Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Chapter 15, part 2: Chi-Square Tests

Jinko Graham

2018-10-17

Chi-square tests of association

- ▶ In chapter 14, we tested for association between two categorical variables by testing for differences between proportions
 - ▶ Call the test from Chapter 14 the *Z* test.
- Applied the Z test to data from the WHI.
 - ▶ Recall that the WHI randomized 16,608 post-menopausal women aged 50-79 years to receive either hormone replacement therapy (estrogen plus progestin EP+; $n_1 = 8506$), or a placebo (EP−; $n_2 = 8102$).
- ▶ We tested for a difference in the proportions of women with invasive breast cancer (BC+) in the hormone replacement therapy (EP+) and placebo (EP−) groups.

The WHI Data click

▶ The first few rows of the dataset are as follows:

```
## FP BC | ## 1 EP BC | ## 2 EP BC | ## 3 EP BC | ## 5 EP BC | ## 6 EP BC | ## 7 EP BC | ## 8 EP BC | ## 8 EP BC |
```

▶ A cross-tabulation of the BC and EP variables in the dataset is:

```
## BC
## EP BC- BC+
## EP- 7980 122
## EP+ 8340 166
```

Association between HRT and breast cancer

- The table of proportions below gives the conditional distributions of BC status given EP status.
 - ▶ The proportions in each row add to 1.
- BC and EP are associated if their conditional distributions are different.

- Previously, we used the Z test for different conditional distributions
- ▶ Looks for differences in the proportion of BC+ in the EP− and EP+ groups.
- ▶ Can be applied to data in 2×2 tables

Chi-square test of association

- ▶ When applied to 2×2 tables, the Z test for a difference in proportions is equivalent to the chi-square (χ^2) test.
- ▶ But the chi-square test has the advantage of generalizing from 2×2 tables to $r \times c$ tables, for $r \ge 2$ rows and $c \ge 2$ columns.
- Compares observed cell counts to expected counts
 - The expected count is the count we would expect if the null hypothesis of no association were true (details deferred).
- The form of the statistic is

$$X^2 = \sum_{\text{cells}} \frac{\left(\text{observed} - \text{expected}\right)^2}{\text{expected}}$$

Sampling distribution of X^2

- ▶ Under the null hypothesis of no association between row and column variables, the test statistic X^2 is approximately distributed as a chi-square distribution with $(r-1) \times (c-1)$ degrees of freedom (df).
- ► Computer software gives upper tail probabilities of different chi-square distributions.

Chi-square test for WHI example

▶ We can perform the chi-square test in R (see demo).

```
##
## Pearson's Chi-squared test
##
## data: wtab
## X-squared = 4.8387, df = 1, p-value = 0.02783
```

▶ At the 5% level, there is statistical evidence of an association between hormone-replacement therapy and invasive breast cancer.

Continuity correction for 2×2 tables

- ▶ The continuity correction to the X^2 test improves the χ^2 approximation for 2 × 2 tables.
- The corrected version of the statistic is:

$$X^{2} = \sum_{\text{cells}} \frac{(|\text{observed} - \text{expected}| - 0.5)^{2}}{\text{expected}}$$

▶ Please note that this correction applies only to 2×2 tables and does not generalize to $r \times c$ tables.

Chi-square test with continuity correction for WHI example

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: wtab
## X-squared = 4.5807, df = 1, p-value = 0.03233
```

▶ At the 5% level, there is still evidence of an association between hormone replacement therapy and invasive breast cancer.

Expected counts

- As mentioned earlier, these are calculated under the null hypothesis of no association between the 2 variables in the table.
- ▶ Let's first discuss expected counts for the WHI example. Later, we'll generalize to arbitrary *r* × *c* tables.

```
## BC
## EP BC- BC+
## EP- 7980 122
## EP+ 8340 166
```

▶ If H₀ holds and HRT has no effect on breast cancer, the proportion of BC+ in each EP group should be the same and can be estimated by pooling:

$$\hat{p} = \frac{122 + 166}{7980 + 122 + 8340 + 166} = \frac{288}{16608} = 0.01734.$$

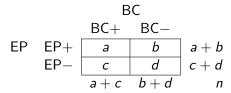
▶ Here is the table of counts with column and row margins added:

```
## BC- BC+ rowTot
## EP- 7980 122 8102
## EP+ 8340 166 8506
## colTot 16320 288 16608
```

- ► Focusing on the 1st row of the table, under H₀ of no association, we expect that, of the 8102 women who are EP-,
 - ▶ $8102 \times \hat{p} = 140.5$ would be BC+, and
 - ▶ $8102 \times (1 \hat{p}) = 7961.5$ would be BC-.
- ▶ Similarly, focusing on the 2nd row, we expect that, under H_0 of no association, of the 8506 women who are EP+,
 - ▶ $8506 \times \hat{p} = 8506 \times \frac{288}{16608} = 147.5$ would be BC+, and
 - ▶ $8506 \times (1 \hat{p}) = 8506 \times \frac{16320}{16608} = 8358.5$ would be BC-.

Expected counts, notation

Notation from the text:



where
$$n = a + b + c + d$$

- ▶ The pooled estimate of the proportion of BC+ women is $\hat{p} = (a + c)/n$.
- Expected count for the EP+ and BC+ cell:
 - ▶ Of the (a + b) women who are EP+, we expect that $(a + b) \times \hat{p} = (a + b)(a + c)/n$ would be BC+
- Notice that the expected count is of the form row total (a + b) times column total (a + c) divided by table total (n). This is a generalizable pattern ...

Expected counts: general formula

▶ For *r* × *c* tables, the expected count for the cell in the *i*th row and *j*th column is the *i*th row total times the *j*th column total divided by table total.

Accuracy of the χ^2 approximation $(r \times c \text{ tables})$

- ▶ The χ^2 approximation for the null distribution of the test statistic is considered accurate when
 - 1. No more than 20% of cells have expected counts < 5, and
 - 2. All expected cell counts are ≥ 1 .
- ▶ Note: These rules-of-thumb are intended regardless of whether or not we use the continuity correction for 2 × 2 tables.

Accuracy of the χ^2 approximation in WHI example

► The expected cell counts are as follows:

► All expected cell counts are greater than 5, and so the χ^2 approximation is considered accurate.

Sampling

- ► The chi-square test is appropriate under different sampling schemes such as:
 - 1. Take simple-random samples (SRSs) from each of c populations and classify individuals in each SRS according to one categorical variable with r levels
 - Take one SRS from a single population and classify individuals according to two categorical variables, one with c levels and the other with r levels
- lacktriangle The first scheme includes case-control sampling (c=2)
 - e.g. an SRS of size $n_1 = 500$ from the case population for non-Hodgkin lymphoma and an SRS of size $n_2 = 500$ from the control population and then classify them according to whether or not they are exposed to some pesticide ingredient.
- ▶ The second scheme pertains to the WHI study
 - A sample of size n=16608 was drawn from the population of post-menopausal women and then cross-classified according to whether or not they were randomized to receive HRT and whether or not they developed invasive breast cancer.