Hypothesis Testing: Categorical Data

Ex. 10.4

- A hypothesis: an important factor for breast cancer is age at first birth.
- An international study was set up to test the hypothesis.
  - Breast cancer cases were identified among women in selected hospitals in the United States, Greece, Yugoslavia, Brazil, and Japan.
  - Controls were chosen from women of comparable age who were in the hospital at the same time as the cases, but who did not have breast cancer.
  - All women were asked about their age at first birth.
  - The set of women with at least one birth was arbitrarily divided into two categories:
    - Women whose age at first birth ≤ 29
    - Women whose age at first birth ≥ 30
- Results among women with at least one birth
  - 683 out of 3220 (21.2%) women with breast cancer had an age at first birth ≥ 30
  - 1498 out of 10,245 (14.6%) women without breast cancer had an age at first birth ≥ 30
- How can we assess whether this difference is significant?

Two-Sample Test for Binomial Proportions

- p_1= the probability that age at first birth is ≥ 30 in case women.
- p_2 = the probability that age at first birth is $\geq$ 30 in control women.
- Whether or not the underlying probability of having an age at first birth of ≥ 30 is different in the two groups.
- H_0: $p_1 = p_2 = p$ versus H_1: $p_1 \neq p_2$

Normal-Theory Method

- Base the significance test on the difference between the sample proportions $\hat{p}_1 - \hat{p}_2$
- Assume samples are large enough

$\hat{p}_1 - \hat{p}_2$ is normally distributed

$$\frac{pq}{n_1} + \frac{pq}{n_2} = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \qquad z = (\hat{p}_1 - \hat{p}_2) \Big/ \sqrt{pq(1/n_1 + 1/n_2)} \doteq N(0,1)$$

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

To better accommodate the normal approximation to the binomial

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

Two-Sample Test for Binomial Proportions (Normal-Theory Test) To test the hypothesis H_0: $p_1 = p_2$ versus 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 $z \le z_{1-\alpha/2}$

then accept H_0.

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

$$p = 2[1 - \Phi(z)]$$

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

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

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

Sample proportion of case women whose age at first birth was ≥ 30 is

$\hat{p}_1 = 683/3220 = .212$

For control women

$\hat{p}_2 = 1498/10,245 = .146$

$\hat{p} = (683 + 1498)/(3220 + 10,245) = .162$

$\hat{q} = 1 - .162 = .838$

$n_1\hat{p}\hat{q} = 3220(.162)(.838) = 437 \geq 5$

$n_2\hat{p}\hat{q} = 10,245(.162)(.838) = 1391 \geq 5$

The test statistic is given by

$$z = \left\{|.212 - .146| - \left[\frac{1}{2(3220)} + \frac{1}{2(10,245)}\right]\right\} \Big/ \sqrt{.162(.838)\left(\frac{1}{3220} + \frac{1}{10,245}\right)}$$

$= .0657/.00744$

$= 8.8$

The p-value $= 2 \times [1 - \Phi(8.8)] < .001$, and the results are highly significant.

Contingency-Table Method

- The data in the previous example can be represented as a 2×2 contingency table.

	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 of *WHO Bulletin*, *43*, 209–221, 1970.

- Row margins
- Column margins
- Grand total

Significance Testing Using Contingency-Table Approach

- Observed contingency table
- Expected table

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	$x_1 + x_2$	$n_1 + n_2 - (x_1 + x_2)$	$n_1 + n_2$

Computation of Expected Values for Contingency Tables

- Under null hypothesis, the expected number of units in the (1, 1) cell is

$$n_1\hat{p} = n_1(x_1 + x_2) / (n_1 + n_2)$$

- For the (2, 1) cell, it is

$$n_2\hat{p} = n_2(x_1 + x_2) / (n_1 + n_2)$$

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 ***i*th** row margin multiplied by the ***j*th** column margin, divided by the grand total.

$$E_{11} = \text{expected number of units in the (1, 1) cell} = 3220(2181)/13{,}465 = 521.6$$

$$E_{12} = \text{expected number of units in the (1, 2) cell} = 3220(11{,}284)/13{,}465 = 2698.4$$

$$E_{21} = \text{expected number of units in the (2, 1) cell} = 10{,}245(2181)/13{,}465 = 1659.4$$

$$E_{22} = \text{expected number of units in the (2, 2) cell} = 10{,}245(11{,}284)/13{,}465 = 8585.6$$

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

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

The best test is base on statistic $(O-E)^2/E,$ where O and E are the observed and expected number of units, respectively, in a particular cell.

Yates-Corrected Chi-Square Test for a 2 × 2 Contingency Table Suppose we wish to test the hypothesis H_0: $p_1 = p_2$ versus 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 $X^2 > \chi_{1,1-\alpha}^2$ and accept H_0 if $X^2 \le \chi_{1,1-\alpha}^2$.

(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.

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

Assess the breast cancer data in Example 10.4 using contingency-table approach

$$X^2 = \frac{(|683 - 521.6| - .5)^2}{521.6} + \frac{(|2537 - 2698.4| - .5)^2}{2698.4}$$
$$+ \frac{(|1498 - 1659.4| - .5)^2}{1659.4} + \frac{(|8747 - 8585.6| - .5)^2}{8585.6}$$
$$= 77.89 \sim \chi_1^2 \text{ under } H_0$$

Because $\chi^2_{1,.999} = 10.83 < 77.89 = X^2$

$p < 1 - .999 = .001$

Short Computational Form for the Yates-Corrected Chi-Square Test for 2×2 Contingency Tables Suppose we have the 2×2 contingency table in Table 10.7. The X^2 test statistic in Equation 10.5 can be written

$$X^2 = n\left(|ad - bc| - \frac{n}{2}\right)^2 \Big/ \left[(a+b)(c+d)(a+c)(b+d)\right]$$

Thus the test statistic X^2 depends only on (1) the grand total n, (2) the row and column margins $a + b$, $c + d$, $a + c$, $b + d$, and (3) the magnitude of the quantity $ad - bc$. To compute X^2,

(1) Compute

$$\left(|ad - bc| - \frac{n}{2}\right)^2$$

Start with the first column margin, and proceed counterclockwise.

(2) Divide by each of the two column margins.

(3) Multiply by the grand total.

(4) Divide by each of the two row margins.

General contingency table

a	b	$a + b$
c	d	$c + d$
$a + c$	$b + d$	$n = a + b + c + d$

Two-Sample Test for Binomial Proportions for Matched-Pair Data (McNemar's Test)

Ex 10.21

- Comparing two different chemotherapy treatments for breast cancer, A and B.
  - The two groups should be as comparable as possible on other prognostic factors.
- A matched study
  - The patients are assigned to pairs matched on age and clinical conditions
  - A random member of each matched pair gets treatment A and the other gets treatment B.
  - The patients are followed for 5 years, with survival as the outcome variable.

A 2 × 2 contingency table comparing treatments A and B for breast cancer based on 1242 patients

Treatment	Outcome: Survive for 5 years	Outcome: Die within 5 years	Total
A	526	95	621
B	515	106	621
Total	1041	201	1242

- Yates-corrected chi-square statistic is 0.59, which is not significant.
- Using this test assumes that the samples are independent.

A 2 × 2 contingency table with the matched pair as the sampling unit based on 621 matched pairs

Outcome of treatment A patient	Outcome of treatment B patient: Survive for 5 years	Outcome of treatment B patient: Die within 5 years	Total
Survive for 5 years	510	16	526
Die within 5 years	5	90	95
Total	515	106	621

- Probability that the treatment B member of the pair survived given that the treatment A member of the pair survived = 510/526 = .970
- Probability that the treatment B member of the pair survived given that the treatment A member of the pair died = 5/95 = .053
- Concordant pair
  - A matched pair in which the outcome is the same for each member of the pair.
- Discordant pair
  - A matched pair in which the outcomes differ for the members of the pair.
- Type A discordant pair
  - Treatment A member of the pair has the event and B does not.
- Type B discordant pair
  - Treatment B member of the pair has the event and A does not.

- Let p = probability that a discordant pair is of type A.
- H_0: p = 1/2 versus H_1: $p \neq 1/2$.

McNemar's Test for Correlated Proportions—Normal-Theory Test

(1) Form a 2 × 2 table of matched pairs, where the outcomes for the treatment A members of the matched pairs are listed along the rows and the outcomes for the treatment B members are listed along the columns.

(2) Count the total number of discordant pairs (n_D) and the number of type A discordant pairs (n_A).

(3) Compute the test statistic

$$X^2 = \left(\left|n_A - \frac{n_D}{2}\right| - \frac{1}{2}\right)^2 \Big/ \left(\frac{n_D}{4}\right)$$

An equivalent version of the test statistic is also given by

$$X^2 = \left(|n_A - n_B| - 1\right)^2 / (n_A + n_B)$$

(4) For a two-sided level α test,

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

then reject H_0;

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

then accept H_0.

(5) The exact p-value is given by p-value = $Pr\left(\chi^2_1 \ge X^2\right)$.

(6) Use this test only if $n_D \geq 20$.