**Clinical Biostatistics**

# Significance Tests

Martin Bland

Professor of Health Statistics

University of York

http://martinbland.co.uk/

## An Example: the Sign Test

Knowledge scores (–18 to +18) from a group of nurses before and after attending a course on systematic reviews.

| 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 | + |

## An Example: the Sign Test

These 10 course attenders are a sample from the population of all course attenders.

Would the other members of this population increase their knowledge score following the course?

In a significance test, we ask whether the difference observed was small enough to have occurred by chance if there were really no difference in the population.

If it were so, then the evidence in favour of there being a difference between scores before and after the course would be weak.

On the other hand, if the difference were much larger than we would expect due to chance if there were no real population difference, then the evidence in favour of a real difference would be strong.

## An Example: the Sign Test

Knowledge scores (–18 to +18) from a group of nurses before and after attending a course on systematic reviews.

| 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 | + |

Is there good evidence that knowledge increases following the course?

Most nurses have higher scores after the course.

## An Example: the Sign Test

To carry out the test of significance we suppose that, in the population, there is no difference between 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 between before and after, in either direction.

We find the probability of getting data as extreme as those observed if the null hypothesis were true.

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.

## An Example: the Sign Test

Knowledge scores (–18 to +18) from a group of nurses before and after attending a course on systematic reviews.

| 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 | + |

The sign test uses the direction of the difference only.

1 negative and 11 positives.

## An Example: the Sign Test

Consider the differences between the knowledge scores before and after for each nurse.

If the null hypothesis were true, then differences in number of attacks would be just as likely to be positive as negative, they would be random.

The probability of a change being negative would be equal to the probability of it becoming positive, 0.5.

Then the number of negatives would behave in exactly the same way as the number of heads if we toss a coin 10 times.

## An Example: the Sign Test

The number of negatives would behave in exactly the same way as the number of heads if we toss a coin 12 times.

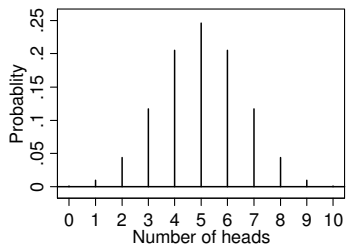
This is quite easy to investigate mathematically. We call it the Binomial Distribution with $n = 10$ and $p = 0.5$.

| Heads | Probability | Heads | Probability |
|---|---|---|---|
| 0 | 0.0009766 | 6 | 0.2050781 |
| 1 | 0.0097656 | 7 | 0.1171875 |
| 2 | 0.0439453 | 8 | 0.0439453 |
| 3 | 0.1171875 | 9 | 0.0097656 |
| 4 | 0.2050781 | 10 | 0.0009766 |
| 5 | 0.2460938 | | |

## An Example: the Sign Test

The number of negatives would behave in exactly the same way as the number of heads if we toss a coin 12 times.

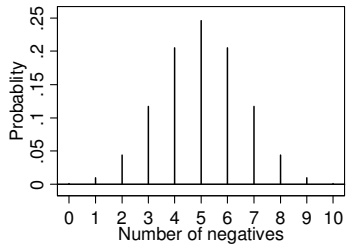
This is quite easy to investigate mathematically. We call it the Binomial Distribution with $n = 10$ and $p = 0.5$.



## An Example: the Sign Test

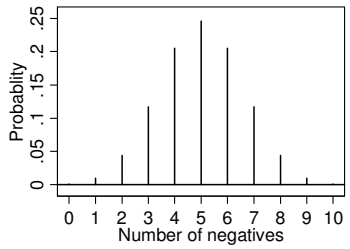
The number of negatives would behave in exactly the same way as the number of heads if we toss a coin 12 times.

This is quite easy to investigate mathematically. We call it the Binomial Distribution with $n = 10$ and $p = 0.5$.




## An Example: the Sign Test

If there were any subjects who had the same number of attacks on both regimes we would omit them, as they provide no information about the direction of any difference between the treatments. In this test, $n$ is the number of subjects for whom there is a difference, one way or the other.




Distribution of number of negatives if null hypothesis were true.

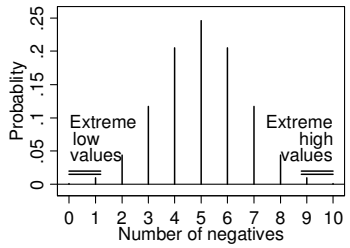
## An Example: the Sign Test

The expected number of negatives under the null hypothesis is 5. The number of negative differences is 1. What is the probability of getting a value as far from this as is that observed?

| -ves | Probability | -ves | Probability |
|---|---|---|---|
| **0** | **0.0009766** | 6 | 0.2050781 |
| **1** | **0.0097656** | 7 | 0.1171875 |
| 2 | 0.0439453 | 8 | 0.0439453 |
| 3 | 0.1171875 | **9** | **0.0097656** |
| 4 | 0.2050781 | **10** | **0.0009766** |
| **5** | **0.2460938** | | |

## An Example: the Sign Test

The expected number of negatives under the null hypothesis is 5. The number of negative differences is 1. What is the probability of getting a value as far from this as is that observed?




## An Example: the Sign Test

The expected number of negatives under the null hypothesis is 5. The number of negative differences is 1. What is the probability of getting a value as far from this as is that observed?

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

## An Example: the Sign Test

The probability of getting as extreme a value as that observed, in either direction, is 0.0214844.

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, one in fifty.

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 strong evidence in favour of a difference between knowledge scores before and after the course.

## The sign test

The sign test is an example of a test of significance.

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

## Principles of significance tests

The general procedure for a significance test is as follows:

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.

## Principles of significance tests

The general procedure for a significance test is as follows:

1. Set up the null hypothesis and its alternative.

Null hypothesis:

'No difference between before and after' OR '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'.

Alternative hypothesis:

'A difference between before and after' OR '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'.

**Principles of significance tests**

The general procedure for a significance test is as follows:

1. Set up the null hypothesis and its alternative.
2. Check any assumptions of the test.

Assumption:

That the subjects are independent.

**Principles of significance tests**

The general procedure for a significance test is as follows:

1. Set up the null hypothesis and its alternative.
2. Check any assumptions of the test.
3. Find the value of the test statistic.

Test statistic:

Number of negatives (= 1).

**Principles of significance tests**

The general procedure for a significance test is as follows:

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.

Known distribution:

Binomial, $n = 10$, $p = 0.05$.

### Principles of significance tests

The general procedure for a significance test is as follows:

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.

Probability:

P = 0.02

### Principles of significance tests

The general procedure for a significance test is as follows:

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.

Conclusion: inconsistent.

### Principles of significance tests

There are many different significance tests, all of which follow this pattern.

### Statistical significance

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**.

It 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.

### Interpreting the P value

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 |

### Significance levels and types of error

How small is small? A probability of 0.02, as in the example above, is clearly small and we have a quite unlikely event. But what about 0.04, or 0.06, or 0.1?

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**.

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.

## Significance levels and types of error

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.

| | 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 |

## Significance levels and types of error

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.

If we decide that the difference is significant, the probability is sometimes referred to as the **significance level**.

## Significant, real and important

If a difference is statistically significant, then may well be real, but not necessarily important.

For example, we may look at the effect of a drug, given for some other purpose, on blood pressure.

Suppose we find that the drug raises blood pressure by an average of 1 mm Hg, and that this is significant.

A rise in blood pressure of 1 mm Hg is not clinically significant, so, although it may be there, it does not matter.

It is (statistically) significant, and real, but not important.

## Significant, real and 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.

These should be given, rather than change them to 'not significant', 'ns' or $P>0.05$.

Similarly, if we have $P=0.0072$, we are wasting information if we report this as $P<0.01$.

This method of presentation arises from the pre-computer era, when calculations were done by hand and P values had to be found from tables.

Personally, I would quote this 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.

## Presenting P values

Sometimes the computer prints 0.0000. This may be correct, in that the probability is less than 0.00005 and so equal to 0.0000 to four decimal places.

The probability can rarely be ***exactly*** zero, so we usually quote this as $P<0.0001$.

## Significance tests and confidence intervals

Often involve similar calculations.

If CI does not include the null hypothesis value, the difference is significant.

E.g. for a difference between two proportion, null hypothesis value = 0.

If 95% CI contains zero, difference is not significant.

If 95% CI does not contain zero, difference is significant.

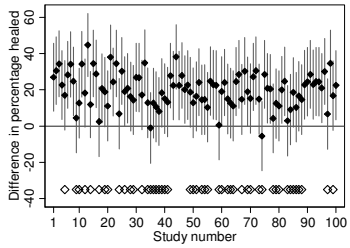
E.g. ulcer healing 63% (31/49) vs. 50% (26/52).

95% CI for difference: –7 to +33 percentage points.

Difference could be zero. Not significant.

## Significance tests and confidence intervals

Ulcer healing simulation:




Open symbols denote no significant differences.

## Significance tests and confidence intervals

The null hypothesis may contain information about the standard error.

E.g. comparison of two proportions, the standard error for the difference depends on the proportions themselves.

If the null hypothesis is true we need only one estimate of the proportion.

This alters the standard error for the difference.

Confidence interval: SE = 0.0977

Significance test: SE = 0.0987

95% CI and 5% significance test 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, we have a probability of 0.95 of coming to a 'not significant' (i.e. correct) conclusion.

If we test two independent true null hypotheses, the probability that neither test will be significant is $0.95 \times 0.95 = 0.90$.

If we test twenty such hypotheses the probability that none will be significant is $0.95^{20} = 0.36$.

This gives a probability of $1 - 0.36 = 0.64$ of getting at least one significant result.

We are more likely to get one than not.

The expected number of spurious significant results is $20 \times 0.05 = 1$.

## Multiple significance tests

Many medical research studies are published with large numbers of significance tests.

These are not usually independent, being carried out on the same set of subjects, so the above calculations do not apply exactly.

If we go on testing long enough we will find something which is 'significant'.

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.

## Multiple significance tests

- Many subgroups.
- Many outcome variables.

## Many subgroups

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 assessed for physical health, disability, and mental state using questionnaire scales.

No significant differences overall between the intervention and control groups.

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.

## Many subgroups

Williams *et al.* (1992)

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.'

## Many subgroups: Bonferroni correction

Multiply the P values by the number of tests.

If any is then significant, the test of the overall composite null hypothesis is significant.

E.g. Williams et al. (1992).

Subjects were cross-classified by age groups, whether living alone, and sex, so 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 $8 \times 0.03 = 0.24$.

Composite null hypothesis: there is a difference between the treatments in at least one group of subjects.

## Many subgroups: Bonferroni correction

Composite null hypothesis: there is a difference between the treatments in at least one group of subjects.

This is ***not*** the same as: the difference between the treatments varies between different group of subjects.

This needs a test of interaction (Week eight).

## Multiple outcome measurements

E.g. Newnham *et al.* (1993) randomized pregnant women to receive a series of Doppler ultrasound blood flow measurements or to control.

They found a significantly higher proportion of birthweights below the 10th and 3rd centiles (P=0.006 and P=0.02).

These were only two of many comparisons. At least 35 were reported in the paper, though only these two were reported in the abstract.

Birthweight was not the intended outcome variable for the trial.

Newnham, J.P., Evans, S.F., Con, A.M., Stanley, F.J., Landau, L.I. (1993) Effects of frequent ultrasound during pregnancy: a randomized controlled trial. *Lancet* **342**, 887-91.

## Multiple outcome measurements

These tests are not independent, because they are all on the same subjects, using variables which may not be independent.

The proportions of birthweights below the 10th and 3rd centiles are clearly not independent, for example.

We can apply the Bonferroni correction.

For the example, the P values could be adjusted by $35 \times 0.006 = 0.21$ and $35 \times 0.02 = 0.70$.

Because the tests are not independent, the adjusted P value is too big.

Test is **conservative**.

### Primary outcome variable and primary analysis

In some studies, 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.

### Primary outcome variable and primary analysis

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.

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.

Any other variables and analyses are **secondary**.

If there is no significant effect for the primary variable these should be treated with great caution.

### One- and two-sided tests of significance

In the knowledge score example, the alternative hypothesis was 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.

**One sided** or **one tailed** test:

Alternative hypothesis: in the population, the knowledge score before the course is less than the knowledge score after.

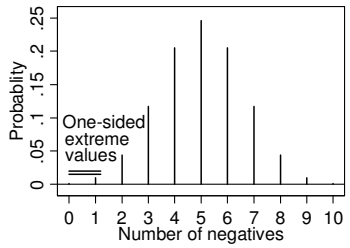
Null hypothesis: in the population, the knowledge score before the course is greater than or equal to the knowledge score after.

P = 0.01, and of course, a higher significance level than the two sided test.

## One- and two-sided tests of significance

One sided null hypothesis: the knowledge score before course is greater than or equal to the knowledge score after course.

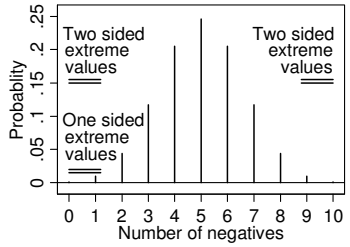
One sided alternative hypothesis: the knowledge score before course is less than the knowledge score after course.




## One- and two-sided tests of significance

Two sided null hypothesis: the knowledge score before course is equal to the knowledge score after course.

Two sided alternative hypothesis: the knowledge score before course is not equal to the knowledge score after course.




## One- and two-sided tests of significance

**One sided** or **one tailed** test:

One sided null hypothesis: the knowledge score before course is greater than or equal to the knowledge score after course.

One sided alternative hypothesis: the knowledge score before course is less than the knowledge score after course.

This implies that a decrease in knowledge in the after direction would have the same interpretation as no change.

Seldom true in health research.

Tests should be two sided unless there is a good reason not to do this.

## One- and two-sided tests of significance

**Example of valid one sided hypothesis:**

Study of occupational health.

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.

## One- and two-sided tests of significance

**Example of valid one sided hypothesis:**

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.

## One- and two-sided tests of significance

**Example of valid one sided hypothesis:**

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.

**Pitfalls of significance tests**

You should never, ever, conclude that there is no difference or relationship because it is not significant.

You should not rely on significance tests alone if you can give confidence intervals. Particularly useful when the test is not significant.

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

You should avoid multiple testing. Be clear what the main hypothesis and outcome variable are.