
XVII. Two Way Tables

Scenario. You will have two *sets* of categories for your population. Each set divides the population into collectively exhaustive and mutually exclusive categories; the sets are defined by nominative variables describing some attribute of the subjects.

Examples of pairs of attributes might be:

- Gender (Male and Female) and political affiliation (Democrat, Independent, Republican). This is a 2×3 design since there are two categories for the first set of attributes and three for the second.
- Ethnicity (White, Black, Hispanic, Indian, Asian, Other) and educational level (Grammar School, High School, Some College, College Graduate, Post Graduate work). This is a 6×5 design.
- Treatments (*experimental and control*) and outcomes (t success and failure). In this case, one way tables are similar to testing for differences in the means between two independent samples. In this case you would be testing:

H_0 : control and experimental proportions are equal

H_A : control and experimental proportions differ

Data and parameters. Typically one set of attributes will have n categories and the other will have m categories. You will have a sample of size N and will partition the sample into $n \times m$ cells according to the categories. The outcomes $o_{i,j}$ for each cell can be tabulated with one set of categories listed as column headings and another listed as row labels giving an *outcomes* table.

| | | | | | | |
|--------------|-----------|-----------|-----|--------------|-----|--------------|
| | Col1 | Col2 | ... | Col <i>j</i> | ... | Col <i>m</i> |
| Row1 | $o_{1,1}$ | $o_{1,2}$ | ... | $o_{1,j}$ | ... | $o_{1,m}$ |
| Row2 | $o_{2,1}$ | $o_{2,2}$ | ... | $o_{2,j}$ | ... | $o_{2,m}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Row <i>i</i> | $o_{i,1}$ | $o_{i,2}$ | ... | $o_{i,j}$ | ... | $o_{i,m}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Row <i>n</i> | $o_{n,1}$ | $o_{n,2}$ | ... | $o_{n,j}$ | ... | $o_{n,m}$ |

For example, a sample of size 86 with gender and political affiliation might look like:

| | | | |
|--------|----------|-------------|------------|
| | Democrat | Independent | Republican |
| Male | 12 | 13 | 20 |
| Female | 19 | 11 | 11 |

Research Objective. To determine if the row and column categories are independent or dependent. In particular, we will test

H_0 : Row and column effects are independent.

against

H_A : Row and column effects are dependent.

Solution Template

Step 1. If the data is not already presented in tabular form, do so now:

| <i>Outcomes Table</i> | | | | | | |
|-----------------------|-----------|-----------|-----|--------------|-----|--------------|
| | Col1 | Col2 | ... | Col <i>j</i> | ... | Col <i>m</i> |
| Row1 | $o_{1,1}$ | $o_{1,2}$ | ... | $o_{1,j}$ | ... | $o_{1,m}$ |
| Row2 | $o_{2,1}$ | $o_{2,2}$ | ... | $o_{2,j}$ | ... | $o_{2,m}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Row <i>i</i> | $o_{i,1}$ | $o_{i,2}$ | ... | $o_{i,j}$ | ... | $o_{i,m}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Row <i>n</i> | $o_{n,1}$ | $o_{n,2}$ | ... | $o_{n,j}$ | ... | $o_{n,m}$ |

Step 2. Add rows for “totals” and “proportions” and a column for “totals.”

| | Col1 | Col2 | ... | Col j | ... | Col m | Total |
|---------|-----------------------|-----------------------|-----|-----------------------|-----|-----------------------|-------|
| Row1 | $o_{1,1}$ | $o_{1,2}$ | ... | $o_{1,j}$ | ... | $o_{1,m}$ | R_1 |
| Row2 | $o_{2,1}$ | $o_{2,2}$ | ... | $o_{2,j}$ | ... | $o_{2,m}$ | R_2 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Row i | $o_{i,1}$ | $o_{i,2}$ | ... | $o_{i,j}$ | ... | $o_{i,m}$ | R_i |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Row n | $o_{n,1}$ | $o_{n,2}$ | ... | $o_{n,j}$ | ... | $o_{n,m}$ | R_n |
| Totals | C_1 | C_2 | ... | C_j | ... | C_m | N |
| Props | $p_1 = \frac{C_1}{N}$ | $p_2 = \frac{C_2}{N}$ | ... | $p_j = \frac{C_j}{N}$ | ... | $p_m = \frac{C_m}{N}$ | 1 |

Step 3. Use the above table to construct an “expectations” table. This is what we would *expect* to happen if the null hypothesis were true.

| | Col1 | Col2 | ... | Col j | ... | Col m |
|---------|---------------------|---------------------|-----|---------------------|-----|---------------------|
| Row1 | $e_{1,1} = p_1 R_1$ | $e_{1,2} = p_2 R_1$ | ... | $e_{1,j} = p_j R_1$ | ... | $e_{1,m} = p_m R_1$ |
| Row2 | $e_{2,1} = p_1 R_2$ | $e_{2,2} = p_2 R_2$ | ... | $e_{2,j} = p_j R_2$ | ... | $e_{2,m} = p_m R_2$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Row i | $e_{i,1} = p_1 R_i$ | $e_{i,2} = p_2 R_i$ | ... | $e_{i,j} = p_j R_i$ | ... | $e_{i,m} = p_m R_i$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Row n | $e_{n,1} = p_1 R_n$ | $e_{n,2} = p_2 R_n$ | ... | $e_{n,j} = p_j R_n$ | ... | $e_{n,m} = p_m R_n$ |

Step 4. Find the value of the test statistic:

$$\text{test statistic} = \sum_{i,j} \frac{(o_{i,j} - e_{i,j})^2}{e_{i,j}}$$

Step 5. Find the degrees of freedom for the test statistic:

$$\text{degrees of freedom} = (\# \text{ of rows} - 1) \times (\# \text{ of cols} - 1)$$

Step 6. Find the cutoff in the chi squared table (Table A-4, page 666 of the text).

Step 7. *Decision Rule.* If the value of the test statistic is *larger than* the cutoff, then *reject* the null hypothesis and accept the alternative. Otherwise, accept the null

hypothesis.

End of Solution Template

Example. A researcher randomly selects 1000 death certificates and, after interviewing the attending physician, records the following information about the deceased:

| | Cancer | Heart Disease | Other |
|-----------|--------|---------------|-------|
| Smoker | 135 | 310 | 205 |
| Nonsmoker | 55 | 155 | 140 |

At a significance of level of 5%, do these data show that smoking and cause of death are dependent?

Note: the data can't show that smoking *causes* death since everyone in the sample is already dead. What the data can show is that dying of cancer or heart disease is related to whether or not the deceased smoked.

Solution.

Step 1. The data are already presented in the proper format.

Step 2. Expand the *observations* table to include totals and proportions:

| <i>Observed Proportions</i> | | | | |
|-----------------------------|--------|---------------|-------|--------|
| | Cancer | Heart Disease | Other | totals |
| Smoker | 135 | 310 | 205 | 650 |
| Nonsmoker | 55 | 155 | 140 | 350 |
| totals | 190 | 465 | 345 | 1000 |
| props | .19 | .465 | .345 | 1 |

Step 3. Build the *expectations* table:

| <i>Expectations</i> | | | | |
|---------------------|--------|---------------|-------|--------|
| | Cancer | Heart Disease | Other | totals |
| Smoker | | | | 650 |
| Nonsmoker | | | | 350 |
| totals | 190 | 465 | 345 | 1000 |

The *cells* of the expectations table are initially blank; you will have to fill them in with computations. You use the proportion row from step two to fill in the cells. The number which goes in the cells is what *you would expect the result to be if the row and column effects were independent*. Since 19% of all deaths were attributable to cancer, if “cancer” and “smoking” were unrelated, we would expect that 19% of all smokers’ deaths would be caused by cancer. Thus, the upper left cell in the expectations table is

$$19\% \text{ of } 650 = 123.5$$

Similarly, the upper middle cell in the expectations table is

$$46.5\% \text{ of } 650 = 302.25$$

and the upper right cell is

$$34.5\% \text{ of } 650 = 224.25$$

More generally, the cells in the expectations table are filled in as follows:

| Expectations Table | | | | |
|--------------------|------------------|-------------------|-------------------|--------|
| | Cancer | ♥ Disease | Other | totals |
| Smkr | $.19 \times 650$ | $.465 \times 650$ | $.345 \times 650$ | 650 |
| NonSmkr | $.19 \times 350$ | $.465 \times 350$ | $.345 \times 350$ | 350 |
| totals | 190 | 465 | 345 | 1000 |

which results in an expectations table which looks like:

| Expectations Table | | | | |
|--------------------|--------|-----------|--------|--------|
| | Cancer | ♥ Disease | Other | totals |
| Smoker | 123.5 | 302.25 | 224.25 | 650 |
| Nonsmoker | 66.5 | 162.75 | 120.75 | 350 |
| totals | 190 | 465 | 345 | 1000 |

Notice that the rows and columns still add up to the marginal totals. This table gives what we would *expect* to observe if the row and column effects were independent. Notice that this differs from our actual observations:

| Observations | | | | |
|--------------|--------|---------------|-------|--------|
| | Cancer | Heart Disease | Other | totals |
| Smoker | 135 | 310 | 205 | 650 |
| Nonsmoker | 55 | 155 | 140 | 350 |
| totals | 190 | 465 | 345 | 1000 |

Step 4. The next step in the process is to compute the test statistic:

$$\chi^2 = \sum \frac{(\text{Observations} - \text{Expectations})^2}{\text{Expectations}}$$

the sum being taken over each data cell in the contingency tables. Thus in our example there are six terms to sum:

$$\begin{aligned} \chi^2 &= \sum \frac{(\text{Observations} - \text{Expectations})^2}{\text{Expectations}} \\ &= \frac{(135 - 123.5)^2}{123.5} + \frac{(310 - 302.25)^2}{302.25} + \dots \\ &\dots + \frac{(205 - 224.25)^2}{224.25} + \frac{(55 - 66.5)^2}{66.5} + \dots \\ &\dots + \frac{(155 - 162.75)^2}{162.75} + \frac{(140 - 120.75)^2}{120.75} \\ &= 1.07 + 0.199 + 1.652 + 1.989 + 0.36 + 3.069 \\ &= 8.349 \end{aligned}$$

Step 5. Compute the degrees of freedom:

$$\text{degrees of freedom} = (\# \text{ of rows} - 1) \times (\# \text{ of cols} - 1)$$

Thus in our problem the degrees of freedom are

$$(2 - 1) \times (3 - 1) = 2$$

Step 6. Find the cutoff in Table A-4, page 666: The degrees of freedom tell you the *row* in the table in which you need to look. The entries across the top correspond

(for this type of problem) to the significance level. Thus the cutoff for this problem is 5.991.

Step 7. The decision rule is:

Reject the null hypothesis if the test statistic is larger than the cutoff.

In our case the test statistic has value 8.349; since this *is larger than* the cutoff (5.991) we *reject* the null hypothesis. This means that the data are statistically significant and we believe that “cause of death” and “smoking” are related. **I**

Problems

1. Given the following two way table, test at the 5% level whether or not the row and column effects are independent.

| | A | B | C | D |
|-----|----|----|----|----|
| I | 15 | 19 | 32 | 12 |
| II | 16 | 22 | 13 | 8 |
| III | 12 | 3 | 5 | 8 |

2. Given the following two way table, test at the 5% level whether or not the row and column effects are independent.

| | A | B | C |
|----|----|----|----|
| I | 45 | 16 | 23 |
| II | 18 | 13 | 4 |

3. “Boot camps” are sometimes proposed as a rehabilitation technique for young offenders. In one study, 149 young offenders who completed boot camps (as opposed to conventional incarceration) were followed; after six months 76 had committed another offense. During the same six month period, 1,360 young offenders who had completed incarcerations in Junville Hall were followed; among this latter group 768 committed another offense. Is this evidence (at the 5% level) that boot camps are more effective than incarceration?

Video Assignment.

View the following program(s) from the series *Against All Odds*:

| Program | Title |
|---------|---------------------------------------|
| 10 | <i>(Inference for two-way tables)</i> |