# Inference for Two-Way Tables

*Chapter 8, Lab 2: Solutions*

*OpenIntro Biostatistics*

**Topics**

- The $\chi^2$ test for independence
- Measures of association in two-by-two tables

This lab generalizes inference for binomial proportions to the setting of two-way contingency tables. Hypothesis testing in a two-way table assesses whether the two variables of interest are associated; this approach can be applied to settings with two or more groups and for responses that have two or more categories. Measures of association in two-by-two tables are also discussed.

The material in this lab corresponds to Sections 8.3 and 8.5 in *OpenIntro Biostatistics*.

## Introduction

*The $\chi^2$ test of independence*

In the $\chi^2$ test of independence, the observed number of cell counts are compared to the number of **expected** cell counts, where the expected counts are calculated under the null hypothesis.

- $H_0$: the row and column variables are not associated
- $H_A$: the row and column variables are associated

The expected count for the $i^{th}$ row and $j^{th}$ column is

$$E_{i,j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{n},$$

where $n$ is the total number of observations.

Assumptions for the $\chi^2$ test:

- *Independence*. Each case that contributes a count to the table must be independent of all other cases in the table.
- *Sample size*. Each expected cell count must be greater than or equal to 10. For tables larger than $2 \times 2$, it is appropriate to use the test if no more than 1/5 of the expected counts are less than 5, and all expected counts are greater than 1.

The $\chi^2$ **test statistic** is calculated as

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}},$$

and is approximately distributed $\chi^2$ with degrees of freedom $(r-1)(c-1)$, where $r$ is the number of rows and $c$ is the number of columns. $O_{i,j}$ represents the observed count in row $i$, column $j$.

For each cell in a table, the **residual** equals

$$\frac{O_{i,j} - E_{i,j}}{\sqrt{E_{i,j}}}.$$

Residuals with a large magnitude contribute the most to the $\chi^2$ statistic. If a residual is positive, the observed value is greater than the expected value; if a residual is negative, the observed value is less than the expected.

*Measures of association in two-by-two tables*

Unit 1 introduced the **relative risk (RR)**, a measure of the risk of a certain event occurring in one group relative to the risk of the event occurring in another group, as a numerical summary for two-by-two ($2 \times 2$) tables. The relative risk can also be thought of as a measure of association.

Consider the following hypothetical two-by-two table. The relative risk of Outcome A can be calculated by using either Group 1 or Group 2 as the reference group:

| | Outcome A | Outcome B | Sum |
|---|---|---|---|
| Group 1 | $a$ | $b$ | $a+b$ |
| Group 2 | $c$ | $d$ | $c+d$ |
| Sum | $a+c$ | $b+d$ | $a+b+c+d=n$ |

Table 1: A hypothetical two-by-two table of outcome by group.

$$RR_{\text{A, comparing Group 1 to Group 2}} = \frac{a/(a+b)}{c/(c+d)}$$

$$RR_{\text{A, comparing Group 2 to Group 1}} = \frac{c/(c+d)}{a/(a+b)}$$

The relative risk is only valid for tables where the proportions $a/(a+b)$ and $c/(c+d)$ represent the incidence of Outcome A within the populations from which Groups 1 and 2 are sampled.

The **odds ratio (OR)** is a measure of association that remains applicable even when it is not possible to estimate incidence of an outcome from the sample data. The **odds** of Outcome A in Group 1 are $a/b$, while the odds of Outcome A in Group 2 are $c/d$.

$$OR_{\text{A, comparing Group 1 to Group 2}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

$$OR_{\text{A, comparing Group 2 to Group 1}} = \frac{c/d}{a/b} = \frac{bc}{ad}$$

## The $\chi^2$ test of independence

1. In resource-limited settings, single-dose nevirapine (NVP) is given to an HIV-positive woman during birth to prevent mother-to-child transmission of the virus. Exposure of the infant to NVP may foster the growth of more virulent strains of the virus in the child.

   If a child is HIV-positive, should they be treated with NVP or a more expensive drug, lopinarvir (LPV)? In this setting, success means preventing a growth of the virus in the child (i.e., preventing virologic failure).

| | NVP | LPV | Total |
|---|---|---|---|
| **Virologic Failure** | 60 | 27 | 87 |
| **Stable Disease** | 87 | 113 | 200 |
| **Total** | 147 | 140 | 287 |

   a) State the null and alternative hypotheses.

      The null hypothesis is that there is no association between treatment and outcome; i.e., treatment and outcome are independent.

      The alternative hypothesis is that there is an association betweeen treatment and outcome; i.e., treatment and outcome are not independent.

   b) Calculate the expected cell counts.

      The expected cell counts are shown in parentheses next to the observed cell counts.

| | NVP | LPV | Total |
|---|---|---|---|
| **Virologic Failure** | 60 (44.56) | 27 (42.44) | 87 |
| **Stable Disease** | 87 (102.44) | 113 (97.56) | 200 |
| **Total** | 147 | 140 | 287 |

```
#use r as a calculator

#set parameters
n = 287
row.1.total = 87
row.2.total = 200
col.1.total = 147
col.2.total = 140

#calculate expected values
exp.1.1 = (row.1.total * col.1.total)/n
exp.1.1
```

```
## [1] 44.56098
```

```
exp.1.2 = (row.1.total * col.2.total)/n
exp.1.2
```

```
## [1] 42.43902
```

```r
exp.2.1 = (row.2.total * col.1.total)/n
exp.2.1
```

```
## [1] 102.439
```

```r
exp.2.2 = (row.2.total * col.2.total)/n
exp.2.2
```

```
## [1] 97.56098
```

c) Check the assumptions for using the $\chi^2$ test.

Independence holds, since this is a randomized study. All expected counts are greater than 10.

d) Calculate the $\chi^2$ test statistic.

$$\chi^2 = \sum \frac{(\text{obs - exp})^2}{\text{exp}} = \frac{(60-44.56)^2}{44.56} + \frac{(27-42.44)^2}{42.44} + \frac{(87-102.44)^2}{102.44} + \frac{(113-97.56)^2}{97.56} = 15.74$$

```r
#use r as a calculator
obs.1.1 = 60
chi.sq.1.1 = ((obs.1.1 - exp.1.1)^2)/exp.1.1

obs.1.2 = 27
chi.sq.1.2 = ((obs.1.2 - exp.1.2)^2)/exp.1.2

obs.2.1 = 87
chi.sq.2.1 = ((obs.2.1 - exp.2.1)^2)/exp.2.1

obs.2.2 = 113
chi.sq.2.2 = ((obs.2.2 - exp.2.2)^2)/exp.2.2

chi.sq = chi.sq.1.1 + chi.sq.1.2 + chi.sq.2.1 + chi.sq.2.2
chi.sq
```

```
## [1] 15.73587
```

e) Calculate the $p$-value for the test statistic using `pchisq()`. The $p$-value represents the probability of observing a result as or more extreme than the sample data.

The $p$-value for the test statistic is $7.28 \times 10^{-5}$.

```r
#use pchisq()
pchisq(chi.sq, df = (2 - 1)*(2 - 1), lower.tail = FALSE)
```

f) Confirm the results from parts c) and d) using `chisq.test()`. Note that the value of the test statistic will be slightly different because R is applying a 'continuity correction'.

From `chisq.test()`, the $\chi^2$ statistic is 14.73 and the associated $p$-value is 0.0001.

```r
#enter the data as a table
hiv.table = matrix(c(60, 27, 87, 113),
              nrow = 2, ncol = 2, byrow = T)

#use chisq.test()
chisq.test(hiv.table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  hiv.table
## X-squared = 14.733, df = 1, p-value = 0.0001238
```

g) Summarize the conclusions; be sure to include which drug is recommended for treatment, based on the data.

There is sufficient evidence at the $\alpha = 0.05$ significance level to reject the null hypothesis and accept the alternative hypothesis that treatment and outcome are associated.

From comparing the expected and observed cell counts (or looking at the residuals), it is possible to determine the direction of the association. When treated with lopinarvir, fewer children than expected experience virologic failure (27 observed versus ~42 expected), and more than expected experience stable disease (113 observed versus ~98 expected). In contrast, when treated with nevirapine, more children than expected experience virologic failure (60 observed versus ~45 expected), and fewer children than expected experience stable disease (87 observed versus ~102 expected).

The data suggest that HIV-positive children should be treated with lopinarvir.

```r
#look at residuals
chisq.test(hiv.table)$resid
```

```
##           [,1]      [,2]
## [1,]  2.312824 -2.369939
## [2,] -1.525412  1.563082
```

```r
#to view the expected values, use $expected
chisq.test(hiv.table)$expected
```

```
##           [,1]     [,2]
## [1,]  44.56098 42.43902
## [2,] 102.43902 97.56098
```

h) Repeat the analysis using inference for the difference of two proportions and confirm that the results are the same.

The proportion of successes on nevirapine is 0.59 and the proportion of successes on lopinarvir is 0.81. The $p$-value is 0.0012; there is sufficient evidence to reject the null of no difference and conclude that stable disease is associated with lopinarvir.

```r
#use prop.test( )
prop.test(x = c(87, 113), n = c(147, 140))
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(87, 113) out of c(147, 140)
## X-squared = 14.733, df = 1, p-value = 0.0001238
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.3251572 -0.1054551
## sample estimates:
##    prop 1    prop 2
## 0.5918367 0.8071429
```

2. In the PREVEND study introduced in Unit 6, researchers measured various features of study participants, including data on statin use and highest level of education attained. From the data in `prevend.samp`, is there evidence of an association between statin use and educational level? Summarize the results.

   Test the null hypothesis that statin use and education level are not associated against the alternative hypothesis that statin use and education level are associated. Let $\alpha = 0.05$.

   The $p$-value of the $\chi^2$ statistic is 0.0003. The results are highly significant, and there is evidence to support accepting the alternative hypothesis that statin use and education level are associated.

   The largest deviations from independence occur in the primary school group and university group. There are more statin users than expected in the primary school group and fewer statin users than expected in the university group. There is an observable overall trend; as highest educational level attained increases, the proportion of statin users goes from higher than expected to lower than expected.

```r
#load the data
library(oibiostat)
data("prevend.samp")

#convert variables to factors
prevend.samp$Statin = factor(prevend.samp$Statin, levels = c(0, 1),
                         labels = c("NonUser", "User"))

prevend.samp$Education = factor(prevend.samp$Education, levels = 0:3,
                            labels = c("Primary", "LowerSec",
                                       "UpperSec", "Univ"))

#create a table
statin.edu.table = table(prevend.samp$Statin, prevend.samp$Education)

#run chi-squared test
chisq.test(statin.edu.table)
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  statin.edu.table
## X-squared = 19.054, df = 3, p-value = 0.0002665
```

```
chisq.test(statin.edu.table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  statin.edu.table
## X-squared = 19.054, df = 3, p-value = 0.0002665
```

**Measures of association in two-by-two tables**

3. Suppose a study is conducted to assess the association between smoking and cardiovascular disease (CVD). Researchers recruited a group of 231 study participants then categorized them according to smoking and disease status: 111 are smokers, while 40 smokers and 32 non-smokers have CVD. Calculate and interpret the relative risk of CVD.

   The relative risk of CVD comparing smokers to non-smokers is 1.35. Smoking is associated with a 35% increase in the probability of CVD. In other words, the risk of CVD is 35% greater in smokers compared to non-smokers.

```
#use r as a calculator
risk.smokers = 40/111
risk.nonsmokers = 32/(231-111)

risk.smokers / risk.nonsmokers
```

```
## [1] 1.351351
```

4. Suppose another study is conducted to assess the association between smoking and CVD, but researchers use a different design: 90 individuals with CVD and 110 individuals without CVD are recruited. 40 of the individuals with CVD are smokers, and 80 of the individuals without CVD are non-smokers.

   a) Is relative risk an appropriate measure of association for these data? Explain your answer.

      No, relative risk should not be calculated for these observations. Since the number of individuals with and without CVD is fixed by the study design, the proportion of individuals with CVD within a certain group (smokers or non-smokers) as calculated from the data is not a measure of CVD risk for that population.

   b) Calculate the odds of CVD among smokers and the odds of CVD among non-smokers.

      The odds of CVD among smokers is the number of smokers with CVD divided by the number of smokers without CVD: 40/50 = 0.80. The odds of CVD among non-smokers is the number of non-smokers with CVD divided by the number of non-smokers without CVD: 30/80 = 0.375.

```r
#use r as a calculator
odds.smokers = 40/(90-40)
odds.nonsmokers = (110 - 80)/80

odds.smokers
```

```
## [1] 0.8
```

```r
odds.nonsmokers
```

```
## [1] 0.375
```

c) Calculate and interpret the odds ratio of CVD, comparing smokers to non-smokers.

The odds ratio of CVD, comparing smokers to non-smokers is 2.13. The odds of CVD in smokers are approximately twice as large as the odds of CVD in smokers. The data suggest that smoking is associated with CVD.

```r
#use r as a calculator
odds.smokers / odds.nonsmokers
```

```
## [1] 2.133333
```

d) What would an odds ratio of CVD (comparing smokers to non-smokers) equal to 1 represent, in terms of the association between smoking and CVD? What would an odds ratio of CVD less than 1 represent?

An odds ratio equal to 1 would represent no association between smoking and CVD. An odds ratio less than 1 would represent an association between not smoking and CVD; i.e., that the odds of CVD in non-smokers were higher than the odds of CVD in smokers.