Topic 21
Goodness of Fit
Contingency Tables

Outline

Introduction

Contingency tables, also known as two-way tables or cross tabulations are a convenient way to display the frequency distribution from the observations of two categorical variables. For an $r \times c$ contingency table, we consider two factors A and B for an experiment. This gives r categories

$$A_1, \ldots A_r$$

for factor A and c categories

$$B_1, \ldots B_c$$

for factor B

Two-way Table

Here, we write O_{ij} to denote the number of occurrences for which an individual falls into both category A_i and category B_j. The results is then organized into a two-way table.

	B_1	B_2	$\cdots$	B_c	total
A_1	O_{11}	O_{12}	$\cdots$	O_{1c}	$O_{1\cdot}$
A_2	O_{21}	O_{22}	$\cdots$	O_{2c}	$O_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$O_{r\cdot}$
total	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot c}$	n

where $O_{i\cdot}, i = 1, \ldots, r$ are the row marginals, $O_{\cdot j}, i = j, \ldots, c$ are the column marginals, and n is the number of observations.

Smoking Habits

Returning to the study of the smoking habits of 5375 high school children in Tucson in 1967, here is a two-way table summarizing some of the results.

	student smokes	student does not smoke	total
2 parents smoke	400	1380	1780
1 parent smokes	416	1823	2239
0 parents smoke	188	1168	1356
total	1004	4371	5375

The Hypothesis

For a contingency table, the null hypothesis we shall consider is that the factors A and B are independent. To set the parameters for this model, we define

$$p_{ij} = P\{\text{an individual is simultaneously a member of category } A_i \text{ and category } B_j\}.$$

Then, we have the parameter space

$$\Theta = \{\mathbf{p} = (p_{ij}, 1 \le i \le r, 1 \le j \le c); p_{ij} \ge 0 \text{ for all } i, j = 1, \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = 1\}.$$

Write the marginal distribution

$$p_{i\cdot} = \sum_{j=1}^{c} p_{ij} = P\{\text{an individual is a member of category } A_i\}$$

and

$$p_{\cdot j} = \sum_{i=1}^{r} p_{ij} = P\{\text{an individual is a member of category } B_j\}.$$

The Test Statistic

The null hypothesis of independence of the categories A and B can be written

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}, \text{ for all } i, j \quad \text{versus} \quad H_1 : p_{ij} \neq p_{i\cdot}p_{\cdot j}, \text{ for some } i, j.$$

The null hypothesis $p_{ij} = p_{i\cdot}p_{\cdot j}$ can be written in terms of observed and expected observations as

$$\frac{E_{ij}}{n} = \frac{O_{i\cdot}}{n}\frac{O_{\cdot j}}{n} \qquad \text{or} \qquad E_{ij} = \frac{O_{i\cdot}O_{\cdot j}}{n}.$$

As before, the appropriate G^2 statistic follows from the likelihood ratio test criterion. The χ^2 statistic is a second order Taylor series approximation to G^2.

$$G^2 = -2\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \ln \frac{E_{ij}}{O_{ij}} \approx \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \chi^2.$$

Smoking Habits

For the data set on smoking habits in Tucson, we find that the expected table is

	student smokes	student does not smoke	total
2 parents smoke	332.49	1447.51	1780
1 parent smokes	418.22	1820.78	2239
0 parents smoke	253.29	1102.71	1356
total	1004	4371	5375

For example,

$$E_{11} = \frac{O_{1\cdot} O_{\cdot 1}}{n} = \frac{1780 \cdot 1004}{5375} = 332.49.$$

Degrees of Freedom

To determine the degrees of freedom, start with a contingency table with no entries but with the prescribed marginal values.

	B_1	B_2	$\cdots$	B_c	total
A_1					$O_{1\cdot}$
A_2					$O_{2\cdot}$
$\vdots$					$\vdots$
A_r					$O_{r\cdot}$
total	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot c}$	n

The degrees of freedom is the number of values that we can place on the table *before* all the remaining values are determined. Note that we can fill $c-1$ values in each of the $r-1$ rows before the remaining values are determined. Thus, the degrees of freedom is $(r-1)\times(c-1)$.

Exercise. Determine the number of degrees of freedom and compute the χ^2 statistic for the example on smoking habits.

Performing the Test

To perform the χ^2 test in R,

```
> smoking<-matrix(c(400,416,188,1380,1823,1168),nrow=3)
> smoking
     [,1] [,2]
[1,]  400 1380
[2,]  416 1823
[3,]  188 1168
> chisq.test(smoking)

Pearson's Chi-squared test

data:  smoking
X-squared = 37.5663, df = 2, p-value = 6.959e-09
```

Introduction

We can look at the residuals

$$\frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

for the entries in the χ^2 test as follows.

```
> smokingtest<-chisq.test(smoking)
> residuals(smokingtest)
            [,1]         [,2]
[1,]  3.7025160 -1.77448934
[2,] -0.1087684  0.05212898
[3,] -4.1022973  1.96609088
```

Exercise. Make three horizontally placed chigrams that summarize the residuals for this χ^2 test in the example above. Use this to explain the sources of the major contribution to the χ^2 statistic.