Significance tests

An example: the sign test

In the evaluation of a course on evidence based health care for nurses, before and after the course the nurses were asked to complete a multiple choice test. This produces a knowledge score between –18 and +18. The knowledge scores and the change following the course are shown in Table 1. Most nurses' knowledge score was higher after the course than before it, though not all were. Is there enough evidence for us to conclude that overall the knowledge of nurses in this population increases following the course?

These 10 nurses are a sample from the population of all nurses who might attend the course. More specifically, they are a sample of the population of nurses with whom they have a common background, e.g. are working in a similar professional environment, health system, etc. Would the other members of this population increase their knowledge after attending the course? Is there good evidence that the knowledge score increases following the course?

When we look at Table 1, what might convince us that knowledge increases is that most of the difference are in the same direction. Only one nurse out of ten had a negative increase in score (i.e. a reduction). But would we be convinced if two out of ten differences were negative? Three out of ten would certainly make us cautious. After all, we might expect five out of ten to be negative if the course had no effect whatsoever. So how many negatives would we allow and still feel able to conclude that there was evidence that knowledge increased following the course?

A significance test is a method for answering this question. We ask: if there were really no difference in the population which our nurses represent, would we be likely to have observed the data we have?

To carry out the test of significance we suppose that, in the population, there would be no difference between knowledge before and after the course. The hypothesis of 'no difference' or 'no effect' in the population is called the **null hypothesis**. We compare this with the alternative hypothesis of a difference in knowledge measured before and after the course. We find how likely it is that we could get data as extreme as those observed if the null hypothesis were true. We do this by asking: if we were to repeat this course over and over again, what proportion of repetitions would give us something as far or further from what we would expect as are the data we have actually observed? We call this proportion of studies which might show such extreme data among the endless repetitions the **probability** of obtaining such extreme data.

If this probability is large the data are consistent with the null hypothesis; if it is small the data are unlikely to have arisen if the null hypothesis were true and the evidence is in favour of the alternative hypothesis.

There are many ways in which we could do this, but for this illustration we shall use a very simple significance test, the **sign test**. This uses the direction of the difference only. In our sample, we have one negative and nine positives.

Table 1. Knowledge scores (–18 to +18) of 10 nurses attending a course on evidence-based health care

Pre-course score	Post-course score	Increase in score	Direction of change
3	8	5	+
6	8	2	+
4	8	4	+
0	4	4	+
−1	1	2	+
1	7	6	+
1	6	5	+
−3	0	3	+
3	0	−3	−
2	4	2	+

Table 2. Probabilities for the number of heads in ten tosses of a coin, or for the number of negative differences out of ten if positive and negative differences are equally likely (Binomial distribution, $n = 10$, $p = 0.5$).

Number of heads or negative differences	Probability
0	0.0009766
1	0.0097656
2	0.0439453
3	0.1171875
4	0.2050781
5	0.2460938
6	0.2050781
7	0.1171875
8	0.0439453
9	0.0097656
10	0.0009766

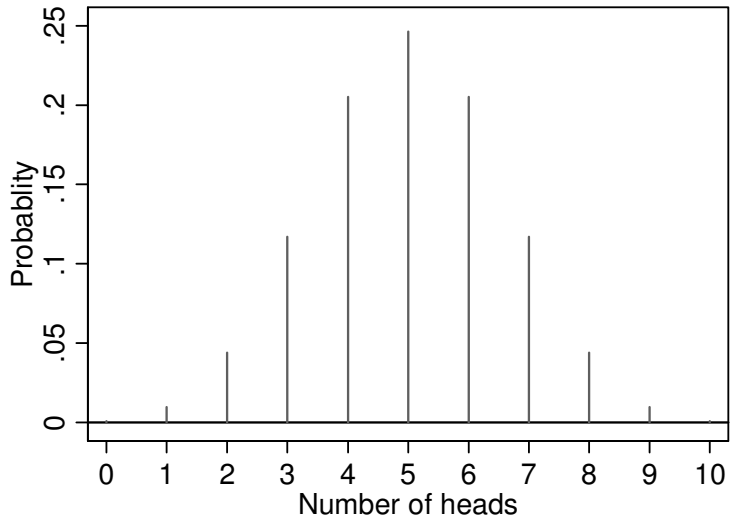

Figure 1. Distribution of the number of heads in ten tosses of a coin, or for the number of negative differences out of ten if positive and negative differences are equally likely (Binomial distribution, $n = 10$, $p = 0.5$)

Consider the differences between the knowledge score before and after the course for each nurse. If the null hypothesis were true, then differences in knowledge score would be just as likely to be positive as negative; they would be random. The probability of a difference being negative would be equal to the probability of it becoming positive. If we ignore for the moment those whose knowledge stays the same, the proportion of nurses whose difference is negative should be equal to the proportion whose difference is positive and so would be one half, 0.5.

Then the number of negatives would behave in exactly the same way as the number of heads which would show if we were to toss a coin 10 times. This is quite easy to investigate mathematically. Table 2 shows a list of the probabilities for zero, one, two, three, . . ., and ten heads showing in ten tosses of a coin. Figure 1 shows a graphical representation of the probabilities in Table 2. Because the only possible values are the whole numbers 0, 1, 2, etc., the probabilities are shown by vertical lines at those points.

The technical name for this distribution is the Binomial distribution with parameters $n = 10$ and $p = 0.5$. The Binomial distribution is a family of distributions, like the Normal (week 2). It has two parameters, the number of coins, n, and the probability that a coin would produce a head, p. It need not be the flip of a coin, we could use anything where the chance of an individual test being a success is the same, such as rolling a die and counting a six as a success. The probability of a success does not need to be a half.

If there were any subjects in Table 1 who had the same scores before and after the course, and hence had difference equal to zero, we would omit them from the calculation of the probabilities. They provide no information about the direction of any difference between the treatments. For the coin analogy, they correspond to the coin falling on its edge. With a coin, we would toss it again. For the sign test we can rarely do that, so we omit them. In this test, n is the number of subjects for whom there is a difference, one way or the other.

If the null hypothesis were true and negative and positive differences were equally likely, we might expect half of the differences to be negative. Indeed, we would expect it in the sense that the average number of negative differences under many repetitions of the course would be five. The number of negative differences is actually observed is one. What is the probability of getting a value as far from what we expect as is that observed? The possible numbers of negatives which would be as far from five (or even further) than is the observed value one are zero, one, nine, and ten (see Figure 2). To find the probability of one of these occurring, we add the probabilities of each one:

-ves	Probability
0	0.0009766
1	0.0097656
9	0.0097656
10	0.0009766
Total	0.0214844

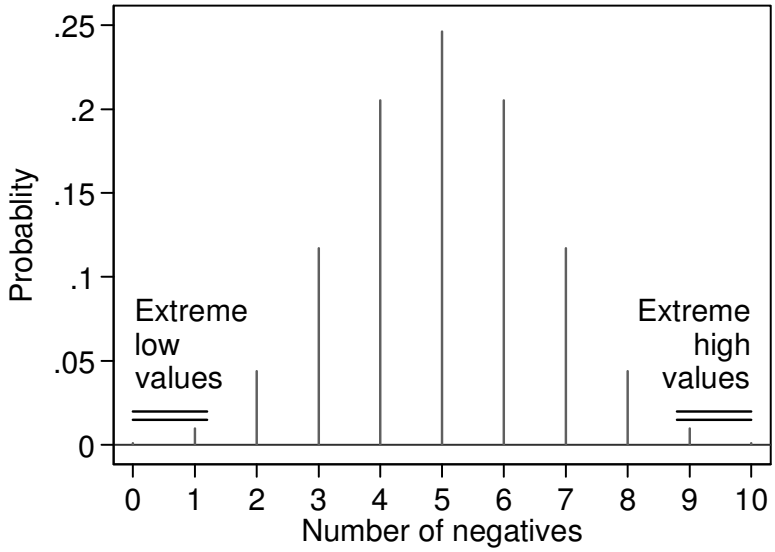

Figure 2 Extreme values for the differences in a sign test with one negative difference out of 10.

Table 3. Errors in significance tests

	Null hypothesis true	Alternative hypothesis true
Test not significant	No error	Type II error, beta error
Test significant	Type I error, alpha error	No error

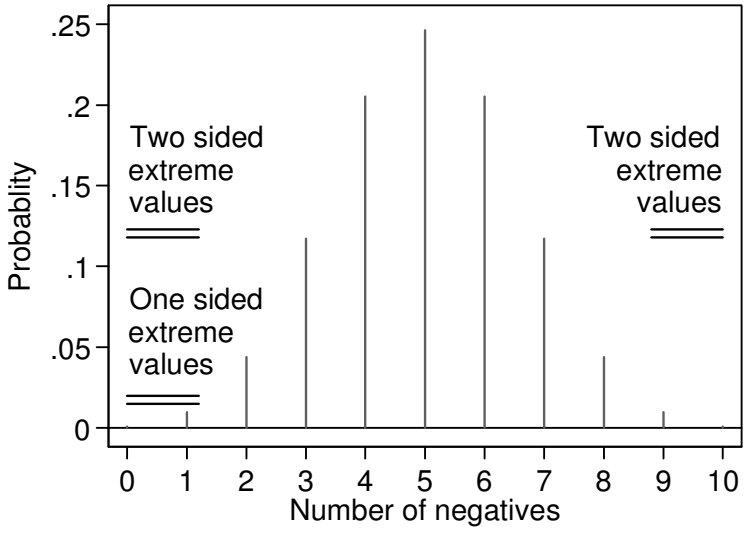

Figure 3. Probabilities for one-tailed and two-tailed sign tests for the nurse data.

The total is 0.02. Hence the probability of getting as extreme a value as that observed, in either direction, is 0.02. If the null hypothesis were true we would have a sample which is so extreme that the probability of it arising by chance is 0.02, two in a hundred. Thus, we would have observed an unlikely event if the null hypothesis were true. The data are not consistent with null hypothesis, so we can conclude that there is evidence in favour of a difference between the treatment periods.

The number of negative changes is called the **test statistic**, something calculated from the data which can be used to test the null hypothesis.

General principles of significance tests

The sign test is an example of a test of significance. There are many of these, but they all follow the same general pattern:

1. Set up the null hypothesis and its alternative.
2. Check any assumptions of the test.
3. Find the value of the test statistic.
4. Refer the test statistic to a known distribution which it would follow if the null hypothesis were true.
5. Find the probability of a value of the test statistic arising which is as or more extreme than that observed, if the null hypothesis were true.
6. Conclude that the data are consistent or inconsistent with the null hypothesis.

How does the sign test follow this pattern? First, we set up the null hypothesis and its alternative. The null hypothesis was:

> 'In this population of nurses, there is no difference between scores before and after the course' OR 'In this population of nurses, the probability of a difference in knowledge score in one direction is equal to the probability of a difference in knowledge score in the other direction'.

There are often several ways in which we can formulate the null hypothesis for a test. Note that the null hypothesis is about the population of nurses, including nurses who have not taken the course. It is not about the sample, the ten nurses who actually took the course. This is often not made explicit when null hypotheses are stated, but it should be.

The alternative hypothesis was:

> 'In this population of nurses, there is a difference between treatments' OR 'In this population of nurses, the probability of a difference in knowledge score in one direction is not equal to the probability of a difference in knowledge score in the other direction'.

Second, we check any assumptions of the test. Most significance tests require us to make some assumptions about the sample and the data. We should always check these as best we can. For the sign test, the only assumption is that the observations are independent, i.e. knowing about one observation would tell us nothing about another. This true here, as the observations are on ten different people. It would not be true, for example, if we had taken each question in the 18 question scale and looked at whether the subject's answer had improved from before the course to after,

then analysed the data as 180 observations. The observations on the same subject would clearly not be independent.

Third, we find the value of the test statistic. We call anything calculated from the data a statistic. A test statistic is something calculated from the data which can be used to test the null hypothesis. For the sign test, the test statistic is the number of negative changes. In our example it was equal to one.

Fourth, we refer the test statistic to a known distribution which it would follow if the null hypothesis were true. For the sign test, the known distribution is that followed by tossing a coin ten times, the Binomial distribution with $n = 12$ and $p = 0.5$.

Fifth, we find the probability of a value of the test statistic arising which is as or more extreme than that observed, if the null hypothesis were true. In our sign test, this was equal to 0.02.

Sixth, we conclude that the data are consistent or inconsistent with the null hypothesis. In the sign test example, the probability of seeing these data was quite small and we were able to conclude that the data were inconsistent with null hypothesis.

There are many different significance tests designed to answer different questions for different types of date, but all of them follow this pattern.

Significant and not significant

If the data are not consistent with the null hypothesis, the difference is said to be **statistically significant**. If the data are consistent with the null hypothesis, the difference is said to be **not statistically significant**. We can think of the significance test probability as an index of the strength of evidence against the null hypothesis. The probability of such an extreme value of the test statistic occurring if the null hypothesis were true is often called the **P value**.

The P value is ***not*** the probability that the null hypothesis is true. The null hypothesis is either true or it is not; it is not random and has no probability. The P value is the probability that, if the null hypothesis were true, we would get data as far from expectation as those we observed.

We said that a probability of 0.02 was small enough for us to conclude that the data were not consistent with the null hypothesi**s.** How small is small? A probability of 0.02, as in the example above, is fairly small and we have a quite unlikely event. But what about 0.06, 0.1, or 0.2? Would we treat these as sufficiently small for us to conclude that there was good evidence for a difference? Or should we look for a smaller probability, 0.01 or 0.001?

Suppose we take a probability of 0.01 or less as constituting reasonable evidence against the null hypothesis. If the null hypothesis is true, we shall make a wrong decision one in a hundred times. Deciding against a true null hypothesis is called an **error of the first kind**, **type I error**, or **α (alpha) error**. Sometimes there will be a difference in the population, but our sample will not produce a small enough probability for us to conclude that there is evidence for a difference in the population. We get an **error of the second kind**, **type II error**, or **β (beta) error** if we decide in favour of a null hypothesis which is in fact false. Table 3 shows these errors.

The smaller we demand the probability be before we decide against the null hypothesis, the larger the observed difference must be, and so the more likely we are

to miss real differences. By reducing the risk of an error of the first kind we increase the risk of an error of the second kind. The smaller we demand the probability be before we decide against the null hypothesis, the larger the observed difference must be, and so the more likely we are to miss real differences. By reducing the risk of an error of the first kind we increase the risk of an error of the second kind.

The conventional compromise is to say that differences are significant if the probability is less than 0.05. This is a reasonable guideline, but should not be taken as some kind of absolute demarcation. For some purposes, we might want to take a smaller probability as the critical one, usually 0.01. For example, in a major clinical trial we might think it is very important to avoid a type I error because after we have done the trial the preferred treatment will be adopted and it would be ethically impossible to replicate the trial. For other purposes, we might want to take a larger probability as the critical one, usually 0.1. For example, if we are screening possible new drugs for biological activity, we might want to avoid type II errors, because potentially active compounds will receive more rigorous testing but those producing no significant biological activity will not be investigated further.

If we decide that the difference is significant, the probability is sometimes referred to as the **significance level**. As a rough and ready guide, we can think of P values as indicating the strength of evidence like this:

P value	Evidence for a difference or relationship
Greater than 0.1:	Little or no evidence
Between 0.05 and 0.1:	Weak evidence
Between 0.01 and 0.05:	Evidence
Less than 0.01:	Strong evidence
Less than 0.001:	Very strong evidence

If a difference is statistically significant, then may well be real, but it is not necessarily important. For example, the UK Prospective Diabetes Study Group compared atenolol and captopril in reducing the risk of complications in type 2 diabetes. 1148 hypertensive diabetic patients were randomised. The authors reported that 'Captopril and atenolol were equally effective in reducing blood pressure to a mean of 144/83 mm Hg and 143/81 mm Hg respectively' (UKPDS 1998). The difference in diastolic pressure was statistically significant, P = 0.02. It is (statistically) significant, and real, but not (clinically) important.

If a difference is not statistically significant, it could still be real. We may simply have too small a sample to show that a difference exists. Furthermore, the difference may still be important. ***'Not significant' does not imply that there is no effect. It means that we have failed to demonstrate the existence of one.***

Presenting P values

Computers print out the exact P values for most test statistics. For example, using Stata 8.0 to do the sign test for the nurses data we get P=0.0215. This is the same as the 0.0214844 calculated above, but rounded to 4 decimal places. Before computers with powerful and easy to use statistical programs were readily available, many P values had to be found by referring the test statistic to a printed table. These often gave only a few P values, such as 0.25, 0.10, 0.05, 0.01, and the best the statistical analyst could do was to say that the P value for the data lay between two of these. Thus it was customary to quote P values as, for example, '0.05>P>0.01', which is how our sign test might have been quoted. This was often abbreviated to 'P<0.05'.

Old habits persist and researchers will often take the computer generated 'P=0.0215' and replace it in the paper by 'P<0.05'. Even worse, 'P=0.3294' might be reported as 'not significant', 'ns', or 'P>0.05'. This wastes valuable information. 'P=0.06' and 'P=0.6' can both get reported as 'P=NS', but 0.06 is only just above the conventional cut-off of 0.05 and indicates that there is some evidence for an effect, albeit rather weak evidence. A P value equal to 0.6, which is ten times bigger, indicates that there is very little evidence indeed. It is much better and more informative to quote the calculated P value.

We do not, however, need to reproduce all the figures printed. 'P=0.0215' is given to four **decimal places**, meaning that there are four figures after the decimal point, '0'. '2', '1', and '5'. It is also given to three **significant figures**. The first 'significant figure' is the first figure in the number which is not a zero, '2' in '0.0215'. The figures following the first non-zero figure are also called significant figures. So in '0.0215' the significant figures are '2', '1', and '5'. For another example, '0.0071056' is given to 7 decimal places and five significant figures. The '0' between '1' and '5' is a significant figure, it is the leading zeros before the '7' which are not significant figures. (The term 'significant figures' is nothing to do with statistical significance.) Personally, I would quote 'P=0.0071056' to one significant figure, as P=0.007, as figures after the first do not add much, but the first figure can be quite informative.

Sometimes the computer prints '0.0000' or '0.000'. The programmer has set the format for the printed probability to four or three decimal places. This is done for several reasons. First. it is easier to read than printing the full accuracy to which the calculation was done. Second, if the P value is very small, it might be printed out in the standard format for very small numbers, which looks like '1.543256E–07', meaning '0.0000001543256'. Third, almost all P values are approximations and the figures at the right hand end do not mean anything. The P value '0.0000' may be correct, in that the probability is less than 0.00005 and so equal to 0.0000 to four decimal places. This is the case for '0.0000001543256'. However, the probability can never be ***exactly*** zero, so we usually quote this as P<0.0001.

Significance tests and confidence intervals

Significance tests and confidence intervals often involve similar calculations and have a close relationship. Where a null hypothesis is about some population value, such as the difference between two means or two proportions, we can use the confidence interval as a test of significance. If the 95% confidence interval does not include the null hypothesis value, the difference is significant.

For example, in the trials of bandages for leg ulcers described in week 2, the differences between the percentages healed, elastic bandages minus inelastic bandages, their confidence intervals, and the P values obtained by significance tests were:

Trial	Estimate	95% confidence interval	P value
Northeast *et al.*	13.3	–5.7 to +32.3	0.2
Callam *et al.*	25.5	+9.3 to +41.7	0.003
Gould *et al.*	20.0	–10.2 to +50.2	0.2.

For the difference between two proportions, the null hypothesis value is zero. This is contained within the confidence intervals for the first and third trials and the difference is not significant in either of these trials. Zero is not contained within the

confidence interval for the second trial and the difference is significant for this trial. If the 95% confidence interval contains zero, the difference is not significant at the 0.05 level (1 – 0.95 = 0.05). If the 95% confidence interval does not contain zero, then the difference is significant.

Although a confidence interval can always be used to carry out a significance test, it may not be the best way. For example, when we calculate the confidence interval for two proportions, we estimate the standard error for the difference and then take 1.96 standard errors on either side of the observed difference. If this interval does not include zero, the difference is statistically significant with a P value less than 0.05. Equivalently, we could divide the observed difference by the standard error. If this ratio were greater than 1.96 or less than –1.96 the confidence interval would not include zero and the difference would be significant. If the null hypothesis were true, the ratio of difference to standard error would be an observation from the Standard Normal distribution and we could use this to give use the actual P value. The standard error of the difference between two proportions depends on the proportions in the two populations which we are comparing. We estimate these proportions from the data and use these estimates to calculate the standard error estimate used for the confidence interval. If the null hypothesis were true, the proportions in the two populations would be the same and we would we need only one estimate of the this proportion. So to test the null hypothesis, we estimate a single common proportion and use that to estimate a slightly different standard error for the difference. We then divide the observed difference by this significance test standard error and use this to get the P value. The single proportion estimate used in this standard error is based on more observations that either of the two proportions used for the confidence interval standard error, so it is closer to the true population proportion if the null hypothesis is true. The standard error using the common proportion will be more reliable if the null hypothesis is true and so we should use it, though of course we can use it only to test the null hypothesis. Using this modified standard error can alter the P value. For the three ulcer healing trials, the two standard errors and corresponding P values are:

Trial	Standard errors for		P values using SE for	
	estimation	testing	estimation	testing
Northeast *et al.*	0.0977	0.0987	0.087	0.090
Callam *et al.*	0.0828	0.0856	0.0010	0.0015
Gould *et al.*	0.154	0.157	0.097	0.102

The differences between results using the two standard error estimates are small, but we can see that the P values are changed slightly. As a result of this, 95% confidence intervals and 5% significance tests sometimes give different answers near the cut-off point.

Multiple significance tests

If we test a null hypothesis which is in fact true, using 0.05 as the critical significance level, the probability that we will come to a 'not significant' (i.e. correct) conclusion is 1.00 – 0.05 = 0.95. The probability that we will come to a 'significant' conclusion (i.e. false) is 0.05. This is the probability of a Type I error. If we test two true null hypotheses, independently of one another, the probability that neither test will be significant is 0.95 × 0.95 = 0.90. The probability that at least one of these tests will be significant is 1.00 – 0.90 = 0.10. If we test three true null hypotheses, independently of one another, the probability that none of the tests will be significant

is $0.95 \times 0.95 \times 0.95 = 0.95^3 = 0.86$. The probability that at least one of these three tests will be significant is $1.00 - 0.86 = 0.16$.

If we test twenty such true null hypotheses the probability that none will be significant is $0.95 \times 0.95 \times 0.95 \times \ldots \times 0.95$ twenty times, or $0.95^{20} = 0.36$. The probability of getting at least one significant result is $1.00 - 0.36 = 0.64$. We are more likely to get one than not. On average we will get one significant result, i.e. one Type I error, every time we do 20 tests of null hypotheses which are true. If we go on testing long enough we can expect to find something which is 'significant', even when all the null hypotheses we test are true and there is nothing to find.

Many health research studies are published with large numbers of significance tests. We must beware of attaching too much importance to a lone significant result among a mass of non-significant ones. It may be the one in twenty which we should get by chance alone.

One way in which we can generate many tests of significance is to test the same hypothesis separately within many subgroups of our study subjects. For example, Williams *et al.* (1992) randomly allocated elderly patients discharged from hospital to two groups: timetabled visits by health visitor assistants versus no visit unless there was perceived need. Patients were assessed for physical health, disability, and mental state using questionnaire scales. There were no significant differences overall between the intervention and control groups. However, the authors reported that among women aged 75-79 living alone the control group showed significantly greater deterioration in physical score than did the intervention group (P=0.04), and among men over 80 years the control group showed significantly greater deterioration in disability score than did the intervention group (P=0.03). The authors stated that 'Two small sub-groups of patients were possibly shown to have benefited from the intervention. . . . These benefits, however, have to be treated with caution, and may be due to chance factors.' We do not know how many groups these authors used to test the treatment difference, but there must have been quite a lot. Even if we take only the divisions they have reported, subjects cross-classified by age group, sex, and living alone, this would give eight subgroups. In addition, they appear to have tested the four groups given by splitting just by age and sex. They may also have tested for all men, all women, all living alone, etc. They did this for three different outcome variables, too: physical health, disability, and mental state. There are a great many possible tests here and a lot of opportunity for Type I errors.

One way to deal with the problem is many significance tests is the **Bonferroni correction**. We multiply all the P values by the number of tests. If any of the tests is significant after this, the test of the overall composite null hypothesis is significant. Thus if we can find any subgroup where for the comparison between intervention and control the P value multiplied by the number of tests the number of tests comparing intervention and control is less than 0.05, then we have evidence that there is a difference between intervention and control in the population from which these subjects come.

The reason we do this is that we want to carry out the test so that the probability that we would get at least one significant difference if all the null hypotheses are true is 0.05. To do this, we must make the probability of a not significant result much bigger than 0.95. We want to multiply together the probabilities of a not significant result for each of the tests and have this total come to 0.95. It turns out that this happens when we set our critical P value for a significant result in any of the individual tests to

be 0.05 divided by the number of tests. If any of the P values were less than this smaller critical value, the difference overall would be significant. It would be significant at the 0.05 level, however, not at the level of the smaller critical value. To avoid the temptation to quote the P value as the smaller one actually observed, Bland and Altman (1995) recommend multiplying the observed P values by the number of tests, The smallest of the resulting P values is the actual P value for the composite Bonferroni test.

In the study of Williams *et al.* (1992) there were at least eight subgroups, if not more. Even if we consider the three scales separately, the true P values are $8 \times 0.04 = 0.32$ and $8 \times 0.03 = 0.24$, both comfortably above 0.05. There is thus no evidence that the intervention and control treatments had different results in this population.

Note that the Bonferroni method here would test the composite null hypothesis that there is a difference between the treatments in at least one group of subjects. This is ***not*** the same as the null hypothesis that the difference between the treatments varies between different group of subjects. We would call this an interaction between treatment and subgroup. I describe methods to test this in Week 7.

Another way in which we can have many tests of the same composite null hypothesis is to test many outcome variables between the same groups. For example, Charles *et al*. (2004) followed up a clinical trial carried out in the 1960s, where pregnant women had been given folate in two different doses and placebo. The authors report tests comparing each dose to placebo, all folate treatments to placebo, and a trend from placebo through lower dose to higher dose (4 tests), unadjusted and adjusted for several risk factors ($4 \times 2 = 8$ tests), each for all cause mortality, cardiovascular mortality, all cancer deaths, and breast cancer mortality ($8 \times 4 = 32$ tests). One of these tests was statistically significant at the 0.05 level, P=0.02 for all cancer deaths versus placebo after adjustment. A simple Bonferroni correction would suggest that the P value for the composite hypothesis that folate increases risk of death is 32 times $0.02 = 0.6$. The authors did not tell us why of all the possible causes of death they picked breast cancer, indeed, they wrote that they 'had no prespecified hypothesis that taking folate supplements in pregnancy would increase the risk of cancer.' One wonders how many other causes they looked at and which have not been mentioned (Bland 2005).

Because the tests in this example are not independent, the adjusted P value is too big. In the absence of any more appropriate analysis by authors it is all the reader can do, however, and it is a very useful brake on the tendency to overestimate the strength of evidence. When a test tends to produce P values which are too large, it is said to be **conservative**. Statisticians prefer their P values to be too large rather than too small; if we are going to be wrong we want to be wrong in the conservative direction.

Researchers sometimes panic about the problem of multiple testing and wonder whether they should apply the correction to all the tests in a paper, or to all the tests in all the papers they have ever written. This would mean that as they did more research the results in their earlier papers would slowly become less significant. There is no need for this. We have to accept that tests of significance will be wrong sometimes, with both Type I and Type II errors, and we must always be cautious in our interpretation. When we have many tests essentially testing the same null hypothesis (e.g. that scheduled visits to the elderly have no effect or that folate supplementation increases the risk of an early death) we must be especially careful and that is when the Bonferroni correction should be used.

Primary outcome variable and primary analysis

In some studies, particularly clinical trials, we can avoid the problems of multiple testing by specifying a **primary outcome variable** in advance. We state before we look at the data, and preferably before we collect them, that one particular variable is the primary outcome. If we get a significant effect for this variable, we have good evidence of an effect. If we do not get a significant effect for this variable, we do not have good evidence of an effect, whatever happens with other variables.

Often, we specify not only the primary outcome variable in advance but also the **primary analysis**, the particular analysis which we intend to carry out. For example, in an asthma trial we might specify the primary outcome variable as being mean peak expiratory flow over a one week diary adjusted for mean peak expiratory flow measured at recruitment to the trial. (Adjustment is done using the regression methods described in Week 7.)

Any other variables and analyses are **secondary**. If there is no significant effect for the primary variable these should be treated with great caution and regarded as generating a hypothesis to be tested in a further trial rather than providing good evidence of a treatment effect in themselves.

One- and two-sided tests

In the sign test for the data from the nurses, the null hypothesis was that, in the population, the knowledge score after the course would be equal to the knowledge score before. The alternative hypothesis was that, in the population, the knowledge score after the course would be different from the knowledge score before, i.e. that there was a difference in one or other direction. This is called a **two-sided** or **two-tailed** test, because we used the probabilities of extreme values in both directions (i.e., the score after the course may be higher or lower than the score before it).

It is also possible to do a **one-sided** or **one-tailed** test, where we consider only the possibility that differences could be in a predefined direction. Here we might make our alternative hypothesis that, in the population, the knowledge score after the course would be greater than the knowledge score before. If we do this, the null hypothesis must change to match. It becomes that in the population, the knowledge score after the course would be equal to or less than the knowledge score before.

For the one-tailed test, we consider only the probability of more extreme events in the direction of the alternative hypothesis (Figure 3). This gives the probabilities of one negative difference and of no negative difference, P = 0.0097656 + 0.0009766 = 0.0107422 or P = 0.01 after rounding. We get a smaller probability and, of course, a higher significance level than the two sided test.

There is a price to pay for this lower P value. Should our course confuse the students so much that they understand less after it than they did before, our test will not detect this. The one-sided test would be not significant, and the more positive differences we had, the larger the P value would be. This implies that a decrease in knowledge score following the course would have the same interpretation as no change. There are few occasions in health research when we want our test to have this property. If treatments do harm, we want to know about it.

There are occasions when we would want a one-tailed test. One might be in the study of occupational health. For example, we might follow up a cohort of people employed in an industry and compare their incidence of cancers to the incidence in

the general population. The general population will include people who could not work in the industry because of their health. For example, people with chromosomal abnormalities might have impairments which prevent them from working in many jobs and may also be at increased risk of cancers. If we were to observe fewer cancers in our industry cohort than in the population a whole, we would not be surprised. We would not ascribe this to the protective effects of working in the industry. It could be the selective effect of comparing employed people to the whole population. We would therefore test the null hypothesis that cancer was no more frequent among people in the industry than it was in the general population, i.e. that the cancer rate in the industry was equal to or less than that in the general population. The alternative hypothesis would be that the cancer rate in the industry was greater than that in the general population. We would have a one-sided test.

It is sometimes said that we do a one-sided test because we are not interested in differences in the other direction. This is an inadequate justification. If we tried a new treatment and discovered that it killed all the patients, we would surely want our statistical methods to say that this was a treatment effect and would want to investigate why it had happened. A one-sided test should only be used if we would take exactly the same action and draw exactly the same conclusions if there was a large difference in the null hypothesis direction as we would if there were no difference at all. Situations where one-sided tests are appropriate are very rare in health research. Tests should be two-sided unless there is a good reason not to do this.

Pitfalls of significance tests

The first pitfall of significance testing is undoubtedly to conclude that there is no difference or relationship because it is not significant. You should never, ever, do this! We can never say that there is no difference, only that we have not detected one, but people do it all the time. For example, in an study of population level screening for abdominal aortic aneurysm Norman *et al.* (2004) concluded that screening 'was not effective in men aged 65-83 years and did not reduce overall death rates.' The difference in mortality was not significant, but the observed mortality among those invited for screening was 61% of the mortality in the control group, with 95% confidence interval 33% to 111%. So it is possible that the screening could reduce mortality by two thirds and this would be compatible with the trial results.

The second is to rely on a significance test alone where we can give a confidence interval. Confidence intervals are particularly useful when the test is not significant. They tell us how big the difference might actually be. It is not always possible, however, and we shall come across some situations where a P value is all that we can produce.

We should give exact P values wherever possible, not $P<0.05$ or P=NS, though only one significant figure is necessary.

We should beware of multiple testing. This does not mean that we should never do it, but we should know when we are doing it and take precautions against interpreting out P values too optimistically. The best way to do this is to be clear what the main hypothesis and outcome variable are before we begin the analysis of the data.

References

Bland JM (2005) Taking folate in pregnancy and risk of maternal breast cancer. What's in a name? *British Medical Journal* **330**, 600.

Bland JM, Altman DG. (1995) Statistics Notes. Multiple significance tests: the Bonferroni method. *British Medical Journal* **310**, 170.

Bland JM, Altman DG. (1998) Statistics Notes. Bayesians and frequentists. *British Medical Journal* **317**, 1151.

Charles D**,** Ness AR**,** Campbell D**,** Davey Smith G**,** Hall MH**.** (2004) Taking folate in pregnancy and risk of maternal breast cancer. *British Medical Journal* **329**, 1375-1376.

Norman PE, Jamrozik K, Lawrence-Brown MM, Le MTQ, Spencer CA, Tuohy RJ, Parsons RW, Dickinson JA. (2004) Population based randomised controlled trial on impact of screening on mortality from abdominal aortic aneurysm. *British Medical Journal* **329**, 1259-1262.

UKPDS Group. Efficacy of atenolol and captopril in reducing risk of macrovascular and microvascular complications in type 2 diabetes. *British Medical Journal* 1998; **317**: 713-720.

Williams, E.I., Greenwell, J., and Groom, L.M. (1992) The care of people over 75 years old after discharge from hospital: an evaluation of timetabled visiting by Health Visitor Assistants. *Journal of Public Health Medicine* **14**, 138-44.