Chapter 9. Inference for categorical data

9.1: Inference for Two-way Tables

Another way to summarize data is through putting the data into *two-way tables*. These tables are labelled as a *r x c table*, r being the number of rows in the table and c for the number of columns. The explanatory variable is always placed as the column variable, whereas the row variable is always the categorical response variable.

Below is an example of a two-way table with the column variable gender and row variable binge drinking. Each combination of both variables is called a cell. In a 2 x 2 table there are 4 cells therefore, and these cells in the table are shaded grey.

Frequent Binge Drinker	**Gender**		**Total**
	Men	**Female**	
Yes	1,630	1,684	3,314
No	5,550	8,232	13,782
Total	7,180	9,916	17,096

There are three distributions in two-way tables. The first is called a *joint distribution*. This is the collection of the proportions each cell is of the total. For example, the proportion for male binge drinkers is 1,630 divided by the total (17,096). This gives a proportion of 0.095. The second type of distribution is a *marginal distribution*, which is the percentage or proportion of a single categorical variable. This means that the row and column totals divided by the table total for each categorical variable are the marginal distributions for those variables. For example, 7,180/17,096 is the marginal distribution of 0.42, or 42% for the variable males. The last type of distribution is a *conditional distribution*, which is when you focus on the value of one variable and compute the distribution of the other variable. For example, when computing the distribution of the variable binge drinking, but only for men, you would get 23% for those who do binge drink and 77% for those who don't binge drink.

The null hypothesis in two-way tables is one of no association between the row and column variables. The alternative hypothesis in two-way tables cannot be one- or two-sided as there are very many possibilities for alternative associations.

Presenting data in bar graphs or mosaic plots

Other than a two-way table, also a bar graph or mosaic plot can be used. All these options can be used to present the variables. A bar graph shows two bars for each variable, of which one bar indicates the percentage of the population that has a positive result for the variable, the other bar indicates the percentage that has a negative result. A mosaic plot consists of four rectangles, also indicating per variable which percentage of the population has positive or negative results.

So the variables can be presented in different ways, but with each way of presenting information the core question remains whether the null hypothesis is true and whether there is an association between the variables.

The Chi-Square test and Chi-Square distribution

A chi-square test can be used to see whether the null hypothesis is true. It is performed as follows:

- Look at the percentages of the rows and columns.
- Look at the expected numbers and use these to calculate the chi-quare test.
- Use the critical vales from the table to determine P.
- Draw a conclusion on the association between the variables of the rows and columns.

When testing the H_o, the observed and the expected cell counts are compared. The formula for expected cell counts is:

$$\text{Expected cell count} = \frac{row\ total\ x\ column\ total}{n}$$

The Chi-square statistic X^2 gives the outcome of this comparison. The formula is:

$$X^2 = \sum \frac{(observed\ count - expected\ count)_2}{expected\ count}$$

When the difference between the observed and expected cell counts is big, then the X^2 shall be large, and this is desirable when rejecting the null hypothesis. You can use this value of X^2 and find the corresponding p-value found in Table F. A *chi-square distribution* χ^2(df) results from this test. This distribution also has degrees of freedom, in this case (r - 1)(c - 1). This you fill into the brackets in χ^2(df). So a chi-square distribution with 4 degrees of freedom is written as: $\chi^2(4)$. The p-value for the chi-square test is: $P(\chi^2 \geq X^2)$.

The higher the cell counts, the more accurate the distribution of χ^2(df). For 2 x 2 tables all four expected cell counts need to be 5 or larger. However for larger tables, the average of the expected cell counts must be 5 or more and the smallest expected cell count is at least 1.

The z test and the chi-square test always give the same result, however there are some differences between the tests. The advantage when using the z test is that you can test both one- and two-sided alternatives, whereas chi-square test only tests the two-sided alternative. The advantage when using the chi-square test is that it is possible to compare more than two populations at one time.

9.2: Goodness of fit test

Data for n number of observations from a categorical variable with k number of results are written as $n_{1,}$ $n_{2,}$ n_2 … n_k observations in k number of cells. The null hypothesis is then concerned with the chances $p_{1,}$ $p_{2,}$ p_3 … p_k for all possible results. For each cell the total number of observations (n) should be multiplied with the chance that is used to calculate the expected cell counts:

- Expected cell counts = $np_{i.}$
- The formula of the chi-square test is:
- x^2 =Σ(observed count – expected count)^2 /expected count
- The degrees of freedom are k-1 and the P-values can be computed from the chi-quare distribution.

The chi-square test can be used to measure the goodness of fit, in how far the actual results differ from the expected results.