Analysing Quantitative Data
Computer Lab 3

Welcome to the third quantitative data analysis lab for Social Research Methods. This workbook will talk you through the activities you should undertake in the third lab session. Please work through these activities at your own pace. Please feel free to work in pairs and to discuss the activities with your classmates. Teaching staff will be on hand to answer any questions, please do not hesitate to ask for assistance.

In this session you will learn how to analyse the association between two categorical variables using the chi-square test.

You can download SPSS to use on your own computer here. If you have any difficulty downloading SPSS you should seek assistance from the IT Services Help Desk.

So far in the quantitative data analysis labs you have been describing patterns in a sample of respondents. The key aim of quantitative data analysis is to understand patterns in the wider population, not only in the sample. This is called making inferences, and the techniques which allow us to do this are called 'inferential statistics'. We can only make inferences successfully if we have good quality, unbiased, samples that is why sampling is so important in quantitative social science research.

There are many different inferential statistics, in this module you will learn how to perform one – the Chi-Square Test. This test is used when you have two categorical variables (but remember you can always recode scale variables into categorical variables). This test investigates if there is a relationship between the two variables which is likely to be seen in the wider population (i.e. statistical significance). If there is a reliable relationship between the two variables we can also determine how strongly the two variables are related (i.e. effect size).

- Open SPSS.
- Open the SRM Practice Data. These data can be downloaded from the VLE. Remember to open data you select: **FILE**, **OPEN**, **DATA**.
- Open the Lab 3 Syntax File. This file can be downloaded from the VLE. Remember that to open a syntax file you select: **FILE**, **OPEN**, **SYNTAX**.

The syntax file contains all the code you will be asked to type throughout this workbook. Work through the syntax file in conjunction with the workbook. Remember to 'run' a piece of code you highlight it and then click on the green triangle. Your results will appear in an Output Viewer window.

Section 1: Chi-Square Test

Example 1

In this example the research question we are interested in is: 'Is being a migrant related to worry about being insulted or pestered?'

- The null hypothesis associated with this research question is: 'There is no association between being a migrant and worry about being insulted or pestered.'
- The alternative hypothesis associated with this research question is: 'There is an association between being a migrant and being worried about being insulted or pestered.'

First we should examine the univariate descriptive statistics for the two variables we are analysing, 'migrant' and 'insult'. This gives you an idea of how many migrants there are in your sample, and how many people reported that they worry about being insulted or pestered.

Examining the variables also lets you see how much missing data there is for a question. If a variable has lots of missing data you should question why this might be, and whether this will introduce bias into your research.

Analysis should always start by examining simple univariate statistics.

The code to examine these variables is shown below, run this code and look carefully at the output.

```
FREQUENCIES migrant.

FREQUENCIES insult.
```

Next you can produce your contingency table and run your chi-square tests. The code to do this is shown below. You will notice that this code is very similar to the code we used last week to produce contingency tables, with the addition of an extra line. Run this code and take a look at the output in the Output Viewer window.

```
CROSSTABS migrant BY insult
  /CELLS count row
  /STATISTICS CHISQ PHI.
```

This piece of code asks SPSS to produce several tables. **SPSS provides far more information than you actually need. You only need to look at the numbers which are highlighted in the workbook, you can ignore the additional information.**

The first table is the 'Case Processing Summary', this summarises how many 'valid' responses there are and how much missing data there is. This information may be slightly different to the information you see when examining the variables individually. If information is missing for one of your variables this respondent will not be included in your chi-square analysis (as information for both variables is required).

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Respondent is a first generation migrant to the UK * How worried are you about being insulted or pestered by anybody?	11657	99.8%	19	0.2%	11676	100.0%

The table above shows that 11,657 respondents are included in the chi-square analysis (i.e. valid) and 19 respondents were excluded from the analysis as they had missing data on at least one of the variables.

The next table SPSS shows is your contingency table. We are interested in whether being a migrant is associated with worry about being insulted or pestered, therefore it makes sense that our first variable forms ('migrant') the rows of the table and our second variable ('insult') forms the columns of the table. We can see that 6% of respondents who are non-migrants are very worried about being insulted or pestered, and 21% are fairly worried. In contrast 10% of respondents who are migrants are very worried about being insulted or pestered, and 28% of migrants are fairly worried.

Respondent is a first generation migrant to the UK * How worried are you about being insulted or pestered by anybody? Crosstabulation

			How worried are you about being insulted or pestered by anybody?				
			1 very worried	2 fairly worried	3 not very worried	4 not at all worried	Total
Respondent is a first generation migrant to the UK	0 not a migrant	Count	576	1950	4218	2557	9301
		% within Respondent is a first generation migrant to the UK	6.2%	21.0%	45.3%	27.5%	100.0%
	1 migrant	Count	238	659	1033	426	2356
		% within Respondent is a first generation migrant to the UK	10.1%	28.0%	43.8%	18.1%	100.0%
Total		Count	814	2609	5251	2983	11657
		% within Respondent is a first generation migrant to the UK	7.0%	22.4%	45.0%	25.6%	100.0%

The contingency table tells us about patterns in the sample of respondents, but we want to know whether there is likely to be a reliable relationship between these variables in the wider population. In order to draw conclusions about the wider population we need to use an inferential test, such as the chi-square test. The chi-square test is used to examine the relationship between two categorical variables.

The 'Chi-Square Tests' table shows the chi-square statistic. You can find the Chi-Square statistic in the first row of the table under the column 'value' (see red box, blue arrow). In this example the chi-square value is 148.351.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	148.351[a]	3	.000
Likelihood Ratio	148.080	3	.000
Linear-by-Linear Association	147.685	1	.000
N of Valid Cases	11657		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 164.52.

The p-value associated with the chi-square statistic is less than 0.05 (see red box, green arrow). Note that the p-value appears here as 0.000, this does not mean that the p-value is zero (it is never zero), it just means that it is too small to appear here. This is not a problem, all you need to know is whether this value is greater than or less than 0.05. We can very clearly see that the value 0.000 is less than 0.05.

As the p-value is less than 0.05 you can conclude that your chi-square test is significant, this means that it is very unlikely that you would see a pattern like this in your data if your null hypothesis was true. Therefore we can reject the null hypothesis and be fairly confident that there is a reliable association between these two variables in the population. In other words you can

conclude that your results suggest that there is an association between being a migrant and worry about being pestered or insulted.

This table also tells us that the degrees of freedom of this table is 3 (see red box, red arrow). This is calculated by multiplying the number of categories in the row variable minus 1 (i.e. 2 - 1) by the number of categories in the column variable minus 1 (i.e. 3 - 1). One multiplied by 3 is 3, so the degrees of freedom is 3.

The chi-square statistic tells us that there is likely to be an association between the two variables, but it does not tell us how strong this relationship is. There could only be a little difference between migrants and non-migrants, or there could be a massive difference. This is an important piece of information when interpreting your research. Examining 'effect size' provides information on how strongly the two variables are associated.

The effect size statistic, Cramer's V, is shown in the table 'Symmetric Measures' (red box, blue arrow). Cramer's V = 0.113 for the association between migrant status and worry about being insulted or pestered. According to Cohen's (1988) criteria, Cramer's V values of less than 0.29 indicate that there is a small effect, values between 0.30 and 0.49 indicate that there is a medium effect, and values of 0.50 or above indicate that there is a large effect.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd Edition. Hillsdale, N.J.: Lawrence Erlbaum.

The Cramer's V value of 0.113 is less than 0.29 which indicates that while there is an association between these two variables this association is quite small, so there are not likely to be major differences between migrants or non-migrants.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.113	.000
	Cramer's V	.113	.000
N of Valid Cases		11657	

Please note that if your Cramer's V value is ever negative, you ignore the negative sign when interpreting the size of the effect.

Based on these results we can make two conclusions about the association between migrant status and worry about being pestered or insulted.

1. Migrant status is significantly associated with worry about being pestered or insulted.
2. The effect size is considered to be small according to Cohen's (1988) guidelines.

We can write this result up formally as follows:

> A chi-square test indicated that there was a significant association between migrant status and worry about being pestered or insulted with a small effect size, $\chi 2$ (3, n=11,657) = 148.351, $p < 0.05$, Cramer's V = 0.113.

- The number highlighted in red is the degrees of freedom.
- The number highlighted in green is the sample size.
- The number highlighted in yellow is the chi-square statistic.
- The section highlighted in blue refers to the size of your p-value.
- The number highlighted in pink is the Cramer's V value.

Example 2

In this example the research question we are interested in is: 'Is sex associated with experience of discrimination?'

- The null hypothesis associated with this research question is: 'There is no association between 'sex' and experience of discrimination.'
- The alternative hypothesis associated with this research question is: 'There is an association between 'sex' and experience of discrimination.'

First we start with examining the two variables. Run the code below. Take a careful look at the variables 'sex' and 'discrim'.

```
FREQUENCIES sex.

FREQUENCIES discrim.
```

Next you can produce your contingency table and run your chi-square test. The code to do this is shown below. Run this code and carefully look at the output in the Output Viewer window.

```
CROSSTABS sex BY discrim
  /CELLS count row
  /STATISTICS CHISQ PHI.
```

We can see from the table below that 11,654 respondents had valid information for the two variables, and 22 respondents (less than 1%) of the respondents had missing data and were therefore excluded from the chi-square analysis.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Respondent's Sex * In the past 5 years, how often do you feel that you have experienced discrimination or unfair treatment?	11654	99.8%	22	0.2%	11676	100.0%

From the contingency table we can see that 6% of male respondents reported that they felt they had experienced discrimination often in the last five years. In contrast, 15% female respondents felt that they had experienced discrimination often in the last five years. The pattern of responses we can see in the contingency table suggest that, in this sample of respondents, women are more likely than men to feel that they have experienced discrimination or unfair treatment in the last five years.

Respondent's Sex * In the past 5 years, how often do you feel that you have experienced discrimination or unfair treatment? Crosstabulation

			In the past 5 years, how often do you feel that you have experienced discrimination or unfair treatment?			
			1 Often	2 Sometimes	3 Rarely	Total
Respondent's Sex	1 male	Count	327	1081	3894	5302
		% within Respondent's Sex	6.2%	20.4%	73.4%	100.0%
	2 female	Count	944	1643	3765	6352
		% within Respondent's Sex	14.9%	25.9%	59.3%	100.0%
Total		Count	1271	2724	7659	11654
		% within Respondent's Sex	10.9%	23.4%	65.7%	100.0%

In order to make inferences from this sample to the wider population we need to examine the chi-square test results. We can see that the chi-square statistic is 325.682 and the degrees of freedom value is 2. The p-value is less than 0.05. Therefore we can conclude that this test is significant. The significant chi-square test result suggests that it is very unlikely that you would see a pattern like this in your data if your null hypothesis was true. Therefore we can reject the null hypothesis and be fairly confident that there is a reliable association between these two variables in the population. In other words you can conclude that your results suggest that there is an association between 'sex' and experience of discrimination.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	325.682[a]	2	.000
Likelihood Ratio	336.784	2	.000
Linear-by-Linear Association	324.330	1	.000
N of Valid Cases	11654		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 578.24.

To examine how strong this association is we should look at an effect size measure. Cramer's V is 0.167, this is less than 0.29 which suggests that this is a small effect size.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.167	.000
	Cramer's V	.167	.000
N of Valid Cases		11654	

We can write this result up formally as follows:

> A chi-square test indicated that there was a significant association between 'sex' and experience of discrimination with a small effect size, χ^2 (2, n=11,654) = 325.682, p < 0.05, Cramer's V = 0.167.

Section 2: Chi-Square Test Assumptions

When we run inferential tests there are rules that must be followed in order to ensure that your results are robust and reliable. These are called 'assumptions'. For the chi-square test there are two assumptions to satisfy:

1. Each respondent should only appear in one cell of your contingency table. You would violate this assumption (i.e. break this rule) if respondents were allowed to provide multiple answers to a question (i.e. tick all that apply). This assumption is known as the 'independence' assumption.
2. You should have a sufficient sample size for your analysis. If you have more cells (i.e. your variables have more categories) you will need a larger sample size. The expected cell count for you contingency table should not be less than 5 for more than 20% of your cells (i.e. you should not have lots of cells where you would expect very few people to appear).

Note that this refers to the **expected** cell count, not the number of people you actually observe. It is ok to have observed cell counts of less than 5.

The first assumption you can check by looking at the survey question. For the second assumption, SPSS will helpfully tell you if this is a problem. In the example below you will see how to check this assumption, and what to do if you violate this assumption.

In this example we are interested in the research question: 'Is ethnicity associated with worry about being pestered or insulted?'

- The null hypothesis associated with this research question is: 'There is no association between ethnicity and worry about being pestered or insulted.'
- The alternative hypothesis associated with this research question is: 'There is an association between ethnicity and worry about being pestered or insulted.'

As usual you should first examine your univariate descriptive statistics. You can use the code below to examine the two variables 'ethgrp' and 'insult'.

```
FREQUENCIES ethgrp.

FREQUENCIES insult.
```

Now you can produce your contingency table and chi-square test. You can use the code below.

```
CROSSTABS ethgrp BY insult
  /CELLS count row
  /STATISTICS CHISQ PHI.
```

If you look at the 'Chi-Square Tests' table you will notice a footnote. This footnote is always included. Here the footnote indicates that 1 cell (5%) has an expected cell count less than 5. This is where SPSS tells you whether you have violated the assumption.

Remember that the assumption is that the expected cell count for you contingency table should not be less than 5 for more than 20% of your cells. Here we have not violated this assumption as only 5% of cells have an expected cell count of less than 5. Therefore we can be confident that we have met all the assumptions (i.e. followed all the rules) and we can proceed with the interpretation of the results.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	171.980[a]	12	.000
Likelihood Ratio	146.655	12	.000
Linear-by-Linear Association	74.982	1	.000
N of Valid Cases	11654		

a. 1 cells (5.0%) have expected count less than 5. The minimum expected count is 3.91.

Imagine that you had violated the assumption, what would you? Would you not be able to examine this research question? No, we can attempt to 'fix' this problem by recoding our variables into a smaller number of categories.

Take a look at the 'ethgrp' variable again using the code below.

```
FREQUENCIES ethgrp.
```

You can see that there are five possible ethnic group categories. We could reduce the number of categories. Plausibly the category 'mixed' is a varied group of respondents so we could theoretically group these with the 'other' respondents (if we had a need to reduce the number of categories).

Respondent's Ethnicity

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 white	10900	93.4	93.4	93.4
	2 mixed	56	.5	.5	93.9
	3 asian or asian british	380	3.3	3.3	97.1
	4 black or black british	203	1.7	1.7	98.9
	5 chinese or other	134	1.1	1.1	100.0
	Total	11673	100.0	100.0	
Missing	999	3	.0		
Total		11676	100.0		

Here we are going to recode the variable 'ethgrp'. If you cannot remember the process of recoding a variable, take a look back at workbook from last week. Here we tell SPSS that we want to take all the respondents currently in category 2 of the 'ethgrp' variable and put them in category 5. Run this code.

```
RECODE ethgrp (2=5).
```

Now we need to update the labels for the categories as category 5 now contains, 'Chinese / Other / Mixed'. Run the code below.

```
VALUE LABELS ethgrp
  1 "White"
  3 "Asian / Asian British"
  4 "Black / Black British"
  5 "Chinese / Other / Mixed".
```

Now you can take a look at the 'ethgrp' variable and see the changes you have made.

```
FREQUENCIES ethgrp.
```

You can see that everyone from category two have now been placed in category five, and the category labels have been updated to reflect this.

Respondent's Ethnicity

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 White	10900	93.4	93.4	93.4
	3 Asian / Asian British	380	3.3	3.3	96.6
	4 Black / Black British	203	1.7	1.7	98.4
	5 Chinese / Other / Mixed	190	1.6	1.6	100.0
	Total	11673	100.0	100.0	
Missing	999	3	.0		
Total		11676	100.0		

Now we can rerun the chi-square test, using the code below.

```
CROSSTABS ethgrp BY insult
  /CELLS count row
  /STATISTICS CHISQ PHI.
```

You will notice that the footnote in the 'Chi-Square Tests' table has changed and there are no longer any cells with expected cell counts of less than 5.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	167.771[a]	9	.000
Likelihood Ratio	143.389	9	.000
Linear-by-Linear Association	75.728	1	.000
N of Valid Cases	11654		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 13.25.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.120	.000
	Cramer's V	.069	.000
N of Valid Cases		11654	

Take a moment to interpret the results of this chi-square test. What do these results tell us about the association between ethnicity and worry about being pestered or insulted?

Section 3: Chi-Square Test (Non-Significant Results)

In this example we are interested in the research question: 'Is social class identity associated with perception of the main cause of crime?'

- The null hypothesis associated with this research question is: 'There is no association between social class identity and perception of the main cause of crime.'
- The alternative hypothesis associated with this research question is: 'There is an association between social class identity and perception of the main cause of crime.'

First you should examine the univariate descriptive statistics for the variables 'class' and 'cause'. You can do this using the below.

```
FREQUENCIES class.

FREQUENCIES cause.
```

Now you can produce your contingency table and run your chi-square test. You can do this using the code below.

```
CROSSTABS class BY cause
  /CELLS count row
  /STATISTICS CHISQ PHI.
```

Looking at this contingency table there are not any clear patterns in the sample. For example 30% of respondents who identify as 'working class' believe drugs are the main cause of crime, this is similar to the proportion amongst 'middle class' respondents (31%), and 'Upper Class' respondents (30%). We cannot see any notable differences in the views of respondents with different social class identities by inspecting the contingency table.

Respondent's Social Class Identity * What is the one main cause of crime in britain today? Crosstabulation

			What is the one main cause of crime in britain today?										
			1 a. too lenient sentencing	2 b. poverty	3 c. lack of discipline from school	4 d. lack of discipline from parents	5 e. drugs	6 f. alcohol	7 g. unemployment	8 h. breakdown of family	9 i. too few police	10 j. do not think there is one main cause	Total
Respondent's Social Class Identity	1 Working Class	Count	621	219	227	1941	1888	372	160	351	210	242	6231
		% within Respondent's Social Class Identity	10.0%	3.5%	3.6%	31.2%	30.3%	6.0%	2.6%	5.6%	3.4%	3.9%	100.0%
	2 Middle Class	Count	137	54	31	359	376	67	38	74	34	39	1209
		% within Respondent's Social Class Identity	11.3%	4.5%	2.6%	29.7%	31.1%	5.5%	3.1%	6.1%	2.8%	3.2%	100.0%
	3 Upper Class	Count	373	141	120	1132	1077	204	83	225	125	144	3624
		% within Respondent's Social Class Identity	10.3%	3.9%	3.3%	31.2%	29.7%	5.6%	2.3%	6.2%	3.4%	4.0%	100.0%
Total		Count	1131	414	378	3432	3341	643	281	650	369	425	11064
		% within Respondent's Social Class Identity	10.2%	3.7%	3.4%	31.0%	30.2%	5.8%	2.5%	5.9%	3.3%	3.8%	100.0%

Now we examine the association using the chi-square test. The chi-square statistics is 16.908 and there are 18 degrees of freedom. The p-value is 0.529 this is greater than 0.05. Therefore we can conclude that this test is not significant. The non-significant chi-square test result suggests that we could see a pattern like this in the data if the null hypothesis was true. Therefore we cannot reject the null hypothesis. In other words the results suggests that there is not likely to be a relationship between social class identity and views about the cause of crime in the population.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	16.908[a]	18	.529
Likelihood Ratio	17.037	18	.521
Linear-by-Linear Association	.029	1	.866
N of Valid Cases	11064		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 30.71.

Effect size statistics tell us about the strength of association between two variables. As we have concluded that there is not a reliable association between these two variables there is no reason to examine the strength of this association, so there is no need to refer to the effect size statistic.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.039	.529
	Cramer's V	.028	.529
N of Valid Cases		11064	

You may be thinking that getting a non-significant result is a bad thing and something you should avoid, this is absolutely not the case. Non-significant results are equally as valuable as significant results and tell as equally important information. Non-significant results should be written up and discussed.

In relation to this example we can conclude that this analysis suggests that social class identity is not associated with perceptions of the main cause of crime. We can still relate this to the literature and discuss why this pattern might be observed.

It is very important that you don't give up on an analysis or change your research question because you find a non-significant result. This practice is described as 'the file drawer problem', and failure to present non-significant results can lead to bias in our understanding of social phenomenon.

We can write this result up formally as follows:

> A chi-square test indicated that there was not a significant association between social class identity and perception of the main cause of crime, χ^2 (18, n=11,064) = 16.908, $p > 0.05$.

Section 4: Independent Practice

Practice running chi-square tests using other variables in the data set. Before you run the test think about the research question you would be asking by investigating the relationship between these two variables.

- Consider whether your variables need recoding.
- Check you have met the assumption of having no more than 20% of your cells with an expected value less than 5.

When interpreting the results of your chi-square test you should examine whether the association is significant or not, and if your result is significant you should consider the effect size of the association.

You have now completed the third SPSS computer lab, well done!

You now have all the skills required to undertake the analysis for your assessment. In next week's lab you will have the opportunity to develop the analysis for your assessment. You will get the most out of next week's session if you practice the activities included in this workbook before attending.

Please remember that additional supporting resources are available on the VLE, and please do not hesitate to ask if you require any additional support or assistance.