

Crosstabulation and chi-square

- Crosstabulation and chi-square 276
- Test procedure: layered 2×3 crosstabulation and chi-square 286
- Chi-square as a 'goodness of fit' test 290

SPSS Explained.
Hinton et al. (2004).

A RESEARCHER PRODUCES a questionnaire to find out about teenagers' leisure activities. There will be a lot of data produced, particularly if there are a lot of questions, and it is given out to teenagers who are 13 to 18 years old. Once the data has been put into the computer the researcher can print out a range of descriptive statistics, such as how many respondents there were of each age, how many males and females, how many people (or what percentage of them) participated in a particular leisure activity. However, the researcher will want to go further than simply summarising the findings of each question and this is where crosstabulation comes in.

We can combine the results of different questions in a table with the results of one question as the rows and the results of another question as the columns. The most common crosstabulation is when the data are frequency counts, so in our example the researcher can crosstabulate the question 'are you male or female?' with the question 'what is your age in years?' This will provide us with a table showing how many males and females there are of each age. We can create a table with more than two variables (questions). We might add the results of another question ('do you have a boy/girlfriend?') to the table. Now this is quite hard to picture as a piece of paper only has two dimensions and, after putting down the first two variables as the rows and columns, the third variable needs a third dimension. There are ways to draw the graph (as we can draw a three-dimensional object like a cube on a piece of paper). One way is to draw one table of age and gender for those who have a boy/girlfriend and a second table of age and gender for those who do not. We can refer to these two tables as 'layers' produced by the third variable. We are producing a crosstabulation with three variables by adding the third variable as a layer after using up the rows and columns. We can add more variables, if we wish, as additional layers. So even though we are creating a table with many variables it is not a problem for SPSS because it adds variables as layers to the table.

Now, why do we go to this trouble to produce a crosstabulation? The answer is that it will allow us to examine the association between the variables. We might have pre-planned predictions or we might simply be examining whether there is evidence of an association. Let us take some different variables from the research questionnaire. One question is 'what is your favourite film?' We want to see if this differs according to age, and to simplify things we might decide that we would add together the results of the 13- to 15-year-olds as younger teenagers and the 16- to 18-year-olds as older teenagers. Imagine that most choices were either the *Harry Potter* films or the *Lord of the Rings* films. So we choose columns as 'Harry Potter', 'Lord of the Rings' and 'other' and rows as 'younger' and 'older'. The data in the table will be the number of people in each group

CROSSTABULATION AND CHI-SQUARE

choosing the film. We might have a prediction that *Harry Potter* will be chosen by a larger proportion of the younger group and *Lord of the Rings* by the older group, but we do not have to.

When we test the association between age and favourite film we are examining the null hypothesis that there is no difference between the age groups in their film choice. So when the null hypothesis is true we should see roughly the same proportion of each age group choosing each of the films. The chi-square test examines these proportions and presents the probability of obtaining this pattern when there is no difference in the choices. A large value of chi-square indicates a large difference between the groups. If the probability is very small ($p < 0.05$) then we can conclude that there is a difference in the choice of film for the different age groups. If we had made a specific prediction that more of the younger group would choose *Harry Potter* then we are making a one-sided (one-tailed or directional) prediction and we would need to check the results to make sure that the predicted pattern is the one that emerged rather than some other difference.

The key assumption of the chi-square test is that there is independence of the observations, i.e. the results in each section (or 'cell') of the table are independent of each other. So if 45 younger teenagers choose *Harry Potter* and 28 choose *Lord of the Rings*, the 45 and 28 are unrelated. We could not allow a respondent to say they liked both films equally and score them in both cells.

There is a second assumption that the calculated chi-square is 'continuous' as the chi-square distribution is continuous. In order to satisfy this assumption it is better to have large numbers in each cell rather than small numbers. When the 'expected frequency' of a cell is lower than five we are concerned that this assumption is violated. Also, when we have a table with only two columns and two rows (2×2) there is a real risk that we will violate this assumption. To compensate for this a 'continuity correction' value is generated in SPSS, so you need to be careful when analysing a 2×2 table. On the positive side, with such a small table, the Fisher's Exact test (also produced by SPSS) can be used for an accurate calculation.

Usually, with tables larger than 2×2 and cell frequencies larger than five (and preferably over ten) we use the 'standard' chi-square value. If the cells have small numbers (expected frequency below five) but the table is large then it might be worth collecting more data or, if this is not possible, combining cells (see Chapter 5 for details on recoding data). If we were looking at favourite films, rather than looking at each film individually, we might group them into 'fantasy', 'romance', etc. to make the numbers in each cell larger. If we don't want to do this we should take the 'continuity correction' value. If we have a 2×2 table we take the Fisher's Exact value, as it is just that, an exact value for our table.

Finally, we have seen that chi-square compares patterns of results to see if they are similar or different. We can also use the chi-square test as a 'goodness of fit test'. This means 'does the pattern of results match (or fit) a predicted pattern?' For example, if we believe that younger teenagers will prefer *Harry Potter* over *Lord of the Rings* by 2 to 1,

then we have a model where we expect there to be two people choosing *Harry Potter* for each one choosing *Lord of the Rings*. We would put the predictions of the model in our SPSS data table. We can then compare the actual findings (the frequencies) to the model to see if it 'fits' the data. In this case a significant chi-square value would tell us that the data is significantly different to the pattern predicted by the model. A non-significant chi-square would indicate that the data does not significantly differ from the model's predicted pattern.

Crosstabulation and chi-square

Scenario

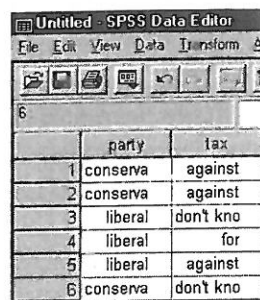
A researcher wanted to test the difference of opinion between conservatives and liberals on some new taxation legislation. In a survey 120 people were identified as conservatives and 80 as liberals. A question on the survey asked whether the respondent agreed with new taxation legislation ('for'), disagreed with it ('against'), or had no opinion or did not know about it ('don't know'). The researcher believed that the respondent's party affiliation would influence their view of the new tax legislation.

Data entry

There are two methods for entering data for a crosstabulation and chi-square.

Data entry procedure: method 1

Each person can be entered separately, using the procedure outlined in Chapter 2, remembering to code the categories appropriately as we will be using nominal data.



	party	tax
1	conserva	against
2	conserva	against
3	liberal	don't kno
4	liberal	for
5	liberal	against
6	conserva	don't kno

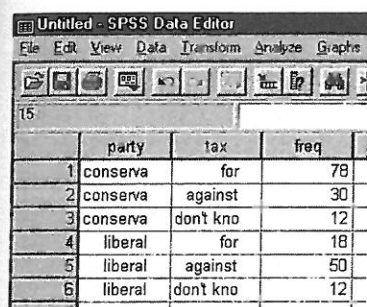
People are entered into the dataset separately, coding their individual preferences for both political party and taxation legislation

- ✓ Remember selecting **View** and **Value labels** can show the value labels given in the data entry procedure.

CROSSTABULATION AND CHI-SQUARE

Data entry procedure: method 2

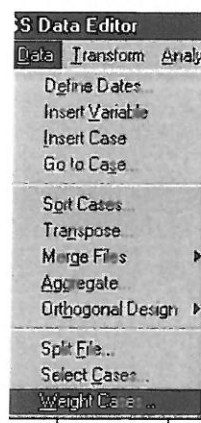
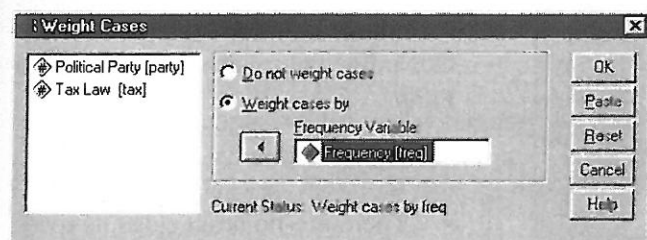
Total frequency counts can be entered by **Weight Cases**. This procedure is particularly useful if you have a large dataset. Each combination of scores is entered along with the total frequency count for that group, as seen in the example below.



	party	tax	freq
1	conserva	for	78
2	conserva	against	30
3	conserva	don't kno	12
4	liberal	for	18
5	liberal	against	50
6	liberal	don't kno	12

In order for SPSS to associate the frequency counts with the variables, the following procedure should be used.

- Go to **Data** and **Weight Cases** to bring up the following window.

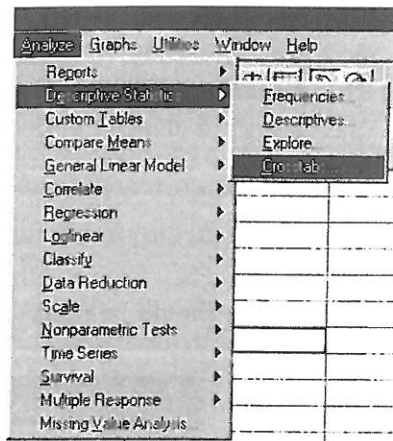


- The variable we want to weight the categories by is frequency, therefore select the **Weight Cases by** option and send the variable 'freq' over to the **Frequency Variable** box.
- Click on **OK**.

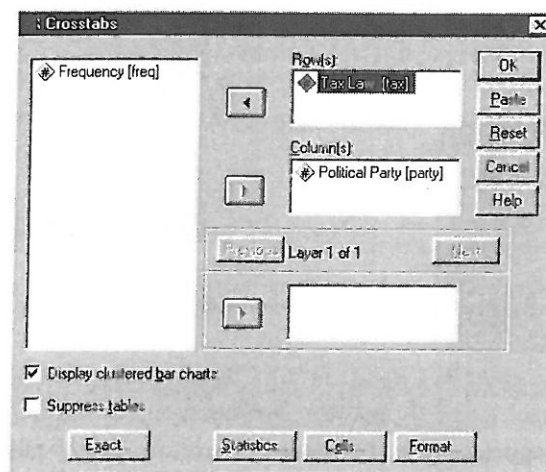
You can be sure that the procedure has been carried out successfully as in the bottom right-hand corner of the screen a message saying '**Weight On**' will be displayed.

Test procedure

To produce the crosstabulation table and chi-square the following procedure should be adopted.



- All inferential statistics are found under the **Analyze** command.
- A crosstabulation is the appropriate descriptive statistic for nominal or category data, so choose **Descriptive Statistics** then **Crosstabs**.
- The **Crosstabs** box will then appear.



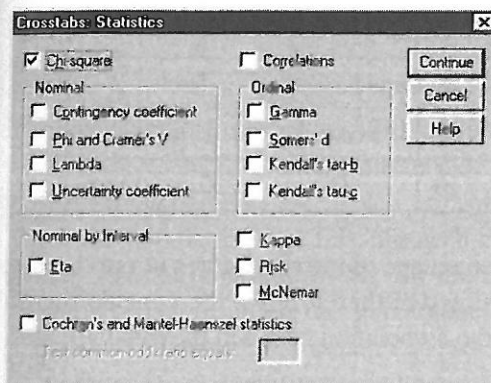
- At this point you need to decide which variable is to be considered your influencing variable and which is the one to be influenced.
- Our influencing variable is 'Political Party'.

✓ There are no strict rules as to which one goes where, although a general rule of thumb is that the influencing variable should go to the **Column(s)**, because when the output appears the convention is to read across the table rather than down a column.

- A clustered bar chart is the most appropriate way to display the findings of your crosstabulation, and this can be generated as an integral part of the procedure. If you would like to create this chart, place a tick against **Display clustered bar charts** in the **Crosstabs** box.

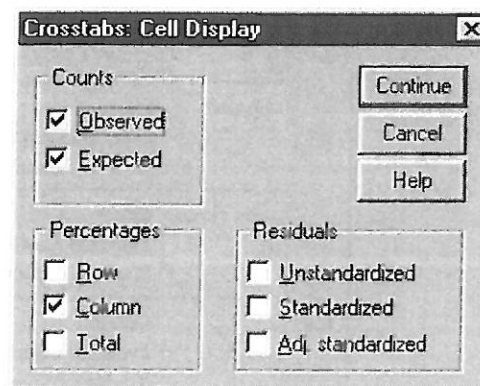
CROSSTABULATION AND CHI-SQUARE

- ✓ If you have set up your dataset by weighting cases, a third variable will appear in the column (frequency). You need to ignore this variable and select the two variables under investigation, the frequency count will automatically be taken into account if the 'Weight On' display is in the bottom right-hand corner of the screen.



- Click on **Statistics** to bring up the box to the left.
- The inferential statistic we want is the **Chi-square**. Place a tick next to this then click on **Continue**.

- Remember that when carrying out a crosstabulation and chi-square we are comparing how many people we have in a category with how many people we would expect to have in that category if the null hypothesis was true.
- In order to generate these we need to click on **Cells** to reveal the following box.
- Place a tick in both the **Observed** and **Expected** counts boxes so the two can be compared.
- A tick also needs to be placed in the **Percentages** box next to where you sent your influencing variable (generally the columns). Percentages give the best descriptor for nominal or category data as we often have unequal numbers in each group and therefore actual counts do not always give a representative picture. Percentages will give a proportion of each group total in each category.
- Click on **Continue** and finally on **OK** to complete the procedure.



SPSS output (2 × 3 crosstabulation and chi-square)

The following list of tables will be produced by SPSS when the above procedure has been completed.

The first table is the **Case Processing Summary**. This gives us a description of our dataset.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Tax Law * Political Party	200	100.0%	0	.0%	200	100.0%

It also gives us an indication of the number and percentage of missing scores in our dataset. As we can see, no scores in our dataset are missing and therefore we are considering all 200 participants when making judgements about the association between the two variables.

Influencing variable

Tax Law * Political Party Crosstabulation

Dependent variable

			Political Party		Total
			conservative	liberal	
Tax Law	for	Count	78	18	96
		Expected Count	57.6	38.4	96.0
		% within Political Party	65.0%	22.5%	48.0%
	against	Count	30	50	80
		Expected Count	48.0	32.0	80.0
		% within Political Party	25.0%	62.5%	40.0%
	don't know	Count	12	12	24
		Expected Count	14.4	9.6	24.0
		% within Political Party	10.0%	15.0%	12.0%
Total	Count		120	80	200
	Expected Count		120.0	80.0	200.0
	% within Political Party		100.0%	100.0%	100.0%

Total number of participants

The number of counts we would expect to find if the null hypothesis was true is shown in this cell

CROSSTABULATION AND CHI-SQUARE

SPSS essential

- The **Count** represents the actual frequency counts obtained in the survey.
- The **Expected Count** represents how many we would expect to find in this combination if the null hypothesis was true, i.e. if the distribution of frequencies was completely random.
- The percentages are a better representation of the data because of the uneven numbers in the categories. All columns add up to a total of 100 per cent as we asked for percentages for the columns, which is where we sent the influencing variable 'political party'.
- The percentages represent what the count is as a proportion of the total number for that column, for example 18 liberal voters who are in favour of the tax law account for 22.5 per cent of the total number of liberal voters, which is 80.
- The total percentages, which appear in the **Total** column, represent the percentage of the population that fell into each category of preference, for example 48.0 per cent of the total population were for the tax law irrespective of voting preference.
- As we have totalled our percentages in the columns because of the placement here of our influencing variable, we read the table across in order to ascertain if there is a difference between the percentages. If there does appear to be a difference, then there may be an association between the variables. The significance of this is assessed by the **chi-square** test statistic.
- In the above example we can see that within the group of people who were in favour of the tax law, 65.0 per cent of them were conservative compared to 22.5 per cent who were liberal. This trend is reversed within the group who are against the tax law, with 25.0 per cent being conservative and 62.5 per cent being liberal. The group who are unsure as to their opinions of the new tax law are reasonably evenly split in their voting preferences (10.0 per cent conservative and 15.0 per cent liberal).

An examination of the **Chi-Square Tests** table will enable us to ascertain whether the patterns identified above are significant.

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	35.938 ^a	2	.000
Likelihood Ratio	37.429	2	.000
Linear-by-Linear Association	22.908	1	.000
N of Valid Cases	200		

^a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 9.60.

SPSS essential

- The test statistic usually chosen to determine whether the association described by the crosstabulation is a significant one, is the **Pearson Chi-Square**. In our example above it can be seen that the value for Pearson chi-square is 35.938, with 2 degrees of freedom, which is significant at the $p < 0.001$ level.
- Results of chi-square are typically reported as follows:

$$\chi^2 = 35.938, df = 2, p < 0.001$$

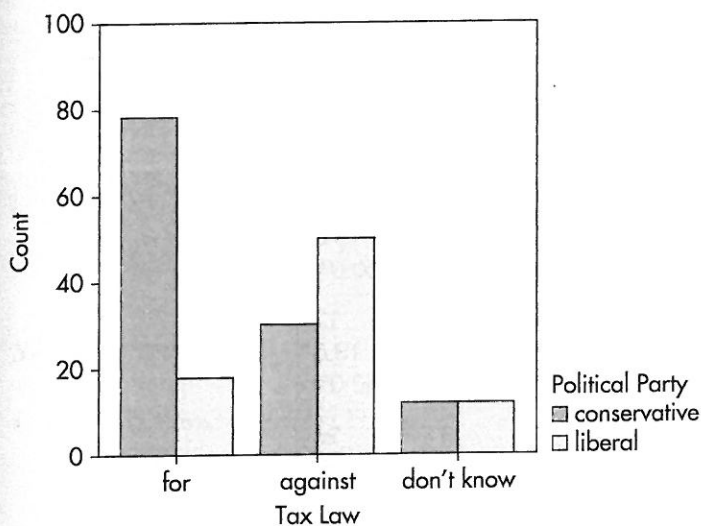
SPSS advanced

- From the above table it can be noted that the chi-square statistic assumes a non-directional hypothesis, which is indicated by the **Asymp. Sig. (2-sided)**. More details of the meaning of the asymptotic significance level can be found in Chapter 6. This refers to the approximations made in our p value calculations.
- The **Likelihood Ratio** is an alternative test to the chi-square employing a different method. Normally we use the chi-square result but this statistic is sometimes preferred when the sample size is small.
- The **Linear-by-Linear Association** is used when the categories are ordinal and we are interested in looking for a trend.
- The note at the bottom of the **Chi-Square Tests** table indicates if any individual cells in the crosstabulation table have expected counts less than five. For a 2×2 table SPSS will automatically perform a continuity correction. For a larger table if you have a

CROSSTABULATION AND CHI-SQUARE

number of values less than five you may wish to combine cells to increase the individual cell size.

The final part of the SPSS output is a summary of the frequency counts in the form of a clustered bar chart. This gives a helpful representation of the data. We can see how the frequencies of the different choices are linked to the political parties.



SPSS output (2×2 crosstabulation and chi-square)

- ✓ The data entry and procedure for a 2×2 crosstabulation are identical to those for the 2×3 crosstabulation detailed earlier so will not be shown.

A survey was carried out to investigate television viewing habits in different parts of the country. The researchers were particularly interested in the viewing figures for the two most popular television soaps at the time *Country Life* and *City Limits*, and whether there was an association between people's preferences and the part of the country the person lived in: in the north or the south.

The first two tables have the same layout as for the 2×3 crosstabulation, indicating the number of responses and any possible missing data, along with the crosstabulation frequencies of the two variables.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
REGION * TV	26	100.0%	0	.0%	26	100.0%

REGION * TV Crosstabulation

			TV		Total
			country life	city limits	
REGION	north	Count	10	3	13
		Expected Count	7.5	5.5	13.0
		% within TV	66.7%	27.3%	50.0%
	south	Count	5	8	13
		Expected Count	7.5	5.5	13.0
		% within TV	33.3%	72.7%	50.0%
Total	Count		15	11	26
	Expected Count		15.0	11.0	26.0
	% within TV		100.0%	100.0%	100.0%

SPSS essential

- It can be seen from the above **Crosstabulation** that there does appear to be an association between what region a person lives in and their preference for television soaps.
- People who live in the north are more likely to watch *Country Life* (66.7 per cent) than *City Limits* (27.3 per cent), with the reverse being true for people from the south.

The table that produces the test statistics is the **Chi-Square Tests** table. The information given in this table will enable us to judge whether the viewing pattern is significantly different from the pattern we would expect by chance.

CROSSTABULATION AND CHI-SQUARE

Test statistic		Chi-Square Tests				p value	
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)		
Pearson Chi-Square	3.939 ^b	1	.047				
Continuity Correction ^a	4.521	1	.112				
Likelihood Ratio	4.057	1	.044				
Fisher's Exact Test				.111	.055		
Linear-by-Linear Association	3.788	1	.052				
N of Valid Cases	26						

^a. Computed only for a 2 × 2 table

^b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.50.

SPSS essential

- The test statistic usually chosen to determine whether the association described by the crosstabulation is a significant one is the **Pearson Chi-Square**. In our example above it can be seen that the test statistic value for Pearson chi-square is 3.939, with 1 degree of freedom, and as the *p* value is smaller than 0.05, we can conclude that there is a significantly different viewing patterns in the two regions ($p < 0.05$).
- Results of chi-square are typically reported as follows:

$$\chi^2 = 3.939, df = 1, p < 0.05$$

- We have a 2 × 2 table with fairly small values, plus the probability of our value is close to the significance level, so you may wish to interpret the result with caution (and read the *SPSS advanced* section below).

SPSS advanced

- The **Continuity Correction** is produced automatically by SPSS for a 2 × 2 table, which performs a Yates' correction. This is a conservative adjustment and the use of this correction is controversial and often not recommended. We advise the use of **Fisher's Exact Test** in analysing a 2 × 2 table with a cell frequency less than five (see below).

- The **Likelihood Ratio** is an alternative test to the chi-square employing a different method. Normally we use the chi-square result but the likelihood ratio is sometimes preferred when the sample size is small.
- **Fisher's Exact Test** calculates the exact probability based on a calculation of all the possible combinations of frequency distributions in the crosstabulation table. This is only calculated for a 2×2 table because of the increase in possible combinations of frequency distributions afforded by a bigger crosstabulation. As the actual probability is calculated, it is often more conservative. As can be seen in the above example, the **Pearson Chi-Square** value = 3.939, which gives a probability of $p = 0.047$. However, when calculations are based on **Fisher's Exact Test** we find $p = 0.111$.
- The **Linear-by-Linear Association** is used when the categories are ordinal. This is also interpreted in the same way as the chi-square test statistic.
- The **Asymp. Sig.** is calculated as the default probability by SPSS. This is an approximation that is calculated for a dataset which is large and adequately distributed.

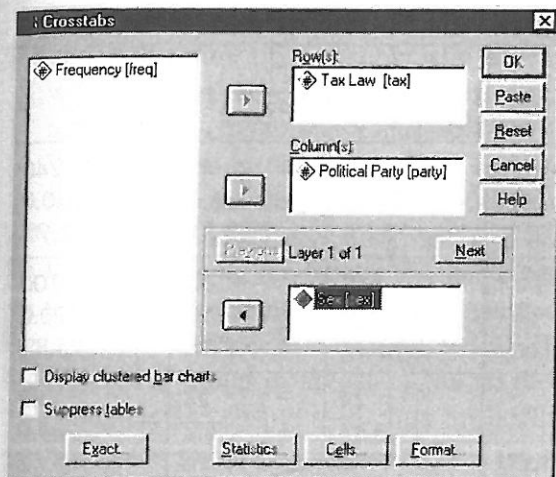
Test procedure: layered 2×3 crosstabulation and chi-square

Sometimes we may want to add a third variable to the crosstabulation calculation in order to obtain a more accurate picture of an association and investigate any possible influences of a third variable. The procedure for setting up the dataset and initial steps of analysis are identical to those previously described. An extra column needs to be added to the dataset for the extra variable and the possible combinations of the groups created.

- ✓ If you are entering data as shown, with the total number of male conservatives who were in favour of the tax law etc., then remember to weight the cases.

	sex	party	tax	freq
1	male	conserva	for	200
2	male	conserva	against	40
3	male	conserva	don't kno	24
4	female	conserva	for	112
5	female	conserva	against	80
6	female	conserva	don't kno	24
7	male	liberal	for	40
8	male	liberal	against	60
9	male	liberal	don't kno	24
10	female	liberal	for	32
11	female	liberal	against	140
12	female	liberal	don't kno	24

CROSSTABULATION AND CHI-SQUARE



- The differences in the procedural element arise when creating layers for our crosstabulation, which are formed by the addition of a third variable.
- In our example we are still examining voting preferences and their relationship to opinions on a tax law. However, we wish to further investigate this relationship by examining the influence that sex may have.
- 'Party' and 'Tax' are entered into the crosstabulation as previously described, and 'Sex' is entered as the first layer.

SPSS output (layered 2 x 3 crosstabulation and chi-square)

The **Case Processing Summary** table again indicates the number of valid and missing cases in our dataset. As before we have no cases missing and are therefore considering all 800 cases when making our judgements about the relationship between the three variables.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Tax Law * Political Party * Sex	800	100.0%	0	.0%	800	100.0%

The **Crosstabulation** table indicates any patterns in our data. As we can see this is layered into our third variable 'sex'.

Tax Law * Political Party * Sex Crosstabulation

Tax Law + Political Party

sex cross-tabulation

sex				Political Party		Total
				conservative	liberal	
male	Tax Law	for	Count	200	40	240
			Expected Count	163.3	76.7	240.0
			% within Political Party	75.8%	32.3%	61.9%
		against	Count	40	60	100
			Expected Count	68.0	32.0	100.0
			% within Political Party	15.2%	48.4%	25.8%
		don't know	Count	24	24	48
			Expected Count	32.7	15.3	48.0
			% within Political Party	9.1%	19.4%	12.4%
Total			Count	264	124	388
			Expected Count	264.0	124.0	388.0
			% within Political Party	100.0%	100.0%	100.0%
female	Tax Law	for	Count	112	32	144
			Expected Count	75.5	68.5	144.0
			% within Political Party	51.9%	16.3%	35.0%
		against	Count	80	140	220
			Expected Count	115.3	104.7	220.0
			% within Political Party	37.0%	71.4%	53.4%
		don't know	Count	24	24	48
			Expected Count	25.2	22.8	28.0
			% within Political Party	11.1%	12.2%	11.7%
Total			Count	216	196	412
			Expected Count	216.0	196.0	412.0
			% within Political Party	100.0%	100.0%	100.0%

SPSS essential

- The **Crosstabulation** table is split between the sexes in order to gain a fuller understanding of the possible relationships between the variables.
- By examining the table above we can see that the relationships between voting preferences and tax law opinions are not the same when sex has been taken into consideration.

CROSSTABULATION AND CHI-SQUARE

- The relationship between voting preferences and tax law identified previously suggested that most of the people who were in favour of the tax law were conservative, while most of those against the law were liberal. Of the group who were unsure of their opinions, there was a reasonably even split in voting preferences.
- When examining the male voters we can see that most male respondents were in favour of the tax law, regardless of their voting preferences (61.9 per cent in favour compared to 25.8 per cent against, and 12.4 per cent of no opinion).
- When examining the female voters we can see that this trend is reversed, with most females being against the tax law regardless of their voting preferences (35.0 per cent in favour, 53.4 per cent against, and 11.7 per cent of no opinion).
- Within the group in favour of the tax law, the majority were conservative in their voting preferences; this was true for both males and females.
- Within the group against the tax law, the majority were liberal in their voting preferences; this was true for both males and females.
- From the **Crosstabulation** we may summarise that the relationship between voting preferences and tax law opinions identified in the 2×3 crosstabulation holds across the sexes. However, the sexes in this sample do differ in their overall opinion about the tax law once voting preferences are separated, with the majority of males being in favour of the new law, and the majority of females being against the new law. This leads to a more complicated relationship between political party affiliation and opinions.

An examination of the **Chi-Square Tests** table will enable us to ascertain whether the patterns identified above are significant.

Chi-Square Tests

Sex		Value	df	Asymp. Sig. (2-sided)
male	Pearson Chi-Square	69.155 ^a	2	.000
	Likelihood Ratio	68.795	2	.000
	Linear-by-Linear Association	48.904	1	.000
	N of Valid Cases	388		
female	Pearson Chi-Square	59.979 ^b	2	.000
	Likelihood Ratio	62.673	2	.000
	Linear-by-Linear Association	33.459	1	.000
	N of Valid Cases	412		

^a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15.34.

^b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 22.83.

SPSS essential

- The test statistic usually chosen to determine whether the association described by the crosstabulation is a significant one is the **Pearson Chi-Square**.
- In the example above we have two values for this as we are examining the relationship across two groups of people, males and females, which creates a layer to our crosstabulation. For the male participants it can be seen that the test statistic value for Pearson chi-square is 69.155, with 2 degrees of freedom and as, the p value is smaller than 0.001, we can conclude that this pattern of scores is significant.
- Results of chi-square are typically reported as follows:

$$\chi^2 = 69.155, df = 2, p < 0.001$$

- For the female participants the relationship is also significant, with a Pearson chi-square value of 59.979, 2 degrees of freedom and, as the p value is smaller than 0.001, we can conclude that this pattern of scores is significant.
- Results of chi-square are typically reported as follows:

$$\chi^2 = 59.979, df = 2, p < 0.001$$

SPSS advanced

- From the above table it can be noted that the chi-square statistic assumes a non-directional hypothesis, which is indicated by the **Asymp. Sig. (2-sided)**.
- The **Likelihood Ratio** is an alternative test to the chi-square employing a different method. Normally we use the chi-square result but the likelihood ratio is sometimes preferred when the sample size is small.
- The **Linear-by-Linear Association** is used when the categories are ordinal and we are interested in looking for a trend.
- The note at the bottom of the **Chi-Square Tests** table indicates if any individual cells in the crosstabulation table have expected counts less than five. For both sexes none of the expected counts are less than five (0 per cent).

Chi-square as a 'goodness of fit' test

As mentioned in the introduction, chi-square can also be used as a 'goodness of fit' test. This assesses whether the pattern of results matches (or fits) a predicted pattern.

CROSSTABULATION AND CHI-SQUARE

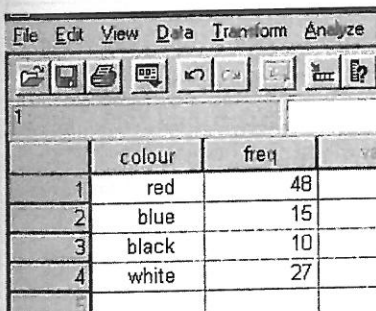
Scenario

An experimenter set out to test whether there is a difference in people's colour preference for cars. One hundred participants were given four pictures of cars, identical but for the colour, and asked to state their preference. The colours presented were red, blue, black and white.

If there was no preference then we would expect each colour to be chosen equally, so we would expect the probability of each category chosen to be a quarter or $p = 0.25$ when the null hypothesis is true. With a total (N) of 100 participants we would expect each category to be chosen by $N \times p$ participants, 100×0.25 , which is 25 participants.

On performing the experiment, the researcher finds 48 participants choose the red car, 15 the blue, 10 the black, and 27 the white. Do these observed frequencies differ significantly from the expected frequencies?

Test procedure

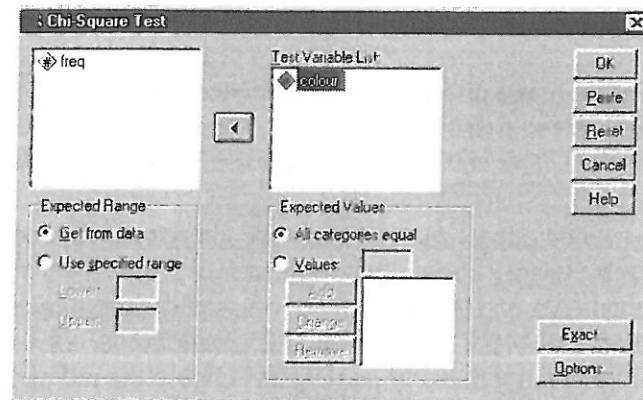


	colour	freq
1	red	48
2	blue	15
3	black	10
4	white	27

- ✓ The data can be entered as frequency counts if the **Weight Cases** command is used. See data entry method 2 earlier in this chapter

- From the drop-down menu, choose **Analyze**, **Nonparametric Tests**, then **Chi-Square**.





- The test variable 'colour' should be sent to the **Test Variable List**. The variable 'freq' can be left because if the data has been entered via the weight cases method, this is no longer necessary.
- As we are testing our null hypothesis that the numbers in each colour category will be equal, we can leave the **Expected Values** box set to the default as **All categories equal**.

- ✓ If we were testing an assumption that certain proportions were not equal but followed a specific pattern, for example 40 per cent of participants would prefer black cars, with the other preferences being equal, then we would enter the respective values by selecting the **Values** option, entering the frequencies, then clicking **Add**.

SPSS output

The first table produced by SPSS shows the observed and expected counts for the four car colours.

COLOUR			
	Observed N	Expected N	Residual
red	48	25.0	23.0
blue	15	25.0	-10.0
black	10	25.0	-15.0
white	27	25.0	2.0
Total	100		

CROSSTABULATION AND CHI-SQUARE

SPSS essential

- The **Observed** counts show us how many people preferred each colour car.
- The **Expected** counts confirm how many people we would expect to choose each colour if the null hypothesis was true. We can see from the table that the observed and expected counts are quite different from each other. This is confirmed by inspection of the **Residual** column. The residual values indicate that only the observed count of preferences for white cars approximates to the expected count. Both blue and black cars fall short of the expected count as indicated by the minus figure, while red cars exceed the expected count by 23.0.

In order to assess whether the distribution found through data collection differs from the expected pattern, the **Test Statistics** table needs to be examined.

Test Statistics

	COLOUR	Test statistic
Chi-Square ^a	34.320	
df	3	<i>p</i> value
Asymp. Sig.	.000	

^a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 25.0.

SPSS essential

- As can be seen from the table above $\chi^2 = 34.320$, $df = 3$, $p < 0.001$.

✓ Remember to change the last zero to a 1 if the *p* value given is 0.000.

- We can therefore conclude that there is a significant difference between the observed and expected counts, and we can therefore reject the null hypothesis; the four colours are not equally preferred.
- The findings can be expressed as follows:

$$\chi^2 = 34.320, df = 3, p < 0.01$$

FAQ

In my dataset one variable is measured on an interval scale and one has dichotomous categories. If I want to look for associations can I use this data in chi-square?

Yes, all you need to do is to convert your interval scale into groups/categories. See Chapter 5 for details on how to recode your data.

I have calculated my chi-square using SPSS, but I think I may have done something wrong in the procedure. The associations that I am describing do not seem to make sense in light of my hypothesis.

Check that you are reading the rows and columns the right way round. Remember that, as a common rule, the influencing variable should be sent to the columns. This will mean that the Crosstabulation table produced by SPSS can be read across the rows. It does not matter however if you have sent your influencing variable to the rows, it just means that your Crosstabulation table should be read down the columns.

Details on how to calculate a chi-square test by hand can be found in Chapter 19 of Hinton (2004).