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

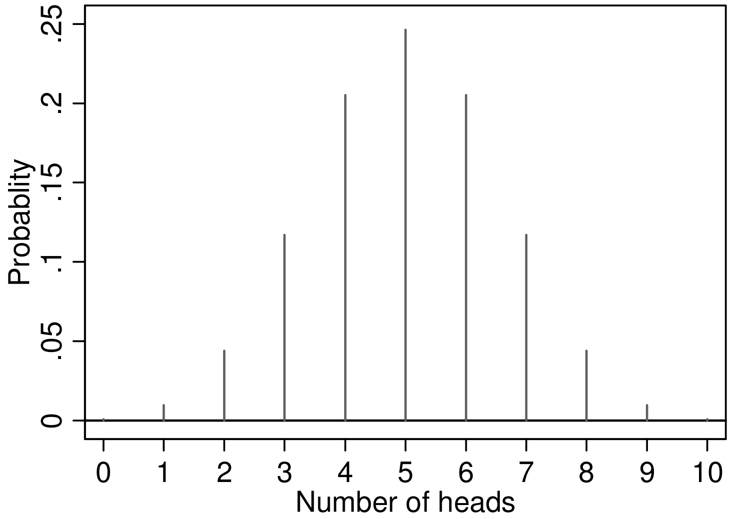

Figure 1. Distribution of the number of heads in ten tosses of a coin, or for the number of negative differences out of ten if positive and negative differences are equally likely (Binomial distribution, $n = 10$, $p = 0.5$)

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
```

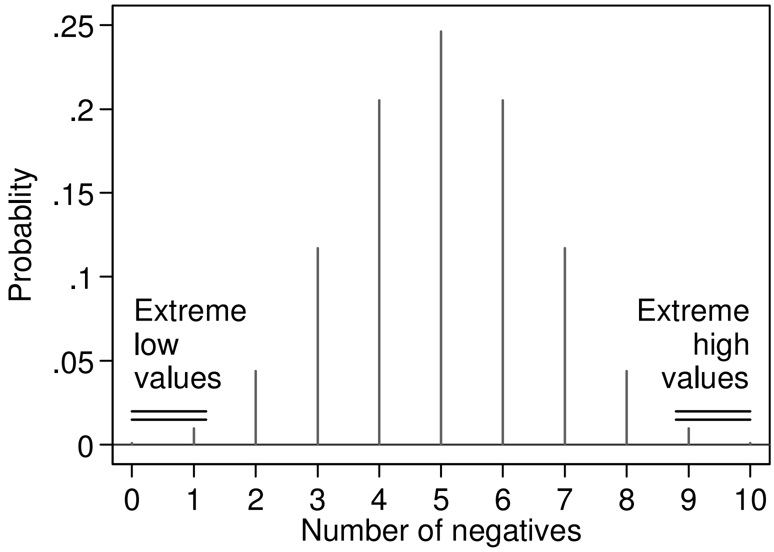

Figure 2 Extreme values for the differences in a sign test with one negative difference out of 10.

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

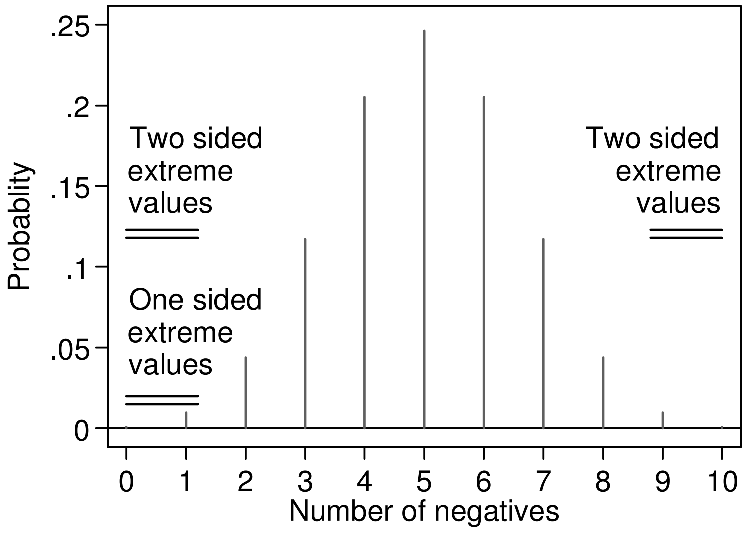

Figure 3. Probabilities for one-tailed and two-tailed sign tests for the nurse data.

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.

References

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.