Chi-squared tests in SPSS (Practical)

Chi-squared test practical

In this practical we are going to investigate how to perform a chi-squared test using SPSS. A chi-squared test is used when we have two categorical variables measured for all observations in a dataset and we want to test if the variables are related or independent. Independent here means that the category observed for one variable does not depend on the category observed for the other variable. To run a single test in SPSS requires that your dataset has two separate columns containing the two (categorical) variables to be tested.

In this example we explore whether there is an association between awareness of an environmental issue and optimism about its development over the next 20 years. The first variable, INFGGAS, is the participant's rating of how informed they feel about the environmental issue of greenhouse gases in the atmosphere. The second variable, OPTGGAS, recorded whether they believed the greenhouse gas issue would improve, stay the same, or get worse, over the next 20 years. In this application the measures will be independent if the level of optimism expressed does not differ between participants who are more or less informed about greenhouse gases.

Below you will see instructions to perform the chisquared test in SPSS. For the chi-squared test we will use SPSS crosstabs options that we have also looked at in our tabulation practical. If you follow the instructions you will see the three tabular outputs that are embedded in the explanations below.

- Select **Descriptive Statistics** from the **Analyze** menu.
- Select **Crosstabs...** from the **Descriptive Statistics** sub-menu.
- Click on the **Reset** button.
- Copy the **OPTGGAS** variable into the **Row(s):** box.
- Copy the **INFGGAS** variable into the **Column(s):** box.
- Click on the **Statistics...** button.
- In the window that appears click on the **Chi-square** tickbox.
- Click on the **Continue** button.
- Click on the **Cells...** button.
- In the window that appears:
- Under counts click on the **Expected** tick box to include it.
- Click on the **Continue** button.
- Click on the **OK** button.

The first table looks at which of the observations have non-missing values for both the two variables to be considered.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Greenhouse gases will improve in 20 years * How informed about greenhouse gases	4721	90.9%	473	9.1%	5194	100.0%

Here we see that there are 5194 observations of which 473 are missing resulting in 4721 that can be used in the test.

The second tabular output contains the cross-tabulation of the two variables. Here different levels of **OPTGGAS** are allocated to different rows whilst different values of **INFGGAS** are allocated to different columns and each observation adds to the count of a particular cell in the table.

Greenhouse gases will improve in 20 years * How informed about greenhouse gases Crosstabulation

			How informed about greenhouse gases				
			Never heard of this	Heard of, but not really able to explain this	Know something, could explain in general	Familiar with this, could explain well	Total
Greenhouse gases will improve in 20 years	Improve	Count	45	98	222	325	690
		Expected Count	22.2	86.7	242.0	339.1	690.0
	Stay about the same	Count	56	200	299	291	846
		Expected Count	27.2	106.3	296.8	415.7	846.0
	Get worse	Count	51	295	1135	1704	3185
		Expected Count	102.5	400.1	1117.2	1565.2	3185.0
Total		Count	152	593	1656	2320	4721
		Expected Count	152.0	593.0	1656.0	2320.0	4721.0

Here we see the observed counts for the different combinations of **OPTGGAS** and **INFGGAS**. So for example there are 325 observations where **OPTGGAS** response is Improve and the **INFGGAS** response is Familiar with this, could explain well. This is out of a total of 690 observations where **OPTGGAS** is Improve and 2320 observations where **INFGGAS** is Familiar with this, could explain well. Under the model of independence we expect to see 339.1 observations where **OPTGGAS** is Improve and **INFGGAS** is Familiar with this, could explain well. This means there are 14.1 fewer observations where **OPTGGAS** is Improve and **INFGGAS** is Familiar with this, could explain well than expected. We can look at similar differences between the other observed and expected counts and the chi-squared test looks at whether these differences are simply chance occurrences or statistically significant.

The third tabular output contains the information on test statistics for testing the hypothesis of no association / relationship between the two variables.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	243.688[a]	6	.000
Likelihood Ratio	226.714	6	.000
Linear-by-Linear Association	104.310	1	.000
N of Valid Cases	4721		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 22.22.

SPSS gives several tests for significance and we will first focus on the Pearson Chi-Square test. This test begins by forming the Pearson test statistic which asymptotically is formed from the observed and expected cell counts. For each cell the difference between the observed and expected counts is found and squared. This positive number is then divided by the expected count to account for different sizes of cells. Having constructed this value for each cell these are summed across all cells to give our test statistic which here is 243.688. This statistic follows a chi-squared distribution under the null hypothesis with degrees of freedom equal to (rows-1)x(columns-1) which in this case equals 6. The statistic is then compared with the appropriate Chi-Squared distribution and this results in an asymptotic (2-sided) p value which has value .000(reported as $p < .001$). Here we see that the p value is less than 0.05 and therefore we can reject the null hypothesis that the two variables are independent and there is therefore some relationship between the variables.

The lines labelled Likelihood Ratio and Linear-by-Linear Association are alternative tests of independence that are less commonly used than the Chi-square test and we will not cover them here.

In conclusion, we could report this to a reader as follows:

The relationship between the categorical variables, **OPTGGAS** and **INFGGAS** was examined to look for associations. A chi-squared test with 6 degrees of freedom was performed resulting in a test statistic of 243.688. This results in an asymptotic p value which is less than 0.001 and therefore we have strong evidence to reject the null hypothesis that **OPTGGAS** and **INFGGAS** are independent and there is therefore some relationship between the variables.

Inspection of the cross-tab shows that levels of optimism about greenhouse gases are strikingly different between those who are more and less informed about the issue. Under independence, we would expect only around 22 of those who had never heard of the issue to expect things to improve in the next 20 years, when in fact 45 did so. Conversely, under independence, around 1565 of those who are familiar with the issue would be expected to predict that things will get worse, when in fact 1704 did so. It appears that scientific understanding is related to more pessimistic judgements about the future of the environment, at least on this issue.