# Chapter 9. Two-Way Tables

**9.1: Inference for Two-way Tables and Goodness of Fit**

This chapter considers another way to summarize data: 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.

Here 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.

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 test

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

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

(Please note that 2 the numerator in the formula should be a $^2$ and so the numerator should be squared!)

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* $\chi^2(df)$ results from this test. To see what this distribution looks like see Figure 9.4 in Introduction to the Practice of Statistics, 7th Ed (Moore, McCabe & Craig). This distribution also has degrees of freedom, in this case (r - 1)(c - 1). This you fill into the brackets in $\chi^2(df)$. 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 $\chi^2(df)$. 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.