M 358K

Supplement to Chapter 26: CONNECTION BETWEEN THE Z- TEST FOR PROPORTIONS AND THE CHI-SQUARED TEST OF HOMOGENEITY

The textbook states (end of p. 643), “The z-test for two proportions generalizes to a chi-square test of homogeneity.” Here is an explanation of why.

The context for doing a chi-square test of homogeneity can be described as follows:

- We have data in which individuals (people, animals, objects, classrooms, hospitals, …) are classified in two ways. (In the example in the text, the two classifications are “college” and “after-graduation plan”)
- In most uses of the test, we can think of one classification as “groups” and the other as “categories”. (In the example, we can think of “college” as “group” and “after graduation plan” as “category”.)
- The question of interest is whether the different groups have the same distribution into the categories. (In the example: Are the after-graduation plans distributed the same from college to college, or do they differ between colleges more than would be expected by chance?)

We need to have more than one group and more than one category, so the simplest situation in which the homogeneity test fits is two groups and two categories. So, for example, consider a simpler version of the example in the book: We are comparing students in the College of Natural Sciences and the College of Liberal Arts (these are the two groups), and categorize their after-graduation plans as “Work full-time” or “Other”. We can arrange the data in a table, with groups in columns and categories in rows. Instead of using actual numbers, we’ll represent the counts by letters, so we can work in general terms:

Counts	Natural Sciences	Liberal Arts
Work full-time	A_1	A_2
Other	B_1	B_2

It will be helpful to have notation for the various row, column, and overall totals, so we’ll use notation as in the following expanded table:

Counts	Natural Sciences	Liberal Arts	Total
Work full-time	A_1	A_2	A
Other	B_1	B_2	B
Total	N_1	N_2	N

So $A_1 + A_2 = A$ and $A_1 + B_1 = N_1$, and similarly for the other rows and columns. In particular, $N = A + B = N_1 + N_2$. (This is just counting the total number of students in the sample two ways: First by after-graduation plan, giving A + B, and second by college, giving $N_1 + N_2$.)

Recall the question we are interested in: Are the after-graduation plans distributed the same from college to college, or do they differ between colleges more than would be expected by chance? Since we have only two groups and two categories, we could address this question by a z-test for two proportions: We ask, "Is the proportion of students in Natural Sciences who plan to work full-time after graduation the same as the proportion of students in Liberal Arts who plan to work full time after graduation?" (Since we have only two categories, if the answer to this question is yes, then the proportion of students in each group who have other plans must also be the same, so we have the same distribution into categories in each group.)

To do the z-test, we consider the proportions

$\hat{p}_1 = A_1/N_1$ (sample proportion of Natural Science students planning to work full time)

$\hat{p}_2 = A_2/N_2$ (sample proportion of Liberal Arts students planning to work full time)

$\hat{p} = \hat{p}_{pooled} = A/N$ (pooled sample proportion = proportion of students in the combined colleges planning to work full time).

The test statistic is (from p. 533 plus a little algebra)

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{1}{N_1}+\frac{1}{N_2}\right)\hat{p}\,(1-\hat{p}\,)}} \qquad (*)$$

To do the chi-square test for homogeneity, we first calculate the expected values for each cell of the table, using the reasoning as described on the bottom of p. 642. For example, the overall proportion of students planning to work full-time is A/N, so the expected number of students in Natural Sciences who plan to work full time is $(A/N)N_1$. Filling in all the expected values gives

Expected	Natural Sciences	Liberal Arts	Total
Work full-time	$(A/N)N_1$	$(A/N)N_2$	A
Other	$(B/N)N_1$	$(B/N)N_2$	B
Total	N_1	N_2	N

(Use the identities N = A + B = N_1 + N_2 to see why the row and column totals are unchanged from the table of observed counts.)

Now we can calculate the chi-square components. For example, the one from Natural Science students who plan to work full time is

$$\frac{(Obs-Exp)^2}{Exp} = \frac{\left(A_1-(\frac{A}{N}N_1)\right)^2}{\frac{A}{N}N_1}$$

Using algebra and the identities N = N_1 + N_2 and A = A_1 + A_2, we can re-express this as

$$\frac{N}{A}\frac{1}{N_1}\left[\frac{NA_1-AN_1}{N}\right]^2 = \frac{N}{AN_1}\left[\frac{NA_1-AN_1}{N}\right]^2$$

$$= \frac{N}{AN_1}\left[\frac{(N_1+N_2)A_1-(A_1+A_2)N_1}{N}\right]^2$$

$$= \frac{N}{AN_1}\left[\frac{N_2A_1-A_2N_1}{N}\right]^2 = \frac{1}{ANN_1}[N_2A_1-A_2N_1]^2$$

A similar calculation (details left to the reader) shows that the chi-square component from the Liberal Arts students who plan to work full time is

$$\frac{(Obs-Exp)^2}{Exp} = \frac{1}{ANN_2}[N_1A_2-A_1N_2]^2 = \frac{1}{ANN_2}[N_2A_1-A_2N_1]^2$$

Adding these first two chi-square components gives

$$\frac{1}{AN}[N_2A_1-A_2N_1]^2\left(\frac{1}{N_1}+\frac{1}{N_2}\right)$$

Using the identities A_1 = $N_1\hat{p}_1$ and A_2 = $N_2\,\hat{p}_2$ shows that this is equal to

$$\frac{(N_1N_2)^2}{AN}[\hat{p}_1-\hat{p}_2]^2\left(\frac{1}{N_1}+\frac{1}{N_2}\right). \qquad (**)$$

Similar calculations (details omitted) using B_1 = $N_1(1-\hat{p}_1)$ and B_2 = $N_2\,(1-\hat{p}_2)$ show that the sum of the remaining two chi-square components is

$$\frac{(N_1N_2)^2}{BN}[\hat{p}_1-\hat{p}_2]^2\left(\frac{1}{N_1}+\frac{1}{N_2}\right). \qquad (***)$$

Adding (**) and (***) then shows that the chi-square statistic for the chi-square test for homogeneity is

$$\chi^2 = \frac{(N_1N_2)^2}{N}[\hat{p}_1 - \hat{p}_2]^2\left(\frac{1}{N_1}+\frac{1}{N_2}\right)\left(\frac{1}{A}+\frac{1}{B}\right).$$

Note that $N_1N_2\left(\frac{1}{N_1}+\frac{1}{N_2}\right) = N_1+N_2 =$ N. Using this, cancelling the N's and adding the fractions $\frac{1}{A}$ and $\frac{1}{B}$ gives

$$\chi^2 = N_1N_2[\hat{p}_1 - \hat{p}_2]^2\left(\frac{A+B}{AB}\right).$$

Using the identities A + B = N, A = N$\hat{p}$, and B = N(1-$\hat{p}$) now gives

$$\chi^2 = N_1N_2[\hat{p}_1 - \hat{p}_2]^2\left(\frac{N}{\mathrm{N}\hat{p}\mathrm{N}(1-\hat{p})}\right)$$

$$= [\hat{p}_1 - \hat{p}_2]^2\left(\frac{N_1N_2}{\mathrm{N}\hat{p}(1-\hat{p})}\right).$$

Since

$$\frac{N_1N_2}{\mathrm{N}} = \frac{N_1N_2}{N_1+N_2} = \frac{1}{\frac{1}{N_1}+\frac{1}{N_2}},$$

we now have

$$\chi^2 = [\hat{p}_1 - \hat{p}_2]^2\left(\frac{1}{\left(\frac{1}{N_1}+\frac{1}{N_2}\right)\hat{p}\,(1-\hat{p})}\right),$$

which is just the square of the z-statistic in (*).

Now reasoning as in the end of the handout *Connection Between the One Sample Test for Proportions and the Chi-Squared Goodness-of-Fit Test*, we see that (in this case of a 2X2 table) the p-value for the chi-square test will be exactly the same as the p-value for the test of two proportions. (Note: Since this case involves a 2X2 table, the degrees of freedom are (R-1)(C-1) = 1X1 = 1.)