# Chapter 11
# Asymptotic Evaluations
## 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

$$A_1, \dots A_r$$

for factor $A$ and $c$ categories

$$B_1, \dots B_c$$

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.

# Blood Type

In addition to blood types $O$, $A$, $B$, and $AB$, the use of blood requires knowledge of the $Rh$ factor. This could bepositive ($Rh+$) or negative ($Rh-$). The data collected from a random sample of 300 from a large population is presented in a contingency table.

| | $O$ | $A$ | $AB$ | $O$ | total |
|---|---|---|---|---|---|
| $Rh+$ | 92 | 89 | 54 | 19 | 244 |
| $Rh-$ | 13 | 27 | 7 | 9 | 56 |
| total | 95 | 116 | 61 | 28 | 300 |

# 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

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

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

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

- For the parameter space $\Theta$, we have $r \times c$ probabilities $p_{ij}$ with the single constraint that their sum is $1$. Thus, $\dim(\Theta) = rc - 1$.
- For the null hypothesis space $\Theta_0$, we have $r$ row probabilities $p_{i\cdot}$ with the constraint that the sum is $1$ and $c$ column probabilities $p_{\cdot j}$ with the constraint that the sum is $1$. Thus, $\dim(\Theta_0) = (r-1) + (c-1)$.

Thus,

$$\begin{aligned} \dim(\Theta) - \dim(\Theta_0) &= rc - 1 - (r-1) - (c-1) \\ &= rc - r - c + 1 = (r-1)(c-1). \end{aligned}$$

# The Test Statistic

The data $\mathbf{n} = \{n_{ij}; 1 \leq i \leq r, 1 \leq j \leq c\}$, the number of observations that lie simultaneously in category $A_i$ and $B_j$ is a sufficient statistic. First we maximize the likelihood

$$L(\mathbf{p}|\mathbf{n}) = \prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{n_{ij}}$$

for $\Theta$. As before, using Lagrange multipliers, we find that the maximum likelihood estimate

$$\hat{p}_{ij} = \frac{n_{ij}}{n}$$

is simply the fraction of observations that are in categories $A_i$ and $B_j$. Here $n$ is the total number of observations. Thus,

$$\log L(\hat{\mathbf{p}}|\mathbf{n}) = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \log \frac{n_{ij}}{n}.$$

# The Test Statistic

To maximize under the null hypothesis note that $p_{0,ij} = p_{i\cdot} \cdot p_{\cdot j}$ and therefore

$$L(\mathbf{p}_0|\mathbf{n}) = \prod_{i=1}^{r}\prod_{j=1}^{c}(p_{i\cdot}p_{\cdot j})^{n_{ij}} = \prod_{i=1}^{r}\prod_{j=1}^{c} p_{i\cdot}^{n_{ij}} p_{\cdot j}^{n_{ij}} = \prod_{i=1}^{r} p_{i\cdot}^{n_{i\cdot}} \cdot \prod_{j=1}^{c} p_{\cdot j}^{n_{\cdot j}}.$$

We now have two maximization problems, for $p_{i\cdot}$ and $p_{\cdot j}$. Again, we return to the Lagrange multiplier strategy to determine the maximum likelihood estimates.

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} \quad \text{and} \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}.$$

Thus, the maximum likelihood estimate under the null hypothesis

$$\hat{p}_{0,ij} = \hat{p}_{i\cdot} \cdot \hat{p}_{\cdot j} = \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}$$

and

$$\log L(\hat{\mathbf{p}}_0|\mathbf{n}) = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \log\left(\frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}\right).$$

# The Test Statistic

Next we subtract to find the logarithm of the likelihood ratio.

$$\begin{aligned}\log \Lambda(\mathbf{n}) &= \log L(\hat{\mathbf{p}}_0|\mathbf{n}) - \log L(\hat{\mathbf{p}}|\mathbf{n}) = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\left(\log\left(\frac{n_{i\cdot}}{n}\cdot\frac{n_{\cdot j}}{n}\right) - \log\frac{n_{ij}}{n}\right).\\ &= \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\left(\log\frac{n_{i\cdot}n_{\cdot j}}{n} - \log n_{ij}\right) = \sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij}\log\frac{E_{ij}}{O_{ij}}\end{aligned}$$

Here, we write $O_{ij} = n_{ij}$ for the observed observations in the $ij$-th cell and

$$E_{ij} = n\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n} = \frac{O_{i\cdot}O_{\cdot j}}{n}.$$

Multiply by $-2$ to obtain the desired expression for $G^2$ as the likelihood ratio test statistic.

# Hemoglobin Data

For the data set on blood type, we find that the expected table is

| | $A$ | $B$ | $AB$ | $O$ | total |
|---|---|---|---|---|---|
| $Rh+$ | 77.27 | 94.35 | 49.61 | 22.77 | 244 |
| $Rh-$ | 17.73 | 21.65 | 11.39 | 5.23 | 56 |
| total | 95 | 116 | 61 | 28 | 300 |

For example,

$$E_{11} = \frac{O_{1\cdot} O_{\cdot 1}}{n} = \frac{244 \cdot 95}{300} = 77.27.$$

# 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 can also be determined by counting 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 $\chi^2$ statistic for the example on blood types.

# Performing the Test

To perform the $\chi^2$ test in R,

```
> blood<-matrix(c(82,13,89,27,54,7,19,9),nrow=2)
> blood
     [,1] [,2] [,3] [,4]
[1,]   82   89   54   19
[2,]   13   27    7    9
> chisq.test(blood)

Pearson's Chi-squared test

data:  blood
X-squared = 8.6037, df = 3, p-value = 0.03505
```

# Introduction

We can look at the residuals

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

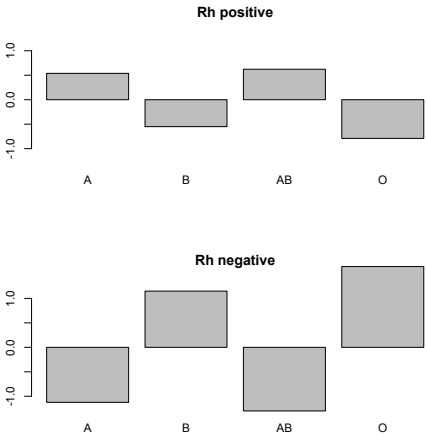
for the entries in the $\chi^2$ test as follows.

```
> bloodtest<-chisq.test(blood)
> resid(bloodtest)
           [,1]       [,2]       [,3]       [,4]
[1,]  0.5384818 -0.5504525  0.6227811 -0.7907002
[2,] -1.1240145  1.1490019 -1.2999790  1.6504895
```

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

# Introduction

```
> residuals<-resid(bloodtest)
> colnames(residuals)
   <-c("A","B","AB","O")
> par(mfrow=c(2,1))
> barplot(residuals[1,],
   ylim=c(-1.2,1.2),main="Rh positive")
> barplot(residuals[2,],
   ylim=c(-1.2,1.2),main="Rh negative")
```



# Applicability and Alternatives to Chi-squared Tests

The chi-square test uses the central limit theorem and so is based on the ability to use a normal approximation. One criterion, the Cochran conditions requires no cell has count zero, and more than 80% of the cells have counts at least 5. If this does not hold, then Fisher's exact test uses the hypergeometric distribution (or its generalization) directly rather than normal approximation.

For example, for the $2 \times 2$ table,

| | $B_1$ | $B_2$ | total |
|---|---|---|---|
| $A_1$ | $O_{11}$ | $O_{12}$ | $O_{1\cdot}$ |
| $A_2$ | $O_{21}$ | $O_{22}$ | $O_{2\cdot}$ |
| total | $O_{\cdot 1}$ | $O_{\cdot 2}$ | $n$ |

# Fisher's Exact Test

The idea behind Fisher's exact test is to begin with an empty table:

| | $B_1$ | $B_2$ | total |
|---|---|---|---|
| $A_1$ | | | $O_{1\cdot}$ |
| $A_2$ | | | $O_{2\cdot}$ |
| total | $O_{\cdot 1}$ | $O_{\cdot 2}$ | $n$ |

and a null hypothesis of equally likely outcomes. We will use as an analogy the model of mark and recapture. Normally the goal is to find $n$, the total population. In this case, we assume that this population size is known and will consider the case that the individuals in the two captures are independent. This is assumed in the mark and recapture protocol. Here we test this independence.

# Fisher's Exact Test

- $A_1$ - an individual in the first capture and thus tagged.
- $A_2$ - an individual not in the first capture and thus not tagged.
- $B_1$ - an individual in the second capture.
- $B_2$ - an individual not in the second capture.

Then, from the point of view of the $A$ classification:

- $O_{1\cdot}$ has the $A_1$ classification ( tagged individuals). This can be accomplished in

$$\binom{n}{O_{1\cdot}} = \frac{n!}{O_{1\cdot}!O_{2\cdot}!}$$

ways. $O_{2\cdot} = n - O_{1\cdot}$ have the $A_2$ classification (untagged individuals).

- From the $O_{\cdot 1}$ belonging to category $B_1$ (individuals in the second capture), $O_{11}$ also belong to $A_1$ ( textcolortealhave a tag). This outcome can be accomplished in

$$\binom{O_{\cdot 1}}{O_{11}} = \frac{O_{\cdot 1}!}{O_{11}!O_{21}!}$$

ways.

# Fisher's Exact Test

- From the $O_{\cdot 2}$ belonging to $B_2$ (individuals not in the second capture), $O_{12}$ also belong to $A_1$ (have a tag). This outcome can be accomplished in

$$\binom{O_{\cdot 2}}{O_{21}} = \frac{O_{\cdot 2}!}{O_{12}!O_{22}!}$$

ways.

Under the null hypothesis that every individual can be placed in any group, provided we have the given marginal information. The probability has a hypergeometric distribution

$$\frac{\binom{O_{1\cdot}}{O_{11}}\binom{O_{2\cdot}}{O_{21}}}{\binom{n}{O_{\cdot 1}}} = \frac{O_{\cdot 1}!/(O_{11}!O_{21}!)\cdot O_{\cdot 2}!/(O_{12}!O_{22}!)}{n!/(O_{1\cdot}!O_{2\cdot}!)} = \frac{O_{\cdot 1}!O_{\cdot 2}!O_{1\cdot}!O_{2\cdot}!}{O_{11}!O_{12}!O_{21}!O_{22}!n!}.$$

Notice that the formula is symmetric in the column and row variables. Thus, if we had derived the hypergeometric formula from the point of view of the $B$ classification we would have obtained exactly the same formula.

# Fisher's Exact Test

To complete the exact test, we rely on statistical software to do the following:

- compute the hypergeometric probabilities over all possible choices for entries in the cells that result in the given marginal values, and
- rank these probabilities from most likely to least likely.
- Find the ranking of the actual data.
- For a one-sided test of too rare, the $p$-value is the sum of probabilities of the ranking lower than that of the data.

A similar procedure applies to provide the Fisher exact test for $r \times c$ tables.

# Fisher's Exact Test

As a test of the assumptions for mark and recapture, we examine a small population of 120 fish. The assumption are that each group of fish are equally likely to be capture in the first and second capture and that the two captures are independent. This could be violated, for example, if the tagged fish are not uniformly dispersed in the pond.
Twenty-five are tagged and returned to the pond. For the second capture of 30, seven are tagged. With this information, given in brown in the table below, we can complete the remaining entries.

| | in 2nd capture | not in 2nd capture | total |
|---|---|---|---|
| in 1st capture | 7 | 18 | 25 |
| not in 1st capture | 23 | 72 | 95 |
| total | 30 | 90 | 120 |

# Fisher's Exact Test

```
> fish<-matrix(c(7,23,18,72),ncol=2)
> fisher.test(fish)

Fisher's Exact Test for Count Data

data:  fish
p-value = 0.7958
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3798574 3.5489546
sample estimates:
odds ratio
  1.215303
```

Fisher's exact test show a much too high $p$-value to reject the null hypothesis.