Inference for Two-Way Tables

Chapter 8, Lab 2

OpenIntro Biostatistics

Topics

- The χ^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 χ^2 test of independence

In the χ^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 χ^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 2×2, 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 χ^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 χ^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 χ^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

Chapter 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 (2×2) 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 χ^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, lopinavir (LPV)? In this setting, success means preventing a growth of the virus in the child (i.e., preventing virologic failure). The following table contains data from a 2012 study conducted in six African countries and India.[1]

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

   a) State the null and alternative hypotheses.

   b) Calculate the expected cell counts.

   c) Check the assumptions for using the χ^2 test.

   d) Calculate the χ^2 test statistic.

   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.

   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'.

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

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

2. In the PREVEND study introduced in Chapter 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.

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.

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.

[1] A. Violari, et al. "Nevirapine versus ritonavir-boosted lopinavir for HIV-infected children." *NEJM* 366: 2380-2389.

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

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

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

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?