Association between Categorical Variables

Two-Way Tables

A two-way table is a table that displays one categorical variable against the other and used to investigate whether the two variables are associated with each other. When we want to examine the relationship between two categorical variables we tabulate one against the other. Two-way tables are alos known as cross-tabulation.

Each of the two variables in a two- way table could have one or more levels. The simplest type of a two-way table is the case where each variable has only two values. This is called a 2X2 table. For example the table opposite shows pre-school children who were burnt in Sulaymaniyah in 2008 by sex and whether they were admitted to hospital or not. The table has two variables, sex and admission status and each of these has only two levels.

Sex	Admission status		Total
	Yes	No	
Male	135	363	498
Female	102	344	446
Total	237	707	944

The variables may have more than 2 levels. For example the table opposite shows FGM data of Kurdistan and displays two variables, the governorate and FGM status. The FGM has two values (levels, strata) and the governorate has 3 values. This is a 2X3 table.

FGM status	Governorate			Total
	Sulaimaniyah	Erbil	Dohuk	
Yes	144	186	18	384
No	359	316	485	1160
Total	503	502	503	1508

Observed frequencies and calculating percentages

These two-way tables are used to check for association between the two variables. An **association** exists between two categorical variables if the distribution of one variable varies according to the value of the other. The initial checking for associations between two variables is done by calculating percentages of each level of the variables. The percentages is calculated from dividing the frequency in each level by total. The frequencies which we find in our study are called observed frequencies because we that is what we have observed in our study. In the table of FGM by governorate we add another column for each level of governorate (i.e. for each of Sulaymaniyah, Erbil, Dohuk) for the percentages of FGM of FGM status. For example in Erbil, we have 186 perosns with FGM and the total persons of Erbil are 502. By dividing 186 on 502 we get the percentage of persons with FGM in Erbil.

FGM status	Governorate						Total	
	Sulaymaniyah		Erbil		Dohuk			
	Number	%	Number	%	Number	%	Number	%
Yes	144	29	186	37	18	4	348	23
No	359	71	316	63	485	96	1160	77
Total	503	100	502	100	503	100	1508	100

What do the percentages mean? The percentage show the percentage of women who have FGM and women who do not have FGM in each governorate. For example the percentage of women with FGM in Sulaymaniyah is 29%, and the women with no FGM is 71%; these two percentages add up to 100% because they comprise the whole sample.

Calculate percentages for other governorates by dividing observed frequencies with FGM on total person in the same governorate.

As we see the percentages of women with FGM are not similar in the three provinces; it is 29% in Sulaymaniyah, 37% in Erbil and 4% in Dohuk. These percentages are not the same; they are different. This is what we mean by association. As the percentage of FGM is not the same in the three governorates, probably there is an association between FGM and governorate. If there were no association between FGM and governorate, we would see similar percentages of FGM in the three governorates.

Expected Frequencies

If there was no association between FGM and governorate, then the percentages for each governorate p would be the same as the Total percentages in the last colum of the table i.e. 23%. If there was no association between FGM and governorate, then the percentage of persons with FGM will be the same as the total percentage for Kurdistan i.e. 23% in each of Erbil, Sulaymaniyah and Dohuk. So if there was no association i.e. percentage of people with FGM was 23% in each governorate, what would be the expected frequency of persons with FGM in each governorate?

The expected frequency of persons with FGM in Sulaymaniya would be
23% of 503= 503X23/100=116

This means that of there was no association between FGM and governorate, we would expect to have 116 persons with FGM rather than 144 persons.

The expected frequencies can be calculated from the colum total, row total and overall total using this formula

$$\text{Expected frequency in each level} = \frac{\text{row total x column total}}{\text{grand total}}$$

For the expected frequency of persons with FGM in Sulaymaniyah, column total (i.e. persons in Sulaymaniyah with FGM) is 503, row total (i.e. all persons with FGM) is 348 and grand total (i.e. the whole sample) is 1508, therefore,

Expected frequency for Sulaymaniyah= 503 X 348/ 1508=116.0.1

FGM status	Governorate						Total	
	Sulaymaniyah		Erbil		Dohuk			
	Observed	Expected	Observed	Expected	Observed	Expected	Number	%
Yes	144	116.1	186		18		348	23
No	359	386.9	316		485		1160	77
Total	503		502		503		1508	100

Now calculate the expected percentages for Erbil and Dohuk

The X^2 (Chi-squared) Test

If there was no association between FGM and governorate, we would get the expected frequencies. The association between categorical variables like FGM and

governorate are based on calculating expected frequencies and the test which does this is called Chi-squared test or X^2 test**.**

The null hypothesis in this test is no association i.e. observed frequencies are similar to expected frequencies. The alternative hypothesis says that there is an association i.e. the observed frequencies are different from the expected frequencies.

The X^2 statistics is calculated by the following formula *(not required)*:

$$X^2 = \sum \frac{(O - E)^2}{E}$$

where *O* = observed cell frequency; *E* = expected cell frequency; Σ = sum over all cells in the table

The result of this calculation which is called X^2 statistic is read from a table of χ^2 distribution to obtain a P-value.

The result of our example using a computer software (STATA) is 171.7 on 2 degrees of freedom.

Note: the degrees of freedom is not derived from the sample size, but from the number of cells in the table using the following formula:

Table A3 Chi-squared distribution.

	Two-tailed *P*-value			
df	0.10	0.05	0.01	0.001
1	2.706	3.841	6.635	10.827
2	4.605	5.991	9.210	13.815
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.466
5	9.236	11.070	15.086	20.515
6	10.645	12.592	16.812	22.457
7	12.017	14.067	18.475	24.321
8	13.362	15.507	20.090	26.124
9	14.684	16.919	21.666	27.877

Degrees of freedom= (number of rows-1)(number of columns-1)

For example in a table with 2 rows and 3 columns
df=(2-1)(3-1)=2

We read this result in the table of If we go to the table χ^2 distribution on 2 df. The biggest number we have in the table against 2 df is 13.8 which corresponds to P value of 0.001. Since our results 171.7 is bigger that 13.8 then our P value will be smaller than 0.001 i.e. probability of obtaining the observed result by chance is less than 0.001. Since this is a small probability, we reject the null hypothesis and accept the alternative hypothesis that the three governorates are different in terms of FGM. We conclude by saying that there is a significant difference between the three governorate of Erbil, Sulaymaniyah and Dohuk in relation to FGM. In other words, governorate is significantly associated with FGM.