Chapter 10

Hypothesis Testing: Categorical Data

Introduction

In this chapter, we will discuss

- Methods of hypothesis testing for comparing two or more binomial proportions

- Methods for testing the goodness-of-fit of a previously specified probability model to actual data

Two-Sample Test for Binomial Proportions

To test whether an important risk factor for breast cancer is age at first childbirth.

p_1 = probability that age at first birth is ≥30 in case women with at least one birth

p_2 = probability that age at first birth is ≥30 in control women with at least one birth

Hypothesis: H_0: $p_1 = p_2 = p$ vs. H_1: $p_1 \neq p_2$ for some constant p.

Two equivalent approaches for testing the hypothesis:

I.Normal-Theory method

II.Contingency-Table method

Normal-Theory method

Significance test is based on the difference between the sample proportions $\hat{p}_1 - \hat{p}_2$.

If this difference is very different from 0 then H_0 is rejected; otherwise, H_0 is accepted.

Dividing $\hat{p}_1 - \hat{p}_2$ by its standard error, $\sqrt{pq(1/n_1+1/n_2)}$ then under H_0

$$z = (\hat{p}_1 - \hat{p}_2) / \sqrt{pq(1/n_1 + 1/n_2)} \sim N(0,1)$$

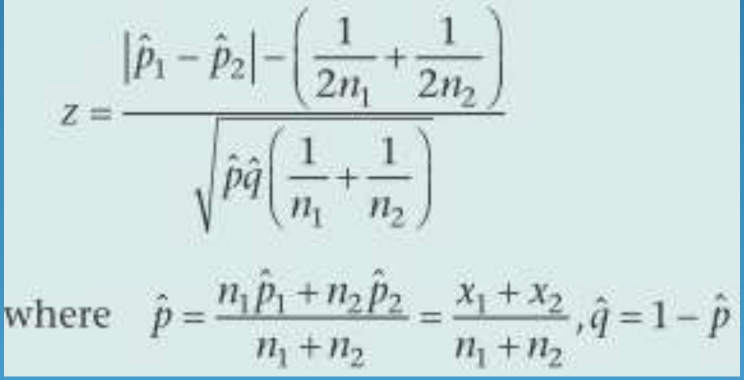
p and q and unknown so denominator of z cannot be determined unless we estimate for p using a weighted average of the sample proportions.

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

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



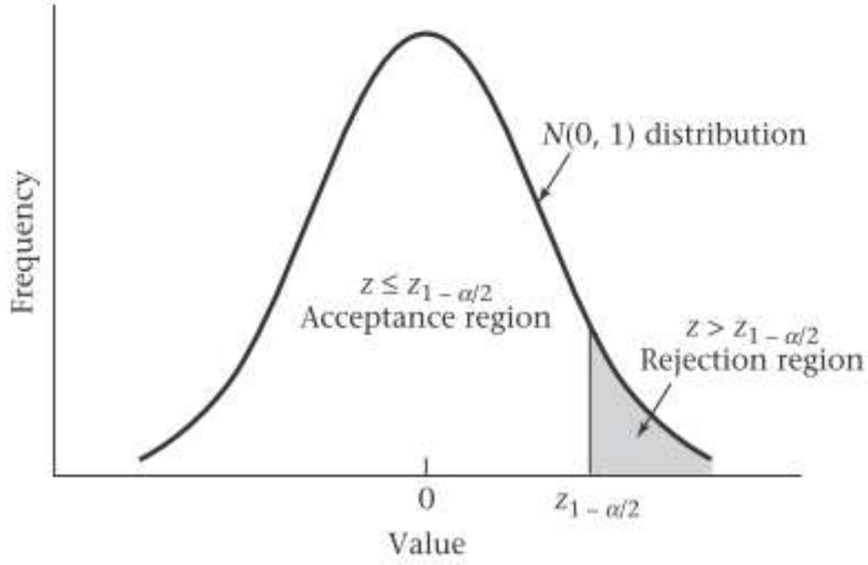
$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

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

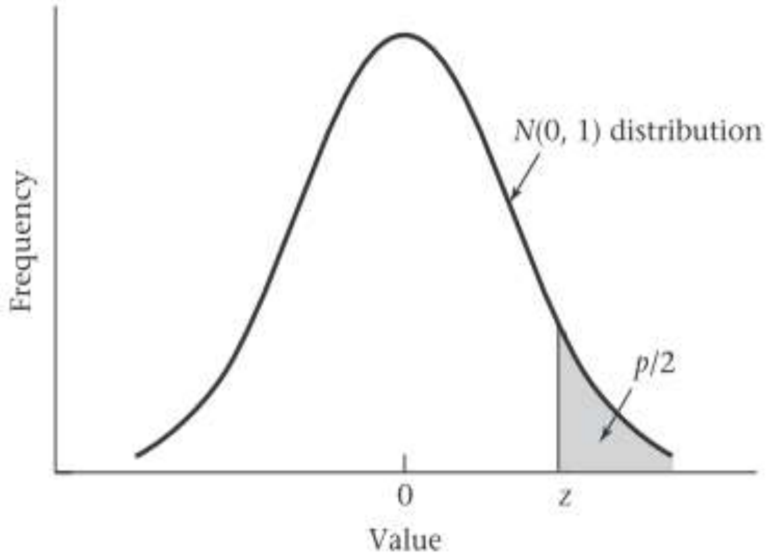
where x_1,x_2 are the number of events in the first and second samples, resp.

2. For a two-sided level α test, if $z > z_{1-\alpha/2}$ then reject Ho
3. If $z \leq z_{1-\alpha/2}$ then accept Ho.
4. The approximate p-value for this test is given by $p = 2[1-\Phi(z)]$
5. 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 $n_2\hat{p}\hat{q} \geq 5$.

Acceptance and rejection regions for the two-sample test for binomial proportions (normal-theory test)



Computation of the exact *p*-value for the two-sample test for binomial proportions (normal-theory test)



Contingency-Table Method

- A 2 × 2 contingency table is a table composed of two rows cross-classified by two columns.

- 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 to the other to the columns.

- Each of the four cells represents the number of units with a specific value for each of the two variables. The cells are sometimes referred to by number, with 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, and the (2,2) cell being the cell in the second column.

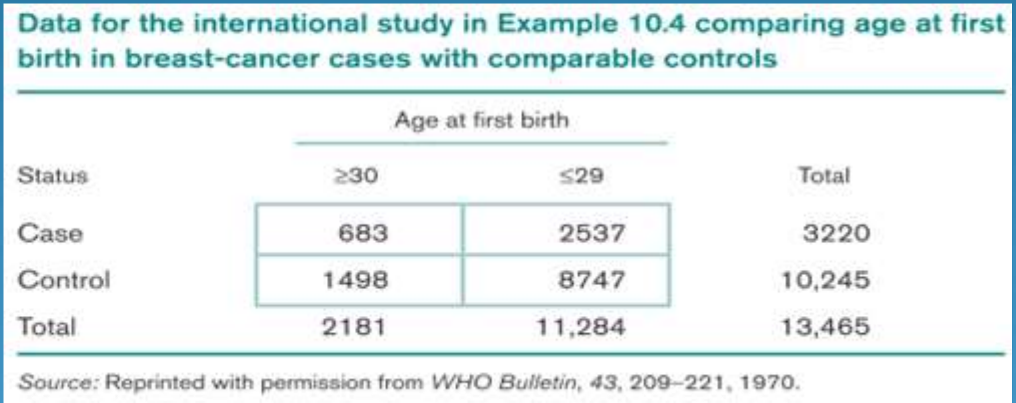
- The observed number of units in the four cells is likewise referred to as O_{11}, O_{12}, O_{21}, and O_{22}, resp.

Furthermore, it is customary to total

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

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

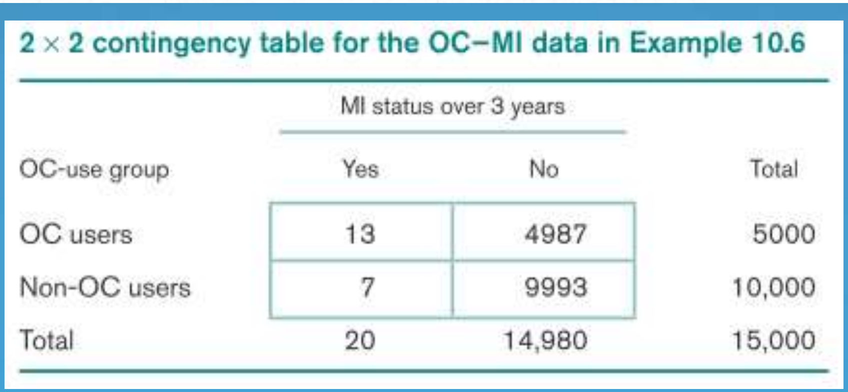
3.The total number of units in the four cells, which is displayed in the lower right-hand corner of the table is called the **grand total**.



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

Source: Reprinted with permission from *WHO Bulletin*, 43, 209–221, 1970.

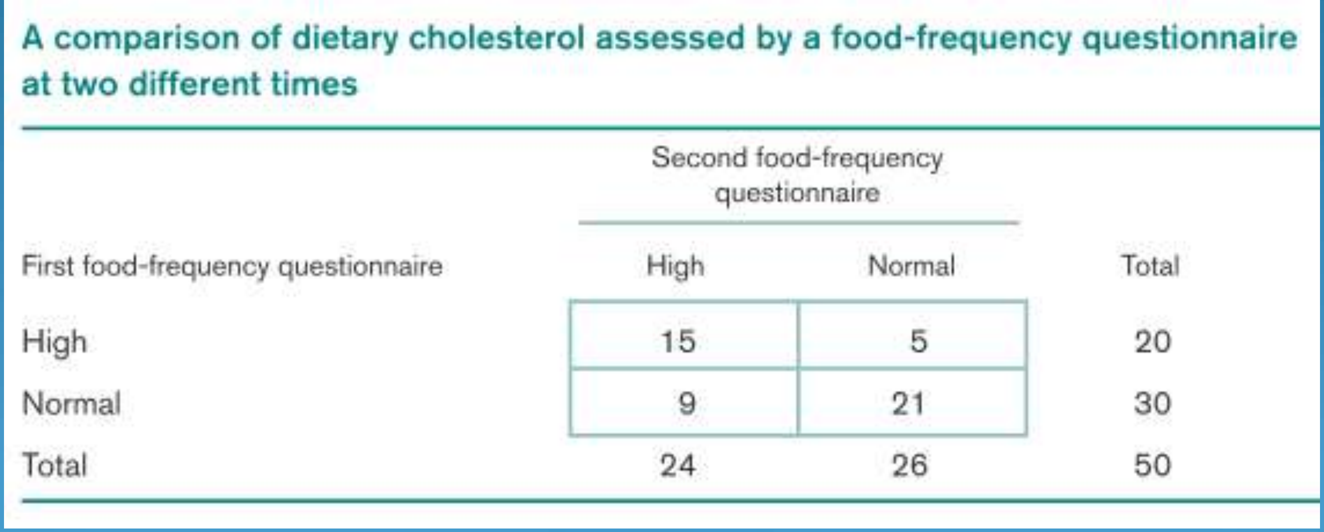


2 × 2 contingency table for the OC–MI data in Example 10.6

	MI status over 3 years		
OC-use group	Yes	No	Total
OC users	13	4987	5000
Non-OC users	7	9993	10,000
Total	20	14,980	15,000

- Test for homogeneity of binomial proportions: tests whether the proportions are the same in two independent samples.

- Test of independence or a test of association: tests whether there is some association between two reported measures of a characteristic.



A comparison of dietary cholesterol assessed by a food-frequency questionnaire at two different times

	Second food-frequency questionnaire		
First food-frequency questionnaire	High	Normal	Total
High	15	5	20
Normal	9	21	30
Total	24	26	50

Significance Testing Using the Contingency-Table Approach

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.

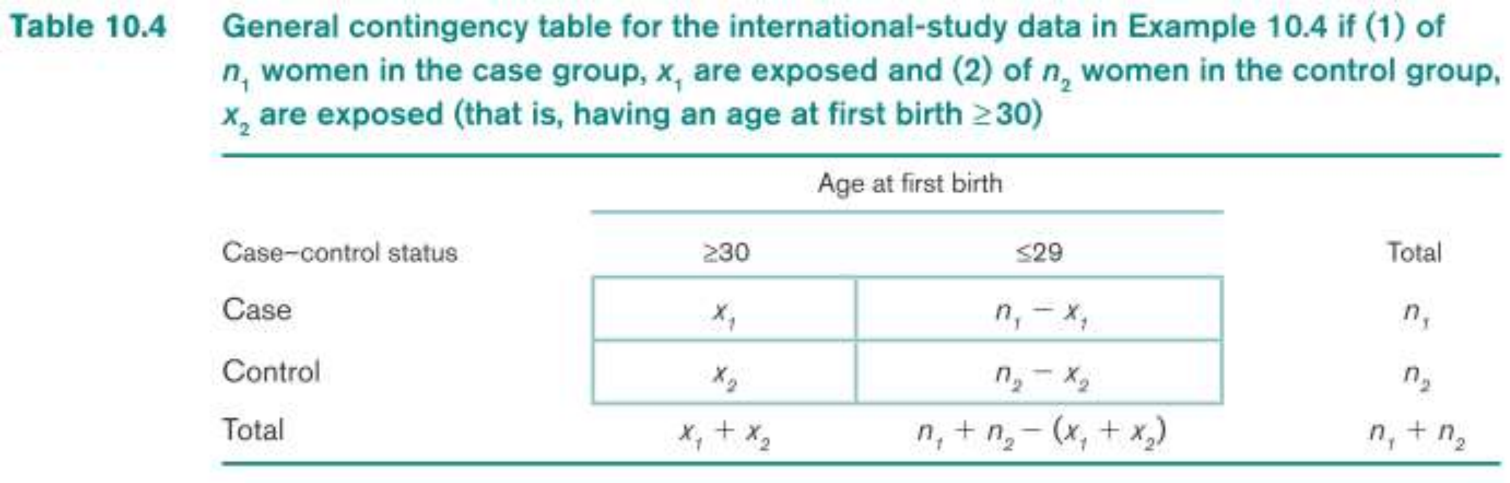


Table 10.4 General contingency table for the international-study data in Example 10.4 if (1) of n_1 women in the case group, x_1 are exposed and (2) of n_2 women in the control group, x_2 are exposed (that is, having an age at first birth ≥30)

	Age at first birth		
Case–control status	≥30	≤29	Total
Case	x_1	$n_1 - x_1$	n_1
Control	x_2	$n_2 - x_2$	n_2
Total	$x_1 + x_2$	$n_1 + n_2 - (x_1 + x_2)$	$n_1 + n_2$

An expected table is a contingency table that would be expected if there were no relationship between parameters that is if H_o: $p_1 = p_2 = p$ were true.

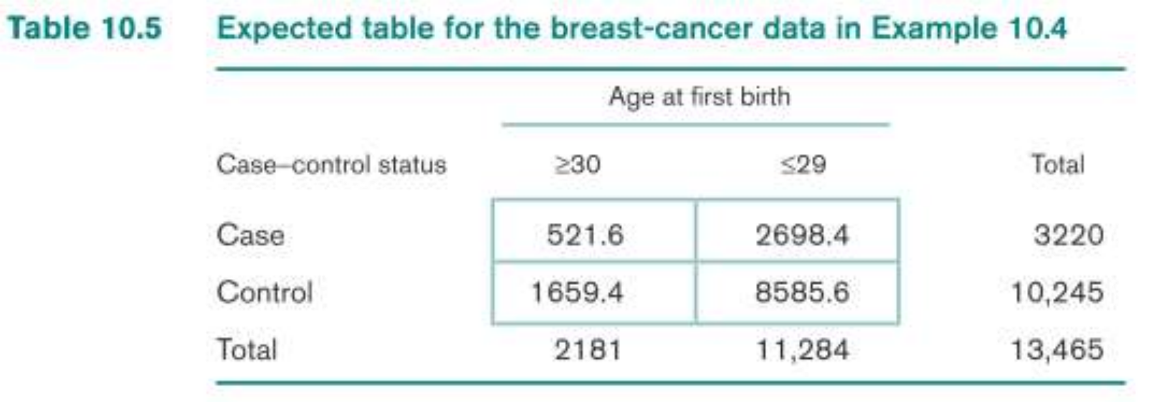


Table 10.5 Expected table for the breast-cancer data in Example 10.4

	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

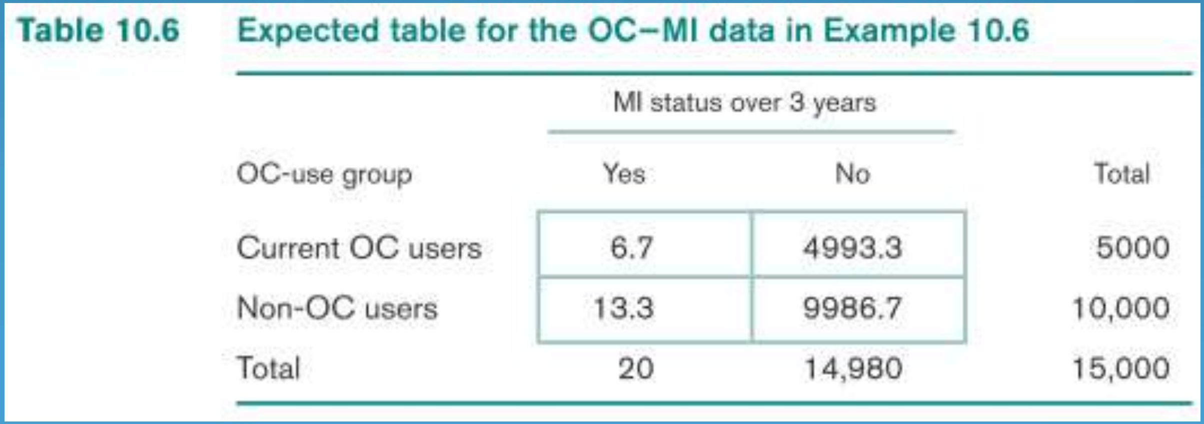


Table 10.6 Expected table for the OC–MI data in Example 10.6

	MI status over 3 years		
OC-use group	Yes	No	Total
Current OC users	6.7	4993.3	5000
Non-OC users	13.3	9986.7	10,000
Total	20	14,980	15,000

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

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 which under H_0 approximately follows a χ_1^2 distribution.

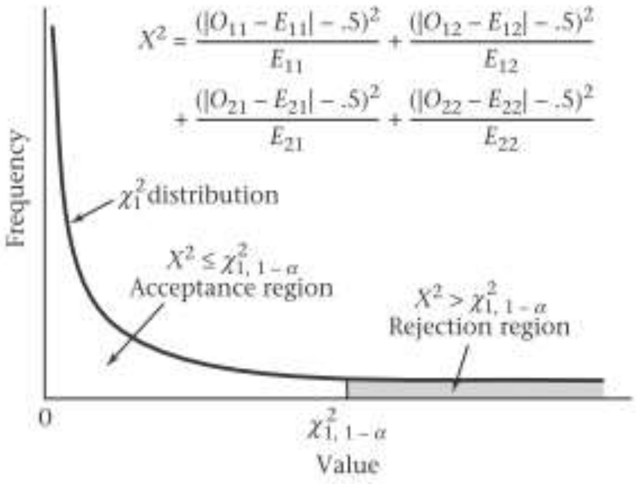
$$X^2 = \left(|O_{11} - E_{11}| - .5\right)^2 / E_{11} + \left(|O_{12} - E_{12}| - .5\right)^2 / E_{12} + \left(|O_{21} - E_{21}| - .5\right)^2 / E_{21} + \left(|O_{22} - E_{22}| - .5\right)^2 / E_{22}$$

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

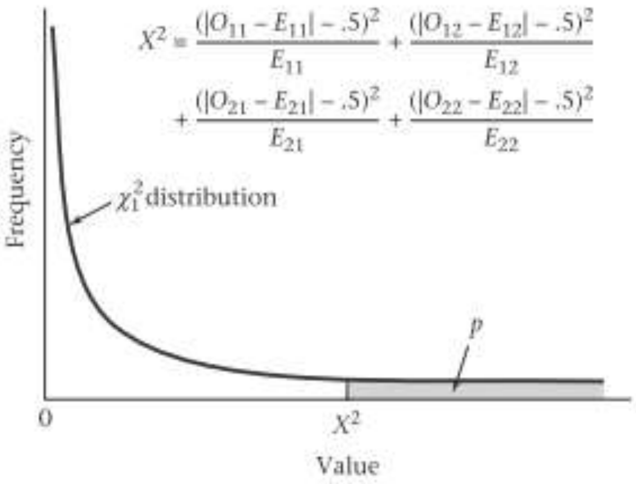
The Yates-corrected chi-square test is a *two-sided* test even though the critical region, based on the chi-square distribution, is one-sided.

Large values of $|O_{ij} – E_{ij}|$ and correspondingly, of the test statistic X^2 will be obtained under H_1 regardless of whether $p_1 < p_2$ or $p_1 > p_2$. Small values of X^2 are evidence in favor of H_o.

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



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



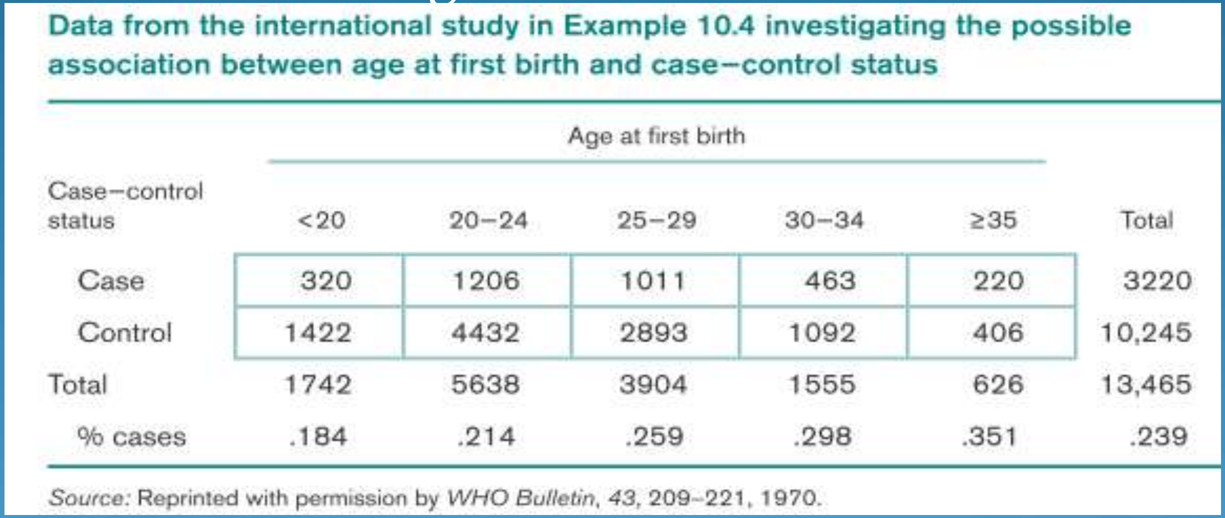
- Contingency table approach is more understandable and results are more frequently reported in this format.

- Use of a continuity correction for the contingency table is a debated subject.

- Generally, p-values obtained using the continuity correction are slightly larger and thus are slightly less significant than comparable results.

R × C Contingency Tables

An **R × 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.

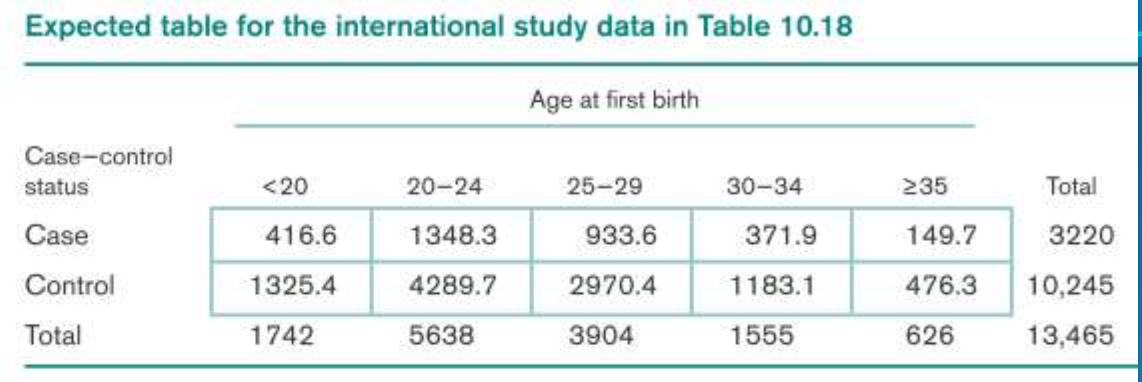


Data from the international study in Example 10.4 investigating the possible association between age at first birth and case−control status

Case−control status	Age at first birth <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

Source: Reprinted with permission by *WHO Bulletin*, 43, 209–221, 1970.

The expected number of units in the (i,j) cell = E_{ij} = the product of the number of units in the *i*th row multiplied by the no. of units in the *i*th column, divided by the total no. of units in the table.



Expected table for the international study data in Table 10.18

Case–control status	Age at first birth					Total
	<20	20–24	25–29	30–34	≥35	
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

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

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

2.Compute the expected table 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_o approximately follows a chi-square distribution with $(R - 1) \times (C - 1)$ *df*.

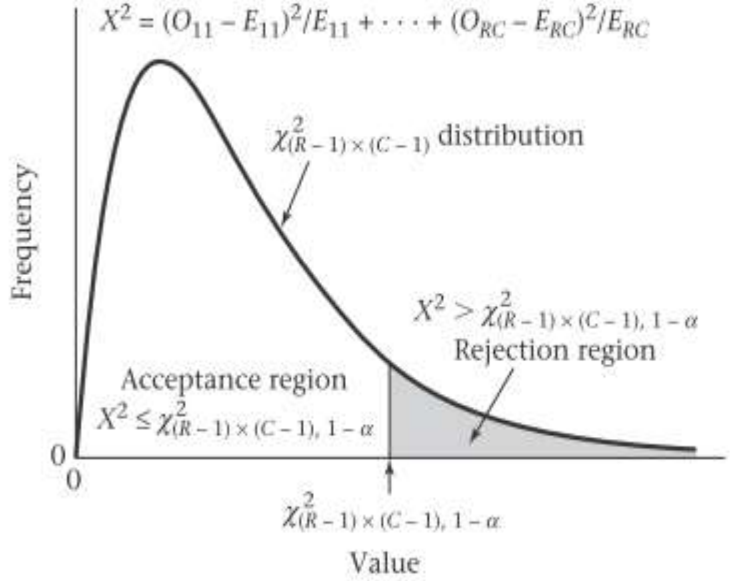
4. For a level α test, if $X^2 > \chi^2_{(R-1)\times(C-1),1-\alpha}$, then reject H_o.
if $X^2 \leq \chi^2_{(R-1)\times(C-1),1-\alpha}$, then accept H_o.

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.

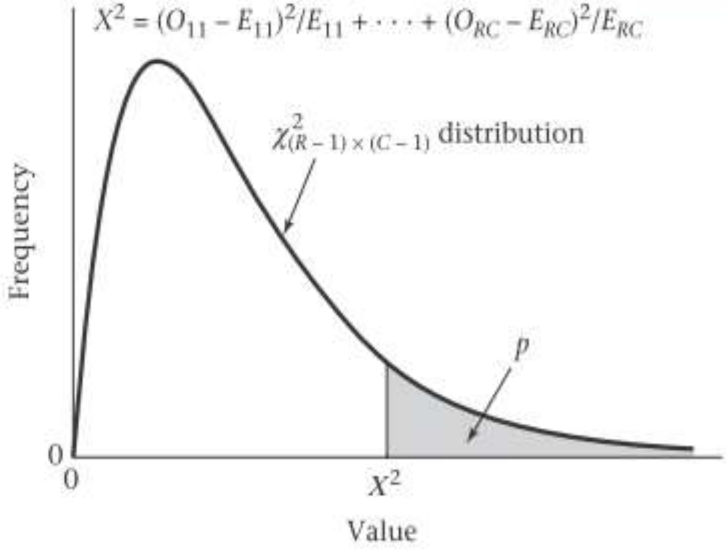
6. Use this test only if both of the following conditions are satisfied:

   a. No more than 1/5 of the cells have expected values <5
   b. No cell has an expected value <1.

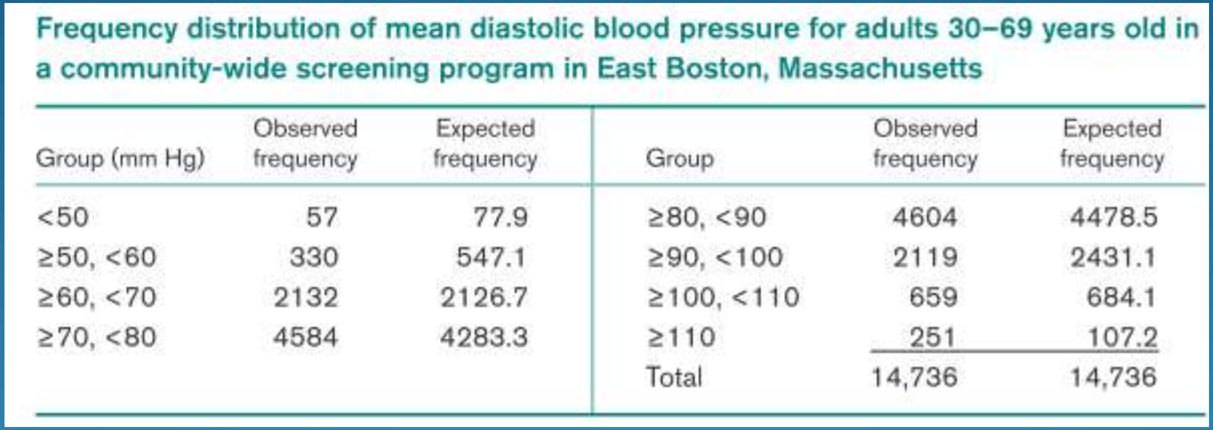
Acceptance and rejection regions for the chi-square test for an $R \times C$ contingency table


$X^2 = (O_{11} - E_{11})^2/E_{11} + \cdots + (O_{RC} - E_{RC})^2/E_{RC}$
$\chi^2_{(R-1)\times(C-1)}$ distribution
$X^2 > \chi^2_{(R-1)\times(C-1),\,1-\alpha}$
Rejection region
Acceptance region
$X^2 \le \chi^2_{(R-1)\times(C-1),\,1-\alpha}$
Frequency
0
0
$\chi^2_{(R-1)\times(C-1),\,1-\alpha}$
Value

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


$X^2 = (O_{11} - E_{11})^2/E_{11} + \cdots + (O_{RC} - E_{RC})^2/E_{RC}$
$\chi^2_{(R-1)\times(C-1)}$ distribution
p
Frequency
0
0
X^2
Value

Chi-Square Goodness-of-Fit Test



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	77.9	≥80, <90	4604	4478.5
≥50, <60	330	547.1	≥90, <100	2119	2431.1
≥60, <70	2132	2126.7	≥100, <110	659	684.1
≥70, <80	4584	4283.3	≥110	251	107.2
			Total	14,736	14,736

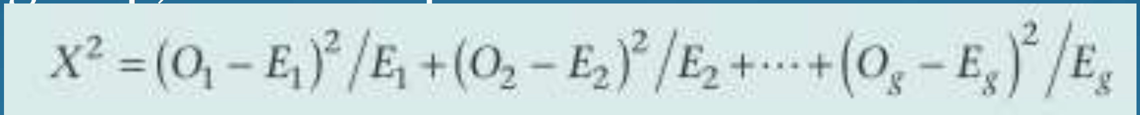
To test the goodness-of-fit of a probability model, use the following procedure:

1.Divide the raw data into groups.

2.Estimate the k parameters f the probability model.

3.Use the estimates in step 2 to compute the probability $\hat{p}$ f obtaining a value within a particular group and the corresponding expected frequency within that group (n), $\hat{p}$ here n is the total no. of data points.

If O_i and E_i are, resp., 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

For a test with significance level α,

if $X^2 > \chi^2_{g-k-1,1-\alpha}$ then reject Ho; if $X^2 \leq \chi^2_{g-k-1,1-\alpha}$ then accept Ho.

The approximate *p*-value for this test is given by $Pr(\chi^2_{g-k-1,1-\alpha} > X^2)$

Use this test only if

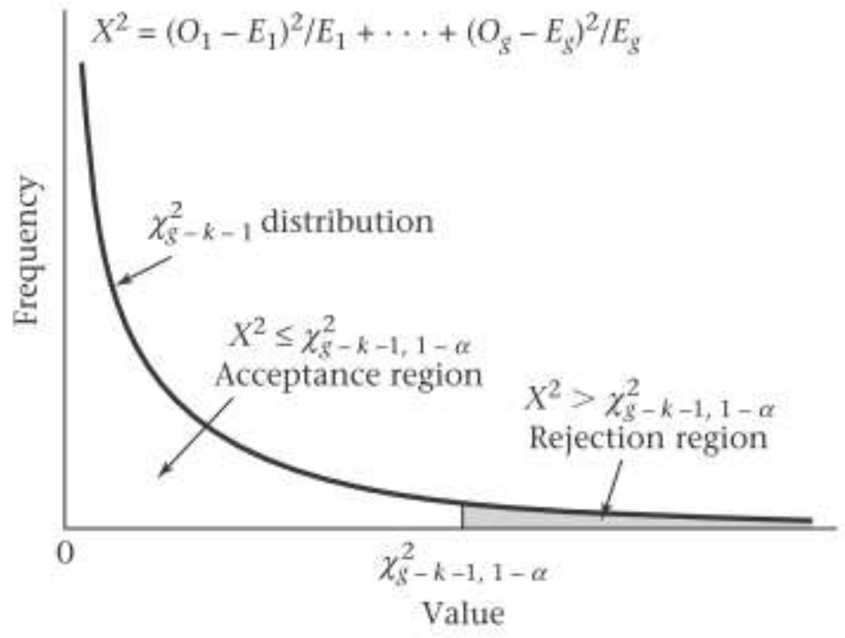
No more than 1/5 of the expected values are <5.

No expected value is <1

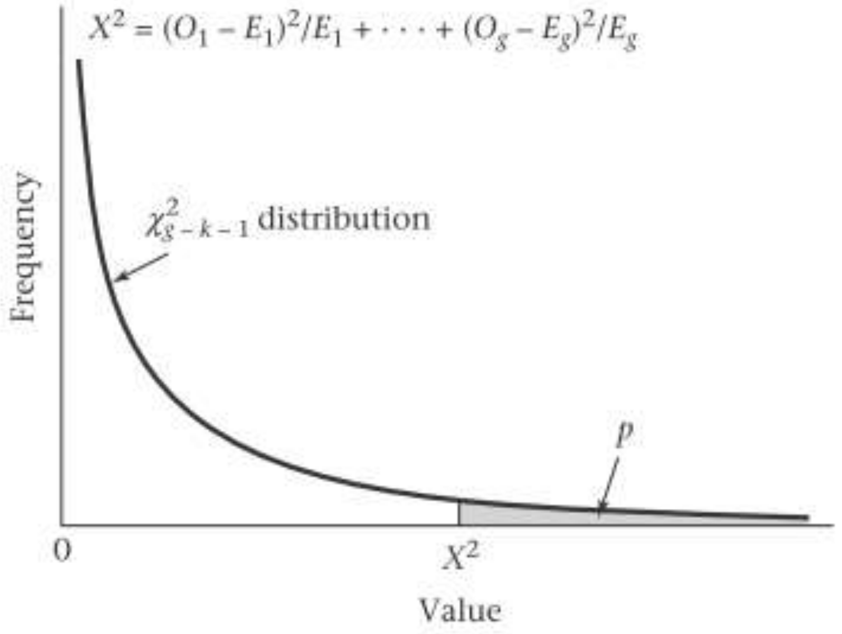
If the parameters of the probability model were specified a priori, without using the present sample data, then k = 0 and $X^2 \sim \chi^2_{g-1}$

Such a model is called an *externally specified model.*

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



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



Summary

In this chapter, we discussed

1.Techniques for analyzing qualitative or categorical data

2.Comparison of binomial proportions from two independent samples using

  i. two-sample test

  ii. chi-square test

  iii. 2×2 contingency tables

3. R×C contingency tables

4. Goodness-of-fit tests.

The End