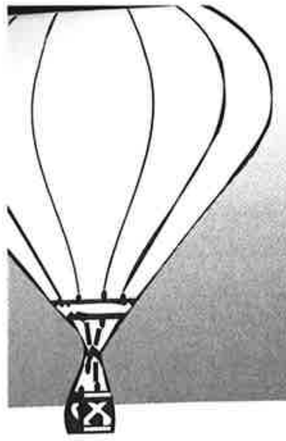


15 Testing for differences between two samples

This chapter introduces statistical significance tests for assessing the significance of differences between two samples and also introduces calculation of *effect size* and *power*.

- The *related t* test is used where data are in pairs, from a repeated measures or matched pairs design. H_0 is that the mean of the population of *difference means* (mean of differences between each pair of values) is zero.
- The t test for *unrelated data* (from independent samples) tests H_0 that the population of differences between two means has a mean of zero; that is, it assumes that the two populations from which the two samples are drawn have identical means.
- The *single sample t* is used to test the hypothesis that a single sample was drawn from a population with a certain mean; we usually want to show that this is unlikely and therefore that the sample is from a different population.
- t tests are a type of *parametric* or *distribution dependent* test that depend on certain data assumptions for their results to be reliable – *homogeneity of variance, interval level data* and a *normally shaped sampling distribution*.
- These tests are considered *robust* and more *power efficient* than their *non-parametric* equivalents, which are also dealt with here – the *Mann-Whitney* for unrelated data and the *Wilcoxon matched pairs* for related data. These *non-parametric tests* use ranks of the data and are considered to have on average 95.5% of the power of their parametric equivalents.
- The *sign test* for related categorical data is described.
- SPSS procedures for all tests in the chapter are provided.
- *Effect size* is introduced as a concept concerned with the size of the effect that the study was investigating whether or not a significant effect was found; if significance was not found, and there *is* an effect, a Type II error has occurred and it is stated that many researchers find the traditional structure of reliance on the significance test too conservative with the fear that many effects are missed through Type II error.
- The likelihood of missing an effect, if it is there, is the probability β and *power* of a specific test is defined as $1 - \beta$. This is the probability of demonstrating a significant effect if an effect really exists. Ways to increase power are discussed.
- Calculations are provided for effect size and power of t tests.

Tests of difference between two conditions or groups

This chapter deals with significance tests on typical data from a two-condition investigation such as those depicted in Table 15.1. We first deal with so-called PARAMETRIC TESTS, or DISTRIBUTION DEPENDENT TESTS, which are the various kinds of t test. These are conducted on data that are at least at *interval level.* We then look at non-parametric equivalents – the Mann-Whitney and the Wilcoxon tests – which are used on data that have been ranked (that is, they are at *ordinal level*). We usually only use these when the data have qualities that make a conclusion from a t test unsafe. The sign test is used on categorical related data.

In all cases we are dealing with the situation where you have two sets of data, typically scores for two conditions of an experiment or scores for two different groups of people. In order to select the appropriate test for your data you also need to decide whether they are *related* or *unrelated* (see Chapter 3). Data from *matched pairs* designs or from two measures of the same participant (*repeated measures* design) are *related. Unrelated* data occur where the two groups of participants providing scores consist of entirely different people; in other words, the two sets of scores to be tested come from two completely different (independent) sources – an *independent samples* design. (See also Chapter 21 and pp. 279–80.) Though it will seem odd, data produced where a single participant provides scores in two conditions of an experiment, several trials in each condition, are treated as unrelated – see p. 76.

Parametric tests

The t test for related data

When to use the related t test

Type of relationship tested	Type of data required	Design of study
Difference between two conditions	At least interval	Within groups: repeated measures matched pairs

Data assumptions: see required data assumptions on p. 363. The data in the example below do not violate any of these assumptions.

Note: If your data do not approximately meet the data assumptions for the test explained on p. 363. you will need to transform your data or use a non-parametric or distribution free test such as the Wilcoxon matched pairs signed ranks test (see p. 368).

Data for a related *t* test

Take a look at the data in Table 15.1, which displays results from an experiment on the improvement in memory recall produced by using imagery. Each participant has been tested in *both* the control condition (no specific instruction) *and* the imagery condition (where they were asked to form vivid images of each item). Hence it is a *repeated measures* or *related* design. The data come in pairs. Let's just think step-by-step through what we expect to happen here, if there is an 'effect' from imagery. Because we argue that imagery should be a memory aid we would expect the imagery scores to be higher than the control condition scores. Note that, in general, they are, but we cannot just say 'it worked'; we need a significance test to demonstrate to the research world that the probability of these differences occurring, if the null hypothesis is true, is less than .05.

	Number of words recalled in:			
Participant number	Imagery condition (*I*)	Control condition (*C*)	Difference	
			d	d^2
1	6	6	0	0
2	15	10	5	25
3	13	7	6	36
4	14	8	6	36
5	12	8	4	16
6	16	12	4	16
7	14	10	4	16
8	15	10	5	25
9	18	11	7	49
10	17	9	8	64
15	12	8	4	16
12	7	8	−1	1
13	15	8	7	49
	$\bar{x}_i = 13.38$	$\bar{x}_c$ 58.85	$\Sigma d = 59$	$\Sigma d^2 = 349$
	$s_i = 3.52$	s_c 51.68	$(\Sigma d)^2 = 3481$	

Mean of differences ('difference mean') $\bar{d} = 4.54$
Standard deviation of differences $s_d = 2.60$

Table 15.1 Number of words correctly recalled under imagery and control recall conditions (columns *I* and *C*) and statistics required to calculate related *t*

Calculating statistical significance tests

Note that, although we will go through the derivation of the *t* test in detail, in order for you to be able to understand what is going on, I will also with each significance test, provide a standard formula so that you can just submit your data to a test by simply following the steps of the calculation. These steps are given in special boxes after the explanation of each test. It is also likely, if you are at a university, that you will use SPSS to do the analysis.

Starting out on the related *t* test

Remember that every inferential statistical test is a test of a null hypothesis. We want to calculate the probability of our result occurring if nothing is really going on. It might help here, then, to think about what kind of results we'd expect if there is no effect – imagery does not help memory.

If we look at the *difference between each person's two scores* we will see whether they improved in the imagery condition or got worse. If the null hypothesis is true then people don't generally improve and all these differences should be close to zero. However, our research argument is that most people should *improve*. In turn this means that differences (imagery score minus control score, shown as *d* in Table 15.1 should generally be *positive*. The larger they are, if positive, the better for our research hypothesis. This is a *directional* approach. In a *non-directional* approach we would simply be saying that the differences should go in one direction, without specifying which.

The null hypothesis in the related *t* test

The null hypothesis here is that the two samples of scores come from populations with the same mean. However, since this is a related design, we can state the null hypothesis in terms of the *difference values*. If there is no imagery effect then the differences should all centre round zero and be relatively small – only the result of random error. Our null hypothesis, then, can be re-stated: the population of differences has a mean of zero (Figure 15.1). We can write this as:

$H_0 : \mu_d = 0$

where μ_d is the mean of the population of differences.

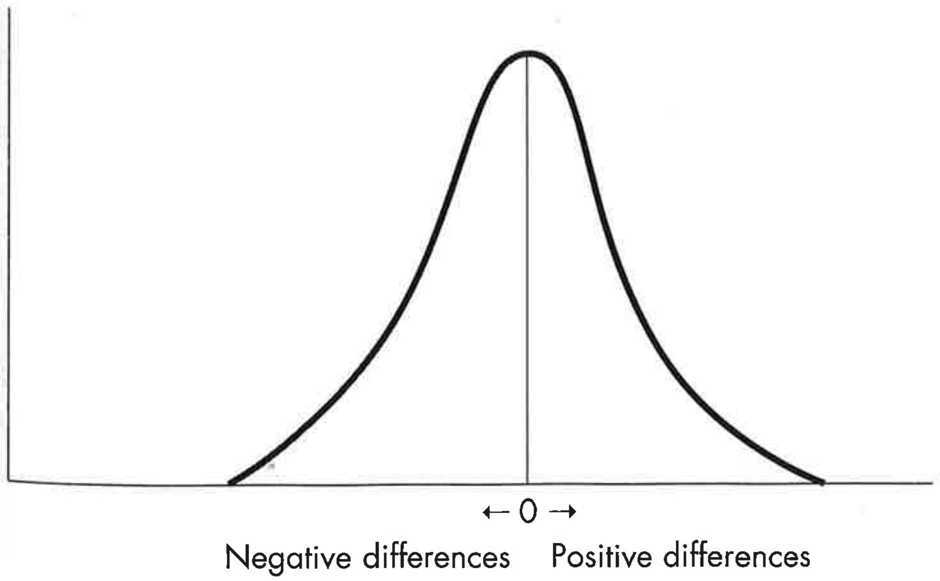


Figure 15.1 Hypothetical distribution of differences under H_0

Testing the null hypothesis

Think of the population of difference values as like the barrel of screws from the last chapter. The set of differences shown in the '*d*' column of Table 15.1 is a *sample* of score differences drawn randomly from this population and it has a mean of 4.54 (see the bottom of the *d* column). This mean of differences is known as a DIFFERENCE MEAN. Hence to test for significance we need to know the probability that a sample of 13 differences with a difference mean as large as, or larger than, 4.54 would be drawn at random from the population in Figure 15.1, which has a mean of zero.

What would be handy would be to know how *other* samples of difference values would be arranged. If, under H_0, the underlying population of differences has a mean of zero then samples taken randomly from it should all *also* have means close to zero. They would differ a bit from zero through *sampling error*, sometimes positive, sometimes negative, sometimes large, mostly small. What would happen if we kept on taking samples of 13 differences from this population? What kind of distribution of difference means would we get? If we knew this, then we could compare *our* difference mean of 4.54 with this distribution and see how unlikely ours would be to occur. Funnily enough, we have already encountered this concept of sampling over and over again in a previous chapter. In Chapter 13 we met the concept of a *sampling distribution*. You might like to re-read the appropriate section of that chapter in order to refamiliarise yourself with the idea.

Figure 15.2 shows the kind of distribution we might obtain if we were to dip into the population of close-to-zero differences taking samples of 13 many times over and recording the difference mean each time. It is called a *sampling distribution of difference means*. Note that it is much narrower than the distribution of differences because it is composed of *samples* of differences taken 13 at a time. The means will not vary as much around zero as the individual differences do.

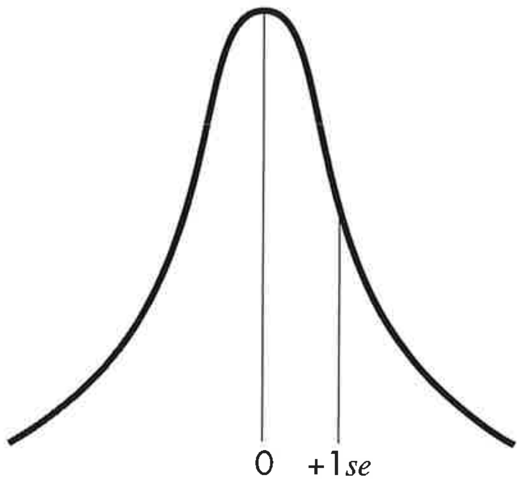


Figure 15.2 Sampling distribution of difference means under H_0

If we knew the statistical properties of this distribution we would be able to say whether our difference mean was a much-to-be-expected one or an extreme one. Trouble is, we don't have those properties ... or don't we? We need the mean and standard deviation of the sampling distribution because we just need to know how many standard deviations our difference mean is from the mean. Well we know the mean. We have already said that if the null hypothesis is true, the difference means will centre around zero. But what about the standard deviation? Well, on p. 316 we saw that statisticians have a formula for *estimating* the standard deviation of a sampling distribution using the sample that you have drawn. Don't forget, though, that the standard deviation is called the *standard error* here. This is because each sample, drawn randomly under H_0, differs from the mean of zero only because of sampling error. Each deviation of a sample from the population mean is an 'error'.

To estimate the standard error of the sampling distribution shown in Figure 15.2 then, we use the *central limit theorem* using standard deviation of the differences (s_d), which gives us

$$se = \frac{s}{\sqrt{N}} = \frac{2.6}{\sqrt{13}} = 0.721$$

Now we can just ask 'How many standard errors is *our* difference mean away from the hypothetical difference mean of zero?' We can get this by dividing our difference mean by the standard error: 4.54/0.721 which gives us 6.297. This value is known as a *t* value.

So our difference mean, if it was sampled at random from all difference means, would be 6.297 standard errors away from the population mean of zero. Is this a long way, making it a very unlikely occurrence? In Chapter 13, we learned that the number of standard deviations a score is from the mean is a *z* value (see p. 312 and a *z* value of over 6 is ever such a long way from the mean on a normal distribution. We might be tempted just to look up our *z* table as we did in Chapter xx and find the probability of a *z* that large occurring. There is just one little snag with this. The distribution of *t* is not normal in shape unless *N* is very large (e.g. 120 or more) when *t* can indeed be treated as a normal *z* value. The lower the value of *N*, though, the more the distribution of *t* would be broader than a normal distribution.

For the theory and mathematical tables associated with *t* distributions, and the *t* tests, we are indebted to William Gossett, who worked for the Guinness organisation. Guinness, at the time he did his stuff, did not permit its workers to publish findings connected with company work. Hence he published under the pseudonym of 'Student' and the distribution statistic is known, therefore, as Student's *t*. As a result of his work, however, we can consult *t* tables and find out whether our obtained value for *t* exceeds the critical value contained in the table.

Consulting critical value tables

We want to know if the probability of obtaining a *t* value of 6.297 is less than α in order to be able to claim our difference between means as significant. We therefore go to the Appendix where we find Table 3 containing critical values for *t*. In order to use this table we need to know a few things.

- First, we need to decide what level of α is appropriate. Usually this is .05 and you should always use this value to start with.
- Next, we have to decide whether we are conducting a one- or two-tailed test. As explained at the end of Chapter 14 it is best to use two-tailed tests and hope that your effect is large enough to show significance with the design you are using.

- Finally, we need to know our *degrees of freedom* (*df*). This term was introduced in Chapter xx. It is the number of items in our sample that are free to vary, given that we know their mean. If we knew the mean of the differences but no individual values for *d*, we could enter any values we liked up to the 12th value but the 13th would then be fixed in order to make the mean what it actually is. Here we have 13 differences so 12 are free to vary and $df = 12$. The **Procedure** for each significance test will provide a simple formula for calculating *df* where needed.

Analysis of our result for *t* – effect size and power

Now we are ready to enter Table x appropriately. We see that for a two-tailed test with α at .05 and with 12 *df*, the critical value for *t* is 2.179 and our value of 6.297 easily beats that. In fact it also beats the value for $p \leq .01$ over to the right. We said in the last chapter that a result significant with $p \leq .01$ is not automatically 'better' than one where *p* is ≤ .05. It all depends on sample size, *effect size* and *power*. *Effect size* is an estimate of the size of effect we appear to have demonstrated. *Power* is the probability of not making a Type II error. If we have low power, then a real existing effect might not show as significant. This issue is discussed more fully later on in this chapter but do be aware that nowadays reports of effect size, along with significance, are becoming more common and may be expected.

Whatever the objections of the statisticians, psychological researchers would tend to report this difference as 'highly significant' and to give the $p < .01$ value. Certainly, based on our sample results, we can confidently reject the null hypothesis (that there is no population difference) and argue that the use of imagery in this experiment appears to improve memory recall for words. However, we should also report on the estimated size of the effect (see p. 385).

Formula for calculation of related *t*

$$(1)\quad t = \frac{\bar{d}\sqrt{N}}{s} \qquad \text{or} \qquad (2)\quad t = \frac{\sum d}{\sqrt{\frac{N\sum d^2 - (\sum d)^2}{N-1}}}$$

Equation (1) is just a slight rearrangement of what we did just above when we divided the difference mean by the estimated standard error of its sampling distribution. This one is the easier calculation if you have a simple statistical calculator that will give you standard deviations. Here are the calculation steps for equation (2) where you only need the differences themselves and *N*. SPSS procedures are given on p. 379.

Hand calculation of related *t*

Procedure	Calculation/result of steps
1 Find the mean for each condition	From Table 15.1 $\bar{x}_i = 13.38$ $\bar{x}_c = 8.85$
2 Arrange columns so that condition with higher mean is to the left of the other condition; this is to make subtraction easier.	Column I before column C in Table 15.1
3 Subtract each score C from score I	See '*d*' column in Table 15.1
4 Square each *d*	See 'd^2' column in Table 15.1
5 Total all *d* (Σd) and all d^2 (Σd^2)	$\Sigma d = 59$ $\Sigma d^2 = 349$ (see Table 15.1)
6 Square Σd to get $(\Sigma d)^2$ Note: This is not the same as Σd^2 – be careful to distinguish between these two terms. $(\Sigma d)^2$ says add the *d*s then square the result. Σd^2 says square each *d* then add the results.	$(\Sigma d)^2$ = 59 x 59 = **3481**
7 Find *N* x Σd^2	13 x 349 = **4537**
8 Subtract $(\Sigma d)^2$ from the result of step 7	4537−3481 = **1056**
9 Divide the result of step 8 by $N-1$	1056/12 = **88**
10 Find the square root of step 9	√88 = **9.38**
11 Divide Σd by the result of step 10 to give *t*	*t* = 59/9.38 = **6.29**
12 Find *df* * In a related design $df = N-1$	$N-1$ = **12**
13 Check *t* for significance in critical value table, finding the *highest* table value of *t* that our obtained *t* is greater than, and make significance decision.	For 12 *df t* must be ≥ 3.055 (two-tailed) for significance with $p \leq .01$ our obtained *t* is greater than 3.055, hence the difference is highly significant and we reject H_0.

* Degrees of freedom: see explanation of this concept on p. 273.

Reporting results of significance tests – what you should actually write

Research psychologists generally employ the conventions laid down by the American Psychological Association (APA) in reporting the results of statistical analysis. You will probably be asked to follow this format in presenting your results section where your assignment is a scientific report of a quantitative psychological investigation. Consequently, after we have looked at the analysis of each test from now on, the APA format for reporting will be given. Most courses will not ask you to report the estimated size of your effect (see p. 385) or confidence limits (see p. 316) but some do, so this information has been included. If you are not asked to report these values then just ignore the last sentence below (before the note) and in future results report examples.

Reporting results of a related t test

The mean number of words recalled in the imagery condition (M = 13.38, SD = 3.52) was higher than the mean for the control condition (M = 8.85, SD = 1.68) resulting in a mean increase (M = 4.54, SD = 2.6) in the number of words recalled per participant. This increase was statistically significant, t (12) = 6.29, p < .001, two-tailed. The mean difference (mean difference = 4.54, 95% CI:2.97 to 6.11) was large (Cohen's d = 2.638).

Note: If this result were not significant do NOT say it was 'insignificant'. Write:

' ... This increase was not significant, t (12) = 1.477, p = .165'

(or you could write 'ns' or 'p > .05')

The t test for unrelated data

When to use the unrelated t test

Type of relationship tested	Type of data required	Design of study
Difference between two conditions or groups	At least interval	Between groups; Independent samples

Data assumptions: see required data assumptions on p. 363. The data in the example below do not violate any of these assumptions.
Note: if your data do not approximately meet the data assumptions for the test explained on p. 363. you will need to transform your data (if skewed) or select the appropriate line in SPSS. The alternative is to use a non-parametric or distribution free test such as the Mann-Whitney U test – see p. 371.

The reasoning for the unrelated t test is similar to that for the related t, the difference being only that the two samples of data have come from independent sources; that is, they are not pairs of scores from the same person or from matched participants. Typically we might have scores in two experimental conditions where each participant has been tested *in one condition only*. Another common source of data for the unrelated t test would be scores on a psychological measure from two different groups of people, e.g. reading scores for dyslexic and non-dyslexic students. Differences between groups of males and females would be unrelated (unless they are brothers and sisters!).

Data for an unrelated t test

Take a look at the data in Table 15.2 where participants have been divided into two groups, those above the median on a measure of disturbed sleep and those below this median. It was proposed that participants with a higher level of disturbed sleep would have higher anxiety levels than participants whose sleep was less disturbed. This does appear to be the case; the mean anxiety score for the higher sleep disturbance group is 12.4 (h = high disturbed sleep) whereas the mean for the lower sleep disturbance (l) group is 10.1 However, we need to know whether this difference between means is significant or not.

Anxiety scores for:

participants above median on disturbed sleep (h)		participants below median on disturbed sleep (l)	
Score x_h (N = 10)	x_h^2	Score x_l (N = 11)	x_l^2
14	196	**8**	64
11	121	**10**	100
9	81	**9**	81
12	144	**11**	121
13	169	**9**	81
15	225	**11**	121
13	169	**8**	64
11	121	**12**	144
17	289	**11**	121
9	81	**13**	169
		9	81
$\Sigma x_h = 124$		$\Sigma x_l = 111$	
$(\Sigma x_h)^2 = 15376$	$\Sigma x_h^2 = 1596$	$(\Sigma x_l)^2 = 12321$	$\Sigma x_l^2 = 1147$
$\bar{x}_h = 12.4$		$\bar{x}_l = 10.1$	
$s_h = 2.55$		$s_l = 1.64$	

Table 15.2 Anxiety scores for high and low sleep-disturbed participants

The null hypothesis for the unrelated t test

The null hypothesis here is that the two populations from which our two samples have been randomly drawn have equal means. We can write this as:

$H_0 : \mu_h = \mu_l$

Testing the null hypothesis

We can think of our two samples of anxiety scores here as like the two samples of screws taken from the two different barrels we encountered in Chapter 14. Because the two samples of scores are independent (from two *different* sets of people) we cannot, as in the related t test, look at pairs of scores and find the difference for each participant. However, what we *can* do is consider what would happen if we took two samples from two identical barrels at random, many times over, and each time recorded the difference between the two sample means. What we would obtain, if we plotted these differences each time, is a distribution looking much like that in Figure 15.3.

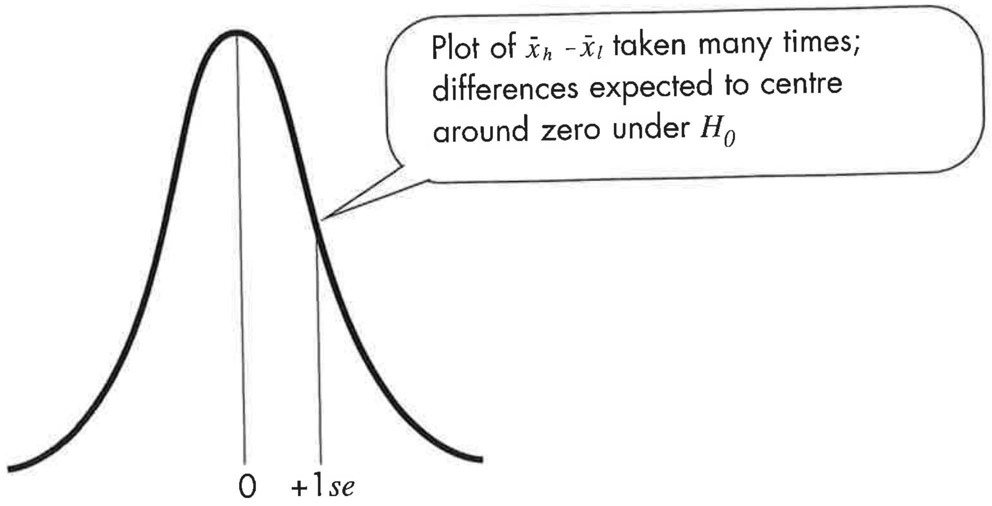


Figure 15.3 Sampling distribution of differences between two sample means

Under H_0, if the two underlying population means are identical, then the mean of this sampling distribution will be zero. That is, if we take two samples from the same population many times over, sometimes the difference between these two means will be positive and sometimes negative; sometimes it will be large but mostly it will be small. Under the laws of random selection the distribution of differences will look like that in Figure 15.3. What we want to know, then, is how far away from zero on this distribution does our *obtained* difference between two means fall? If we know this we can easily find the probability of our difference occurring under H_0.

Again, we need the properties of the sampling distribution of differences between two means. We know its mean is zero so what will be its standard error? Well, unfortunately this is not as easy a question to answer as it was for the related *t* test. In keeping with the philosophy that psychology students need only understand the basic principles of statistics rather than appreciate the finer points of derivation, I will simply explain what the equation *does* rather than produce a comprehensive explanation. Remember that all we want to do is to estimate where our obtained difference between two means falls on the distribution expected under H_0 and shown in Figure 15.3. To do this, as for a *z* value, we divide the difference by the standard error of the distribution. This answers the question 'How many standard errors is our obtained difference from zero?' In the related *t* test we used the central limit theorem to estimate easily the population variance from the sample variance. Trouble is, on this occasion we have *two* variances from two samples. What we do in the unrelated *t* test is estimate the variance of the distribution shown in Figure 15.3 from the POOLED VARIANCE of the two samples. Now, if you can bear it, take a peek at equation 3 below, but please do not panic! Yes, it is nasty, but it involves no more arithmetic than can be done on the simplest of calculators – there's just a lot of it!

The nasty bit on the bottom is the pooled variance used in the estimate of standard error of the sampling distribution of differences between two means (i.e. the standard deviation in Figure 15.3). On top of the nasty equation below, then, is our obtained difference between two means. Below is the estimated standard error. The equation will give us *t*, which will be the number of standard errors we estimate our difference to be from the difference of zero expected under the null hypothesis. This estimate takes into account the fact that sample sizes may be different.

$$\frac{|\bar{x}_a - \bar{x}_b|}{\sqrt{\left(\frac{\left(\sum x_a^2 - \frac{(\sum x_a)^2}{N_a}\right) + \left(\sum x_b^2 - \frac{(\sum x_b)^2}{N_b}\right)}{(N_a + N_b - 2)}\right)\left[\frac{N_a + N_b}{N_a N_b}\right]}}$$

By-hand calculation of unrelated *t*

Obviously, equation (3) looks pretty complex but in fact it just involves a lot of basic steps with the emphasis on *a lot*! If you want to calculate *t* by hand, the following procedure box will take you through these steps. Computer packages like SPSS will of course whisk you through it in a jiffy ... but you won't have the satisfaction of cracking this monster! SPSS procedures appear on p. 381.

Please note that, from the general equation above, in this example *a* has been substituted by the *h* or *high sleep deprivation* scores, whereas *b* is substituted by the *l* or *low sleep deprivation* scores.

	Procedure	**Calculation/result of steps** See Table 15.2 for all summary statistics
1	Add the scores in the first group	Σx_h = **124**
2	Add all the *squares* of scores in the first group	Σx_h^2 = **1596**
3	Square the result of step 1; always be careful here to distinguish between Σx_h^2 and $(\Sigma x_h)^2$	$(\Sigma x_h)^2$ = **15376**
4	Divide the result of step 3 by N_h	15376 ÷ 10 = **1537.6**
5	Subtract result of step 4 from result of step 2	1596 – 1537.6 = **58.4**
6	Steps 6–8: Repeat steps 1 to 3 on the scores in the second group	Step 6: Σx_l = **111** Step 7: Σx_l^2 = **1147** Step 8: $(\Sigma x_l)^2$ = **12321**
9	Divide the result of step 8 by N_l	12321 ÷ 11 = **1120.1**
10	Subtract result of step 9 from result of step 7	1147 – 1120.1 = **26.9**
11	Add the results of steps 5 and 10	58.4 + 26.9 = **85.3**
12	Divide the result of step 11 by $N_h + N_l - 2$	85.3 (10 + 11 – 2) = **4.49**
13	Multiply the result of step 12 by: $\frac{N_h + N_l}{N_h N_l}$	4.49 ÷ 21/110 = **0.85**
14	Find the square root of the result in step 13	√ 0.85 = **0.92**
15	Find the difference between the two means	$(x_h - x_l)$ 12.4 – 10.1 = **2.3**

16 Divide the result of step 15 by the result of step 14 to give t	$t = 2.3 \div 0.92 = \mathbf{2.5}$
17 Find degrees of freedom (df) where $df = N_h + N_l - 2$	10 + 11 − 2 = **19**
18 Consult Appendix Table 3 and decide upon significance	For a two-tailed test with $df = 19$, t must be ≥ 2.093 for significance with $p \leq .05$; hence the difference between means here is significant.

We can reject the null hypothesis that people with high sleep disturbance do not differ from people with low sleep disturbance on anxiety. It remains open to question, since this was *not* an *experiment*, whether disturbed sleep is a cause of anxiety or whether anxiety causes sleep disturbance. The issues of effect size and power should also be considered in reporting this result (p. 385).

Reporting results of an unrelated *t* test

High sleep-disturbance participants produced higher anxiety scores (M = 12.4, SD = 2.55) than did the low sleep-disturbance participants (M = 10.1, SD = 1.64). The difference between means was significant, t (19) = 2.5, p < .05, two-tailed.

The difference between means (difference = 2.3, 95% CI: 0.37 to 4.25) was large (Cohen's d = 1.08)

The single sample *t* test

Data requirements: Interval level data and normal distribution

The tests so far covered are the most common type of t tests, those where we do not know the features of the appropriate underlying population and where we are testing the difference between two samples. Usually we have a second sample, which is a *control* group, because we need to make a comparison with what would happen if no treatment were applied. In some cases, however, we *do* know the features of a population, in which case our significance testing is made easier. In this case we do not need a control group because we already know the population mean for the condition in which no specific treatment is applied.

Let's return to the spinach-eating example in Chapter 13. There we found that the mean reading score for a population of eight-year-olds was 40, with a standard deviation of 10. We said there that significance tests would normally be carried out on a *sample* of children's scores, not just one. Suppose we identified a sample of 20 spinach-eating children whose combined average reading score was 43 with a standard deviation of 6. This is in favour of our hypothesis that spinach eating enhances reading. However, we need to test our difference for significance.

The null hypothesis for a single sample *t* test

In a single sample t test we know (or can argue for) the population mean for the variable under investigation. In the spinach case we assume: H_0: $\mu = 40$

What we don't know is what the sampling distribution for samples of 20 children at a time would look like, so we can again use the central limit theorem to estimate this for us. The standard error of the sampling distribution will be

$$se = \frac{s}{\sqrt{N}} = \frac{6}{\sqrt{20}} = 1.34$$

Our sample mean is 3 points away from the assumed mean of 40 under H_0 (which assumes that our spinach-eating kids have been randomly sampled from the normal reading population). How many standard errors is this from 40? We must divide the difference between our mean and 40 by the standard error. This will be our obtained value for t, so let's find it:

$$t = \frac{3}{1.34} = 2.23$$

df here, as usual, are one fewer than the number of data points we have, so $df = 19$. From Appendix Table 3, t must be ≥ 2.093 for significance with $p \leq .05$ (two-tailed). Hence this difference is significant and we may reject the null hypothesis that spinach-eating children do not differ from other children on reading (I do stress this is fictitious; don't rush to the greengrocer's!). The issue of effect size should also be considered in reporting this result (see p. 385).

Reporting the result of a single sample *t* test

The spinach-eating group produced higher reading scores (M = 43, SD = 6) than the known mean for the normal population (M = 40, SD = 10). This difference was significant, t (19) = 2.23, p < .05.

The difference between the sample mean and the population mean (3) was medium (95% CI: 0.188 to 5.812. Cohen's d = 0.5).

(**Note:** The data set for these calculations is provided on the companion website at: www.hodderplus.com/psychology/

Data assumptions for *t* tests

The t tests are often referred to as being in a class known as *parametric tests*. Actually, they are more appropriately known as *distribution dependent tests* and this is because, as we saw above, they make estimations of underlying distributions. These estimations will be seriously distorted if the data we have gathered do not conform to certain criteria. In turn, any inference we make from our data may be suspect and we may need to rethink the analysis of our data in order to be more confident of the decision we make about them. There are three basic considerations about the data for our dependent variable that we must make, and these are described below.

1 The level of measurement must be at interval level

This is a much-debated point but there is an easy rule of thumb. As we saw in Chapter 11, interval-level data have equal amounts for equal measures on the scale. Hence, distance (which is also a ratio measure) increases in equal amounts for every increase in unit. We know that this can hardly be true for many psychological measures such as intelligence, extroversion, and so on, although it is the psychometrist's dream to create such scales. However, if a scale has been *standardised* (see p. 201) it is usually safe to treat its scores as interval-level data. Scores on a memory test or errors on some other cognitive task represent sensible ratios – six words recalled is twice as good as three words recalled – but we would not claim therefore that Harry with six has twice as good a memory overall as Ron. We are only measuring performance on an isolated task. Hence one need not get into this kind of metaphysical argument in order to be satisfied that these kinds of measures produce data we can safely treat as being at interval level.

Where we run into difficulty is with invented scales that rely on human judgement. Typically, we might ask people to 'rate your level of confidence on a scale of 1 (= not at all confident) to 10 (= highly confident)'. Here we know that your 8 might be my 6 and we have no independent way of knowing that people using this scale are separating themselves from others by equal amounts. When using this kind of scale to produce data which we then want to test for significance there are two main options: (1) subject the data to normality checks as described below and in more advanced textbooks, transforming the data if necessary, or (2) choose one of the *non-parametric tests* to be described in the next section.

2 Data from a normal distribution

Our data should have been drawn from an underlying normal distribution. For a standardised psychological measure this should be true since this is what is created when the test is first developed. However, you may be drawing data from a somewhat different population from that on which the test was standardised. For any data set it is worth checking that there is not too large a skew and that the data are distributed as would be expected if drawn from a normal distribution (see p. 308). Your data are not normally distributed if they do not fit the criteria outlined on p. 319. In addition you can inspect several kinds of plot using SPSS, especially the Normal Q-Q plots, as described in Pallant (2007).

What do you do if you find the data really are a long way out from normal? Chapter 13 tells us that skew is too much when the value for skew is twice its standard error. There are two common solutions, the first of which is simply to use a non-parametric test (see below). The other is to stay with the *t* test and use, instead of actual scores, a TRANSFORMATION of these scores. This sounds like a glorious cheat but in fact we can 'normalise' our data this way and then legitimately carry on with the *t* test, so long as the skew is now acceptable. Among the transformations you can perform are:

- square
- square root
- log (to base 10 or other bases).

Be careful if you use square root since the square root of zero is an impossible calculation and SPSS will exclude all scores of zero from the analysis. The answer is to first add 1 to every score. Getting the log of each score can be done quite easily using a computer spreadsheet (such as Microsoft Excel or SPSS) or using a simple scientific calculator. You can use log10, $\log_e$ (as **LN** does in the SPSS instructions below), or any other log base, so long as it has the effect of normalising your data (reducing the skew).

In SPSS the commands are: **transform/compute**. You see a box, top left, labelled **Target Variable** which asks you to name a new variable that will be the log of your raw scores (e.g. ***loghits***). You then select **Sqrt LOG10** or **LN** after clicking **Arithmetic** on the right-hand side (scroll down alphabetically) and click that upwards into the box labelled **Numeric Expression**. Place the cursor between the two brackets in the new expression and double-click the variable to be logged (e.g. ***hits***). It should now appear between the two bracket signs (e.g. you should see **LN(*hits*)**, and you're ready. Click **OK** and a new logged variable will appear in your datasheet; check that its skew is not more than twice its standard error and use this instead of ***hits*** as the dependent variable in conducting your *t* test. If this doesn't work try another transformation. Note that, although you use this transformed variable in your analysis (e.g. the *t* test), having explained to your reader that this is what you have done, you should from then on refer to ***hits*** and not ***loghits*** when discussing your findings.

3 Homogeneity of variance

For two condition tests, this check need only be made where the design is unrelated and the sizes of N_a and N_b are very different, e.g. 7 and 23. Homogeneity of variance requires that the variances in the two populations are equal and we check this by showing that the two sample variances are *not* significantly different. The reason for this lies in the estimation of pooled variance in the formula on p. 360. The null hypothesis assumes that both samples are drawn from similar distributions. Each individual sample variance is an estimate of the underlying population variance and we average these two to get a better estimate. This averaging makes no sense if the variances are very different.

To test for homogeneity of variance you can do the following:

1. If using SPSS, you will automatically be given the result of a Levene's test when you conduct an independent samples *t* test. You just need to note that the significance value for this is *not* under .05; if it is, the variances are not homogenous. If this happens then you just have to consult the line '***Equal variances not assumed***' in the *t* test results table.
2. Use a rough guide – if you cannot use these methods you can at least decide that a *t* test is unsafe if one variance is more than four times the value of the other (for small *N*, i.e. 10 or fewer) or more than twice the value for larger *N*.
3. Consult more advanced texts – such as Howell (2001) – for Levene's and O'Brien's tests, calculated by hand.

If you do not think that the homogeneity of variance assumption is satisfied you can decide to use a non-parametric test, probably the Mann-Whitney, which is described later in this chapter. Alternatively you can add participants to the condition with lower *N* until numbers in each condition are equal (of course, you don't then have random allocation).

Why not wing it? – The robustness of *t* tests – the alternatives and their power efficiency

A certain amount of leeway is tolerated with these assumptions. If you are in a position where you wish to use a *t* test but have violated one of the conditions a bit, then you can draw your reader's attention to this but also hope that the significance level is so high (i.e. *p* is so low) that it is likely that you are still making the correct decision about the existence of the effect you are investigating. For this reason the *t* tests are called 'robust' (you can violate the assumptions a bit and still trust your result). However, if you really are in doubt there is simply no big problem with switching to one of the NON-PARAMETRIC or DISTRIBUTION-FREE equivalent tests, which we will move on to below (and they are a *lot* easier to calculate by hand). These tests (usually the Mann-Whitney *U* and the Wilcoxon *T* tests) will give you the same significance decision as the *t* test on the large majority of occasions. The reason they don't always do so is because they deal with less of the information in the data than do interval-level tests. The non-parametric tests reduce data to ordinal level thus losing the distance between individual positions of scores (see p. 255).

Because rank tests do not always detect significance when a *t* test would, they are sometimes described as being less POWER EFFICIENT. Power efficiency is determined by comparing one type of test with another in terms of their ability to avoid Type II errors. With rank tests, then, we are somewhat more likely to retain H_0 when it is false than with the *t* tests. However, if the research study is well designed, with appropriate and large enough samples, an effect should be detected with either test. The issue of statistical power is dealt with later in this chapter.

Exercises

1 What precautions need to be taken before carrying out a *t* test on each of the following two sets of data:

(a)	17	23		(b)	17	23	
	18	9			18	11	
	18	31			18	24	
	16	45	(unrelated data)		16	29	(related data)
	16				12	19	
	18				15	16	
	17						
	6						

2 Brushing caution aside, calculate the *t* values for the data in 1a and 1b above any way you like.

3 A report claims that a *t*-value of 2.85 is significant ($p < .01$) when the number of people in a repeated measures design was 11. Could the hypothesis tested have been non-directional?

4 At what level, if any, are the following values of *t* significant? The last two columns are for you to fill in. Don't forget to think about degrees of freedom.

	t =	*N*	Design of study	One- or two-tailed	*p* ≤	Reject null hypothesis?
a)	1.750	16	related	2		
b)	2.88	20	unrelated	2		
c)	1.70	26	unrelated	2		
d)	5.1	10	unrelated	1		
e)	2.09	16	related	2		
f)	2.76	30	related	2		

5 Two groups of children are observed for the number of times they make a generous response during one day. The researcher wishes to conduct a *t* test for differences between the two groups on their 'generosity response score'. A rough grouping of the data shows this distribution of scores:

	Number of generous responses						
	0–3	4–6	7–9	10–12	13–15	16–19	20–22
Group							
A	2	16	24	3	1	0	1
B	5	18	19	4	5	1	3

(a) Why does the researcher's colleague advise that a *t* test on the raw data might be inappropriate?

(b) What are the options for the researcher?

Answers

1 (a) Variances not at all similar, unrelated design and very different sample numbers. If using SPSS use 'unequal variances line' (see p. 382) or carry out the non-parametric equivalent.

(b) Lack of homogeneity of variance but related design. Therefore, safe to carry on with *t*.

2 (a) *t* (10) = 2.06; (b) *t* (5) = 1.57

3 No. *df* = 10. Critical value (two-tailed) at $p \leq .01$ = 3.169

4 (a) NS, keep NH
(b) .01, reject NH
(c) NS, keep NH
(d) .005, reject NH
(e) NS, keep NH
(f) .01, reject NH

5 (a) Distributions are skewed. As samples are large, the whole population may well be skewed too, and this is contrary to normal distribution assumption.

(b) Try to get rid of skew by transformation of the data or switch to a Mann-Whitney.

Non-parametric tests of difference

We have seen that there are some restrictions on the type of data that are suitable for a safe significance assessment using *t* tests. Sometimes your data just won't be suitable. You may have a scale that certainly isn't interval – it does not have equal intervals for equal amounts of the variable measured. Typical here are those invented assessment scales that ask you to 'Assess ... on a scale from 1 to 10...'. You may have severely skewed data that won't go away with a transformation. This isn't as big a problem as it might seem. You can use a non-parametric test (usually easier to calculate if doing it by hand) and still get the significant result that a *t* test would give. Non-parametric tests are estimated to be 95% power efficient as compared with *t* tests. That is, on 95 occasions out of 100 they will give you significance if the *t* test does.

The Wilcoxon (*T*) matched pairs signed ranks test

When to use the Wilcoxon

Type of relationship tested	Type of data required	Design of study
Difference between two conditions	At least ordinal	Within groups: Repeated measures Matched pairs

Data assumptions: Data at least at ordinal level.
Note: When *N* is > 20 and/or if the Wilcoxon critical values table does not include your size of *N*, please see p. 375

The Wilcoxon is one of two major tests used at the ordinal level for testing differences. It is used with related data (from a repeated measures or matched pairs design). One initial word of warning: the Wilcoxon statistic is *T* and this is very easy to confuse with the (little) '*t*' test we met in the previous section. SPSS does not help by referring to *t* as *T*! Just be aware of which test you are in fact using. There is also a rarely encountered Wilcoxon's rank sum test for unrelated samples.

Data for Wilcoxon's *T*

Suppose we ask students to rate two methods of learning which they have experienced on two different modules. Method A is a traditional lecture-based approach while method B is an active assignment-based method. We might hypothesise that students would be very likely to prefer a more active, involved approach. If you look at the data in Table 15.3 you'll see that, for each student, we know which they preferred by looking at the sign of the difference between their two ratings (column C). If the sign is positive then their rating for Method B was higher than their rating for Method A. In column D, the sizes of the differences have been ranked, ignoring the sign of the difference. This converts the differences in column C into *ordinal data*. Just three students prefer the lecture method to the assignment method, and this is shown by the fact that their differences are negative.

	Rating of traditional lecture	Rating of assignment-based method	Difference (B–A)	Rank of difference
Student (N = 15)	A	B	C	D
Griffiths	23	33	+10	12
Ashford	14	22	+8	9.5
Woodlock	35	38	+3	3
Jamalzadeh	26	30	+4	5
Manku	28	31	+3	3
Masih	19	17	—2	1
Salisbury	42	42	0	
Maman	30	25	—5	6
Quinliven	26	34	+8	9.5
Blay	31	24	—7	8
Harrison	18	21	+3	3
Ramakrishnan	25	46	+21	14
Apostolou	23	29	+6	7
Dingley	31	40	+9	11
Milloy	30	41	+11	13

Table 15.3 Student ratings of a lecture-based and an assignment-based module

If we are to convince ourselves and others that the preference for an assignment approach is real, and that we can dismiss the idea that the ratings fluctuate only randomly, we need more positive than negative differences. However, it would also be much more convincing if those negative differences (i.e. the 'unwanted' ones) were small compared with the others. In a sense, if we were arguing for the assessment-based method, we could say 'Sure, there were a couple of people who voted in the opposite direction, *but not by much*.' The way we show that the unwanted differences are not large is by looking not at the actual difference (as we did in the *t* test) but at the *ranks* of the differences. We want the negative ranks to be small. *T* is simply the smaller of the two sums of ranks – the sum for the positive differences and the sum for the negative differences. If we have a significant difference then *T* will be very small because differences that went in one of the two possible directions are also very small.

For any fixed value of N there is a fixed sum of ranks for the differences. In Table 15.3 there are 15 people. One of these does not have a difference since they rated both methods the same. For the purposes of the Wilcoxon analysis we ignore any ties like this. Hence, the 14 remaining participants must receive the ranks 1 to 14. These ranks add up to 105 (that's $1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 11 + 12 + 13 + 14$). We will have two sums of ranks, one for negative and one for positive differences. Hence, these two sums must add to 105. This will be true *whenever* there are 14 data pairs no matter what was measured on whatever scale. Ordinal data are just the ranks – the raw scores are not used in calculations. If there is no difference between ratings of the two teaching methods (for the 'population') then every time we sample 14 people's pairs of ratings we should get a T close to 52.5. T is the smaller sum of ranks but if there is no difference between the methods then most of the time the two sums of ranks will be equal and T will be half of the total sum of 105. In other words, if we obtained a sampling distribution of Ts by taking 14 pairs of scores over and over again, the mid-point of this distribution would be where $T = 52.5$.

The null hypothesis for the Wilcoxon matched pairs signed ranks tests

It is conventional to state the null hypothesis as the claim that the two populations from which scores are sampled are identical. Most of the time this is more specifically that the two medians are equal (not *means* because we are working at the ordinal level).

Testing the null hypothesis

T values of around 52 and 53 will occur most frequently, then, if we persist with the drawing of 14 random score pairs. Values for T will range either side of this but very few will be close to zero or to 105. In other words, we are sampling under H_0 and we could work out how likely it is to get any value for T. Fortunately, those who have gone before us have developed tables of critical values. These will tell us what value of T will occur less than 5% of the time if the null hypothesis is true. If our T is lower than or equal to the critical value, then the probability of our T occurring under H_0 is $\leq .05$.

By-hand calculation of Wilcoxon's (related) T

Procedure	Calculation/result of steps
1 Find the difference between each pair of scores; it makes things easier to subtract in the direction that differences were expected to go or, anyway, smaller from larger	See Table 15.3 column C
2 Rank the differences, *ignoring their sign*; see p. 253 for ranking method; omit any zero differences from the analysis*	See Table 15.3 column D Note we drop Salisbury from the analysis
3 Find the sum of ranks of positive differences and the sum of ranks of negative differences; the smaller of these is T	Since the sum of ranks for negative differences will obviously be smaller we need only add these, so: $T = 1+6+8 = 15$
4 Consult Appendix Table 5 to find the critical value required; use N, which doesn't include any zero difference scores already discarded	Relevant line is $N = 14$
5 Using two-tailed test values, for significance T must be less than or equal to the table value for α i.e. usually $p \leq .05$	Critical value of T when $p \leq .05$ is 21; our obtained T is less than this, so the difference is significant; we may reject H_0
NOTE: In the t test our value for t had to be greater than the crucial value; here T must be *lower*	Note that the obtained T also equals the critical value for $p \leq .02$ so we know that p was in fact as low as .02

Note: *Almost all writers tell you to ignore zero differences so you'll be in safe company if you do, and you can certainly ignore them when there are only two or three. However, with larger numbers, a small bias is incurred and Hays (1973) advises the following: with even numbers of zero differences, give each the average rank that all the zeros would get – they get the lowest ranks, before you move on to values of 1 and 2 so four zeros would get the ranks 1, 2, 3 and 4, and each would receive the average of these which is 2.5. Arbitrarily give half of the zero score ranks a negative sign. Do the same if there is an odd number of zeros, but randomly discard one of them first. This might make some results significant that wouldn't otherwise be. Notice, this has no effect on our calculation above because, with one zero difference, the methods are the same.

It appears that there is a real overall preference among students for the assignment-based teaching method. Effect size and power should be considered (see p. 385).

Reporting the results of a Wilcoxon matched pairs signed ranks test

One student showed no preference for either method and this result was discarded from the analysis. The remaining 14 students were rank ordered by the size of their preference for one teaching form over the other. A Wilcoxon T was used to evaluate these differences. A significant preference was shown for the assignment-based method, $T = 15$, $p \leq .05$; the total of the ranks where students were in favour of the assignment-based method was 90 and the total for the traditional method was 15.

The effect size was medium to large, $r = .43$ (see p. 393 for effect size calculation)

The Mann-Whitney U test

When to use the Mann-Whitney

Type of relationship tested	Type of data required	Design of study
Difference between two conditions	At least ordinal	Between groups: Independent samples

Data assumptions: Data at least at ordinal level.

Note: When N is > 20 and/or the Mann-Whitney critical values table does not include your size of N, please see p. 375.

Data for the Mann-Whitney U test

In order to understand how the Mann-Whitney test works have a look at the data in Table 15.4. Imagine that children's tendency to stereotype according to traditional sex roles has been observed. The children have been asked questions about several stories. The maximum score was 100, indicating extreme stereotyping. Two groups were observed, one with mothers who had full-time paid employment and one whose mothers did not work outside the home.

Stereotype scores for children whose mothers had:

Full-time jobs		No job outside home	
Score	**Points**	**Score**	**Points**
17	9	19	6
32	7	63	0
39	6.5	78	0
27	8	29	4
58	6	39	1.5
25	8	59	0
31	7	77	0
		81	0
		68	0
Totals:	**51.5** U_1		**11.5** U_2

U is the lower of 51.5 and 11.5, so U is 11.5

Table 15.4 Stereotyping scores for children with employed and unemployed mothers

It looks as though the stereotyping scores for children of employed mothers are far lower than those for the other group. It is true that there are two fewer employed than non-employed mothers. However, this doesn't matter and the statistical test will take this into account. Never worry about slight disparities between participant numbers in two conditions (though it is a good idea to plan on getting even numbers if you can). Certainly never use this as a critical point when discussing a research study unless the disparity is very large. The statistical procedures reported in this book all take into account such disparities and nevertheless calculate the value of p under H_0 in all cases.

The Mann-Whitney test, like the Wilcoxon, is based on rank order, though you will not need to do any ranking in order to perform the test. Imagine that the values in columns 1 and 3 of Table 15.4 were the scores obtained by members of team A and team B respectively, each throwing three darts at a dartboard. Because we are only working at ordinal level the information that 81 in the B team is far higher than the highest score of 58 in the A group is not used. All we use is the information that 81 is better than the highest team A score; we don't take into account *how much* better it is. What we do, in fact, is to find out, for each person in a group, how many people in the other group beat that person's score. We do this by allotting points according to the following simple system:

- each time a score X is beaten by one in the other group award a point to score X
- each time a score X equals a score in the other group award ½ a point to score X.

If you look at columns 2 and 4 of Table 15.4 this has been done. The first score in the first group is 17. This is beaten by every score in the other group so 17 is awarded 9 points. You'll see that in this (rather odd) scoring system the higher your points total the more people have beaten your score. The third score in the first group is 39. This is beaten only by the scores of 63, 78, 59, 77, 81 and 68 in the other group, so 6 points are awarded. However 39 is also equalled by the fifth score in the second group so a half point is awarded here also, giving 39 a total of 6.5 points altogether. We proceed in this way through both groups, although if it is obvious which group has the higher scores you need only award points for that group. The total of points for each group is found and the lower of these two totals is the statistic U.

There is a simple rationale to this. Suppose each person in each group has played each person in the other group just once, each throwing the three darts. There will be 7 x 9 contests altogether, giving 63. For each of these contests a point is awarded, either one to the winner or a ½ each in the case of a draw. This is precisely what we just did in awarding our points. Hence we must have awarded 63 points altogether, and you can tell this by adding the two values of U. We know, then, that:

$$N_1 N_2 = U_1 + U_2$$

and you can use this in future just to check you haven't made an error.

The null hypothesis for the Mann-Whitney U test

In general, H_0 is that the populations from which the two samples have been randomly selected are identical. In most cases it is specifically that the two population medians are equal.

Testing the null hypothesis

Remembering that points awarded here are like penalty points, if the members of team B are really brilliant at darts they'll have very few points awarded against them and team A will amass a large score. If, on the other hand, the two teams are equally matched, then each time they play it is like drawing samples under the null hypothesis. The most either team can get is 63 and the least zero. Under H_0 we would expect 31 or 32 to occur most frequently (with equal team numbers) and smaller values of U to occur relatively less frequently. For each combination of N_1 and N_2 there will be a value of U where, if our obtained value of U falls below this, the probability of the difference occurring (under H_0) is $\leq .05$. This will be our critical value, then, and our statistical train spotters have of course devised tables for us to consult (see the Appendix, Table 6).

By-hand calculation of Mann-Whitney U

Procedure	Calculation/result of steps
1 For each score in each group give a point each time it is beaten by a score in the other group and a ½ point for a tie; if one group obviously has higher scores you need only do this for that group	See columns 2 and 4 of Table 15.4
2 Add up the points for each group and find the lower of two values; this value is U	In Table 15.4 we have $U_1 = 51.5$ and $U_2 = 11.5$ (Check: $N_1 N_2 = U_1 + U_2$: 7 x 9 = 51.5 + 11.5 = 63 so we have not made an error) $U =$ **11.5**
3 Consult Table 6 for critical values with a two-tailed test and α at .05	Critical value for $N_1 = 9$ and $N_2 = 7$ is 12 11.5 is less than 12 so we have a significant result (just!) and may reject H_0.

We have support for the hypothesis that children of working mothers are less likely to use sex-role stereotypes. Effect size and power should be considered (see p. 393).

Reporting the results of a Mann-Whitney U test

The children's stereotyping scores were each allocated points when they were exceeded by or equalled each score in the other group. The lower points total was taken as a Mann-Whitney U value for $N_1 = 7$ and $N_2 = 9$. The results indicated lower stereotyping scores for the children of full-time employed mothers than for the other children. This difference was significant, $U = 11.5$, $p < .05$, with 51.5 points for the employed mother group and 11.5 for the non-employed mother group. The effect size was large, $r = .53$.

Note: Where the formula approach is used (see below), and scores are rank ordered, you would include the rank totals for each group rather than the points total.

Formula for U

Most texts ask you to rank all the scores *as one group* then apply two formulae to find U_1 and U_2. The original procedure is that just described but statisticians like to encapsulate procedures in a formula. Some argue that the points method is unwieldy for large N but my view would be that the ranking method is even more frustrating and error prone for large numbers, where many ties occur and where the student inevitably finds they have to restart at least once. Even with large samples, if I had to calculate by hand, I would always choose the points method. To calculate with formulae, first, rank all 16 scores as one group. Then use the ranks in the following formulae:

$$U_a = N_a N_b + \frac{N_a (N_a + 1)}{2} - R_a \qquad U_b = N_a N_b + \frac{N_b (N_b + 1)}{2} - R_b$$

where R_a is the sum of ranks for group A and R_b is the sum for group B. Again you select the lower value of U_a and U_b as your observed U.

Non-parametric tests and z values – Effect size and large N

Both U and T can be converted to a z value. This is particularly useful in calculating effect sizes and we will do this on p. 395.

It is also useful when N is large and the critical values only go up to a modest sample size of 20 or 25. The value of z has to be large enough to cut off less than the final 5% of the normal distribution at the predicted end (one-tailed tests) or less than 2.5% at either end (two-tailed tests). From the normal distribution table in the Appendix, Table 2 I hope you'll agree that a z score of 1.96 is the critical value for a two-tailed test and that 1.65 is the critical value for a one-tailed test, where α is .05. The relevant formulae are:

Mann-Whitney

$$z = \frac{U - \frac{N_a N_b}{2}}{\sqrt{\left(\frac{N_a N_b}{N(N-1)}\right)\left(\frac{N^3 - N}{12} - \sum \frac{t^3 - t}{12}\right)}} \quad \text{where } N \text{ is } N_a + N_b$$

t accounts for tied scores. Each time you find a tied value in your data set you count up how many times the value occurs and this value is t. Remember though that you have to do this for each value that is tied and add up the results. For instance, for the data in Table 15.2 the score 11 appears five times so $t = 5$ and you then put this into the $\frac{t^3 - t}{12}$ formula and record your result. Then you do the same again for 9, which also occurs five times, 12 which appears twice, 13 which appears three times and 8 which appears twice. Finally you add the results of these five calculations. If there are no ties then $\frac{t^3 - t}{12}$ is just ignored.

Wilcoxon signed ranks T

$$z = \frac{N(N+1) - 4T}{\sqrt{\frac{2N(N+1)(2N+1)}{3}}} \quad \text{where the } T \text{ is the observed Wilcoxon's } T$$

The (binomial) sign test for related data (S)

When to use the binomial sign test

Type of relationship tested	Type of data required	Design of study
Difference between two conditions	In categorical form – may be reduced from interval or ordinal level	Within groups: Repeated measures Matched pairs

Data assumptions: Measures of the dependent variable have two equally likely values under H_0, e.g. negative or positive, correct or incorrect and so on.

Data for the sign test

The sign test works on a very simple kind of categorical data. When we have interval-like data on each participant taken under two related conditions we may feel that the *difference* between the two values cannot be taken as a meaningful interval measure. For instance, if you rate two modules on a scale of 1 to 10, giving one 8 and the other 4, we cannot claim that the difference of 4 is an interval measure. However, what we can say pretty confidently is that you preferred the first module to the second. We can take as data the *sign* of the difference. Often, all we *have* is one of two possible outcomes.

Suppose that, in order to assess the effectiveness of therapy, a psychotherapist investigates whether or not, after three months of involvement, clients feel better about themselves or worse. If therapy improves people's evaluation of themselves then we would expect clients' self-image ratings to be higher after three months' therapy than they were before.

Take a look at the data in Table 15.5 showing clients' self-image ratings before and after three months' therapy on a scale of 1–20, where a high value signifies a positive self-image. Here we would expect the scores to be higher in column C than they are in column B as we do in the related *t* and Wilcoxon tests. Therefore we would expect positive differences in column D. Unlike the *t* and Wilcoxon tests, here we ignore the *size* or *rank* of each difference, and simply put the *sign* (or direction) of each difference into column E. If the therapy is working, we would hope to obtain a large number of positive signs and a small number of negative signs, if any. The SIGN TEST gives us the probability of finding this number of negative signs (or fewer), given that the null hypothesis is true. That is, it tells us how likely it is that such a large (or even larger) split between positive and negative signs would be drawn 'by chance' under the null hypothesis where even splits are expected. This is just what we looked at with the glove drawer problem in Chapter 14.

DATA

A	B	C	D	E
Client	**Self-image rating before therapy**	**Self-image rating after 3 months' therapy**	**Difference (C–B)**	**Sign of difference**
a	3	7	4	+
b	12	18	6	+
c	9	5	−4	−
d	7	7	0	
e	8	12	4	+
f	1	5	4	+
g	15	16	1	+
h	10	12	2	+
i	11	15	4	+
j	10	17	7	+

Table 15.5 Self-image scores before and after three months' therapy

The null hypothesis in the binomial sign test

We assume that there are equal numbers of positive and negative signs in the 'population' we have sampled from – and we assume we have sampled from that population at random. This is exactly the position we were in with the baby-sexing result on p. 331 and, in effect, we went through the details of a sign test there. Here, we simply present the 'cookbook' method of conducting a sign test on a set of this kind of paired data.

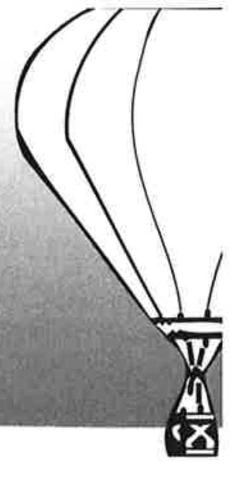
Procedure	Calculation on our data
1 Calculate the difference between columns B and C, always subtracting in the same direction. If a directional prediction has been made, it makes sense to take the expected smaller score from the expected larger one. Enter difference in column D.	Find difference between scores in columns B and C of Table 15.5. We expect column C scores to be higher. Hence we take C–B in each case.
2 Enter the sign of the difference in column E. Leave a blank where the difference is zero and ignore these in the analysis.	See column E of Table 15.5. *N* becomes 9 because the difference for client d is zero. This case is dropped from any further analysis.
3 Count the number of times the less frequent sign occurs. Call this *S*.	Negative signs occur less frequently, so $S = 1$
4 Consult Table 7 in the Appendix. a) Find the line for *N* (the total number of signs not including zeros). b) Consult one- or two-tailed values.	 a) *N*=9 (see step 2, above). b) We would be interested if the therapy made people worse so stick with two-tailed test.
5 Compare *S* with the critical value for the significance level set. For significance, *S* must be equal to or less than the critical value.	Our *S* is 1. The critical value under the column headed $p \leq .05$ (two-tailed) is 1. Therefore, our result is not greater than the appropriate critical value and meets the criteria for significance.
6 Make statement of significance	Our result is significant with $p \leq .05$. We may reject the null hypothesis.

Reporting the result of a sign test

For each client an improvement in self-image score after three months' therapy was recorded as a positive, whereas a deterioration was recorded as a negative. One client's self-image score did not change and this result was omitted from the analysis. The remaining nine results were submitted to a binomial sign test and the rate of improvement over deterioration was found to be significant, $S = 1$, $p \leq .05$.

Glossary

Binomial sign test (S)	Nominal-level test for difference between two sets of paired/related data using *direction* of each difference only
d, Cohen's	Measure of effect size; used here in calculating *power*
Delta (δ)	Statistic used to estimate power using the effect size
Difference mean	Mean of differences between pairs of scores in a related design
Distribution dependent test	Significance test making estimations of population parameters
Distribution free test	Significance test that does not depend on estimated parameters of an underlying distribution
Eta-squared $\eta 2$	Measure of effect size
Homogeneity of variance	Situation where sample variances are similar
Mann-Whitney U test	Ordinal-level significance test for differences between two sets of unrelated data
Non-parametric test	Significance test that does not make estimations of parameters of an underlying distribution; also known as a *distribution free test*
Parametric test	Relatively powerful significance test that makes estimations of population parameters; the data tested must usually therefore satisfy certain assumptions; also known as a *distribution dependent test*
Pooled variance	Combining of two sample variances into an average in order to estimate population variance
Power efficiency	Comparison of the *power* of two different tests of significance.
Related t test	Parametric difference test for related data at interval level or above.
Robustness	Tendency of test to give satisfactory probability estimates even when data assumptions are violated.
S	See binomial sign test
Sign test	See binomial sign test
t	See related and unrelated t test
T	See *Wilcoxon test*
Ties (tied ranks)	Feature of data when scores are given identical rank values.
U	See *Mann-Whitney test*
Unrelated t test	Parametric difference test for unrelated data at interval level or above
Wilcoxon's T – matched pairs signed ranks test	Ordinal-level significance test for differences between two related sets of data



Tests for categorical variables and frequency tables

16

The *chi-square* (χ^2) test presented in this chapter is concerned entirely with categorical variables; those producing nominal data, that is, frequencies by categories.

Chi-square is first used to analyse a simple division of one variable into two levels of frequencies.

- The concept of *expected frequencies* under the null hypothesis is introduced.
- *Cross-tabs tables* are then introduced and chi-square used to analyse for *association* between two categorical variables with two levels each (a 2 x 2 analysis).
- The generalised form of chi-square testing r x c tables (those with any number of rows and columns) is then covered, e.g. three types of training by pass/fail.
- Chi-square can also be used as a *goodness of fit* test to check whether a distribution of frequencies in categories is a close fit to a theoretical distribution (e.g. whether a college's pattern of degree classifications match the average pattern for the country).
- There are *limitations* on the use of chi-square: data must be *frequencies*, not ratios, means or proportions, and must belong exclusively to one or another category, i.e. the same case (person) must not appear in more than one 'cell' of the data table.
- There is statistical debate about *low expected cell frequencies*. It is advisable to avoid these where possible. If low expected frequencies *do* occur, sample sizes above 20 make the risk of a Type I error acceptably low.
- *Power* and *effect size* calculations for chi-square are given. SPSS procedures for chi-square analyses are described.

The chapter then moves on to the analysis of *multi-way* (i.e. not just two-way) tables using *log-linear analysis*. The *likelihood ratio chi-square* is used to investigate higher-order interactions for significance, then proceeds hierarchically downwards from the initially saturated model, to one-way effects. SPSS analysis is described.

Tests on two-way frequency tables

Very often the design of our research study entails that we gather data that are *categorical* in nature. Have a look at the data in Table 16.1 and Table 16.2 which are frequency tables for people assessed on two categorical variables. Such tables are called CROSS-TABULATION (or CROSS-TABS) TABLES. In the first of these, the (fictitious) data have been gathered by observing whether a car is new or old, and whether its driver does or does not obey the amber signal at a pedestrian crossing. The hypothesis is that drivers of newer cars conform more often to the traffic regulation of stopping on amber. The rationale might be that drivers of newer cars are likely to be older and more experienced.

The numbers in the cells of this kind of table are *frequencies* – they are just a *count* of the number of cases (people in this case) observed in each cell of the table. Notice from the table that, of those driving a new car, far more stopped at amber than drove on, whereas for the drivers of older cars, the frequencies of stopping and not stopping were almost equal. Our statistical test will tell us whether this difference in stopping proportions between the two sets of drivers can be considered significant or not. We will see whether stopping is *associated* with age of car.

We *could* of course have gathered *measured* (not categorical) data by measuring speed or obtaining the exact year of manufacture of a car and by stopping drivers and interviewing them about conformity or giving them a questionnaire. This might be possible in a shopping area car park but it would be time-consuming and our drivers would be susceptible to several of the forms of bias involved when participants know they are being studied. The observation method has the great advantage of gathering data on naturally occurring behaviour but often has the disadvantage, in this case, of having the independent variable (age of car) and the dependent variable (stopped or not) *both* assessed only at a categorical level.

	Age category of car		
	New	Old	Total
Behaviour at amber light			
Stopped	90 (a)	88 (b)	178
Did not stop	56 (c)	89 (d)	145
Total	146	177	323

Table 16.1 Frequencies of drivers by age of car and whether they stopped at an amber light or not

The data in Table 16.2 are the actual results of the study by Cialdini *et al.* (1990) mentioned in Chapter 3 where people were observed on a path after they had been handed a leaflet. The researchers varied the number of pieces of litter already present and observed whether each person dropped their leaflet or not. Here the independent variable is *not* originally categorical (it had the measured values 0, 1, 2, 4, 8, 16), but since the dependent variable had to be categorical (they either dropped their leaflet or they didn't) it was simplest to treat both variables as categorical by reducing the independent variable values to three categories (0/1, 2/4, 8/16) as shown in the table.

	Amount of existing litter		
Observed person	**0 or 1 piece**	**2 or 4 pieces**	**8 or 16 pieces**
Dropped litter	17	28	49
Didn't drop litter	102	91	71

Table 16.2 Number of pieces of existing litter and consequent littering (Cialdini *et al.*, 1990)

Cialdini, R.B., Reno, R.R. and Kallgren, C.A. (1990) A focus theory of normative conduct: Recycling the concept of norms to reduce litter in public places. *Journal of Personality and Social Psychology, 58,* 1015–26. Copyright © 1990 American Psychological Association.

A measured dependent variable in any study can always be reduced to categorical level in a similar way where this is useful. We may, for instance, split a group of extroversion scores at the median value and refer to those above the median as 'extroverts' and those below as 'introverts'. We may reduce smoking information down to the categories: 'non-smoker', '1 to 5 a day', '6–20 a day' and 'over 20 a day'. The codes 1 to 4 given to these four categories could at a stretch be treated as ordinal-level data but with the practical problem that too many people would be tied at each rank. Instead, we can treat the codes as category names and simply count the number of people in each category.

Unrelated data – the Chi-square test of association

When to use chi-square or χ^2

Type of relationship tested	Level of data required	Design of study
Association between two variables	Nominal/categorical	Between groups: Independent samples

Data assumptions: Each observed person (or case) must appear in *one only* of the frequency cells. It must be impossible for them to appear in more than one cell.

Notes: No more than 20% of the expected frequency cell counts should be less than 5.

'chi' is pronounced 'kye' in English. It is an approximation to the name for the Greek letter χ which starts with 'ch' as in the Scottish pronunciation of 'loch' and is the symbol for the statistic in this test.

Chi-square is the test to use when we are looking for an *association*, or a difference in proportions, as in the examples above, and where the variables concerned are both categorical. The design will be *between groups*. To move towards the thinking behind chi-square I would like to start with one of those situations I like to use that demonstrate the value of statistical competence in protecting us against the outlandish claims of some advertisers. Have a look at Box 16.1.

The marketing survey results described in Box 16.1 might seem, at first sight, very impressive (one colleague I spoke to about this said, 'Never mind the stats, Hugh, where do I get hold of the stuff?'!). We learn that of 550 women provided with a free sample and using it for one month, 56% reported a loss of up to one inch from their thighs, and 52% reported the same for their hips. However, let's think what we would expect if the null hypothesis were true. H_0 would be based on the concept of the women choosing one answer or the other ('gained' or 'lost') entirely at random. On this basis, then, we would expect, from 550 choices, 275 'gains' and 275 'losses'. In chi-square terminology, these frequencies predicted under the null hypothesis are known as EXPECTED FREQUENCIES – they are what we typically expect to occur with our overall frequencies *if H_0 is true.* The frequencies we *actually* obtain from our study are referred to as OBSERVED FREQUENCIES.

Info Box 16.1 Does the magic gel really work?

Some years ago Christian Dior ran an advert in a colour supplement claiming that, of 550 women asked to use a fat-reducing gel (Svelte) for one month, 52%, in a later survey, claimed they had lost 'up to one inch' off their hips during that period, while 56% had lost the same amount off their thighs. Now this might sound very impressive indeed, except that we do not know what questions were asked in the survey. This is a perfect example of the need to know what question was asked before being able to interpret fully an apparently strong piece of evidence. It is unlikely that the women were simply asked to give open-ended responses. It is very likely indeed that they were asked to respond to multiple-choice items, such as 'Over the last month did you:

(a) lose up to one inch off your hips

(b) gain up to one inch on your hips

(c) notice no change at all on your hips?'

For simplicity's sake let's ignore the last alternative since there would always be some, perhaps very tiny, change over one month. In fact, the Dior marketing people might have only asked each woman to measure their hips at the start and at the end of the one-month trial and to take the difference. There will always be a small difference between two measures of the same thing (random error) so each woman could then have recorded either 'increased' or 'decreased'.

Here, then, let's imagine we have 52% of the sample of 550 saying 'lost' and 48% saying 'gained' in reference to their hips. That's 286 positive and 264 negative outcomes from the Dior marketing perspective. Questions for you to ponder are:

1 How many of the 550 women would respond 'positive' and how many 'negative' if they were simply tossing a coin (i.e. selecting an alternative at random)?

2 On the basis of your answer to the question above, are the 286 vs 264 results impressive (i.e. will we consider them to be a *significant* difference?) or are they within the range we might reasonably expect 'by chance' if the women are selecting their response at random?

Taking the slightly more impressive 56% losing up to one inch from their thighs, you might ponder the same questions.

Calculating expected frequencies in a one-row chi-square analysis

To make things formal (and for more complex examples) we calculate the expected frequencies for a single-row analysis using *N/k* where *N* is the total number of cases (550 in this case) and *k* is the number of cells to average across. Hence, here, 550/2 = 275.

	Women reporting a loss of up to 1″	**Women reporting a gain of up to 1″**	**Total**
Observed frequencies (obs)	(a) 286	(b) 264	550
Expected frequencies (exp)	(a) 275	(b) 275	550

Table 16.3 Observed and expected frequencies for women reporting losses or gains after using gel for one month

Data for a one-row chi-square analysis

The data for our first simple chi-square test on the hip data, then, would appear as in Table 16.3.

The null hypothesis for chi-square

The null hypothesis in a chi-square analysis is always that the population is distributed in the pattern of proportions shown by the expected frequencies. Our alternative hypothesis (or rather Christian Dior's) is that more people (in the population) report a loss than report a gain after one month's use. Referring to Table 16.3, we need therefore to see whether our *observed frequencies* of 286 ('loss') and 264 ('gain') differ significantly from the *expected frequencies* of 275 and 275, which would occur under H_0.

Testing the null hypothesis

The chi-square statistic gets larger as the observed cell frequencies depart from what is expected under H_0 – that is, from the expected frequencies. We can see in Table 16.3 that we would be more convinced of the effectiveness of Svelte gel the further *cell* a_{obs} rises above *cell* a_{exp}. We calculate chi-square using:

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

To calculate, we take each set of cells in turn (in Table 16.3, *cell a* then *cell b*) and perform the calculation shown after the Σ symbol above. As in the past, the Σ symbol means 'add up the results of each of what follows'.

Calculation of chi square using the data in Table 16.3

	$O - E$	$(O - E)^2$	$\Sigma\,(O - E)^2/E$	**Result**
Cell *a*	286-275 = **11**	11^2 = **121**	121/275 =	**0.44**
Cell *b*	264-275 = **–11**	-11^2 = **121**	121/275 =	**0.44**
			$\chi^2 = \Sigma\,(O - E)^2/E$ =	**0.88**

In this calculation we find that χ^2 is 0.88. We need to check this value for significance. χ^2 uses degrees of freedom. For a one-row analysis *df* are *k*–1 where *k* is the number of cells, so here *df* are 2–1 = 1.

Consulting Appendix Table 8 we find that we require a χ^2 value of at least 3.84 for $p \leq .05$ with a two-tailed test. Hence our difference is not significant. The conclusion here would be that use of Svelte gel has not resulted in a significant proportion of women reporting a loss of up to one inch from around their hips.

What about the result for thighs?

You might think we cheated a little there by dealing only with the less impressive hip data. OK. Let's look at the thigh data, then.

Pause for thought

Calculate the chi-square value for the thigh data following the example above for hips. Dior said that 56% of the sample reported a loss of up to one inch from their thighs; 56% of 550 is 308 so we have *observed frequencies* of 308 lost up to one inch and 242 gained up to one inch, by our definitions. Expected frequencies are again 275 in each cell.

You should find that χ^2 is 7.92. This value is well above the required critical value of 3.84 so we certainly have a significant result here. This appears to *support* the effect of Svelte gel.

The 2 x 2 chi-square

A 2 x 2 arrangement is the simplest of cross-tabs tables and is really what we need for a fair scientific test of the gel data. Of course I wasn't going to accept that the gel worked on thighs. What any scientist worth their salt would have immediately asked on hearing that results is 'Well where was the control group?'. We can't really assume that of 550 women, using the gel for a month, just half would report a loss and half a gain. This is what we might assume if we could know no better – if we had no chance of determining what would happen under a free choice. But we *can* find out what *would* happen. What we need is a *control group* with whom to compare our 'experimental' group, the one whose results Dior reported. On one occasion I did informally ask all the women in a lecture audience to answer the thighs-larger-or-smaller-after-a-month question 'cold' (with no prior information about the gel advert, but with assurances that the purpose was statistical demonstration); 53% reported a loss and 47% reported a gain, when forced to choose between these two alternatives, even though they had not used any gel. Let's just suppose that these same percentages would be found in a formal and well-designed study using a control group of 550 women, equal in number to the Dior survey group. If 53% chose loss and 47% chose gain, then we would obtain the (rounded) figures shown in Table 16.4.

	Participant reports:		
	Lost up to one inch	**Gained up to one inch**	**Total**
Gel use group	(a) 308	(c) 242	550
Control group	(b) 292	(d) 258	550
Total	600	500	1100

Table 16.4 Fictitious observed frequencies of gel-using and control group participants reporting loss or gain of up to one inch from thighs

What we have in Table 16.4 is a classic form of data table for which we would calculate a 2 x 2 CHI-SQUARE in order to discover whether there is an association between using gel and losing fat ('2 x 2' because there are two columns and two rows). Note that we are assuming that the independent variable (gel use) is having a causal effect on a dependent variable (loss or gain of fat). Note also that these two variables are both at a categorical level because they are not measured on any sort of scale and each has just two qualitatively separate levels. It is not necessary, however, for there to be an experimental independent variable and dependent variable. We could be interested, for instance, in whether introverts are more likely to feel awkward on a nudist beach than extroverts (see Table 16.5). Introversion need not *cause* introverts to feel awkward; awkwardness may be related to or simply a part of the overall introverted personality characteristic.

	Extrovert	Introvert	Total
Would feel comfortable	(a) 40	(b) 10	50
Would not feel comfortable	(c) 10	(d) 40	50
Total	50	50	100

Table 16.5 Observed frequencies of introverts and extroverts who report that they would or would not feel comfortable on a nudist beach

Expected frequencies for the new gel data

The null hypothesis for the new (fictitious) gel study is based on the assumption that there is absolutely no association, in the population as a whole, between using gel and changes in fat. More technically, it assumes that frequencies in the population are arranged as are the frequencies in the 'total' columns in Table 16.4; that is, we assume that frequencies of people reporting a loss of up to one inch would be equally split between those using the gel and those not using the gel. In other words, whether you use gel or not, you have the same chance of appearing in the 'lost' column.

Pause for thought

In our fictitious study, 550 women used gel and 550 women did not. *If there is no association between using gel and losing fat*, how many of the 600 shown in Table 16.4 who 'lost up to one inch' would you expect to find in the 'gel use' row, that is, in *cell a*?

I hope you decided that just half the fat losers (i.e. 300) should be gel users and half should be in the control group. There were equal numbers of users and non-users and, if gel use has nothing to do with fat loss, then about half those who lose weight would be from each group. The expected frequencies are shown in Table 16.6.

	Participant reports: Lost up to one inch	Gained up to one inch	Total
Gel use group	(a) 300	(c) 250	550
Control group	(b) 300	(d) 250	550
Total	600	500	1100

Table 16.6 Fictitious expected frequencies of gel-using and control group participants reporting loss or gain of up to one inch from thighs

In the frequency tables above, the cells under the title 'Total' are known as MARGINALS; that is, they are the margins of all the rows, showing how many in each row altogether, and the margins of all columns showing how many altogether in each column. All expected frequencies are calculated based on the reasoning for the gel table above. We assume that the total for each column will be divided according to the proportions of the row marginals; 600 will be divided in the ratio 550 to 550. The formula for calculating expected cell frequencies is:

$$E = \frac{RC}{T}$$

where R is the total of the *row* in which the cell is situated, and C is the total of its appropriate *column*. T is the overall total (1100 in the gel table example). However, you already did in fact use a version of this in your head in deciding that 550/1100 (R/T) of the 600 fat losers (C) would be expected in *cell a*. The general formula is used because, in most cases, the numbers are not quite as simple as the ones I've partly invented here. It is important to remember that 'expected frequencies' are those 'expected' under the null hypothesis, not those (in fact the opposite of those) that the researcher usually expects (or would like) to occur in the research study.

Let me try to outline a visual example of what a 2 x 2 chi-square does (roughly speaking) by referring to the extrovert/introvert nudist data in Table 16.5. 50% of all participants reported feeling comfortable on a nudist beach. Hence, half the introverts and half the extroverts *should*, in turn, report feeling comfortable, *if* there is no link between feeling comfortable and extroversion. The expected frequencies for this example, then, are 25 in each cell, as shown in Table 16.7. For a significant result, indicating an association between extroversion and feeling comfortable, we would want the observed frequencies in cells *a* and *d* to be much higher than 25 and for the frequencies in cells *b* and *c* to be much lower.

	Extrovert	Introvert	Total
Would feel comfortable	(a) 25	(b) 25	50
Would not feel comfortable	(c) 25	(d) 25	50
Total	50	50	100

Table 16.7 Expected frequencies of introverts and extroverts who report that they would or would not feel comfortable on a nudist beach

In Figure 16.1 we have an imaginary box with four compartments into which we 'drop' the observations in a random manner. Imagine each one of the 100 observations is a little ball dropped on to the centre spot and bouncing randomly into one of the four equal-sized compartments. There is a limitation to the randomness here – when any row or column adds up to 50 we stop permitting balls into that row or column. If we dropped the 100 balls many many times then, roughly speaking, the results would vary around those in Table 16.7, mostly by only a little but sometimes (less frequently) by quite a lot. If we calculate chi-square for every drop of 100 balls, then through this random process we will create a distribution of chi-square values. For a significant result, what we are interested in is obtaining a chi-square value that is in the top 5% of this distribution of randomly produced values – that is, we want a chi-square value that would occur less than 5 times in 100 if the null hypothesis were true (if the balls were bouncing randomly).

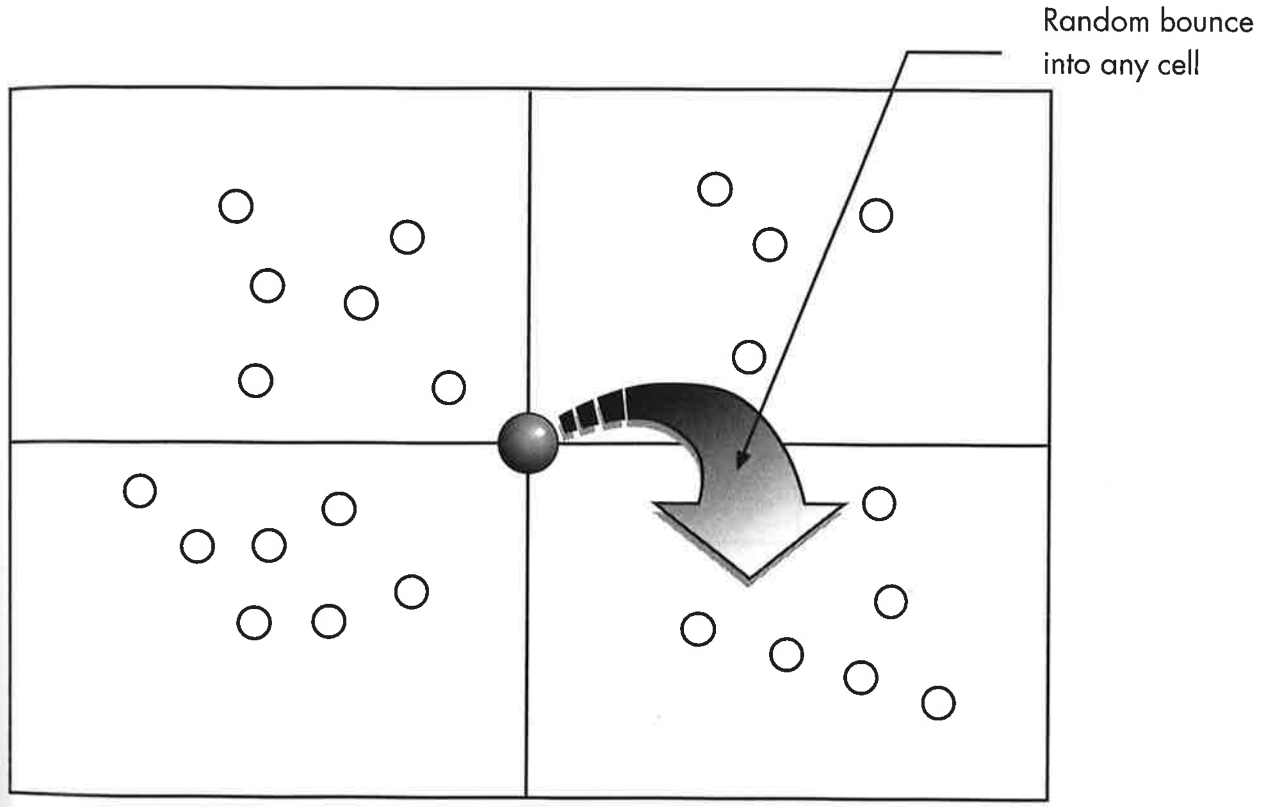


Figure 16.1 Chi-square assumes random bouncing into cells up to the row and column totals

Data for regular 2 x 2 chi-square test

Now let's calculate a 2 x 2 chi-square on the data in Table 16.1. Assume these were data gathered in a psychology practical workshop and the proposal is that drivers of new cars are more law abiding, as explained at the start of this chapter. Here, then, if car age is not related to stopping at the amber light (the null hypothesis), the expected frequencies would be that just 80 of the 146 new car drivers should stop (calculated below) whereas, in fact, 90 did so. Only 50% of the drivers of old cars stopped, compared to the greater proportion of new car drivers. The observed frequencies vary quite far from the expected frequencies, so chi-square will be high and perhaps significant.

Calculation of a 2 x 2 chi-square

Procedure	Calculation/result of steps
1 Give each corresponding observed and expected cell a letter	See letters a, b, c, d in Table 16.1
2 Calculate expected frequencies using $E = \frac{RC}{T}$ (see p. 404)	Cell a: 146 x 178/323 = **80.46** Cell b: 177 x 178/323 = **97.54** Cell c: 146 x 145/323 = **65.54** Cell d: 177 x 145/323 = **79.46**
3 Calculate χ^2 according to the χ^2 formula given on p. 401	See the calculation table below:

	$O-E$	$(O-E)^2$	$(O-E)^2/E$	Result
Cell a	90-80.46 = **9.54**	9.54^2 = **91.01**	91.01/80.46 =	**1.13**
Cell b	88-97.54 = **–9.54**	-9.54^2 = **91.01**	91.01/97.54 =	**0.93**
Cell c	56-65.54 = **–9.54**	-9.54^2 = **91.01**	91.01/65.54 =	**1.39**
Cell d	89-79.46 = **9.54**	9.54^2 = **91.01**	91.01/79.46 =	**1.15**
			$\chi^2 = \Sigma(O-E)^2/E$ =	**4.6**

4 Calculate degrees of freedom (df) according to the formula: $df = (R-1)(C-1)$	df = (2–1)(2-1) = 1

Note that, for all χ^2 tests, df are the number of cells you would need to know, *given you already know the marginal values*, in order to calculate all the rest of the cell values; in a 2 x 2 table, once we know one cell we can calculate all the rest if we already know the column and row totals. Hence df = 1

5 Using Table 8 in the Appendix, check that χ^2 reaches the appropriate critical value for df and alpha (usually set at .05) and decide upon significance.	For df = 1 and $p \leq$ .05 and a two-tailed test, χ^2 must be greater than or equal to 3.84, hence, our χ^2 is significant and we may reject the null hypothesis.

Interpreting and reporting the result

Around half the old cars didn't stop, whereas only around a third of new cars failed to stop. Our observed frequencies differ significantly from what we'd expect if H_0 is true where the proportion stopping and not stopping would be the same for new and old cars. We have therefore provided evidence that drivers of new cars are more law-abiding at traffic lights than drivers of old cars.

Effect size

For a general introduction to the importance of estimating effect size and checking power please see p. 385. Effect size for 2 x 2 chi-square analyses can be estimated using the PHI COEFFICIENT (which we will meet again in Chapter 17 for different but related reasons). Phi is pronounced as in the English word 'fie'. The more general term is CRAMER'S PHI (also called Cramer's V) when analysing cross-tab tables where at least one variable has more than two levels or categories. The formula we require is:

$$\text{Cramer's } \Phi = \sqrt{\frac{\chi^2}{(N)\ df_{smaller}}}$$

$df_{smaller}$ means either (rows -1) or (columns -1), whichever is smaller.

For our 2 x 2 case, where (rows -1) *or* (columns -1) = 1, the expression reduces to = $\sqrt{\frac{\chi^2}{N}}$ which is the phi coefficient. Our Φ will be $\sqrt{4.6/323} = \sqrt{0.014} = 0.118$.

Cohen (1988) produced some effect size conventions for Cramer's Φ that depend on the df for the smaller side of the contingency table, i.e. $df_{smaller}$ (see Table 16.8).

		Effect size	
$df_{smaller}$	Small	Medium	Large
1	.10	.30	.50
2	.07	.21	.35
3	.06	.17	.29

Table 16.8 Cohen's effect size definitions for Cramer's Φ

Our effect size of 0.118 would therefore be designated as 'small' since our $df_{smaller}$ is 1.

Reporting the result of a chi-square analysis

> Some 50.3% of drivers in old cars (89/177) failed to stop at an amber traffic light, whereas only 38.4% of drivers in new cars (56/146) failed to stop. A χ^2 analysis of the difference between stop/didn't stop frequencies across drivers of new and old cars was significant, χ^2 (1, N = 323) = 4.6, p < .05. The effect size was small with phi = 0.118.

Quick 2 x 2 formula

This can be used only where there are two columns and two rows, as in the example above. It saves the labour of calculating expected frequencies and, if you're handy with a calculator, you'll find this can be done in one move from the observed cell totals:

$$\chi^2 = \frac{N\,(ad-bc)^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)}$$

where N is the total sample size.

Exercises

A researcher suggested that graduate extroverts/introverts should not show the divisions previously shown by *non*-graduates about feeling comfortable or not on a nudist beach. The (fictitious) data below were gathered. Analyse the data using hierarchical log-linear analysis and check whether the hypothesis is supported.

	Graduates		Non-graduates	
	Comfortable	Not comfortable	Comfortable	Not comfortable
Extrovert	39	11	40	10
Introvert	24	26	10	40

Answers

The three-way interaction is significant; χ^2 (1) = 4.63, p = .031. Hence the final model had the generating class extrovert/introvert × graduation status × comfort level. Two-way interactions were also significant; χ^2 (3) = 47.234, p = .001. One-way (main) effects were not significant. The researcher appears vindicated in the proposition that graduates do not show the extremes that non-graduates show but the effect comes almost entirely from graduate introverts being less likely to feel uncomfortable than non-graduate introverts.

Glossary

Chi-square (χ^2)	Statistic used in tests of association between two unrelated categorical variables. Also used in goodness of fit test, log-linear analysis and several other tests
Chi-square change	Change in chi-square as items are removed from the saturated model in log-linear analysis
Cramer's phi or V	General statistic used to estimate effect size in chi-square analyses
Cross-tabs table	Term for table of frequencies on levels of a variable by levels of a second variable
Expected frequencies	Frequencies expected in table if no association exists between variables – i.e. if null hypothesis is true
Goodness of fit	Test of whether a distribution of frequencies differs significantly from a theoretical pattern
Hierarchical loglinear analysis	Removing items from a saturated log-linear model moving downwards towards one-way effects
Likelihood ratio chi-square	Type of chi-square statistic used in log-linear analysis
Log-linear analysis	Analysis similar to chi-square but which will deal with three-way tables or greater
Log-linear model	A theoretical and statistical structure proposed to explain cell frequency variation in a multi-way frequency table
Marginals	The total of columns and rows, and the overall total of frequencies, in a cross-tabs table
Observed frequencies	Frequencies obtained in a research study using categorical variables
Phi coefficient (Φ)	Statistic used for effect size estimate in a 2 x 2 table after χ^2 analysis
Saturated model	Model in log-linear analysis that explains all variation in a multi-way frequency table so that chi-square is zero and expected frequencies are the same as observed frequencies