Chapter 23. Two Categorical Variables: The Chi-Square Test

Two-Way Tables

Note. We quickly review two-way tables with an example.

Example. Exercise 23.2a page 550.

Expected Counts in Two-Way Tables

Definition. The **expected count** in any cell of a two-way table when H_0 is true is

$$\text{expected count } = \frac{\text{row total } \times \text{ column total}}{\text{table total}}.$$

Example. Exercise 23.6 page 554.

The Chi-Square Test

Definition. The **chi-square test** is a measure of how far the observed counts in a two-way table are from the expected counts. The formula for the statistic is

$$\chi^2 = \sum \frac{(\text{observed count } - \text{ expected count})^2}{\text{expected count}}.$$

The sum is over all cells in the table.

Cell Counts Required for the Chi-Square Test

Note. You can safely use the chi-square test with critical values from the chi-square distribution when no more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater. In particular, all four expected counts in a 2×2 table should be 5 or greater.

Uses of the Chi-Square Test

Note. Use the chi-square test to test the null hypothesis:

H_0: there is no relationship between two categorical variables

when you have a two-way table from one of these situations:

- Independent SRSs from each of two or more populations, with each individual classified according to one categorical variable. (The other variable says which sample the individual comes from.)
- A single SRS, with each individual classified according to both of two categorical variables.

The Chi-Square Distributions

Note. The **chi-square distributions** are a family of distributions that take only positive values and are skewed to the right. A specific chi-square distribution is specified by giving its **degrees of freedom**. The chi-square test for a two-way table with r rows and c columns uses critical values from the chi-square distribution with $(r-1)(c-1)$ degrees of freedom. The P-value is the area to the right of χ^2 under the density curve of this chi-square distribution. Table E gives the relationships between the degrees of freedom, χ^2, and P-values.

Example. Exercise 23.40 page 577.

Example S.23.1. χ^2-Stooges.
We now consider all 190 Three Stooges films and two categories. One category is "the role of third stooge" (Curly/Shemp/Joe) and the other is "number of slaps in the film" (which we break into intervals as $[0, 10]$, $[11, 20]$, $[21, 30]$, $[31, 40]$, $[41, \infty)$). Notice that both of these are in fact categorical variables, even though "number of slaps in the film" could be dealt with as a quantitative variable. The data can be put in a two-way table as follows. (This is similar to Example S.6.1.)

	Curly	Shemp	Joe	TOTAL
0 to 10 slaps	49	34	10	93
11 to 20 slaps	36	21	5	62
21 to 30 slaps	7	14	1	22
31 to 40 slaps	3	2	0	5
more than 40 slaps	2	6	0	8
TOTAL	97	77	16	190

Calculate the χ^2 statistic and perform a χ^2 test on H_0: there is no relationship between two categorical variables.

Solution. First observe that the number of degrees of freedom is $df = (r-1)(c-1) = (5-1)(3-1) = 8$. Using the first formula of this chapter, we find the following expected counts:

	Curly	Shemp	Joe	TOTAL
0 to 10 slaps	47.48	37.69	7.83	93
11 to 20 slaps	31.65	25.13	5.22	62
21 to 30 slaps	11.23	8.92	1.85	22
31 to 40 slaps	2.55	2.03	0.42	5
more than 40 slaps	4.08	3.24	0.67	8
TOTAL	97	77	16	190

We now sum over the 15 table entries to calculate the χ^2

statistic:

$$\frac{(49-47.48)^2}{47.48}+\frac{(34-37.69)^2}{37.69}+\frac{(10-7.83)^2}{7.83}+\frac{(36-31.65)^2}{31.65}+\frac{(21-25.13)^2}{25.13}+\frac{(5-5.22)^2}{5.22}+$$
$$\frac{(7-11.23)^2}{11.23}+\frac{(14-8.92)^2}{8.92}+\frac{(1-1.85)^2}{1.85}+\frac{(3-2.55)^2}{2.55}+\frac{(2-2.03)^2}{2.03}+\frac{(0-0.42)^2}{0.42}+$$
$$\frac{(2-4.08)^2}{4.08}+\frac{(6-3.24)^2}{3.24}+\frac{(0-0.67)^2}{0.67}=11.754.$$

We now find $\chi^2 = 11.754$ in Table E in the row containing $df = 8$. We see that 11.754 lies in Table E between $p = 0.20$ and $p = 0.15$. Software (Minitab, say) gives $p = 0.1625$. The p value is not small enough for us to reject the null hypothesis that there is not relationship between two categorical variables.

The Chi-Square Test for Goodness of Fit

Note. A categorical variable has k possible outcomes with probabilities $p_1, p_2, 0_3, \ldots, p_k$. That is, p_i is the probability of the ith outcome. We have n independent observations from this categorical variable. To test the null hypothesis that the probabilities have specified values

$$H_0 : p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0}$$

use tje **chi-square statistic**

$$\chi^2 = \sum \frac{(\text{count of outcome } i - np_{i0})^2}{np_{i0}}.$$

The P-value is the area to the right of χ^2 under the density curve of the chi-square distribution with $k-1$ degrees of freedom.

Example. Exercise 23.16 page 568.

rbg-4-4-2009