CHI-SQUARE TESTS:

When to use a Chi-Square test:

Usually in psychological research, we aim to obtain one or more ***scores*** from each subject. However, sometimes data consist merely of the ***frequencies*** with which certain ***categories*** of event occurred. Instead of each subject providing a score, subjects contribute to a "head count": they fall into one category or another. In these circumstances, where one has **categorical** data, one of the family of Chi-Squared tests is the appropriate test to use.

To make this clear, let's consider a concrete example. Imagine that you were interested in the type of transport used by people to arrive at the university. Do people prefer some forms of transport rather than others? One way to tackle this question might be to provide people with a set of alternatives (car, bus, motorcycle, bicycle, etc) and ask them which method of transport they habitually used. Out of 100 people, you might end up with 60 who said they used their car, 20 who said they came by bus, 10 who came by motorcycle and 10 who came by bicycle. Each person asked would therefore fall into one of these categories. We would now have a set of ***observed*** frequencies of certain things happening.

We might now ask: are these observed frequencies similar to what we might expect to find by chance, or is there some non-random pattern to them? In the example above, do people show no particular preference for one form of transport over another, or are some forms of transport preferred and others disliked? In this particular case, it's pretty clear-cut: just looking at the frequencies would enable us to claim that most people come by car and that other forms of transport are relatively unpopular. But what if the frequencies had been 30, 22, 23, 25? Given these obtained frequencies, we might not be as confident that cars were the most popular form of transport.

The Chi-Square test helps us to decide whether or not our observed frequencies are due to chance, by comparing our observed frequencies to the frequencies that we might ***expect*** to obtain purely by chance.

Chi-Square is a very versatile statistic that crops up in lots of different circumstances. We will concentrate on two applications of it.

(a) Chi-Square "Goodness of Fit" test:

This is used when you want to compare an observed frequency-distribution to a theoretical frequency-distribution. The most common situation is where your data consist of the observed frequencies for a number of mutually-exclusive categories, and you want to know if they all occur equally frequently (so that our theoretical frequency-distribution contains categories that all have the same number of individuals in them).

Our example above would be appropriate for this test: we have one independent variable ("type of transport habitually used"), and a number of different levels of it (car, bus, bike, etc). We wanted to know if some forms of transport were preferred over others, or whether there were no preferences. In this context, "no preferences" means "all categories occurred with equal frequencies".

Another example might involve applications to study psychology at Sussex: we might want to see if certain schools of study are more popular with applicants than others. The independent variable might be "school of study", and the levels of this variable would be the different schools - SOC, BIOLS, COGS, AFRAS and CCS. Our data would consist of the number of applicants to each school. Again, we would be comparing these observed frequencies to those that would occur if all schools were equally popular.

A third example: we might be interested in the relationship between age of driver and likelihood of being involved in an accident. We could obtain the records of 1000 accidents, and see how many of the drivers involved fell into each of the following age-categories: 21-30, 31-40, 41-50, 51-60, 61-70, and 71-80. If there is no relationship between accident-rate and age, then there should be similar numbers of drivers in each category (all categories should occur with equal frequency). If on the other hand, younger drivers are more likely to have accidents, then there would be a large number of accidents in the younger age-categories and a low number of accidents in the older age-categories.

Chi-Square Goodness of Fit test - a worked example:

We select a random sample of 100 UCAS psychology applicants to Sussex, and find that they are distributed across five Schools of Study in the following way (fictional data, I hasten to add!)

	School of Study:					
	CCS	**SOC**	**AFRAS**	**COGS**	**BIOLS**	**total**
Observed frequency:	40	35	5	10	10	100

Are all Schools equally popular, as far as psychology applications are concerned? Or do some Schools attract more applications than others?

Looking at the observed frequencies, it appears that CCS and SOC are much more popular than the other Schools. We can perform a Chi-Square test on these data to find out if this is true, i.e. to see the pattern of applications deviates significantly from what we might expect to obtain if all schools were equally popular with applicants.

Our statistical hypothesis:

As with most statistical tests, we are faced with two possibilities. One is that there is *no* difference between the Schools in terms of application rate; any differences between our observed frequencies and those which we would expect to obtain, are just minor discrepancies which are due to chance. The alternative possibility is that there *is* a real difference between Schools in the application rates; the observed discrepancies are so large that they are unlikely to have arisen merely by chance - it is more likely that they reflect some "real" phenomenon, in this case the fact that some Schools are truly preferred over others.

We start off by assuming that the former possibility is true: this is the so-called "null hypothesis", that any discrepancies between our observed frequencies and those frequencies which we would expect to get, are due merely to chance. Only if the evidence is strong enough, will we reject the null hypothesis in favour of the alternative hypothesis - that our data really *are* different from the expected pattern of frequencies. What we mean by "evidence" in the context of the Chi-Square test, is that the discrepancies between the observed and expected frequencies are so large that they cannot be "explained away" as having occurred merely due to random variation.

Here is the Chi-Square formula:

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

Step 1: calculate the "expected frequencies" (the frequencies which we would expect to obtain if there were no School preferences).

$$\text{Expected frequency} = \frac{\text{total number of instances}}{\text{number of categories}}$$

We have 100 applications and 5 categories, so the Expected Frequency for each category is 20 (i.e., 100/5):

	CCS	SOC	AFRAS	COGS	BIOLS	total:
Observed frequency:	40	35	5	10	10	100
Expected frequency:	20	20	20	20	20	100

Step 2: subtract, from each observed frequency, its associated expected frequency (i.e., work out (O-E)):

O - E : 20 15 -15 -10 -10

Step 3: square each value of (O-E):

$(O - E)^2$: 400 225 225 100 100

Step 4: divide each of the values obtained in step 3, by its associated expected frequency:

$\frac{(O-E)^2}{E}$: 20 11.25 11.25 5 5

Step 5: add together all of the values obtained in step 4, to get your value of Chi-Square:

χ^2 = 20 + 11.25 + 11.25 + 5 + 5 = **52.5**

This is our obtained value of Chi-Square; it is a single-number summary of the discrepancy between our obtained frequencies, and the frequencies which we would expect if all of the categories had equal frequencies. The bigger our obtained Chi-Square, the greater the difference between the observed and expected frequencies.

How do we assess whether this value represents a "real" departure of our obtained frequencies from the expected frequencies? To do this, we need to know how likely it is to obtain various values of Chi-Square by chance.

Assessing the size of our obtained Chi-Square value:

(1) What you do, in a nutshell...

(a) Find a table of "critical Chi-Square values" (at the back of most statistics textbooks).

(b) Work out how many "degrees of freedom" (d.f.) you have. For the Goodness of Fit test, this is simply the number of categories minus one.

(c) Decide on a probability level.

(d) Find the critical Chi-Square value in the table (at the intersection of the appropriate d.f. row and probability column).

If your obtained Chi-Square value is **bigger** than the one in the table, then you conclude that **your obtained Chi-Square value is too large to have arisen by chance**; it is more likely to stem from the fact that there were real differences between the observed and expected frequencies. In other words, contrary to our null hypothesis, the categories did *not* occur with similar frequencies.

If, on the other hand, your obtained Chi-Square value is **smaller** than the one in the table, you conclude that there is no reason to think that the observed pattern of frequencies is not **due simply to chance** (i.e., we retain our initial assumption that the discrepancies between the observed and expected frequencies are due merely to random sampling variation, and hence we have no reason to believe that the categories did not occur with equal frequency).

For our worked example...

(a) We have an obtained Chi-Square value of 52.5.

(b) We have five categories, so there are 5-1 = 4 degrees of freedom.

(c) Consult a table of "critical values of Chi-Square". Here's an excerpt from a typical table:

	probability level:		
d.f.	**0.05**	**0.01**	**0.001**
1	3.84	6.63	10.83
2	5.99	9.21	13.82
3	7.81	11.34	16.27
4	**9.49**	**13.28**	**18.46**
5	11.07	etc.	etc.
6	12.59		
7			

(d) The values in each column are "critical" values of Chi-Square. These values would be expected to occur by chance with the probabiity shown at the top of the column. We are interested in the values in the 4 d.f. row, since our obtained Chi-Square has 4 d.f..(If we had a different number of d.f., we would use a different row of the table: so, with 5 d.f., you use the 5 d.f. row, 10 d.f. you use the 10 d.f. row, and so on)).

With 4 d.f., obtained values of Chi-Square as large as 9.49 will occur by chance with a probability of 0.05: in other words, you would expect to get a Chi-Square value of 9.49 or more, only 5 times in a hundred Chi-Squared tests. Thus, values of Chi-Square that are this large do occur by chance, but not very often. Looking along the same row, we find that Chi-Square values of 13.28 are even more uncommon: they occur by chance once in a hundred times (0.01). Values of Chi-Square as large as 18.46 are extremely unlikely to crop up by chance: the probability of them doing so is less than once in a thousand Chi-Square tests.

(e) We compare our obtained Chi-Square to these tablulated values. If our obtained Chi-Square is larger than a value in the table, it implies that our obtained value is even less likely to occur by chance than is the value in the table. Our obtained value of 52.5 is much bigger than the critical value of 18.46; and so we can conclude that the chances of our obtained Chi-Squared value having occurred by chance are less than one in a thousand (written formally, "p<0.001"). We can therefore be relatively confident in concluding that our *observed* frequencies are significantly different from the frequencies that we would *expect* to obtain if all categories were equally popular. In other words, not all Schools are equally popular with UCAS applicants.

(2) Why we do it - the Sampling Distribution of Chi-Square:

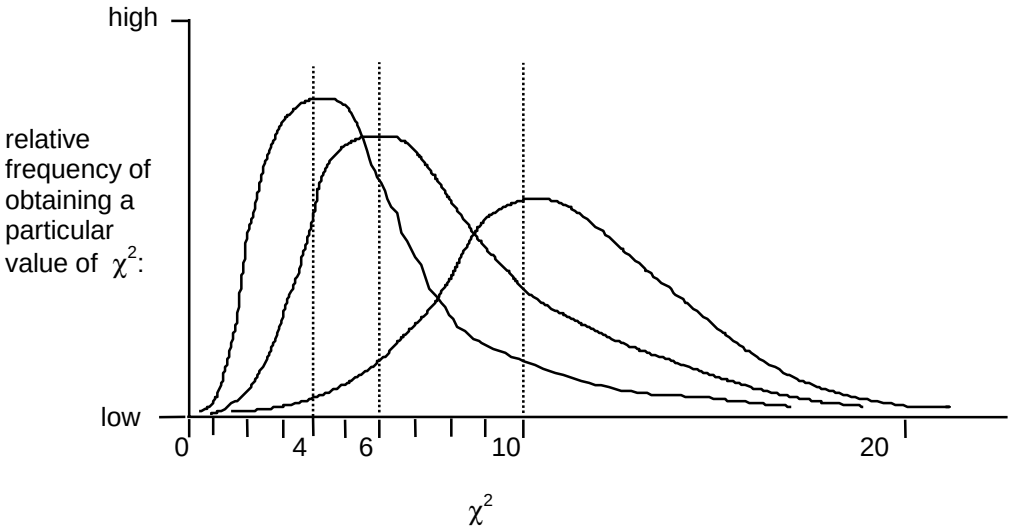
In essence, our obtained Chi-Square value represents the sum of the discrepancies between the obtained and expected frequencies. In interpreting it, our problem is that our data come from a random sample which has been taken from a population. Random variation in our sample means that our observed pattern of category frequencies may or may not be a good reflection of what's going on in the parent population. In the present example, we have a sample of 100 UCAS applicants, and the pattern within that sample may or may not be a good reflection of the pattern of preferences within the whole population of UCAS applicants to Sussex.

In short, there is a "true" state of affairs within the population. We are trying to work out on the basis of our sample, what that state of affairs is - but our sample is all we actually have to work with.

Suppose for the moment that the expected frequencies are the true ones within the population - in the present example, that there really is no difference between the Schools in terms of application rates. In practice, even if this were so, this state of affairs is unlikely to be reflected perfectly accurately in our sample. Imagine that we took lots of different random samples from the population of UCAS applicants to Sussex. Due to random variation from sample to sample, the observed and expected frequencies would rarely match exactly; more often, they would be fairly close; and occasionally, we would get a weird sample and the observed and expected frequencies would not appear to match at all. Although the true state of affairs was that all of the categories occurred with equal frequency, purely by chance a sample would sometimes give rise to an odd-looking frequency distribution and hence produce a large value of Chi-Square. If we took this large value of Chi-Square at face value, we might falsely conclude that there was a "real" difference in the frequencies of our categories, when in fact our obtained pattern of frequencies was distorted and really *there was no difference in the parent population*.

We clearly need to know how likely it is to obtain small, moderate and large values of Chi-Square from sample data, when the true state of affairs in the parent population is that all categories occur with equal frequencies. Fortunately, we don't need to actually do this for ourselves: statisticians have worked it all out for us.

The following graph shows what values of Chi-Square you get, when you take repeated random samples from a population in which the expected frequencies are true (i.e., when in fact there is no preference at all in the parent population for some categories rather than others).



The first thing to note is that the sampling distribution of Chi-Square looks rather different depending on the number of degrees of freedom. When we have 4 d.f., we are most likely to get a value of Chi-Square that is close to 4; sometimes Chi-Square will be much larger than 4 and sometimes it will be much smaller. The larger (or smaller) the value of Chi-Square, the less likely it is to occur. When we have 10 d.f., we are most likely to get a value of Chi-Square that is about 10; again, smaller and larger values are progressively less likely to be obtained. These distributions reflect the fact that very small and very large discrepancies between the observed and expected frequencies are unlikely to occur by chance. Note that, with small d.f., the Chi-Square distribution becomes progressively more asymmetrical, or "skewed": for example, with 4 d.f., large values of Chi-Square are much easier to obtain than very small values.

The graph shows that in assessing the size of our obtained Chi-Square, we need to take account of how many d.f. we have. The graph also shows that, for a given value of d.f., large Chi-Square values are unlikely to arise by chance - although the possibility always remain that they can do so! In other words, if we get a large value of Chi-Square, there are two possibilities:

(a) it reflects the true state of affairs in the parent population: hence there *is* a difference between the categories in terms of how many instances occur in each, and our theoretical frequency distribution is not a good description of the population;

(b) the observed discrepancies between the observed and expected frequencies in our sample are just due to sampling variation; we have no reason to suspect that the theoretical frequency distribution is not a good description of the population. It seems more likely that our sample poorly reflects the characteristics of the population from which it was taken.

We can never entirely be sure which of these two possibilities is true. However, the larger the value of Chi-Square, the more likely (a) becomes and the less likely (b) becomes. Very large Chi-Square values (and hence large discrepancies between obsrved and expected frequencies) are highly unlikely to have arisen by chance, although the possibility always remains that they have done so and are merely freak results. This fits in with our intuitions: the more marked the discrepancy betwen the frequencies which we *actually* obtain (from our sample), and the frequencies we *expected* to obtain (based on our ideas of what the parent population is like), then the more confident we would be in concluding that the discrepancy was not simply due to chance. If the discrepancies were small, we could dismiss them as being due to random variation, and maintain a belief that the charactristics of the populationn were being imperfectly reflected in our particular sample ; but there comes a point at which the discrepancies are so great that we cannot explain them away in this manner. Instead we have to accept that our sample *does* reflect the characteristics of the parent population; it's just that the population has different characteristics (a diffrent frequency distibruiotn) to those we originally thought it had.

(b) Chi-Square Test of Association between two variables:

This is appropriate to use when you have nominal (categorical) data for two independent variables, and you want to see if there is an association between them.

(b) Chi-Square Test of Association - worked example:

We could take our hypothetical 100 UCAS applicants and categorise them on two independent variables: educational background (two levels: Science versus Arts) and School choice (five levels: COGS, SOC, CCS, BIOLS or CCS). What we want to know is: is there a non-random association between these two variables: for example, do students with an Arts background show a different pattern of applications to that shown by students with a Science background?

Our observed frequencies can be put into a ***contingency table***, as follows:

	SCHOOL OF STUDY:					
	COGS	SOC	AFRAS	BIOLS	CCS	column total:
ARTS	10	12	5	3	40	70
SCIENCE	6	2	1	20	1	30
row totals:	16	14	6	23	41	100

These are our observed frequencies. In this particular case, simply looking at the observed frequencies gives us an idea that there is a systematic relationship between our two independent variables: CCS seems very popular with Arts students and unpopular with Science students, and the opposite seems to be true for BIOLS. In real life, however, things are often not as clear-cut as this. As with the one-variable Chi-Square test, our aim is to see if the pattern of observed frequencies is signficantly different from the pattern of frequencies which we would exect to see by chance - i.e., would expect to obtain if there were no relationship between the two variables in question. With respect to the example above, "no relationship" would mean that the pattern of school preference shown by Arts students was no different to that shown by Science students.

The Chi-Square formula is exactly the same as for the one-variable test described earlier; the only difference is in how you calculate the expected frequencies.

Step 1: calculate the "expected frequencies" (the frequencies which we would expect to obtain if there were no association between School preference and educational background).

$$\text{Expected frequency} = \frac{\text{row total} * \text{column total}}{\text{grand total}}$$

Do this for each cell in the table above. Here is that table again, with the expected frequencies in brackets below the obtained frequencies:

	COGS	SOC	AFRAS	BIOLS	CCS	row total:
ARTS	10	12	5	3	40	70
	(11.2)	(9.8)	(4.2)	(16.1)	(28.7)	
SCIENCE	6	2	1	20	1	30
	(4.8)	(4.2)	(1.8)	(6.9)	(12.3)	
column totals:	16	14	6	23	41	100

Thus the expected frequency for Arts students applying to COGS is 70 (the row total representing the total number of Arts students) times 16 (the column total representing the total number of applications to COGS), divided by the grand total (100). (70 * 16)/100 = 11.2.

The expected frequency for Science students applying to BIOLS is 30 (the row total corresponding to the total number of Science students) times 23 (the total number of students

applying to BIOLS) divided by the grand total (100). (30 * 23)/100 = 6.9. And so on for the rest of the table.

Note that the expected frequencies should add up to the same total as the observed frequencies (100 in this case), give or take a small rounding error; if they do not, you have made a mistake in the arithmetic somewhere.

Step 2: subtract, from each observed frequency, its associated expected frequency (i.e., work out (O-E)):

		COGS	SOC	AFRAS	BIOLS	CCS
ARTS	O:	10	12	5	3	40
	E:	11.2	9.8	4.2	16.1	28.7
	O-E:	**-1.2**	**2.2**	**0.8**	**-13.1**	**11.3**
SCIENCE	O:	6	2	1	20	1
	E:	4.8	4.2	1.8	6.9	12.3
	O-E:	**1.2**	**-2.2**	**-0.8**	**13.1**	**-11.3**

Step 3: square each value of (O-E):

ARTS **$(O-E)^2$: 1.44 4.84 0.64 171.61 127.69**

SCIENCE **$(O-E)^2$: 1.44 4.84 0.64 171.61 127.69**

(NB: It's just a coincidence that for this particular example, the $(O-E)^2$ values are identical for the Arts and Science students; there's no reason why they should be).

Step 4: divide each of the values obtained in step 3, by its associated expected frequency:

ARTS:

$\frac{(O-E)^2}{E}$ **1.44/11.2 4.84/9.8 0.64/4.2 171.61/16.1 127.69/28.7**

= 0.123 = 0.494 = 0.152 = 10.659 = 4.449

SCIENCE:

$\frac{(O-E)^2}{E}$ **1.44/4.8 4.84/4.2 0.64/1.8 171.61/6.9 127.69/12.3**

= 0.300 = 1.152 = 0.356 = 24.871 = 10.381

Step 5: add together all of the values obtained in step 4, to get your value of Chi-Square:

χ^2 = 0.123 + 0.494 + 0.152 + 10.659 + 4.449 + 0.300 + 1.152 + 0.356 + 24.871 + 10.381

χ^2 **= 52.94**

This is our obtained value of Chi-Square; it is a single-number summary of the discrepancy between our obtained frequencies, and the frequencies which we would expect if there was no association between our two variables.The bigger our obtained Chi-Square, the greater the difference between the observed and expected frequencies.

Interpreting Chi-Square:

To assess whether this value represents a "real" departure of our obtained frequencies from the expected frequencies, we compare our obtained Chi-Square value to the appropriate "critical" value of Chi-Square obtained from the same table as we used before. To do this, we need to know how many degrees of freedom we have. The d.f. are given by the following formula:

d.f. = (number of rows - 1) * (number of columns - 1)

In the present example, we have two rows and five columns, so we have (2-1) * (5-1) = **4 d.f.**

If our obtained Chi-Square is larger than a value in the table, it implies that our obtained value is even less likely to occur by chance than is the value in the table. Our obtained value of 52.94 is much bigger than the critical value of 18.46; and so we can conclude that the chances of our obtained Chi-Squared value having occurred by chance are less than one in a thousand (written formally, "p<0.001"). We can therefore be relatively confident in concluding that our *observed* frequencies are significantly different from the frequencies that we would *expect* to obtain if there were no association between the two variables. In other words, the pattern of applications to various Schools is different for Arts and Science students.

Note that the Chi-Square test merely tells you that there is *some relationship* between the two variables in question: it does not tell you what that relationship is, and most importantly, it does not tell you anyhting about the causal relationship between the two variables. In the present example, it would be tempting to interpret the results as showing that possession of an Arts or Science background *causes* people to apply to different Schools of Study. However, the direction of causality could equally well go the other way: possibly students *begin* by choosing a preferred School of Study, and *then* pick A-levels which will be most apporopriate for that School. Chi-Square merely tells you that the two variables are associated in some way: precisely what that association is, and what it means, is for you to decide - the test does not do it for you.

Another problem in interpreting Chi-Square is that the test tells you only that the observed frequencies are *different* from the expected frequencies in some way: the test does not tell you where this difference comes from. Sometimes, it might be that all of the observed frequencies are significantly discrepant from the expected frequencies, whereas on other occasions it might be that the significant Chi-Square value has arisen because of one particular marked discrepancy between the observed and expected frequencies. Usually, you can get some idea of why the Chi-Square was significant by looking at the size of the discrepancies between the observed and expected frequencies. In the present example, the discrepancies between observed and expected frequencies are relatively minor for Arts and Science students applying to COGS, SOC and AFRAS; about as many students apply to these Schools as you would expect if there were no differences between Arts and Science students in application-preference. The major discrepancies occur for CCS and BIOLS: inspection of the O-E discrepancies suggests that Arts students apply to CCS much more often than one would expect, and apply to BIOLS much less often than one would expect.For example, one would expect 16 applications to BIOLS by the Arts students, but in practice there wre only three. The Science students show the opposite pattern with their CCS and BIOLS preferences. In many cases, however, Chi-Square contingency tables are not so easy to interpret!

Assumptions of the Chi-Square test:

For a Chi-Square test to be used, the following assumptions must hold true:

1. Your data are a random sample from the population about which inferences are to be made.

2. Observations *must* be independent: each subject must give one and only one data point (i.e., they must contribute to one and only one category).

3. Problems arise when the expected frequencies are very small. As a rule of thumb, Chi-Square should not be used if more than 20% of the expected frequencies have a vlaue of less than 5. (It does not matter what the observed frequencies are). Note that both of the examples used in this handout violate this rule: 2 out of 5 of the expected frequencies in the goodness of fit example, and 4 out of the 10 expected frequencies in the contingency table example, are less than 5! You can get around this problem in two ways: either combine some categories (if this is meaningful, in your experiment), or obtain more data (make the sample size bigger).