01$.

10.6 R × C Contingency Tables

Tests for Association for R × C Contingency Tables

In this section of this chapter, methods of analyzing data that can be organized in the form of an R × C contingency table—that is, one or both variables under study have more than two categories —were studied.



DEFINITION 10.7 An $R \times C$ **contingency table** is a table with R rows and C columns. It displays the relationship between two variables, where the variable in the rows has R categories and the variable in the columns has C categories.

Example

Suppose we want to study the relationship between age at first birth and development of breast cancer, as in Example 10.4 (p. 373). In particular, we would like to know whether the effect of age at first birth follows a consistent trend, that is:

(1) More protection for women whose age at first birth is < 20 than for women whose age at first birth is 25−29, and

(2) Higher risk for women whose age at first birth is ≥ 35 than for women whose age at first birth is 30−34.

The data are presented in Table 10.16, where case−control status is indicated along the rows and age at first birth categories are indicated along the columns. The data are arranged in the form of a 2 × 5 contingency table because case−control status has two categories (R = 2) and age at first birth has five categories (C = 5):

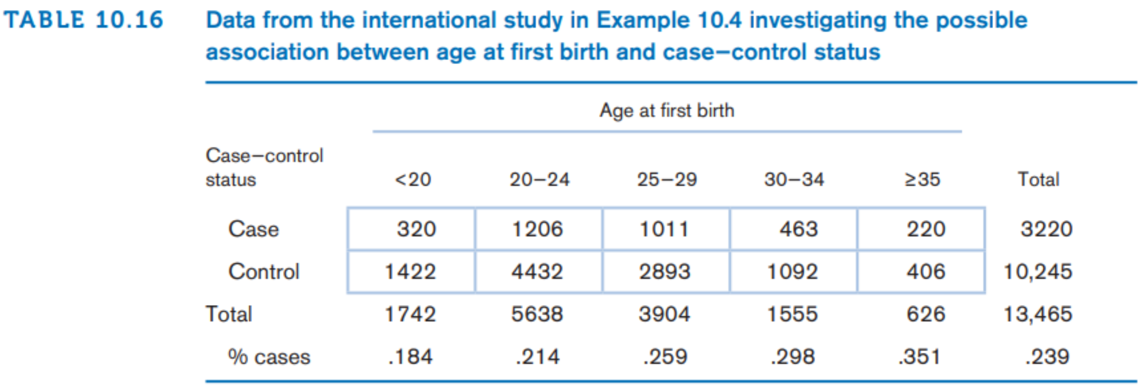


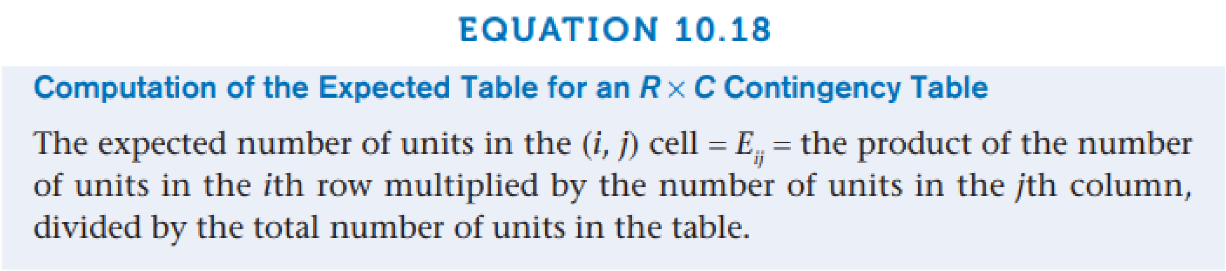
TABLE 10.16 Data from the international study in Example 10.4 investigating the possible association between age at first birth and case−control status

	Age at first birth					
Case−control status	<20	20−24	25−29	30−34	≥35	Total
Case	320	1206	1011	463	220	3220
Control	1422	4432	2893	1092	406	10,245
Total	1742	5638	3904	1555	626	13,465
% cases	.184	.214	.259	.298	.351	.239

Objective of the Study: We want to test for a relationship between age at first birth and case−control status.

Question: How should this be done?

Now, the expected table for an R × C contingency table under H_o can be formed in the same way as for a 2 × 2 contingency table as follows:



EQUATION 10.18

Computation of the Expected Table for an $R \times C$ Contingency Table

The expected number of units in the (i, j) cell $= E_{ij} =$ the product of the number of units in the ith row multiplied by the number of units in the jth column, divided by the total number of units in the table.

Example

Cancer Compute the expected table for the data in Table 10.16?

Solution

$$\begin{array}{l}\text{Expected value}\\ \text{of the (1,1) cell}\end{array} = \frac{\text{first row total} \times \text{first column total}}{\text{grand total}} = \frac{3220(1742)}{13{,}465} = 416.6$$

$$\begin{array}{l}\text{Expected value}\\ \text{of the (1,2) cell}\end{array} = \frac{\text{first row total} \times \text{second column total}}{\text{grand total}} = \frac{3220(5638)}{13{,}465} = 1348.3$$

$$\vdots$$

$$\begin{array}{l}\text{Expected value}\\ \text{of the (2,5) cell}\end{array} = \frac{\text{second row total} \times \text{fifth column total}}{\text{grand total}} = \frac{10{,}245(626)}{13{,}465} = 476.3$$

All 10 expected values are given in Table 10.17:

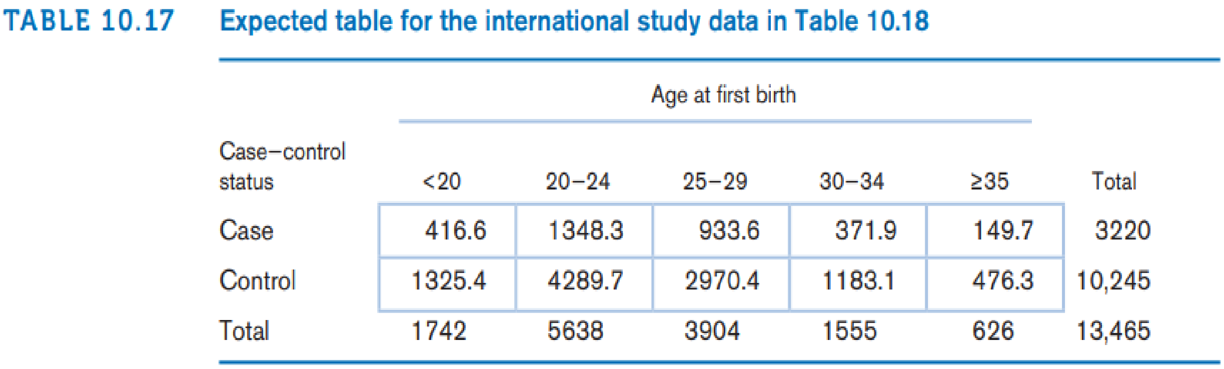


TABLE 10.17 Expected table for the international study data in Table 10.18

	Age at first birth					
Case–control status	<20	20–24	25–29	30–34	≥35	Total
Case	416.6	1348.3	933.6	371.9	149.7	3220
Control	1325.4	4289.7	2970.4	1183.1	476.3	10,245
Total	1742	5638	3904	1555	626	13,465

Note that: The sum of the expected values across any row or column must equal the corresponding row or column total, as was the case for 2 × 2 tables. This fact provides a good check that the expected values are computed correctly. The expected values in Table 10.17 fulfill this criterion except for round off error.

Notations

- We again want to compare the observed table with the expected table.

- The more similar these tables are, the more willing we will be to accept the null hypothesis H_0: that there is no association between the two variables. The more different the tables are, the more willing we will be to reject H_0.

- Again the criterion $(O - E)^2/E$ is used to compare the observed and expected counts for a particular cell.

- Furthermore, $(O - E)^2/E$ is summed over all the cells in the table to get an overall measure of agreement for the observed and expected tables.

- Under H_0, for an R × C contingency table, the sum of $(O - E)^2/E$ over the RC cells in the table will approximately follow a chi-square distribution with $df = (R - 1) \times (C - 1)$.

- H_0 will be rejected for large values of this sum and will be accepted for small values.

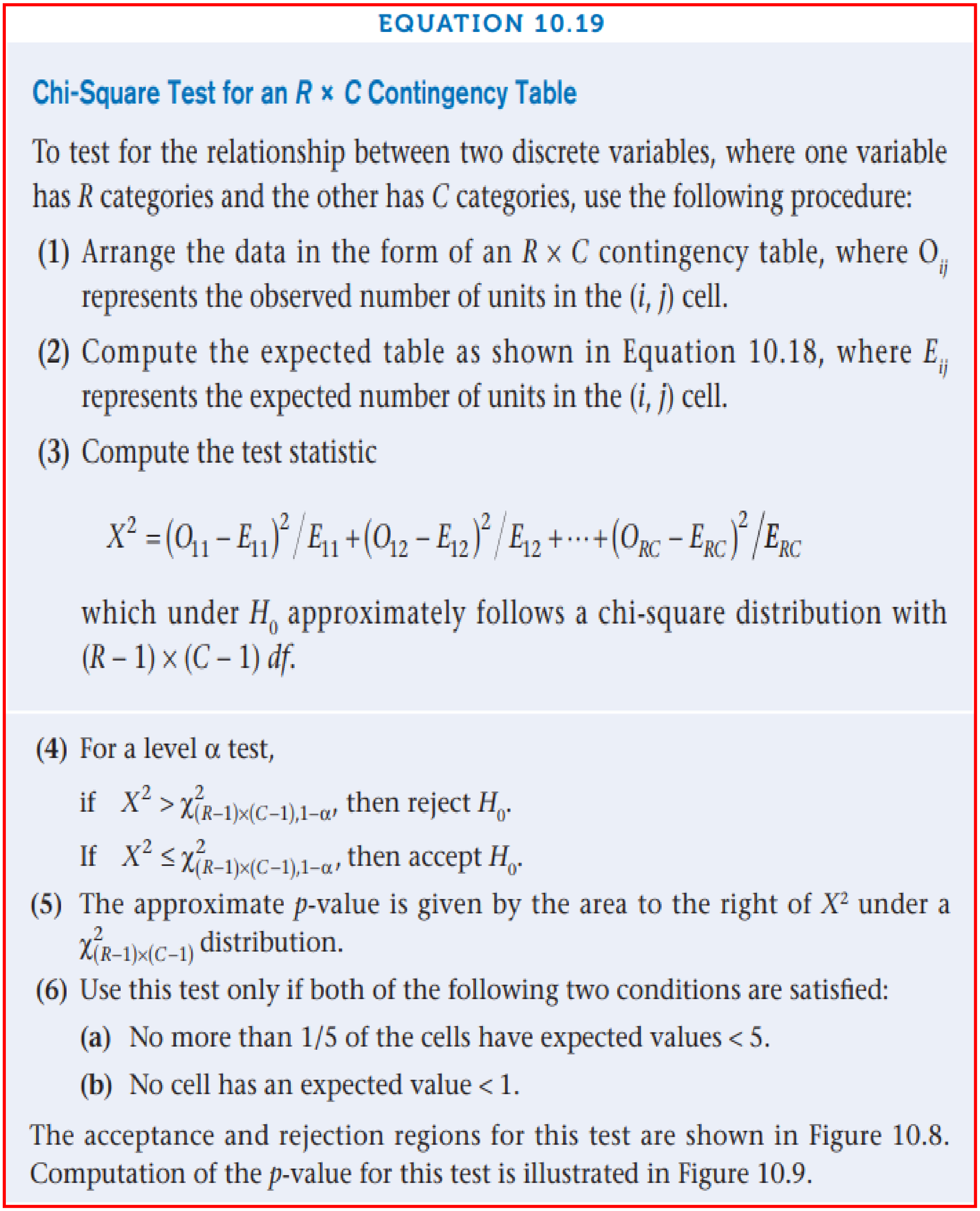
- Generally speaking, the continuity correction is not used for contingency tables larger than 2 × 2 because statisticians have found empirically that the correction does not help in the approximation of the test statistic by the chi-square distribution.

- As for 2 × 2 contingency tables, this test should not be used if the expected values of the cells are too small.

- Cochran [4] has studied the validity of the approximation in this case and recommends its use if:

  (1) No more than 1/5 of the cells have expected values < 5, and
  (2) No cell has an expected value < 1.

Now, the test procedure for an R × C contingency table can be summarized as follows:



EQUATION 10.19

Chi-Square Test for an $R \times C$ Contingency Table

To test for the relationship between two discrete variables, where one variable has R categories and the other has C categories, use the following procedure:

(1) Arrange the data in the form of an $R \times C$ contingency table, where O_{ij} represents the observed number of units in the (i, j) cell.

(2) Compute the expected table as shown in Equation 10.18, where E_{ij} represents the expected number of units in the (i, j) cell.

(3) Compute the test statistic

$$X^2 = (O_{11} - E_{11})^2 / E_{11} + (O_{12} - E_{12})^2 / E_{12} + \cdots + (O_{RC} - E_{RC})^2 / E_{RC}$$

which under H_0 approximately follows a chi-square distribution with $(R-1) \times (C-1)$ *df*.

(4) For a level α test,

if $X^2 > \chi^2_{(R-1)\times(C-1),1-\alpha}$, then reject H_0.

If $X^2 \le \chi^2_{(R-1)\times(C-1),1-\alpha}$, then accept H_0.

(5) The approximate p-value is given by the area to the right of X^2 under a $\chi^2_{(R-1)\times(C-1)}$ distribution.

(6) Use this test only if both of the following two conditions are satisfied:

(a) No more than 1/5 of the cells have expected values < 5.

(b) No cell has an expected value < 1.

The acceptance and rejection regions for this test are shown in Figure 10.8. Computation of the p-value for this test is illustrated in Figure 10.9.

FIGURE 10.8 Acceptance and rejection regions for the chi-square test for an $R \times C$ contingency table

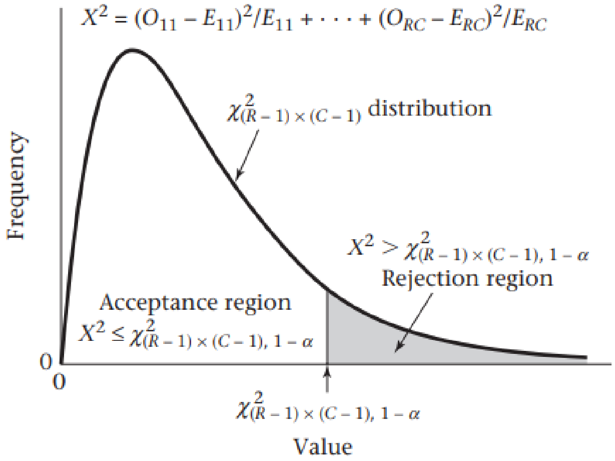
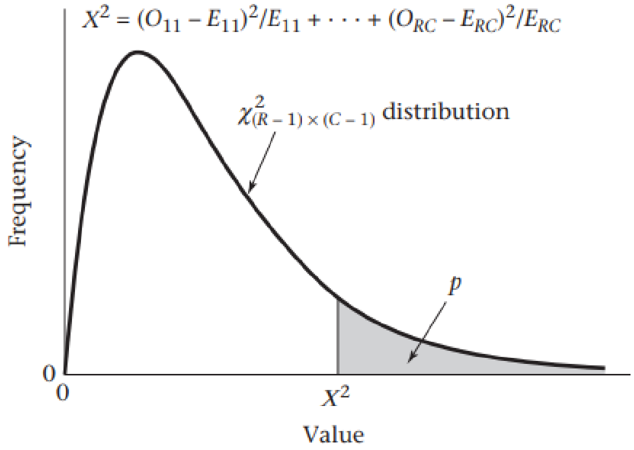


FIGURE 10.9 Computation of the p-value for the chi-square test for an $R \times C$ contingency table



Example

Cancer Assess the statistical significance of the data in Example 10.38 between the two variables using $\alpha = 0.05$:

H_0: There is no association between the age at first birth and prevalence of breast cancer (*the two variables are independent*).

$$vs$$

H_1: There is association between the age at first birth and prevalence of breast cancer (*the two variables are not independent or dependent*).

Solution

From Table 10.17 we see that all expected values are ≥ 5, so the test procedure in Equation 10.19 can be used. From Tables 10.16 and 10.17, we have the following:

$$X^2 = \frac{(320-416.6)^2}{416.6} + \frac{(1206-1348.3)^2}{1348.3} + \quad + \frac{(406-476.3)^2}{476.3} = 130.3$$

Under H₀, X^2 follows a chi-square distribution with df = (2 − 1) × (5 − 1) = 4.

Decision and Conclusion

Because we get:

$$\chi^2_{(4,\ \ 0.95)} = 9.49 < X^2 = 130.30, \text{ we have } p < 1 - 0.95 = 0.05$$

Therefore, H_0 is rejected and H_1 is accepted, then the results are very highly significant. Thus, we can conclude that there is a significant relationship (*not independent*) between the two variables under study age at first birth and prevalence of breast cancer. However, although this result shows some relationship between breast cancer and age at first birth, it does not tell us specifically about the nature of the relationship.

Notation

In this section, we have discussed tests for association between two categorical variables with R and C categories, respectively, where either R > 2 and / or C > 2. If both R and C are > 2, then the chi-square test for R × C contingency tables is used.

10.7 Chi-Square Goodness-of-Fit Test

In our previous work on estimation and hypothesis testing, we usually assumed the data came from a specific underlying probability model and then proceeded either to estimate the parameters of the model or test hypotheses concerning different possible values of the parameters. This section presents a general method of testing for the *goodness-of-fit of a probability model*. Consider the problem in Example 10.46 given below:

EXAMPLE 10.46

Hypertension Diastolic blood-pressure measurements were collected at home in a community-wide screening program of 14,736 adults ages 30−69 in East Boston, Massachusetts, as part of a nationwide study to detect and treat hypertensive people. The people in the study were each screened in the home, with two measurements taken during one visit. A frequency distribution of the mean diastolic blood pressure is given in Table 10.20 in 10-mm Hg intervals.

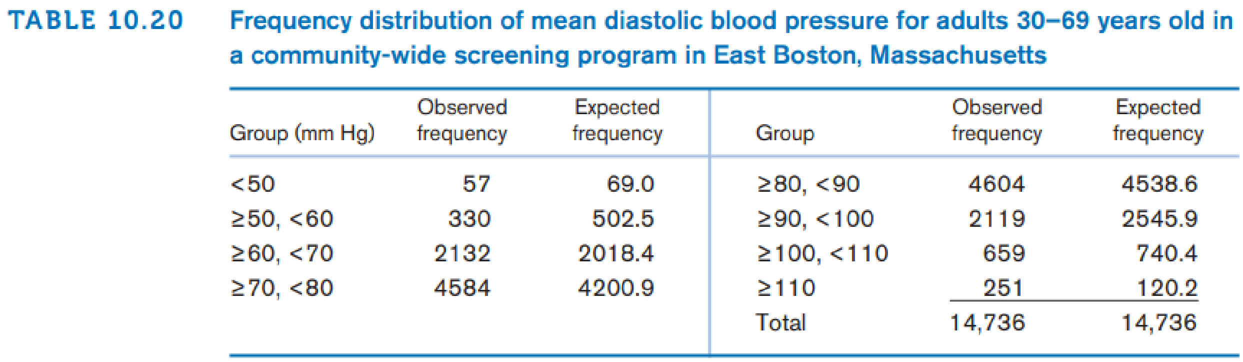


TABLE 10.20 Frequency distribution of mean diastolic blood pressure for adults 30–69 years old in a community-wide screening program in East Boston, Massachusetts

Group (mm Hg)	Observed frequency	Expected frequency	Group	Observed frequency	Expected frequency
<50	57	69.0	≥80, <90	4604	4538.6
≥50, <60	330	502.5	≥90, <100	2119	2545.9
≥60, <70	2132	2018.4	≥100, <110	659	740.4
≥70, <80	4584	4200.9	≥110	251	120.2
			Total	14,736	14,736

We would like to assume these measurements came from an underlying normal distribution because standard methods of statistical inference could then be applied on these data as presented in this text.

Question: How can the validity of this assumption be tested?

Answer: This assumption (*measurements came from an underlying normal distribution*) can be tested by:

- First computing what the expected frequencies would be in each group if the data did come from an underlying normal distribution.
- Then comparing these expected frequencies with the corresponding observed frequencies.

Computation of the Expected Frequencies

The expected frequency can be calculated using three rules as follows:

(1) The expected frequency within a group interval from a to b can be given by:

$$14,736\left\{\Phi\left[\left(b+\frac{1}{2}-\mu\right)\Big/\sigma\right]-\Phi\left[\left(a-\frac{1}{2}-\mu\right)\Big/\sigma\right]\right\}$$

(2) The expected frequency less than a can be given by:

$$14,736\left\{\Phi\left[\left(a-\frac{1}{2}-\mu\right)/\sigma\right]\right\}$$

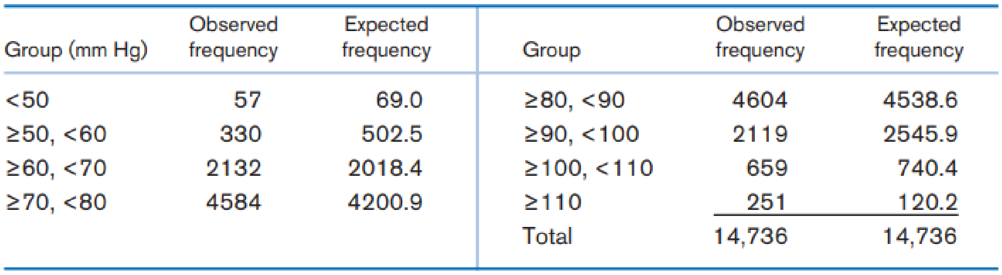
(3) The expected frequency greater than or equal to b can be given by:

$$14,736\left\{1-\Phi\left[\left(b-\frac{1}{2}-\mu\right)/\sigma\right]\right\}$$

EXAMPLE 10.47

Hypertension Compute the expected frequencies for the data in Table 10.20, assuming an underlying normal distribution:

TABLE 10.20 Frequency distribution of mean diastolic blood pressure for adults 30–69 years old in a community-wide screening program in East Boston, Massachusetts



Group (mm Hg)	Observed frequency	Expected frequency
<50	57	69.0
≥50, <60	330	502.5
≥60, <70	2132	2018.4
≥70, <80	4584	4200.9
≥80, <90	4604	4538.6
≥90, <100	2119	2545.9
≥100, <110	659	740.4
≥110	251	120.2
Total	14,736	14,736

Solution

- Assume the mean and standard deviation of this hypothetical normal distribution are given by the sample mean ($\bar{x} = 80.68$) and the sample standard deviation ($s = 12.00$).
- The expected frequency within the (≥ 50 , < 60) group would be computed as follows:

$$14{,}736 \times \{\Phi[(59.5 - 80.68)/12] - \Phi[(49.5 - 80.68)/12]\}$$
$$= 14{,}736 \times [\Phi(-1.765) - \Phi(-2.598)]$$
$$= 14{,}736 \times (.0388 - .0047) = 14{,}736(.0341) = 502.5.$$

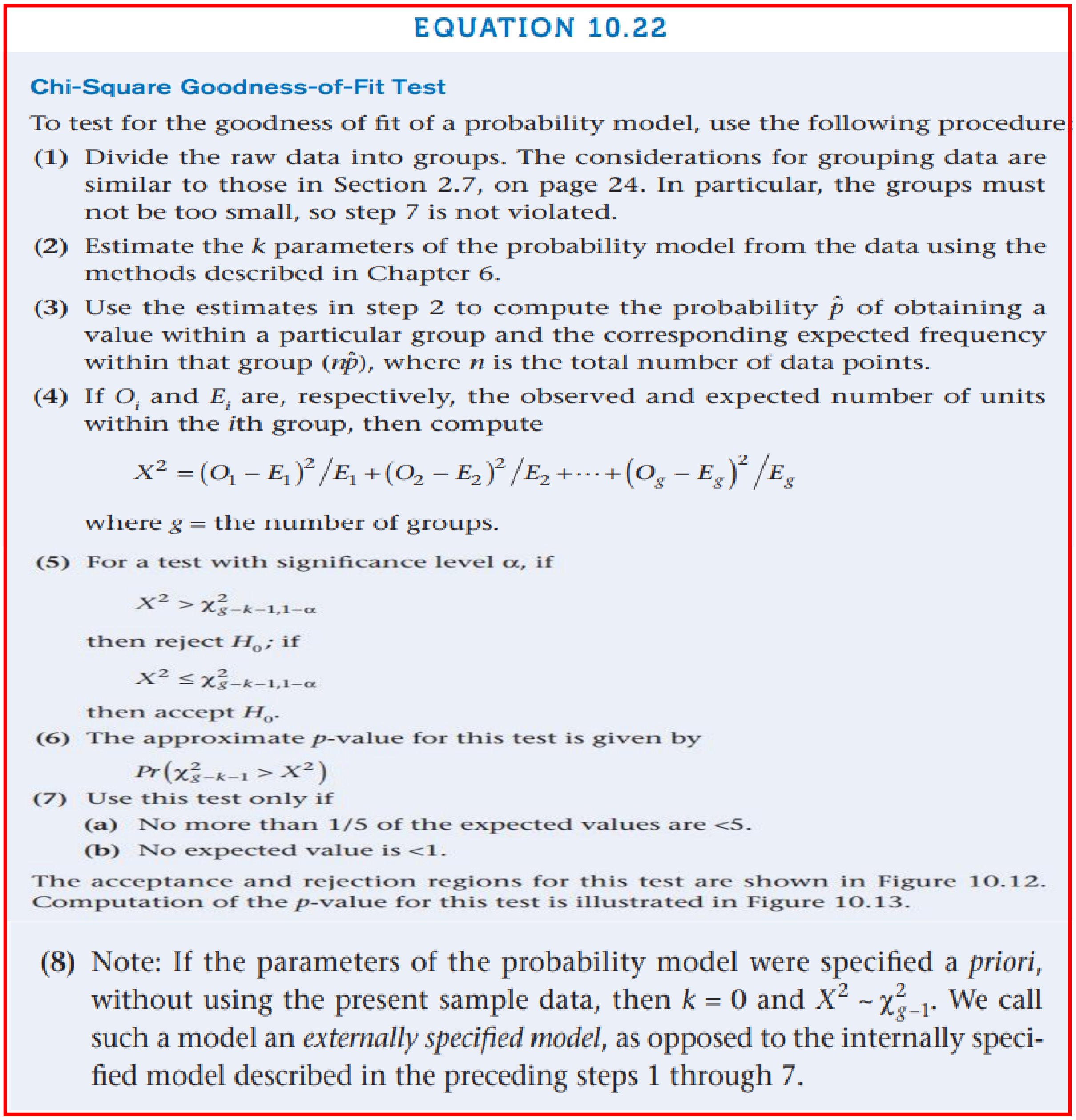
- The expected frequencies for all the groups in Table 10.20 are computed and given also in Table 10.20.

Notations

- We use the same measure of agreement between the observed and expected frequencies in a group that we used in our work on contingency tables, namely, $(O - E)^2/E$.
- The agreement between observed and expected frequencies can be summarized over the whole table by summing $(O - E)^2/E$ over all the groups.
- If we have the correct underlying model, then this sum will approximately follow a chi-square distribution with ($df = g - 1 - k$), where:
    - g = the number of groups.
    - k = the number of parameters estimated from the data used to compute the expected frequencies.
- This approximation will be valid only if the expected values in the groups are not too small.
- In particular, the requirement is that no expected value can be < 1 and not more than 1/5 of the expected values can be < 5.

- If there are too many groups with small expected frequencies, then some groups should be combined with other adjacent groups so the preceding rule is not violated.

The test procedure for the Chi-Square Goodness-of-Fit Test can be summarized as follows:



EQUATION 10.22

Chi-Square Goodness-of-Fit Test

To test for the goodness of fit of a probability model, use the following procedure:

(1) Divide the raw data into groups. The considerations for grouping data are similar to those in Section 2.7, on page 24. In particular, the groups must not be too small, so step 7 is not violated.

(2) Estimate the k parameters of the probability model from the data using the methods described in Chapter 6.

(3) Use the estimates in step 2 to compute the probability $\hat{p}$ of obtaining a value within a particular group and the corresponding expected frequency within that group ($n\hat{p}$), where n is the total number of data points.

(4) If O_i and E_i are, respectively, the observed and expected number of units within the ith group, then compute

$$X^2 = (O_1 - E_1)^2/E_1 + (O_2 - E_2)^2/E_2 + \cdots + (O_g - E_g)^2/E_g$$

where g = the number of groups.

(5) For a test with significance level α, if

$$X^2 > \chi^2_{g-k-1,1-\alpha}$$

then reject H_0; if

$$X^2 \le \chi^2_{g-k-1,1-\alpha}$$

then accept H_0.

(6) The approximate p-value for this test is given by

$$Pr(\chi^2_{g-k-1} > X^2)$$

(7) Use this test only if

(a) No more than 1/5 of the expected values are <5.

(b) No expected value is <1.

The acceptance and rejection regions for this test are shown in Figure 10.12. Computation of the p-value for this test is illustrated in Figure 10.13.

(8) Note: If the parameters of the probability model were specified a *priori*, without using the present sample data, then $k = 0$ and $X^2 \sim \chi^2_{g-1}$. We call such a model an *externally specified model*, as opposed to the internally specified model described in the preceding steps 1 through 7.

FIGURE 10.12 Acceptance and rejection regions for the chi-square goodness-of-fit test

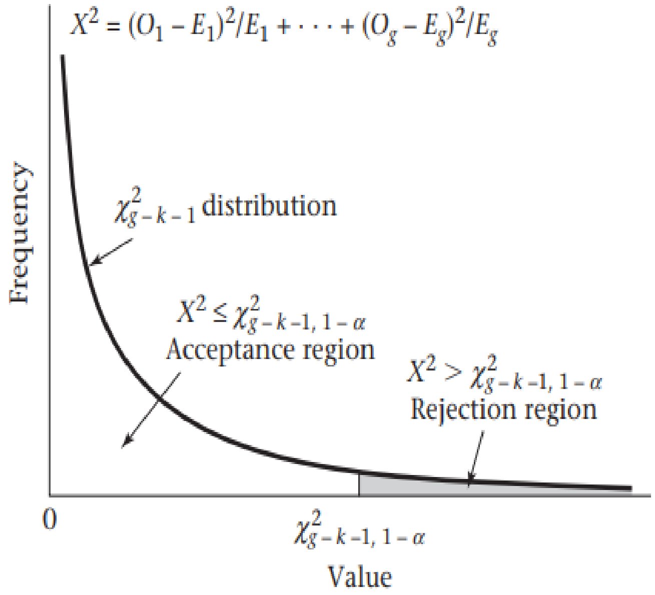
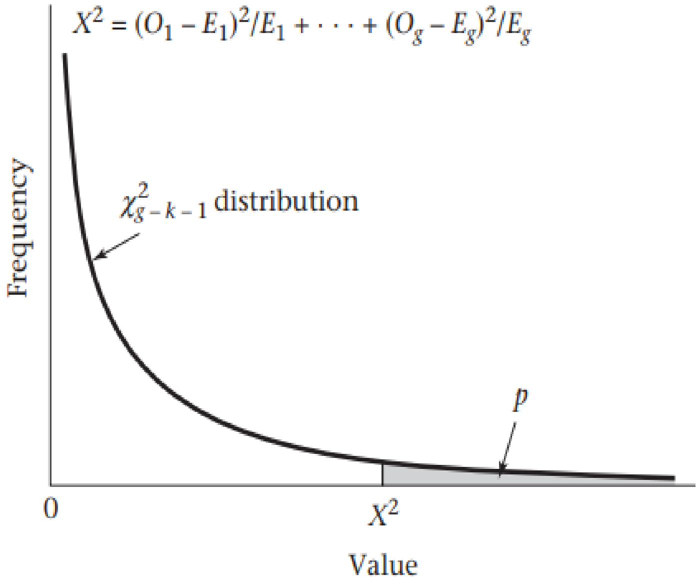


FIGURE 10.13 Computation of the p-value for the chi-square goodness-of-fit test



EXAMPLE 10.48

Hypertension Test for goodness of fit of the normal-probability model using the data in Table 10.20 given as follows:

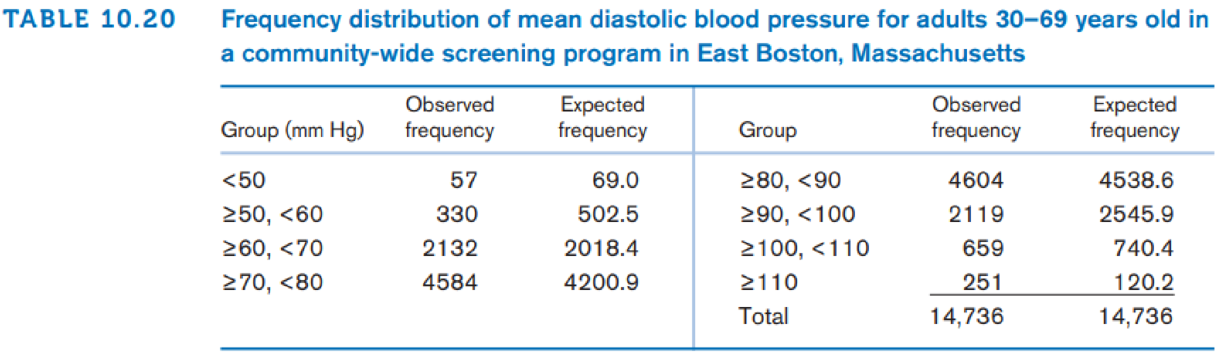


TABLE 10.20 Frequency distribution of mean diastolic blood pressure for adults 30–69 years old in a community-wide screening program in East Boston, Massachusetts

Group (mm Hg)	Observed frequency	Expected frequency	Group	Observed frequency	Expected frequency
<50	57	69.0	≥80, <90	4604	4538.6
≥50, <60	330	502.5	≥90, <100	2119	2545.9
≥60, <70	2132	2018.4	≥100, <110	659	740.4
≥70, <80	4584	4200.9	≥110	251	120.2
			Total	14,736	14,736

That is, test the following hypothesis:

H$_0$: The normal model (distribution) provide an adequate fit to the data.

$$vs$$

H$_1$: The normal model (distribution) does not provide an adequate fit to the data.

Solution

- Two parameters have been estimated from the data (μ, σ^2), and there are 8 groups. Therefore, k = 2, g = 8.
- Under H$_0$, X^2 follows a chi-square distribution with df = 8 − 2 − 1 = 5.
- The test statistic (X^2) can be calculated as follows:

$$X^2 = (O_1 - E_1)^2/E_1 + \cdots + (O_8 - E_8)^2/E_8$$
$$= (57 - 69.0)^2/69.0 + \cdots + (251 - 120.2)^2/120.2 = 326.2 \sim \chi_5^2 \text{ under } H_0$$

Decision and Conclusion

Because we get:

$$\chi^2_{(5,\ \ 0.999)} = 20.52 < X^2 = 326.2\text{, we have } p < 1 - 0.999 = 0.001$$

Therefore, H$_0$ is rejected and H$_1$ is accepted, then the results are very highly significant. Thus, the normal model does not provide an adequate fit to the data. The normal model appears to fit fairly well in the middle of the distribution (between 60 and 110 mm Hg) but fails badly in the tails, predicting too many blood pressures below 60 mm Hg and too few over 110 mm Hg.

Notation

The test procedure in Equation 10.22 can be used to assess the goodness of fit of any probability model, not just the normal model. The expected frequencies would be computed from the probability distribution of the proposed model, and then the same goodness-of-fit test statistic as given in Equation 10.22 would be used. Also, the test procedure can be used to test for the goodness of fit of both a model in which the parameters are estimated from the data set used for testing the model as described in steps 1 through 7 and a model in which the parameters are specified a priori as in step 8.

Exercises

Exercise (1)

A sample of 150 people from a certain industrial community showed that 80 people suffered from a lung disease. A sample of 100 people from a rural community showed that 30 suffered from the same lung disease. At $\alpha = 0.05$, is there a difference between the proportion of people who suffer from the disease in the two communities? Use normal theory test?

Answer

$\hat{p}_1 = 0.533$, $\hat{p}_2 = 0.3$, $\hat{p} = 0.44$, $\hat{q} = 1 - \hat{p}$ = 0.56

Hypotheses to be tested are:

$$H_0: p_1 = p_2 \text{ vs. } H_1: p_1 \neq p_2$$

Test Statistic: Z = 3.64

Decision: Reject H_0

Conclusion: There is enough evidence to support the claim that there is a significant difference in the two proportions.

Exercise (2)

A recent study showed that in a sample of 80 surgeons, 45 smoked. In a sample of 120 general practitioners, 63 smoked. At $\alpha = 0.05$, by using a 2 x 2 contingency-table approach is there a difference in the two proportions?

Answer

Observed Table

Practitioner	Smoking Status		Total
	Smoked	Not Smoked	
Surgeons	45	35	80
Non- Surgeons	63	57	120
Total	108	92	200

Expected Table

Practitioner	Smoking Status		Total
	Smoked	Not Smoked	
Surgeons	43.2	36.8	80
Non- Surgeons	64.8	55.2	120
Total	108	92	200

$\hat{p}_1 = 0.5625$, $\hat{p}_2 = 0.525$, $\hat{p} = 0.54$, $\hat{q} = 1 - \hat{p}$ = 0.46

Hypotheses to be tested are:

$$H_0: p_1 = p_2 \text{ vs. } H_1: p_1 \neq p_2$$

Test Statistic: Z = 0.521

Decision: Do Not Reject (Accept) H_0

Conclusion: There is not enough evidence to support the claim that there is a significant difference in the two proportions.

Exercise (3)

A researcher wishes to determine whether there is a relationship between the gender (sex) of an individual and the amount of headache medications consumed. A sample of 69 people is selected, and the data in the following contingency table are obtained:

Contingency Table

Gender	Headache Consumption			Total
	Low	Moderate	High	
Male	10	9	8	27
Female	13	16	12	41
Total	23	25	20	68

At $\alpha = 0.10$, can the researcher conclude headache consumption is related to gender?

Answer

H_0: The amount of headache medications consumes is independent of the individual's gender.

$$vs$$

H_1: The amount of headache medications consumes is not independent (dependent) of the individual's gender.

We have the following:

Expected Table

Gender	Headache Consumption			Total
	Low	Moderate	High	
Male	9.13	9.93	7.94	27
Female	13.87	15.07	12.06	41
Total	23	25	20	68

X^2 = 0.283 follows a chi-square distribution with df = (2 − 1) × (3 − 1) = 2.

Decision and Conclusion

Because we get:

$$\chi^2_{(2,\ \ 0.95)} = 4.605 > X^2 = 0.283, \text{ we have p} < 1 - 0.95 = 0.05$$

Therefore, H_0 is not rejected (accepted) and H_1 is rejected, then the results shows that there is not enough evidence to support the claim that the amount of headache a person consumes is dependent on the individual's gender.
