The University Of Sheffield.

The Higher Education Academy

4. CHI-SQUARE: INTRODUCING THE 'GOODNESS OF FIT' TEST AND THE 'TEST OF ASSOCIATION'

Dr Tom Clark & Dr Liam Foster

Department of Sociological Studies | University of Sheffield

CONTENTS

4.

Chi-square

*So now you should be able to undertake a descriptive analysis of your data. However, what if you want to do something more with your data than describe the properties of your sample? What if you want to infer something about a population based on that sample?

4.1 Statistical tests

Statistical tests concerned with effects, associations, differences, and relationships, allow us to infer answers to our research questions. This is where the null and alternative hypotheses become important. Whereas descriptive techniques enable us to describe data, inferential statistics allow us to infer answers from that data using the hypotheses that emerge from our research rationales and research questions. In essence, the job of an inferential statistical test is to allow us make as assessment of which hypothesis is more likely to be the 'way things really are' in a population by looking at a smaller part of it. Indeed, inferential statistics are employed to make judgments concerning the probability that our observations concerning a sample would hold true with respect to the wider population. That is, they help us to assess whether what we think is going on in our sample has happened due to chance or whether it is indicative of something more meaningful.

This workbook will introduce you to a common inferential statistic that is sometimes used by social researchers and one that you are also likely to find useful in a variety of different projects; the chi-square statistic.

- Correctly identify when to use the chi-square of goodness of fit test and when to use the chi-square test of association
- Check that your data satisfies the assumptions of the chi-square test
- Calculate the chi-square statistic
- Assess the significance of the result
- Calculate strength of the association using the appropriate test
- Interpret, report, and critically understand the implications of your results
- Present this material in the appropriate form

4.2 The chi-square statistic

The chi-square statistic is represented by $\chi 2$. For those of you have forgotten their maths (or Greek), this means 'the square of chi' pronounced 'ki'. The tests associated with this particular statistic are used when your variables are at the nominal and ordinal levels of measurement – that is, when your data is categorical. Briefly, chi-square tests provide a means of determining whether a set of observed frequencies deviate significantly from a set of expected frequencies.

Chi-square can be used at both univariate and bivariate levels. In its univariate form –the analysis of a single variable – it is associated with the 'goodness of fit' test. For instance, if you wanted to explore whether the frequency of weddings across months of the year were evenly distributed, a 'goodness of fit' test would be appropriate as would help you to assess whether the number of weddings you observed was a close fit with what you would expect, or whether those observations deviated significantly from those expectations.

When used for bivariate analysis – the analysis of two variables in conjunction with one another – it is called the chi-square test of association, or the chi-square test of independence, and sometimes the chi-square test of homogeneity. Generally speaking, this type of test is useful when you are dealing with cross tabulations or contingency tables. Whilst

the process of working out the statistic is essentially the same in both the independent and associative variations of the test, there are some subtle differences between the two that are worth noting. To take one of our earlier examples, if your research rationale was concerned with whether there were differences between different ethnic groups and the rate of pension provision, then we might want to call the test the 'chi-square test of independence' or the 'chi-square test of homogeneity' as we are concerned with whether the groups in our sample come from different populations or not. That is, the ethnic groups in our sample either are different to each other with respect to pension provision (independence) or they are actually the same (homogeneity).

However, suppose our rationale was broader and wanted to find out whether gender was generally associated with fear of crime. That is, whether levels in the measurement fear of crime are likely to vary according to the gender of the person in question. In this case we would be more likely to call the test a 'chi-square test of association' as the test can also tell us whether there is an association between two variables. Whether it is used to test for independence or association, however, the chi-square test can be used to help us evaluate differences between many different demographic (and related characteristics) with respect to a target measure. To name just a few, we can explore gender, ethnic group, social class, age, household occupancy, occupation, etc. etc. with reference to employment status, fear of assault, perceived level of health, etc. etc. For ease of presentation, we shall just refer to the test as 'the chi-square test of association' in the rest of this workbook as an association implies difference/similarity as well as relationship.

Indeed, regardless of what you actually call your test, as a general rule, if you can present your data in a contingency table that is 2x2 or bigger (2x3, 3x2, 3x3 etc.), then a chi-square test might be useful. Indeed, chi-square tests are very flexible and can be used with many different configurations of variable(s). In part, this is what has made them so popular for social research. It should also be noted, however, that there are always alternatives and, like most things in life, there are other always options of analysis which will have their own advantages and disadvantages. Unfortunately, these are beyond the scope of this workbook. If you can get your hands on a copy, Dean Champion's (1970) classic text 'Basic statistics for social scientists' is very readable and provides a good introduction to some of the alternatives.

All that said:

- If you are exploring the distribution of a single variable that is measured at the categorical level, a chi-square goodness of fit test might be appropriate.
- If you have two variables that you want to explore in relation to each other and they have been measured at a categorical level - that is, data at the nominal and ordinal levels - then a chi-square 'test of independence' or 'test of association' may be appropriate.

Fortunately, the process of determining the chi-square statistic is quite similar whether you are using it to determine the distribution of a single variable or testing for an association, independence, or homogeneity.

We'll begin by looking at the less complicated version of the test: the chi-square goodness of fit test.

4.3. The chi-square 'goodness of fit' test

Whenever we collect data, it will usually vary from what is expected. Role a dice 600 times and it is unlikely that you will role 100 1's, 100 2's, 100 3's, etc. like we would expect - but you are likely to role something around those scores. However, sometimes we need to know whether our results are just a bit of expected random variation around these values, or whether our observed scores are due to something other than chance (for instance, if the dice are loaded).

Here's a more sociological example. Suppose I want to find out whether more burglaries are committed in Sheffield on particular days of the week.

T Remembering that the chi-square statistic is typically concerned with associations, write an alternative and null hypothesis for this research question.

The null hypothesis should be something like:

- There is no significant association between days of the week and burglaries committed

And the alternative hypothesis:

- There is a significant association between days of the week and burglaries committed

Looking at the regional crime statistics I find out the following:

Table 4.1. The frequency of burglaries committed in Sheffield by days of the week for 2010

Days of the week	Burglaries committed
Sunday	44
Monday	54
Tuesday	61
Wednesday	65
Thursday	52
Friday	95
Saturday	98

How can I tell whether this distribution is due to some random variation around what might be expected according to a theoretical model? How do I know which hypothesis to accept and which hypothesis to reject?

A chi-square goodness of fit test - sometimes called the chi-square one-sample test - can help us to do this as it tells us whether there is a difference between what was actually observed in our results and what we might expect to observe by chance. So, the test is able to determine whether a single categorical variable fits a theoretical distribution or not. It enables us to make an assessment of whether the frequencies across the categories of the variable are likely to distributed according to random variation or something more meaningful.

4.3.1. Assumptions of the chi-square goodness of fit test

For the chi-square goodness of fit test to be useful, a number of assumptions first need to be met.

As an absolute requirement, your data must satisfy the following conditions:

- The variable must be either nominal or ordinal and the data represented as counts/frequencies.
- Each count is independent. That is, one person or observation should not contribute more than once to the table and the total count of your scores should not be more than your sample size: one person = one count.

If your data do not satisfy these conditions then it is not possible to use the test and it should not be used.

However, your data should also typically conform to the following:

- None of the expected frequencies in the cells should be less than 5.
- The sample size should be at least 20 – but more is better.

If the data in your sample does not satisfy these two criteria, the test becomes unreliable. That is, any inferences that you may make about your data have a significantly higher likelihood of error. In such instances of low sample size or very low expected frequencies, it has been repeatedly demonstrated by statisticians that the chi-square statistic becomes inflated and no longer provides a useful summary of the data. If your expected frequencies are less than 5, it is probably worth considering collapsing your data into bigger categories[1] or using a different test[2].

1 See the previous workbooks for instructions on how to do this.
2 In instances where expected frequencies are lower than 5, then the Kolmogorov-Smirnov one-sample test can be employed as an alternative if you are using ordinal data.

Many textbooks will also suggest that the sample needs to be random. However, you can still use the test even if your sample is not random, but you do need to examine your sample carefully to make sure that it is not susceptible to systematic error – that is, that your sample is not biasing the data in some way. Let's return to the previous example of reported crime in Sheffield by days of the week. Although our data in this example are taken from a fixed record (the 'official' number of reported crimes), think about all the little tiny variations in how that data was originally recorded and in how the burglaries were actually counted (or not). There are many ways in which the actual number of burglaries may vary a little bit. This is not usually a problem as long as that variation is random: that is, that these little variations are all evenly distributed throughout the week. However, in some instances these variations may not be random. For example, let's suppose that the sergeant who regularly worked on Sunday diverted crimes that should be recorded as 'burglary' into a different category – mugging perhaps - whereas the sergeant who worked Saturday did the exact opposite. If this happened, there would be systematic bias in the sample and our data would become invalid.

Therefore for our purposes this assumption is probably better expressed as:

- You need to make sure your sample is free from systematic error

4.3.2. Calculating the χ2 test statistic

The χ2 goodness of fit test statistic is derived from the following formula:

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

This might look a little imposing when presented in this manner, but it is really quite simple.

- χ2 = Chi-square
- $\sum$ = The sum of
- O = Observed Scores
- E = Expected Scores

The chi-square statistic is effectively a function of the sum of the difference between the observed and the expected scores and once we have worked it out, we can make an assessment concerning the likelihood of whether to accept or reject our alternative hypothesis.

Before we go through the formula, however, we need to specify what is meant by an 'expected score'. Indeed, the only tricky bit in the entire formula is in working out the expected scores.

Briefly, the expected scores are what we would expect each cell count to be 'in theory'. This 'theoretical distribution' is actually a measure of proportional distribution and it allows us to see whether our observed scores vary according to chance. To see what we would expect if the distribution was approaching 'random', all we need to do is add together the total number of counts in each cell and divide by the number of categories in the variable. We will see this 'in action' as we continue through our burglaries example.

4.3.2.1. Calculating the goodness of fit statistic: Burglaries and days of the week

Fortunately, once we have calculated our expected scores, the rest of the equation can be explained rather easily.

The Chi-square value is:

- The observed score minus the expected score for each cell $O - E$
- The answers of which are then squared $(O - E)^2$
- These numbers are then divided by the respective expected score $(O - E)^2/E$
- These totals are then summed to give the value of our chi-square statistic $\sum(O - E)^2/E$

Let us return to our days of the week by burglaries committed example.

T Check that the data does not violate the assumptions of a chi-square analysis.

Table 4.1. The frequency of burglaries committed in Sheffield by days of the week for 2010

Days of the week	Burglaries committed
Sunday	44
Monday	54
Tuesday	61
Wednesday	65
Thursday	52
Friday	95
Saturday	98

Everything seems to be in order (at least to the best of our knowledge). Remember that our hypotheses ran:

Null:

- There is no significant association between days of the week and burglaries committed

Alternative:

- There is a significant association between days of the week and burglaries committed

We are looking to accept one of these and reject the other – the chi-square goodness of fit test will allow us to do this with some degree of confidence.

So, we already have our observed scores, now we need to work out what our expected values are. To do this, we need to take a measure of the expected proportional distribution so we first need to calculate the total amount of burglaries.

T Calculate this value.

Adding together the values in our observed column we get a total of 469. There are seven days in the week, so if these burglaries were distributed evenly throughout the week, we would divide 469 by 7 – this calculation will provide us with our 'theoretical distribution'.

T Calculate this value.

The answer is 67. The 'thought process' associated with Table 4.1. now looks like this:

Table 4.1a. The frequency of burglaries committed in Sheffield by days of the week for 2010 with expected scores

Days of the week	Observed scores	Expected scores
Sunday	44	67
Monday	54	67
Tuesday	61	67
Wednesday	65	67
Thursday	52	67
Friday	95	67
Saturday	98	67
Total	**469**	**469**

Now we have a series of scores that we could expect to see if our results were due to chance (the expected values) and our 'real' distribution (the observed values). Our next job is to subtract the expected scores from the observed scores to see what the difference is.

T Calculate these values.

They should be:

Table 4.1b. The frequency of burglaries committed in Sheffield by days of the week for 2010 with difference between observed and expected scores

Days of the week	Observed scores	Expected scores	$O - E$
Sunday	44	67	-23
Monday	54	67	-13
Tuesday	61	67	-6
Wednesday	65	67	-2
Thursday	52	67	-15
Friday	95	67	28
Saturday	98	67	31
Total	**469**	**469**	**0**

Now we need to square the values. This needs to be done as squaring the values removes the negative scores. Note how our total column now adds up to 0? We need to change this as it fails to summarise the distribution appropriately. Squaring our values will allow us to work with positive numbers as it essentially gets rid of the minus signs from our values. Remember from your maths classes how two minuses make a positive when multiplying? If we multiply the Sunday value (-23) by itself we get 529: **-23×-23 = 529.**

To do this on an ordinary calculator, all we have to do is multiply the number by itself ignoring any minus signs.

T Calculate these values.

Table 4.1c. The frequency of burglaries committed in Sheffield by days of the week for 2010 with differences between observed and expected scores squared

Days of the week	Observed scores	Expected scores	$O-E$	$(O-E)^2$
Sunday	44	67	-23	529
Monday	54	67	-13	169
Tuesday	61	67	-6	36
Wednesday	65	67	-2	4
Thursday	52	67	-15	225
Friday	95	67	28	784
Saturday	98	67	31	961
Total	**469**	**469**	**0**	**2708**

Now we have some positive numbers to work with that reflect the differences in our table better. The next task is to divide each number by its expected value.

T Calculate these values to two decimal places.

Table 4.1d. The frequency of burglaries committed in Sheffield by days of the week for 2010 with differences between observed and expected scores squared and divided by expected score.

Days of the week	Observed scores	Expected scores	$O-E$	$(O-E)^2$	$(O-E)^2/E$
Sunday	44	67	-23	529	7.90
Monday	54	67	-13	169	2.52
Tuesday	61	67	-6	36	0.54
Wednesday	65	67	-2	4	0.06
Thursday	52	67	-15	225	3.36
Friday	95	67	28	784	11.70
Saturday	98	67	31	961	14.34
Total	**469**	**469**	**0**	**2708**	

Nearly there! Now we just need to add these values together – this gives us our chi-square goodness of fit statistic.

T Calculate the chi-square statistic.

Table 4.1e. The frequency of burglaries committed in Sheffield by days of the week for 2010 showing process of calculation for the chi-square statistic.

Days of the week	Observed scores	Expected scores	$O-E$	$(O-E)^2$	$(O-E)^2/E$
Sunday	44	67	-23	529	7.90
Monday	54	67	-13	169	2.52
Tuesday	61	67	-6	36	0.54
Wednesday	65	67	-2	4	0.06
Thursday	52	67	-15	225	3.36
Friday	95	67	28	784	11.70
Saturday	98	67	31	961	14.34
Total	**469**	**469**	**0**	**2708**	**40.42**

See – it wasn't so hard was it? But now what do we do? How do we find out whether our value is significant or not? Do our observed scores differ significantly from what we would have expected if the scores were distributed evenly? And which hypothesis can we reject?

Fortunately, all we have to do is compare our value against a chart of pre-determined critical values. These figures were originally calculated by a famous statistician called Karl Pearson - this is why the test is sometimes known as Pearson's chi-square test. If our number exceeds the pre-determined value, then our statistic is judged to be significant as it is deemed to be a long way from what we would expect to happen by chance. This is the main reason why we must make sure that our data satisfies the assumptions of the chi-square test as the critical values that Pearson worked out become invalid if the conditions are not met.

However, Pearson demonstrated that these pre-determined critical values largely depend upon the dimensions of our data. Put rather crudely, different sizes of table have different critical values. Therefore we need to determine the critical value for our particular table. This is achieved through something called the degrees of freedom (df). Fortunately, this is easy to work out too. For the goodness of fit test it is the total number of rows in our table minus one[3].

This is expressed as a formula:
d.f. = r - 1

T Using this formula, calculate the degrees of freedom for our original days of the week by burglaries table.

T Now compare our chi-square value against the relevant critical value (use p = 0.05). If our value is bigger than the one in the table then our result is significant – is it significant? Which hypothesis can we reject?

3 Depending upon how you have composed your table – either horizontally or vertically - this could just as easily be c -1 (columns -1).

Table 4.2. Critical values for the chi-square 'goodness of fit' test

df	p = 0.05	p = 0.01
1	3.84	6.84
2	5.99	9.21
3	7.82	11.35
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.73
12	21.03	26.22
13	22.36	27.69
14	23.69	29.14
15	25.00	30.58
16	26.30	32.00
17	27.59	33.41
18	28.87	34.81
19	30.14	36.19
20	31.41	37.57
21	32.67	38.93
22	33.92	40.29
23	35.17	41.64
24	36.42	42.98
25	37.65	44.31
26	38.89	45.64
27	40.11	46.96
28	41.34	48.28
29	42.56	49.59
30	43.77	50.89

Our degrees of freedom are 6 (7-1). Using the table of critical values we can see that the critical value when df=6 is 12.59. Therefore, at the .05 level of significance, our chi square value of 40.42 exceeds the critical value and our results are significant at the .05 level. This means that there is only a 5% probability that our results are due to chance and we can reject the null hypothesis with some confidence. As a result, we can conclude that based on this sample there is an association between days of the week and burglaries committed.

We would commonly report this along these lines:

> *The study found a significant association between days of the week and burglaries committed ($\chi 2 = 40.42$, d.f. = 6, $p<.05$).*

The numbers in the brackets are a way of reporting the 'vital statistics' of our test that we've just worked out. Basically it refers to:

- $\chi 2$ = our chi-square value = 40.42
- d.f. = degrees of freedom = c-1 = 7-1 = 6
- $p<.05$ = the probability of our results being due to chance (p) are less (<) than 5% (.05).

However, what this actually means probably needs explaining a little. According to Bryman (2008, p. 33), the level of statistical significance is a measure of the level of risk that your findings are due to random probability and not the meaningful or lasting association that we are suggesting. The value summarises the level of confidence you have in accepting the alternative hypothesis based on this particular sample. There's always a chance that our results might be due to something other than the association we are suggesting and there is always a chance we might be jumping to the wrong conclusion - the level of significance simply allows us to measure the chances of us doing that according to chance. In this case, the risk is sufficiently low for us to say with some confidence that the data in our sample is likely to represent a meaningful association in the wider population. Hence we can make the suggestion that the results that we have found are likely to be generalizable.

So, if our results are significant at the .05 level, this suggests there is a 5% probability that our results are significant by chance alone. Let us suppose that we record the amount of burglaries by days of the week for 100 weeks. If there is a 5% probability that our results are due to chance alone, we would expect 5 (false) positives across those 100 weeks – even if our results are actually insignificant. Therefore the chances that burglaries and days of the week are not related and we've accidently hit upon one of those freak events are actually pretty low.

So low in fact, that statistical wisdom suggests that if the level of significance is below 5%, we can reject the null hypothesis and accept the alternative hypothesis. So we can conclude that our results are significant. In fact, our results go a bit further than this. Look at the significance table again. Remember that our degrees of freedom are 6 (7-1). The critical value at the .01 level of significance is 16.81. Our chi square value of 40.42 also exceeds this critical value. As a result, our results are significant at the .01 level ($\chi 2 = 40.42$, d.f. = 6, $p<.01$). This means the probability of our results being due to chance is actually less than 1%.

WARNING: This does not mean that our results are any better, stronger, or more reliable than they were at the 5%. It just means that the probability of our results being due to chance is lower at the 1% level than they are at the 5% level. The chi-square statistic tells us nothing about how important our results are, how strong they might be, or how reliable they are. The statistic simply allows us to make an assessment as to the probability of something happening or not. In turn, this allows us to make an assumption with some degree of confidence concerning whether to accept the null or alternative hypothesis based on the data in our sample.

Statistical convention is to report levels of significance at the 5% ($p<.05$), the 1% ($p<.01$), or the non-significant level ($p<.n.s$). Although some print-outs will often report significance at smaller levels, this is likely to over-estimate the importance of our results and can often lead to a spurious level of accuracy that is not helpful and it should not be reproduced. We want to avoid the kind of thinking that goes along the lines of: 'WOW, our results are significant at the .0000001 level, this means our results are really, really, important'. They may be statistically significant, but if you've made a mistake somewhere along the line, your results will be invalid anyway. Similarly, the chi-square statistic doesn't tell us anything about the strength of a result, so reporting probability at lower levels is a bit dubious as it might be seen to imply stronger results. Therefore, we should err on the side of caution and stick to statistical convention.

4.3.3. Interpreting the chi-square test

So we've found a significant result – once we've stopped celebrating, now what? The problem with the chi-square test is that it is a summary of the difference between observed and expected scores. It does tell us whether there is a difference between what has been observed and what we would have expected, but it doesn't tell us where those differences might be. There is still more work to do after we've found a significant result – we have to 'eyeball' the table. This means that we have to look at our table, and our calculations, in order to describe where those major differences might be. Data never speaks for itself, it needs to be interpreted.

So, we now need to work out where the major discrepancies are in our table and attempt to explain what they might indicate. A good method of doing this is to return to the differences between our observed scores and our expected scores.

Table 4.1b. The frequency of burglaries committed in Sheffield by days of the week for 2010 with difference between observed and expected scores

Days of the week	Observed scores	Expected scores	O – E
Sunday	44	67	-23
Monday	54	67	-13
Tuesday	61	67	-6
Wednesday	65	67	-2
Thursday	52	67	-15
Friday	95	67	28
Saturday	98	67	31
Total	**469**	**469**	**0**

By looking at the O – E column, it's fairly easy to see how we might go about this. All those scores that have minus values are lower than expected, and all those positive scores are higher than expected. As all of these scores come from a fixed point of reference – 67 – it's easy to spot where the differences actually are and what parts of the table contribute most to the test statistic.

T Describe the distribution.

The biggest discrepancies are on Sunday, Friday and Saturday. On Sunday, the observed scores are at their lowest and much lower than what we should expect, but on Friday and Saturday they are much higher. Similarly, the burglaries committed on Monday are also lower, but not as low as those observed on Sunday. There is little difference from what we would expect in the rates for Tuesday and Wednesday, but on Thursday burglaries are again as low as they are on Monday. However, on Friday and Saturday, the observed values are much higher with Saturday being the most common for burglaries being committed.

We can summarise the table thus:

> *The study found a significant association between days of the week and burglaries committed (χ^2= 40.42, d.f. = 6, p<.01). Examination of the frequency distribution reveals that the most common day for burglaries in this study is Saturday, closely followed by Friday. The lowest amount of burglaries committed is on Sunday. Similarly, burglaries committed on both Monday and Thursday are also somewhat lower than should be expected. However burglaries on Tuesday and Wednesday are very similar to the expected scores and unlikely to contribute much to the chi-square statistic.*

Q Can you account for this distribution? Try to briefly explain why the data might be distributed in this way.

There are many possible answers, but given that burglars tend to target places that are unoccupied, perhaps the most probable is that the distribution reflects trends in the daily occupation of places. Friday and Saturday are likely to be the days where people are most likely to leave places unoccupied for extended periods of time giving burglars the opportunity to go to their work uninterrupted.

In terms of housing, this would probably help to account for the lower rate on Sunday too – and perhaps even the lull on a Thursday. Furthermore, burglars may not like to work on Sunday because that is the day when houses etc are occupied for the longest. Perhaps the noise that is associated with Friday and Saturday, particularly at night, acts as good 'cover', with the sedate nature of a Sunday doing the opposite. Of course, this would assume that most of these burglaries occurred during the night, which certainly is not the case with residential burglary; hence a more fine grained analysis of burglary by time and day would be helpful to examine these trends further.

Perhaps the rise on Friday and Saturday can be partly accounted for by commercial burglary with people leaving places of works such as schools and factories unoccupied for the weekend. Again, the general lack of noise and activity on a Sunday might put burglars off attempting to burgle these premises. Indeed, our data may not actually be sensitive enough here because we have lumped all burglaries into one category. There may be different patterns between commercial burglary and residential burglary (which may in turn be divisible into house burglary and burglary associated with general unoccupied property such as sheds and garages). Temporal patterns of residential, non-residential, and commercial burglary might be better examined separately.

Other factors may include: perhaps some burglars have 'day jobs'; perhaps some burglaries are even 'spur of the moment' events conducted under the influence of alcohol on the way home from a Friday or Saturday night out.

As you can probably see by now, finding a significant chi-square result is only the first step in analysing your data. Indeed, it's easy to see how just one 'answer' has provided many more questions and many more routes for analysis. Although research reports and papers often present the quantitative analysis of data as a very systematic and linear affair, it is often much more complex. One research question often leads to more research questions, which in turn lead to more and more.

4.4. The chi-square 'test of association': Gender and perceived safety

Now you should be able to conduct the chi-square goodness of fit test, compute the chi-square statistic, and most crucially of all, interpret the results. The good news here is that if you can understand all that, then you can also understand the chi-square test of association. We use this test when we want to see whether two categorical variables are associated with each other.

Besides the number of variables, the only real difference between the two tests is in the computation of the expected scores and degrees of freedom. But it is still fairly straight-forward to work out. To demonstrate how to conduct the chi-square test of association, we are going to explore the association between gender and perceived safety when walking alone in the dark using data from the British Crime Survey of 2000 (n=19319). This particular survey was conducted with a cross-section of the public in England and Wales between January 1999 and February 2000.

T Write a research rationale, formulate a research question, and identify a suitable null and alternative hypothesis.

It might look something like this:

> *Despite official statistics demonstrating that young men are the most likely to be the victims of crime, studies have consistently demonstrated that women are more likely to be fearful of crime. However, few studies have explored how that fear is expressed in specific contexts and whether these trends in the fear of crime are reproduced in relation to how safe people feel in their particular environments. Therefore, this project aims to further develop the literature on the gendered nature of the fear of crime by examining whether the perceived safety of an individual walking alone at night is associated with gender.*

Null hypothesis:

- There is no significant association between men and women in their perception of safety when walking home at night

Alternative hypothesis:

- There is a significant association between gender and the perception of safety when walking home alone at night

T Now see if you can construct a variable that measures 'gender' and one that measures the 'perceived safety walking alone after dark'.

In this particular instance, the gender variable was measured with a standard gender variable:

Q4.1. Please state your gender: *e*

☐ Male ☐ Female

The perceived safety walking alone after dark variable was measured using the following question:

Q4.2. How safe do you feel walking home alone in this area after dark? Would you say you feel? *e*

☐ Very safe ☐ Fairly Safe
☐ A Bit Unsafe ☐ Very Unsafe

Q Can you critique the perceived safety walking alone after dark variable?

Whilst the gender variable is fairly conventional, the 'perceived safety walking home after dark' variable can certainly be questioned in terms of its reliability and validity. To begin with, we are quantifying safety. This is a complex emotional experience - to what extent can we capture that experience in a four-point scale? Imagine if you were asked the question without the four point scale - it's likely that an unstructured answer would be a lot longer and much more complex than the options you have been given. Secondly, interpretations of safety might not be consistent across individuals - people may experience safety differently; what is 'fairly safe' to one person, might be 'very unsafe' to another. We can't be sure that the measure is consistently interpreted and understood across our sample. Thirdly, what people report may be different to their actual experience. That is not to say that people are particularly likely to lie, but we shouldn't assume that the self-reported measure directly matches their experience. Many people will not actually walk alone after dark but are likely to answer just because they have been asked. Other people may have forgotten their reaction(s) and are just summarising as best they can. Fourthly, experiences also vary across time. It's perfectly possible to experience the same walk differently from week to week. Think about what might happen if a story appears in the local news about a seemingly

unprovoked violent attack in the neighbourhood - can this variable account for this context? There are also questions about the geographical element of the question. It is fairly imprecise - what does 'this area' refer to? The street? The community? The neighbourhood? The city? Finally, can this variable act as a proxy for fear of crime more generally? That is, can we generalise any association between 'perceived safety walking home alone after dark' and 'gender' to the 'fear of crime' and 'gender'? We'd certainly have to be careful as the question does not actually make any reference to crime itself, yet alone particular crimes. It may be more indicative of feelings of safety, but given the specific nature of 'walking alone after dark' it might not generalise that well. For instance, it doesn't measure feelings of safety in your own home - no doubt a crucial factor in a more general feeling of safety.

We could go on[4], but we'll stop here. Now we do have to ask the question - so what now? Do we plough on regardless or give up because the limitations of the variables are too great? Neither - we make sure that we recognise their limitations in our methods and discussion sections of our report, and we pay careful attention to how we phrase any findings and conclusions to make sure that we don't go beyond what our data is actually suggesting. Given the flaws in the variable, it could also be worth considering whether we might amend our research design to incorporate a mixed methods approach. Some in-depth qualitative interviews could provide some of the valuable context that we are losing in the survey.

In 1999, the British Crime Survey (published in 2000) asked 19319 people these questions. Having considered the limitations of the variables we can now look at the frequency distribution in the form of a contingency table. The questions were answered thus:

Table 4.3. Gender by perceived fear of walking alone after dark (BCS, 2000)

	Very safe	Fairly safe	A bit unsafe	Very unsafe
Male	3475	3930	1031	365
Female	1262	4242	2999	2015

The assumptions for the chi-square test of association are the same as they are for the chi-square goodness of fit test:

- The variable must be either nominal or ordinal and the data represented as counts/frequencies[5]
- Each count is independent. That is, one person or observation should not contribute more than once to the table and the total count of your scores should not be more than your sample size: one person = one count.
- None of the expected frequencies in the cells should be less than 5.
- The sample size should be at least 20 – but more is better.

Q Does the data in the table satisfy the assumptions of the chi-square test?

Everything seems to be in order.
Remember our chi-square formula? It's pretty much exactly the same for a test of association as it is for a goodness of fit test.

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

The processes we go through to calculate the statistic are the same too. So first we have to work out the expected counts. To do this in a contingency table we need to multiply the row total by the column total and divide by the total number of scores in the total. As a formula it looks like this:

$$\frac{Row\ Total \times Column\ Total}{Overall\ Total}$$

4 If you want to read more about the problematic nature of measuring 'fear of crime' a good place to start is Ferraro and LaGrange's (1987) paper 'The measurement of the fear of crime' available in the journal 'Sociological Inquiry' (issue 57, pp 70-101).

5 It is probably worth noting that if you are exploring the relationship between two variables at the ordinal level of measurement , there are other options that might be 'better' suited to your data. However, these are beyond the scope of this workbook. Fielding and Gilbert's (2009) 'Understanding Social Statistics', published by Sage, is a good place to learn about these techniques.

T Calculate the row, column and overall totals.

Table 4.3a. Frequencies of gender by perceived fear of walking alone after dark with column and row totals (BCS, 2000)

	Very safe	Fairly safe	A bit unsafe	Very unsafe	
Male	3475	3930	1031	365	8801
Female	1262	4242	2999	2015	10518
Total	4737	8172	4030	2380	19319

For the first cell (Male/Very Safe) the row total is 8801, the column total is 4737 and the overall total is 19319. So using the calculation to find the expected count for this cell we find that the answer for this cell is 2158.

$$\frac{Row\ Total \times Column\ Total}{Overall\ Total} \qquad \frac{8801 \times 4737}{19319} = 2158$$

T Now calculate the expected frequency for each cell.

Table 4.3c. Frequencies of gender by perceived fear of walking alone after dark with expected cell counts (BCS, 2000)

	Very safe	Fairly safe	A bit unsafe	Very unsafe	
Male	3475	3930	1031	365	8801
(expected)	2158	3723	1836	1084	
Female	1262	4242	2999	2015	10518
(expected)	2579	4449	2194	1296	
Total	4737	8172	4030	2380	19319

Now we are all ready to work out the chi-square value.

First we have to calculate the $O - E$ part of the formula, and we have to do this for each cell. So for the first cell (Male/Very Safe) the observed count was 3475 and expected count 2158. Therefore we need to calculate 3475 – 2158, which equals 1317.

T Calculate $O - E$ for each cell.

Table 4.3d. Gender by perceived fear of walking alone after dark showing differences between observed and expected scores (BCS, 2000)

	Very safe	Fairly safe	A bit unsafe	Very unsafe
Male	1317	207	-805	-719
Female	-1317	-207	805	719

Not too difficult so far.

Now we need to calculate $\frac{(O-E)^2}{E}$ for each cell (to 2 decimal places).

For the first cell (Male/Very Safe) this would mean:

$$\frac{(3475-2158)^2}{2158} = 803.75$$

To break this down a little more, first we need to square 1317. Remember, this enables us to get rid of any minus signs in order that the expected frequencies don't just cancel themselves out. So 1317 multiplied by 1317 is 1734489. We now need to divide this by 2158, which is the expected count. This gives us 803.75. We need to conduct this same procedure for each cell in the table.

T Calculate $(O-E)^2/E$ for each cell (to 2 decimal places).

Your table should now look something like this:

Table 4.3e. Gender by perceived fear of walking alone after dark showing differences between observed and expected scores dived by expected scores (BCS, 2000)

	Very safe	Fairly safe	A bit unsafe	Very unsafe
Male	803.75	11.51	352.95	476.90
Female	672.54	9.63	295.36	398.89

Now all that is left to do to calculate the chi-square statistic is to add all of these values together.

The chi-square statistic is:
χ2 =3021.53

Just like the goodness of fit test, the final step is to check this number against the critical value to see if our statistic is significant. Again like the goodness of fit test, we also need to calculate the degrees of freedom. This is done slightly differently for the chi-square test of association – but it is still fairly straight-forward. The formula looks like this:

d.f. = (rows -1)×(columns -1)

The degrees of freedom in this example
d.f. = (2-1)×(4-1) = 3

T Now check whether our value exceeds the critical value.

Table 4.5. Critical values for the chi-square test of association

df	p = 0.05	p = 0.01
1	3.84	6.84
2	5.99	9.21
3	7.82	11.35
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.73
12	21.03	26.22
13	22.36	27.69
14	23.69	29.14
15	25.00	30.58
16	26.30	32.00
17	27.59	33.41
18	28.87	34.81
19	30.14	36.19
20	31.41	37.57
21	32.67	38.93
22	33.92	40.29
23	35.17	41.64
24	36.42	42.98
25	37.65	44.31
26	38.89	45.64
27	40.11	46.96
28	41.34	48.28
29	4256	49.59
30	43.77	50.89

Our chi-square value exceeds the critical value and as a result we can accept the alternative hypothesis and reject the null hypothesis.

T Report this result

> *The study reports a significant association between gender and perceived safety walking alone after dark (χ2 = 3021.53, d.f. = 3, p<.01).*

We don't necessarily have to report that we have rejected the null hypothesis and accepted the alternative hypothesis as it is implied in the statistic. The 'vital statistics' that we have reported contain all the necessary information that we need to deduce this so there is no need to repeat ourselves when reporting the result.

4.5. Assessing the strength of an association: Phi and Cramer's V test

The chi-square statistic looks pretty impressive doesn't it? At least it looks like a high number. However, remember that the chi-square test tells us nothing about the strength of the association between our variables – it just tells us that there is one. Unfortunately, no matter how high the chi-square value, the test cannot tell us anything about how strong the association is. Luckily enough, there are some post-hoc tests we can conduct to determine this and they are really easy to work out.

4.5.1 The Phi Coefficient

If we have a 2x2 table, then we can use something called phi, which is usually represented by the symbol φ. Where we already know χ2, the formula looks like this:

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

If we haven't already calculated χ2, then the formula is a bit more complex:

$$\varphi = \frac{(ad-bc)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

However, both methods should give the same answer so it's useful to know how to do both so you can check your answer.

Let us suppose we have conducted a project that aimed to explore student experiences of crime. As

part of this project we have been investigating whether being student increases the risk of being a victim of crime. Using the British Crime Survey 2007-2008 dataset, we have found the following:

Table 4.5. Student status by experience of crime in the last 12 months (BCS, 2009)

		Experience of crime in the last 12 months		
		Not a victim of crime	Victim of crime	Total
Student status	**Yes**	182	114	**296**
	No	5147	1622	**6769**
Total		**5329**	**1736**	**7065**

T Conduct a chi-square test of association on the table. Report the result.

The study reports significant association between student status and experience of crime in the last 12 months (χ2 = 32.4, d.f. = 1, p<.01).

To calculate φ where we already know χ2, all we need to do is divide the chi-square statistic (χ2) by the number of counts in the sample (n); and then calculate the square root of the answer.

So in this case:

$$\varphi = \sqrt{\frac{32.4}{7065}}$$

Therefore, the phi coefficient is 0.068.

The useful thing about this statistic is that when it is used in 2x2 tables, it should always give a value between 0 and 1. The closer to 1, the stronger the association.

As a general rule of thumb anything between 0 and .3 is weak to moderately weak; anything between .3 and .6 is moderate to moderately strong; and anything above that is strong to very strong. Unfortunately, this is a very general rule of thumb, and in some cases, it can be misleading. We'll find out why in a short while, but, for the time being we'll stick with the methods and it looks like our phi coefficient is relatively weak in strength. So, if we were to report this result, we would state something along the lines of the following:

The study reports significant association between student status and experience of crime in the last 12 months (χ2 = 32.4, d.f. = 1, p<.01). Inspection of the phi coefficient, however, suggests that the strength of this association is moderately weak (n = 7065, φ = .068).

However, in order to check that we have made the right calculation, we'll see if we get the same answer using the longer method.

To do this, we need to return to our table and label the cells so we can apply the formula. Working from left to right, we get the following:

Table 4.5a. Student status by experience of crime in the last 12 months with cell labels (BCS, 2009)

		Experience of crime in the last 12 months		
		Not a victim of crime	Victim of crime	Total
Student status	**Yes**	*a*	*b*	**296**
		182	114	
	No	*c*	*d*	**6769**
		5147	1622	
Total		**5329**	**1736**	**7065**

Now we just need to insert the corresponding numbers into the formula, so:

$$\varphi = \frac{(182 \times 1622) - (114 \times 5147)}{\sqrt{(182 + 114)(5147 + 1622)(182 + 5147)(114 + 1622)}}$$

As is almost always the case with equations, we need to work out the values contained within the brackets first, so it's time to get the calculator our again:

$$\varphi = \frac{295204 - 586758}{\sqrt{(296 \times 6769 \times 5329 \times 1736)}}$$

And again:

$$\varphi = \frac{-291554}{4305323.93}$$

Therefore the phi coefficient is -0.068.

Well, the numbers are at least the same, but this time the value is negative: what does that mean? Unlike the chi-square version of the test where the minus signs are purposefully removed, the longer version of phi is calculated as a product of the diagonal cells (*axd*) and the off-diagonal cells (*bxc*), hence scores of -1 to +1 are possible: the sum of b*c* can be larger than the sum of *ad*. Indeed, the longer version of the test gives us a little more information about the nature of the association than the shorter version does. As a rule of thumb, variables within 2x2 tables are considered positively associated if *a* and *d* are larger than *b* and *c*. In contrast, the variables are considered to be negatively associated if *b* and *c* are larger than *a* and *d*.

Actually, it is possible to tell this from just looking at the data table and working out the difference between the observed and expected values $(O - E)^2$.

Table 4.5b. Student status by experience of crime in the last 12 months – observed and expected values with cell labels (BCS, 2009)

		Experience of crime in the last 12 months		
		Not a victim of crime	Victim of crime	Total
Student status	**Yes**	*a*	*b*	**296**
		-43.3	41.3	
	No	*c*	*d*	**6769**
		41.3	-43.3	
Total		**5329**	**1736**	**7065**

If you look at our data table, all that is being implied by the negative value is that *b* and *c* are larger than we would expect them to be if the results were due to random variation. In this case, the negative association means that students are more likely to have been a victim of crime in the last 12 months than the general populace who are not students.

Evidently, whether the phi coefficient is positive or negative is entirely dependent on how you construct your table and in a 2x2 table, and you can always do this in one of four different ways. I tend to find it easier to read tables in a left to right fashion, so I always try to position the variable I suspect is the one driving the effect on the horizontal axis, and the variable that I think is associated with its impact on the vertical axis. Similarly, I like to place zero values on the left on the horizontal axis, and at the bottom on the vertical axis. In this case, the zero values are 'not' and 'no' respectively. If we do this, the construction of a table often reveals the direction of our alternative hypothesis. So, in the above example, we are hypothesising that student status is having some impact on the measurement of experience of crime. However, it should be noted that statistical convention does not always agree in this regard and you will often see tables presented with the driving variable on the vertical axis, and the measurement of the horizontal. Effectively, the position of the cells is quite arbitrary. This means you need to take care when interpreting your own tables, and when you are reading tables in research papers.

It is also worth noting that the construction of our table should also influence the way we present our results and vice versa. Indeed, there are at least two different slants we could place on our results here:

- People who are not students are less likely to have been a victim of crime in the last 12 months than those who are students.
- People who are students are more likely to have been a victim of crime than those who are not students.

Q Which interpretation better represents our hypothesis?

As implied by the construction of our table, our focus here is on whether someone is a student, and whether that particular occupation status is associated with experience of crime. Therefore, the second is a better interpretation given the (implied) purposes of our study.

The moral of the story is that a positive or negative phi coefficient is not, in itself, terribly meaningful if we don't interpret the results properly: statistics never speak for themselves. So, it might be better to report the results with a description of the vital statistics without the arbitrary negative phi value and instead provide an accompanying commentary with an indication of the strength of the association.

> *The study reports a significant association between student status and experience of crime in the last 12 months ($\chi 2 = 32.4$, d.f. = 1, $p<.01$). Based on this sample, it appears that students who are in full time education are more likely to have been a victim of crime in the last 12 months than those who are not students. Inspection of the phi coefficient, however, suggests that the strength of this association is weak ($n= 7065, \varphi= 0.068$).*

4.5.1.2. Calculating the significance of the phi coefficient

Of course, none of this helps with the fact that our phi coefficient appears to be quite low. However, in itself, the value of the phi coefficient does not tell us anything about whether the strength of this association is significant or not. It is fortunate, therefore, that we can also work out the significance of phi with relative ease, and given the dimensions of our table (2×2), we do so using the critical values of $\chi 2$ with 1 degree of freedom.

To test for the significance of phi, the formula is as follows:

$$\chi 2 = n\varphi^2$$

So, if we are to insert our values into the equation, we get the following:

$$\chi^2 = 7065 \times (0.068)^2$$

Which resolves to be:
32.67$= n\varphi^2$

Q Is this a significant result?

If we check the critical values in Table 4.5., we can see that 32.67 exceeds the critical value at the p<.01 level. So, although it may be weak, it is significant.

Given this finding, we might want to amend how we have reported the results.

> *Based on BCS 2008-2009 data, the study reports a significant association between student status and experience of crime in the last 12 months ($\chi 2 = 32.4$, d.f. = 1, $p<.01$). It appears that students who are in full time education are more likely to have been a victim of crime in the last 12 months than those who are not students. Whilst inspection of the phi coefficient suggests that the strength of this association is weak, given the very large sample size, it is one that is significant ($\varphi= 0.068, n= 7065, p<0.01$).*

It's quite easy to see why, even with an apparently fairly weak association, we find the phi coefficient to be significant as the higher the number of respondents, the more important small differences become. In this case, we have a very large number of respondents (n = 7065), and actually, given the lengths that the ONS go to in order to collect this sample, it is one that is estimated to follow the population of the country very closely. Hence, the relatively small phi co-efficient gains significance because so many people have been included in the sample and the significance of phi is, therefore, sensitive to quite small effects. Indeed, this is why it is important to the report the sample size when reporting the phi co-efficient as it impacts on the way that we interpret the phi value. Although there is disagreement when working with samples that are random, anything under 30 is a small sample, anything over 100 is moving toward large, and anything over 1000 is very large.

4.5.2. Cramer's V

Unfortunately, if we use phi in tables larger than 2x2, phi can return figures of more than 1. This makes the test much more difficult to interpret in larger tables and so phi should not be used with tables other than 2×2.

Luckily, a Swedish mathematician and statistician called Harald Cramer invented another post-hoc test that can help us. Even better still, it's pretty simple to work out too.

Like phi, Cramer's V is a way of calculating the strength of the association, but it can be used in tables which have more than 2×2 rows and columns. It is used as post-test to determine strengths of association after chi-square has determined significance. Again, Cramer's V varies between 0 and 1: where V is close to 0, it shows little association between variables; and where it is close to 1, it indicates a strong association. So, we can follow the same general rule of thumb as before: anything between 0 and .3 is weak to moderately weak; anything between .3 and .6 is moderate to moderately strong, and anything above that is strong to very strong.

V is calculated by first calculating chi-square, then using the following calculation:

$$\sqrt{\frac{\chi^2}{(n) \times (min.of\ r-1\ or\ c-1)}}$$

This looks a little more complicated, but it isn't really. The only tricky part is the "(n) x (min. of r-1 or c-1)". This basically means that we need to work out which is lower: the number of rows -1, or the number of columns -1. Once we do that everything slips neatly in place and we just need to multiply that figure by our sample size (n).

So, if we return to our gender by fear of walking home alone in the dark example, we can calculate the level of the association between the two variables. Remember that our table and our findings looked like this:

Table 4.3a. Frequencies of gender by perceived fear of walking alone after dark with column and row totals (BCS, 2000)

	Very safe	Fairly safe	A bit unsafe	Very unsafe	
Male	3475	3930	1031	365	8801
Female	1262	4242	2999	2015	10518
Total	**4737**	**8172**	**4030**	**2380**	**19319**

> *The study reports a significant association between gender and perceived safety walking alone after dark (χ2= 3021.53, d.f. = 3, p<.01).*

T Calculate V for our table.

$$V = \sqrt{\frac{3021.53}{19319 \times 1}}$$

V=.3954

Remember our findings? We can now add to them to produce a more rounded result.

T Try to report the result.

It should look something like this:

> *The study reports a significant association between gender and perceived safety walking alone after dark (χ2 = 3021.53, d.f. = 3, p<.01). Using Cramer's V, this was found to be an association that was moderate in strength (n= 19319, V=.395).*

4.6. Interpreting tables

So we have a significant result – and we've discovered that the association is moderate in strength.

So what now?

Well, in themselves, those figures do not tell us a huge amount about what is actually going on in the table and where. Indeed, chi-square, the phi coefficient, and Cramer's V are all statistical attempts to summarise a range of data according to a specific set of parameters. As useful as these summary statistics are, in attempting to reduce data down to just a few values, however, we inevitably lose some clarity about the full range of our data. Indeed, statistics do not speak for themselves and we cannot rely on those statistics to tell the whole story of our data. They do help us tell part of the story but they are not, in themselves, sufficient to tell it all.

So, in order to interpret our results we need to return to the contingency table to provide the important context that our results are missing.

Here's table 3c from our gender by perceived fear of walking home alone after dark example, and the findings we generated from it.

Table 4.3c. Frequencies of gender by perceived fear of walking alone after dark with column and row totals (BC, 2000)

	Very safe	Fairly safe	A bit unsafe	Very unsafe	
Male	3475	3930	1031	365	8801
Female	1262	4242	2999	2015	10518
Total	**4737**	**8172**	**4030**	**2380**	**19319**

> *The study reports a significant association between gender and perceived safety walking alone after dark (χ2 = 3021.53, d.f. = 3, p<.01). Using Cramer's V, this was found to be an association that was moderate in strength (n= 19319, V=.395).*

T Interpret the result.

The most common answer for both male and females is 'Fairly Safe' and there is little difference between the observed and the expected scores in this category. Most people feel fairly safe when walking alone at night. However, the least common answer for males is 'very unsafe' whereas it is 'very safe' for females. Indeed, the vast majority of males feel 'very safe' or 'fairly safe' when walking alone at night and the frequencies are larger than would be expected for these cells. This is especially the case for the 'very safe' category where the difference between observed and expected frequencies is 1317 – the largest difference in the table. Equally, men score much less than we would expect in the 'bit unsafe' and 'very unsafe' categories. Again, this is likely to account for a good proportion of the chi-square statistic. Men tend to feel safer than we would expect when walking alone at night.

Women, on the other hand, are much less likely to report feeling 'very safe' than would be expected. Similarly, they are much more likely to report feeling very unsafe than would be expected. These differences are similarly likely to account for much of the chi-square statistic. Women tend to feel less safe than we would expect when walking home alone at night.

In summary, most people feel fairly safe when walking alone at night. However, whereas men tend to feel safer than we would expect, women tend to feel less safe than we would expect when walking home alone at night.

4.6.1. Using residuals to help you interpret tables

Simply 'eye-balling' the crosstab is one way to interpret what's going on in the table. The problem of 'eye-balling', however, is that it can often be quite difficult to make sense of all that information, particularly where our crosstab is large.

Fortunately, there is a sneaky trick that will help you understand your data in a slightly more systematic way. Ladies and gentlemen, let me introduce you to residuals.

Trust me, it's less complicated than it sounds. Having a working knowledge of residuals is very, very useful and the best thing is that you've been using them already.

Residual basically means something that is left over. Indeed, for our purposes, a residual is basically a number that expresses the difference between our observed score and the expected score.

Essentially, it's this part of the chi-square calculation:

$$(\boldsymbol{O} - \boldsymbol{E})^2$$

The residual is effectively the value that is 'left over' after we have subtracted the expected score from the observed score.

If you think about it, the chi-square statistic is actually based on the residuals from all of the cells in the table, so it makes a good deal of sense to use them to help us to tell us where the 'big' differences are in

the table. Indeed, an examination of the unstandardised residuals can sometimes help us to do this.

Table 4.3d. Gender by perceived fear of walking alone after dark showing unstandardized residuals (BCS, 2000)

	Very safe	Fairly safe	A bit unsafe	Very unsafe
Male	1317	207	-805	-719
Female	-1317	-207	805	719

When we present the data in this format, it is easy to see where the big differences between what we have observed, and what we would expect occur in the table. Unfortunately, there can be a problem with this method if we're not careful.

Quite clearly, the biggest residuals in our table are those in the 'very safe' column, and the smallest differences in the 'fairly safe' column. As we stated when we simply eye-balled the data, men are more likely than we would expect to feel 'very safe' or 'fairly safe' than women are - and it looks like there is a bigger effect for 'very safe'.

The problem, however, is that when we use the unstandardised residuals in this raw fashion, we're not necessarily comparing 'like with like'. This is easily demonstrated if we return to our original crosstab:

Table 4.3a. Frequencies of gender by perceived fear of walking alone after dark with column and row totals (BC, 2000)

	Very safe	Fairly safe	A bit unsafe	Very unsafe	
Male	3475	3930	1031	365	8801
Female	1262	4242	2999	2015	10518
Total	**4737**	**8172**	**4030**	**2380**	**19319**

If you look at the column totals, it's easy to see that there are a lot more 'fairly safes' than there are 'very safes'. Therefore differences are proportionally bigger in the 'very safe' category than they are in the 'fairly safe' category because there are far fewer counts in the 'very safe' category.

e For example, a change of 1000 would be proportionally much bigger for the 'very safes' than the 'fairly safes'. Indeed, expressed as a percentage, a change of 1000 would be a shift of 20% in the 'very safe' category, and just 12% in the 'fairly safe' category.

If we return to our table of unstandarised residuals again, we might actually be in danger of under-estimating the effect of the 'very safe' category because, as a proportion of the total, a difference of 1317 is very, very large (1317/4737 = 27.8%) in comparison to the difference of 207 (207/8172 = 2.5%) in the 'fairly safe' category.

So, we need to find a way of comparing 'like with like' - essentially, we need to find a way of standardising our residuals.

Thankfully, this is fairly straight-forward to do. The **standardized residual** for each cell is found by dividing the difference of the observed and expected values by the square root of the expected value. Expressed as an equation, this looks like:

$$\frac{(O-E)^2}{\sqrt{E}}$$

Remember, the expected score is worked out by looking at both the column total and the row total, so by further including them in our calculations of residuals, we are now being much more sensitive to the relative proportions of the totals in our table.
So, for the 'male' x 'very safe' cell, the calculation is thus:

$$\frac{(3475-2158)^2}{\sqrt{2158}}$$

The answer is 28.4.

T Calculate the standardised residuals for the rest of the table.

The table of standardised residuals should now look something like this:

Table 4.3e. Gender by perceived fear of walking alone after dark showing standardized residuals (BCS, 2000)

	Very safe	Fairly safe	A bit unsafe	Very unsafe
Male	28.4	3.4	-18.8	-21.8
Female	-25.9	-3.1	17.2	20.0

Now we have a better idea of the relative effect of each of the residuals on our chi-square statistic – and it's much easier to see 'what's going on' in the table.

As a general rule, anything over 2.0 can be taken to be a major contributor to the chi-square statistic[6]. Although some contribute much, much more than others, it's fairly easy to see that all of the cells in our table are major contributors. So if we return to our original analysis of the table, we can check whether our interpretation is the correct one.

T Check whether our original interpretation is correct

Yay! Looks like we've nailed it – and we can support our conclusions statistically if we ever need to.

6 This is not uncommon when using large datasets because the sample size is usually very large (see conclusion for some further discussion of this).

However, you don't very often see standardised residuals presented in tables that appear in journal articles etc. This is mainly because they would quickly complicate the presentation of our data, and if the reader wasn't familiar with the idea of a standardised residual, they wouldn't have the first clue of what they were looking at - and, more than likely, they would quickly give up on reading any more of our paper. So, whilst they are a valuable tool for interpreting tables, as a general rule don't present them unless you're specifically asked to. As an extra bonus however, most statistical analysis software packages will do all the calculations so you don't have to!

4.7. Rounding up….

Just like any other research technique, the popularity of the chi-square statistic has risen and fallen. Despite achieving considerable popularity in the 50's, 60's and 70's, the frequency of its use has probably declined in recent years - mainly due to the advancement of more complex and refined techniques.

The social world is a complex place. Think about our last example - we've only looked at gender in isolation, but gender doesn't live in a vacuum: all women aren't the same, and neither are men. How do women in different ethnic groups, different ages, different social classes, perceive their safety walking alone in the dark? Unfortunately, as we begin to ask these more complex questions, and build more and more complex theories to answer them, the chi-square statistic becomes less and less useful. We could actually build a contingency table to look at all these variables in conjunction with each other and calculate a chi-square statistic, but the table would be huge. Unfortunately, the bigger the table, the harder it becomes to interpret the chi-square statistic.

Another problem is the general lack of sensitivity of the test - it just isn't powerful enough to pick up discrete differences and similarities in the data. It is a blunt instrument as we still have to analyse the content of the table ourselves to work out where the differences and similarities lay. The test only tells us that there is a significant association (or not) - it doesn't tell us where or how strong this is. Indeed, the more astute student may have noticed that when using the chi-square statistic with the cat-

egorical data contained within large datasets, significant results are more common than not. This is because the statistic is simply unable to cope with small differences in large datasets. This problem is partly due to the statistical limitations of categorical data - interval and ratio data are much more sensitive than categorical data because the measurements are necessarily much more refined. This is why interval and ratio data are often described as being statistically 'better' and the associated tests more powerful than their categorical counter-parts.

However, part of the attraction of chi-square is its simplicity. One of the dangers of building more and more complex theories is that we become too precise and begin to get lost in the detail of the data and the method. To paraphrase C Wright Mills (1959, p 83): "Precision is not the sole criterion for choice of method; certainly 'precise' ought not to be confused, as it so often is with 'empirical' or 'true'." For Mills, the point of sociological method is to expose the private and public troubles of society and place them in their historical context; it is not to show off our mathematical knowledge. In part, this is why it is important to remember the role of theory in data analysis. Indeed, the role of a research rationale is to direct your attention to area of the social world where there is some reason to think that something interesting might be going on. We can then employ tests like the chi-square statistic to find out if our hunches are borne out by the data. It would be easy to trawl datasets for statistically significant tests of association, but this, in itself, doesn't tell us very much. It is also why we shouldn't get too carried away when we find statistically significant results. Theory drives data analysis, not the other way around. In any case, whether you regard statistical significance as relevant is not always the point. Understanding the mechanics of a test like chi-square actually helps us to interpret the data with greater efficacy than we would be able if we were to simply eyeball data tables.

Indeed, not only is chi-square a test that is easy to understand and calculate, it is a test that is able to reveal the private and public troubles of society. It has a very useful ability to clearly and concisely identify trends in social data that are necessarily messy and not amenable to precise interval measurements. As a result, whilst the chi-square will not explain all the messiness and complexity of society, it can help you to see through it more clearly if you use it appropriately and critically.

You should now be able to:

- Correctly identify when to use the chi-square goodness of fit test and when to use the chi-square test of association
- Check that your data satisfies the assumptions of the chi-square test
- Calculate the chi-square statistic
- Assess the significance of the result
- Calculate strength of the association using the appropriate test
- Interpret, report, and critically understand the implications of your results
- Present this material in the appropriate form