Applied Categorical & Nonnormal Data Analysis

Contingency Tables

A table that cross classifies two variable is called a two-way contingency table. It is also known as a cross tabulation or crosstabs for short. If each of the two variables has two levels then the table is a 2x2. If there are three levels of one variable and 5 of the other, it would be a 3x5 table. We will start off by looking at a 2x2 table.

Observed Frequencies

The following table gives a representation of the observed frequencies of a 2x2 contingency table.

row var	column variable col 1	col 2	Total
row 1	n11	n12	n1+
row 2	n21	n22	n2+
Total	n+1	n+2	n

Here is what the observed frequencies look like for an example using myocardial infarction and the use of aspirin.

group	myocardial infarction yes	no	Total
placebo	189	10845	11034
aspirin	104	10933	11037
Total	293	21778	22071

The values in the body of the table represent the joint distribution and the values around the edges represent the marginal distributions.

Observed Proportions

Here is a representation of the observed proportions which can also be treated as probabilities.

row var	column variable col 1	col 2	Total
row 1	p11	p12	p1+
row 2	p21	p22	p2+
Total	p+1	p+2	1.0

The observed proportions for our example look like this:

group	myocardial infarction yes	no	Total

```
   placebo |     .0086    .4914     |    .4999
   aspirin |     .0047    .4954     |    .5001
-----------+----------------------+--------
     Total |     .0133    .9867     |   1.0000
```

Relative Risk

The relative risk in a 2x2 table is the ratio of "success" probabilities for the two groups. For the MI example, it would look like this.

RR = p11/p21 = .0086/.0047 = 1.82

In this example, the sample proportion of myocardial infarction was 82% higher for the placebo group. If you take the reciprocal of the relative risk the value is .55. The proportion of myocardial infarction was 45% lower for the aspirin group.

Odds Ratio

Before we can talk about odds ratios we need to define odds.

odds = p/(1 - p)

Theoretically, odds can run from 0 to positive infinity. When the odds equal one, the probability of success is equal to the probability of failure. When the odds are less than one, the probability of success is less than the probability of failure. And, when the odds are greater than one, the probability of success is greater than the probability of failure. An odds ratio is exactly what it seems, the ratio of two odds.

$$OR = \frac{odds_1}{odds_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{.0086/.9914}{.0047/.9954} = 1.832$$

This is not the only way to compute the odds ratio. It is easier to compute it as a ratio of the cross products of either the frequencies or the proportions.

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{189 * 10933}{10845 * 104} = 1.832$$

$$OR = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{.0086 * .4954}{.4914 * .0047} = 1.832$$

When the odds ratio equal 1, the odds for group 1 are the same as the odds for groups 2. When the odds ratio is greater than 1, the odds for group 1 are greater than the odds for groups 2. When the odds ratio is less than 1, the reverse is true. The farther odds ratio goes in either direction, the stronger the association among the variables.

In this example, the odds of a myocardial infarction are 83% higher for the placebo group. If you take the reciprocal of the odds ratio the value is .546. Thus, the odds of myocardial infarction was about 45% lower for the aspirin group than for the placebo group.

Odds ratios are invariant when the orientation of the rows and columns reversed. The odds ratios are relatively invariant to changes in the marginal frequencies. For example, if you were to multiply each of the frequencies in the table by a constant, c, the odds ratio would remain unchanged.

$$OR = \frac{cn_{11}cn_{22}}{cn_{12}cn_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

The same is true if you multiply the frequencies for one row by one constant and the frequencies in the other row by a different constant.

$$OR = \frac{cn_{11}dn_{22}}{dn_{12}cn_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Relation of Relative Risk to Odds Ratio

When p1 and p2 are both very small, the value of the odds ratio is close to that of the relative risk, In any case, the odds ratio can be obtained from the relative risk by the following formula.

$$OR = RR * \frac{(1-p_2)}{(1-p_1)}$$

This is useful because there are times when it isn't possible to estimate relative risk directly.

Conditional Probabilities

Conditional probabilities are the probabilities of an event given that some other event has occurred. In our MI example, the conditional probabilities for the groups are:

group	myocardial infarction yes	no	Total
placebo	.0171	.9829	1.0000
aspirin	.0094	.9906	1.0000
Total	.0133	.9867	1.0000

Recall that,

group	myocardial infarction yes	no	Total
placebo	.0086	.4914	.4999
aspirin	.0047	.4954	.5001
Total	.0133	.9867	1.0000

Thus, the conditional probability of a myocardial infarction for the placebo group is .0171, while for the aspirin group in is .0094.

Two variables are said to be independent when the conditional distributions of one are identical for each level of the other. In this example, the conditional distributions are not identical.

Expected Frequencies

Here are the expected frequencies for our example given independence of group and myocardial infarction.

group	myocardial infarction yes	no	Total
placebo	146.480	10887.520	11034
aspirin	146.520	10890.480	11037
Total	293	21778	22071

Note that the marginal frequencies are that same as in the table of the observed frequencies. This is the case because the expected frequencies are obtained form the marginal distribution of the observed frequencies. For example, the expected frequency of 146.480 is obtained as follows:

$\mathbf{e_{ij} = (n_{i+})(n_{+j})/n_{++}}$ = 11034*293/22071 = 146.480

where $\mathbf{n_{i+}}$ is the frequency for the ith row, $\mathbf{n_{+j}}$ is the frequency for the jth column, and $\mathbf{n_{++}}$ is the total frequency for the entire table.

What this means, is that, the joint distribution is determined by the marginal distribution of the variables when the two variables are independent.

This property is just a variation of the rule for the joint probability of independent events **P(A & B) = P(A)*P(B)**.

Chi-Squared Statistic

In two-way contingency tables chi-squared is used to test the independence of the two marginal variables. The chi-squared test is often called a goodness-of-fit test but is perhaps better thought of as a badness-of-fit test, because a large value of chi-squared is indicative of a bad fit between the observed and expected frequencies.

There are two commonly computed chi-squared statistics; the Pearson chi-squared (χ^2) and the likelihood ratio chi-squared ($\mathbf{G^2}$)

$$\chi^2 = \sum \frac{\left(x_{ij} - e_{ij}\right)^2}{e_{ij}}$$

$$G^2 = 2\sum x_{ij} \ln\left(\frac{x_{ij}}{e_{ij}}\right)$$

with degrees of freedom = (I-1)(J-1)

Asymptotically, χ^2 and $\mathbf{G^2}$ are equivalent. However, in finite samples there can be a considerable difference the estimates of these two statistics.

Stata Examples

```
use http://www.gseis.ucla.edu/courses/data/hsb2

tabulate ses prog, all
```

```
           |          type of program
       ses |   general   academic   vocation |     Total
-----------+---------------------------------+----------
       low |        16         19         12 |        47
    middle |        20         44         31 |        95
      high |         9         42          7 |        58
-----------+---------------------------------+----------
     Total |        45        105         50 |       200

          Pearson chi2(4) =  16.6044   Pr = 0.002
 likelihood-ratio chi2(4) =  16.7830   Pr = 0.002
```

```
               Cramer's V =   0.2037
                    gamma =   0.0109  ASE = 0.097
          Kendall's tau-b =   0.0069  ASE = 0.062
```

tabulate ses prog, cell nofreq

```
           |          type of program
       ses |   general   academic   vocation |     Total
-----------+---------------------------------+----------
       low |      8.00       9.50       6.00 |     23.50
    middle |     10.00      22.00      15.50 |     47.50
      high |      4.50      21.00       3.50 |     29.00
-----------+---------------------------------+----------
     Total |     22.50      52.50      25.00 |    100.00
```

tabulate ses prog, row nofreq

```
           |          type of program
       ses |   general   academic   vocation |     Total
-----------+---------------------------------+----------
       low |     34.04      40.43      25.53 |    100.00
    middle |     21.05      46.32      32.63 |    100.00
      high |     15.52      72.41      12.07 |    100.00
-----------+---------------------------------+----------
     Total |     22.50      52.50      25.00 |    100.00
```

tabchi ses prog

```
          observed frequency
          expected frequency

----------------------------------------
          |       type of program
      ses |  general  academic  vocation
----------+-----------------------------
      low |       16        19        12
          |   10.575    24.675    11.750
          |
   middle |       20        44        31
          |   21.375    49.875    23.750
          |
     high |        9        42         7
          |   13.050    30.450    14.500
----------------------------------------

          Pearson chi2(4) =  16.6044   Pr = 0.002
 likelihood-ratio chi2(4) =  16.7830   Pr = 0.002
```

tabchi ses prog,raw pearson cont adjust noo noe

```
          raw residual
          Pearson residual
          contribution to chi-square
          adjusted residual

----------------------------------------
          |       type of program
      ses |  general  academic  vocation
----------+-----------------------------
      low |    5.425    -5.675     0.250
```

```
          |    1.668    -1.142     0.073
          |    2.783     1.305     0.005
          |    2.167    -1.895     0.096
          |
   middle |   -1.375    -5.875     7.250
          |   -0.297    -0.832     1.488
          |    0.088     0.692     2.213
          |   -0.466    -1.666     2.371
          |
     high |   -4.050    11.550    -7.500
          |   -1.121     2.093    -1.970
          |    1.257     4.381     3.879
          |   -1.511     3.604    -2.699
----------------------------------------

          Pearson chi2(4) =  16.6044   Pr = 0.002
 likelihood-ratio chi2(4) =  16.7830   Pr = 0.002
```

A note about tetrachoric correlations

Tetrachoric correlations measure the association between two dichotomous variables by estimating the correlation between their associated latent variables.

The **tabulate** command includes an estimate of phi, a measure of association between dichotomous variables. Stata, in the 2x2 case, labels phi as "Cramer's V." The same coefficient can be obtained by computing a standard correlation correlation between the two variables.

The **tetrac** command (**findit tetrac**) available from ATS uses an approximation of the tetrachoric correlations due to Edwards (1957).

let $\alpha = ad/bc$
then $r = (\alpha^{\pi/4}+1)/(\alpha^{\pi/4}-1)$

The tetrachoric correlations are often larger than the phi coefficients for the same variables.

```
use http://www.gseis.ucla.edu/courses/data/tetra

tabulate hon sci, all

           |          sci
       hon |         0          1 |     Total
-----------+----------------------+----------
         0 |       111         36 |       147
         1 |        22         31 |        53
-----------+----------------------+----------
     Total |       133         67 |       200

          Pearson chi2(1) =  20.2150   Pr = 0.000
 likelihood-ratio chi2(1) =  19.4693   Pr = 0.000
               Cramer's V =   0.3179
                    gamma =   0.6258  ASE = 0.103
          Kendall's tau-b =   0.3179  ASE = 0.072

corr  hon sci
(obs=200)

             |      hon      sci
-------------+------------------
         hon |   1.0000
```

```
         sci |   0.3179   1.0000
```

tetrac hon sci

```
(obs=200)

Approximate Tetrachoric Correlations

        hon     sci
hon  1.0000
sci  0.5204  1.0000
```

tetrac female schtyp ses hon sci

```
(obs=200)

Approximate Tetrachoric Correlations

         female   schtyp      ses      hon      sci
female   1.0000
schtyp  -0.0331   1.0000
   ses  -0.2844  -0.5840   1.0000
   hon   0.2504   0.0365   0.0837   1.0000
   sci  -0.2616  -0.1434   0.2996   0.5204   1.0000
```