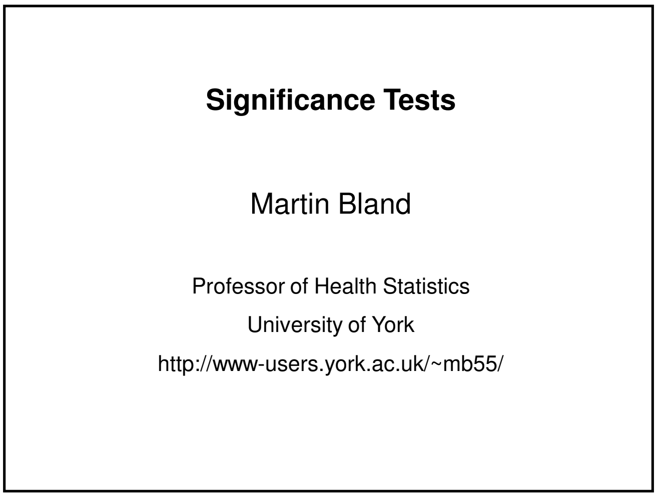


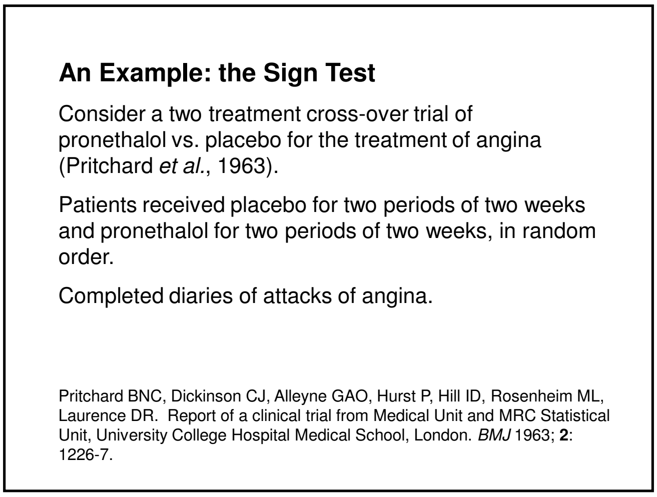
Significance Tests

Martin Bland

Professor of Health Statistics

University of York

http://www-users.york.ac.uk/~mb55/



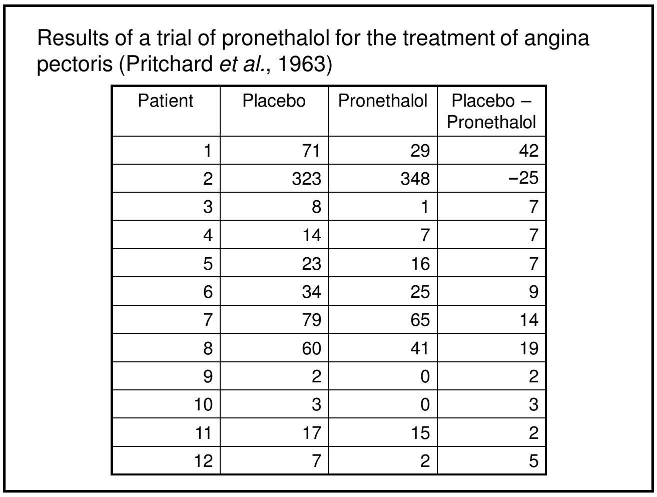
An Example: the Sign Test

Consider a two treatment cross-over trial of pronethalol vs. placebo for the treatment of angina (Pritchard *et al.*, 1963).

Patients received placebo for two periods of two weeks and pronethalol for two periods of two weeks, in random order.

Completed diaries of attacks of angina.

Pritchard BNC, Dickinson CJ, Alleyne GAO, Hurst P, Hill ID, Rosenheim ML, Laurence DR. Report of a clinical trial from Medical Unit and MRC Statistical Unit, University College Hospital Medical School, London. *BMJ* 1963; **2**: 1226-7.



Results of a trial of pronethalol for the treatment of angina pectoris (Pritchard *et al.*, 1963)

Patient	Placebo	Pronethalol	Placebo – Pronethalol
1	71	29	42
2	323	348	−25
3	8	1	7
4	14	7	7
5	23	16	7
6	34	25	9
7	79	65	14
8	60	41	19
9	2	0	2
10	3	0	3
11	17	15	2
12	7	2	5

These 12 patients are a sample from the population of all patients.

Would the other members of this population experience fewer attacks while using Pronethalol?

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 the treatment periods 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.

Results of a trial of pronethalol for the treatment of angina pectoris (Pritchard *et al.*, 1963)

Patient	Placebo	Pronethalol	Placebo – Pronethalol
1	71	29	42
2	323	348	−25
3	8	1	7
4	14	7	7
5	23	16	7
6	34	25	9
7	79	65	14
8	60	41	19
9	2	0	2
10	3	0	3
11	17	15	2
12	7	2	5

Is there good evidence that Pronethalol reduces the number of attacks?

Most patients experience fewer attacks on Pronethalol.

To carry out the test of significance we suppose that, in the population, there is no difference between the two treatment periods.

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 the treatments, 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.

Results of a trial of pronethalol for the treatment of angina pectoris (Pritchard *et al.*, 1963)

Patient	Placebo	Pronethalol	Placebo – Pronethalol	Sign
1	71	29	42	+
2	323	348	−25	–
3	8	1	7	+
4	14	7	7	+
5	23	16	7	+
6	34	25	9	+
7	79	65	14	+
8	60	41	19	+
9	2	0	2	+
10	3	0	3	+
11	17	15	2	+
12	7	2	5	+

The sign test uses the direction of the difference only.

1 negative and 11 positives.

The sign test

Consider the differences between the number of attacks on the two treatments for each patient.

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.

If we kept on testing patients indefinitely, the proportion of changes which were negative would be equal to the proportion which were positive,

OR 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 12 times.

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.

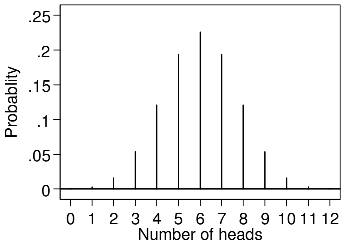
This is quite easy to investigate mathematically. We call it the Binomial Distribution with $n = 12$ and $p = 0.5$.

```
Heads Probability      Heads Probability
---------------------------------------
  0    0.00024           7    0.19336
  1    0.00293           8    0.12085
  2    0.01611           9    0.05371
  3    0.05371          10    0.01611
  4    0.12085          11    0.00293
  5    0.19336          12    0.00024
  6    0.22559
---------------------------------------
```

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.

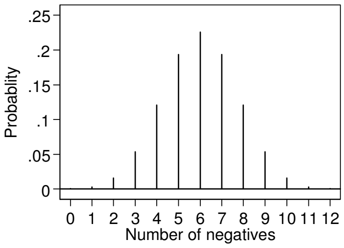
This is quite easy to investigate mathematically. We call it the Binomial Distribution with $n = 12$ and $p = 0.5$.



The sign test

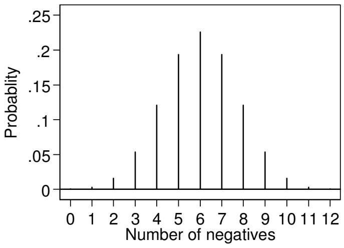
The number of negatives would behave in exactly the same way as the number of heads if we toss a coin 12 times.

This is quite easy to investigate mathematically. We call it the Binomial Distribution with $n = 12$ and $p = 0.5$.



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.

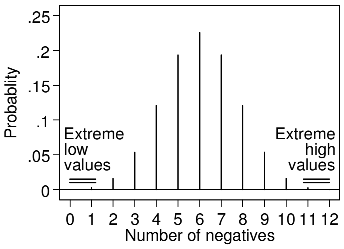
The sign test

The expected number of negatives under the null hypothesis is 6. 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	**0.00024**	7	0.19336
1	**0.00293**	8	0.12085
2	0.01611	9	0.05371
3	0.05371	10	0.01611
4	0.12085	**11**	**0.00293**
5	0.19336	**12**	**0.00024**
6	**0.22559**		

The sign test

The expected number of negatives under the null hypothesis is 6. The number of negative differences is 1. What is the probability of getting a value as far from this as is that observed?



The sign test

The expected number of negatives under the null hypothesis is 6. 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	**0.00024**
1	**0.00293**
11	**0.00293**
12	**0.00024**
Total	**0.00634**

The sign test

The probability of getting as extreme a value as that observed, in either direction, is 0.00634.

If the null hypothesis were true we would have a sample which is so extreme that the probability of it arising by chance is 0.006, less than one in a hundred.

Thus, we would have observed a very 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 the treatment periods.

(Since this was a double blind randomized trial, it seems reasonable to suppose that this was caused by the activity of the drug.)

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 treatments’ OR ‘Probability of a difference in number of attacks in one direction is equal to the probability of a difference in number of attacks in the other direction’.

Alternative hypothesis:

‘A difference between treatments’ OR ‘Probability of a difference in number of attacks in one direction is not equal to the probability of a difference in number of attacks 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 patients 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, $n = 12$, $p = 0.5$.

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:

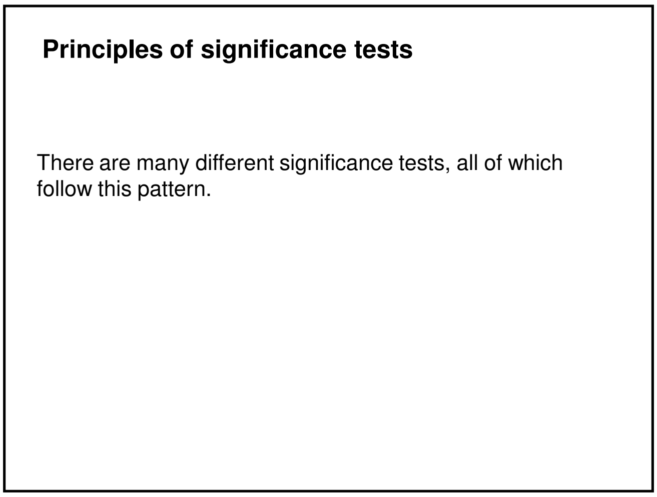
P = 0.006

Principles of significance tests

The general procedure for a significance test is as follows:

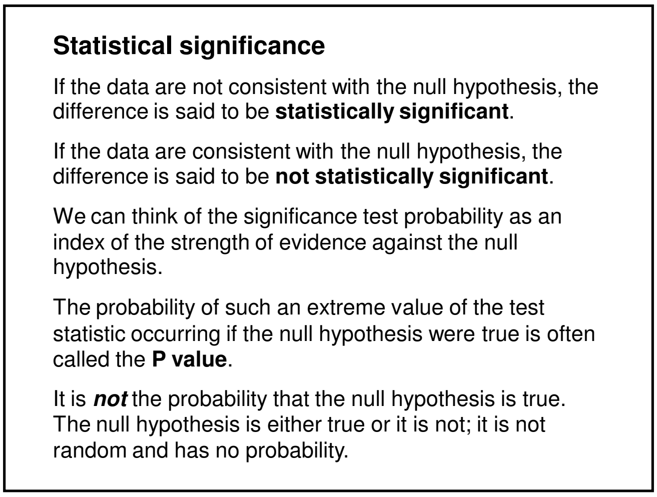
1. Set up the null hypothesis and its alternative.
2. Check any assumptions of the test.
3. Find the value of the test statistic.
4. Refer the test statistic to a known distribution which it would follow if the null hypothesis were true.
5. Find the probability of a value of the test statistic arising 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

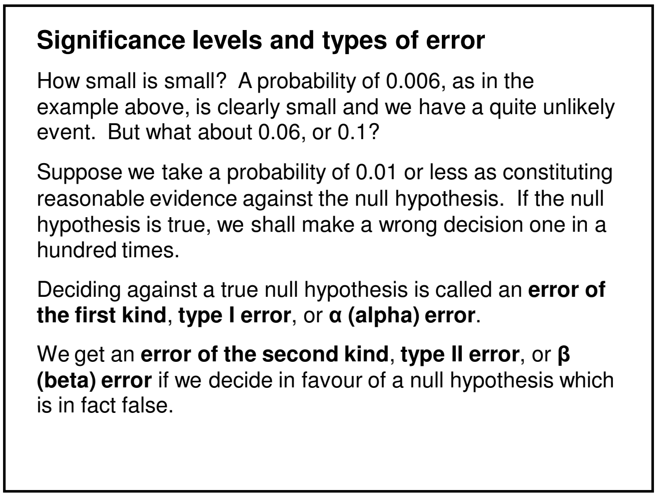
If the data are not consistent with the null hypothesis, the difference is said to be **statistically significant**.

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.



Significance levels and types of error

How small is small? A probability of 0.006, as in the example above, is clearly small and we have a quite unlikely event. But what about 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**.

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

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 never be ***exactly*** zero, so we usually quote this as P<0.0001.

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 × 0.95 = 0.90.

If we test twenty such hypotheses the probability that none will be significant is 0.95 × 0.95 × 0.95 … × 0.95 = 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.

We expected to get one spurious significant result.

Multiple significance tests

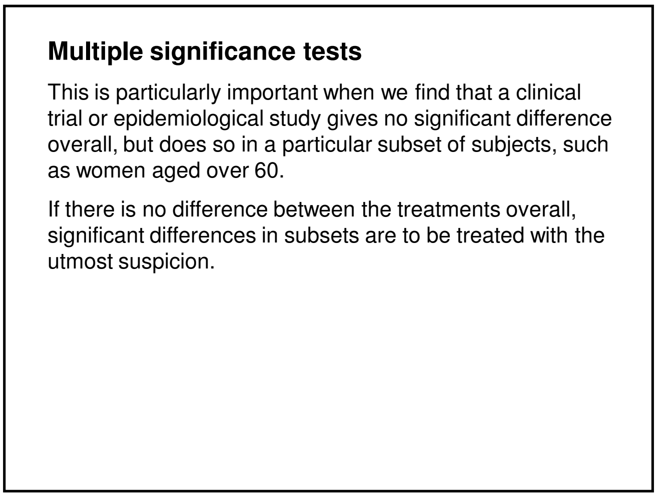
Many medical research studies are published with large numbers of significance tests.

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.

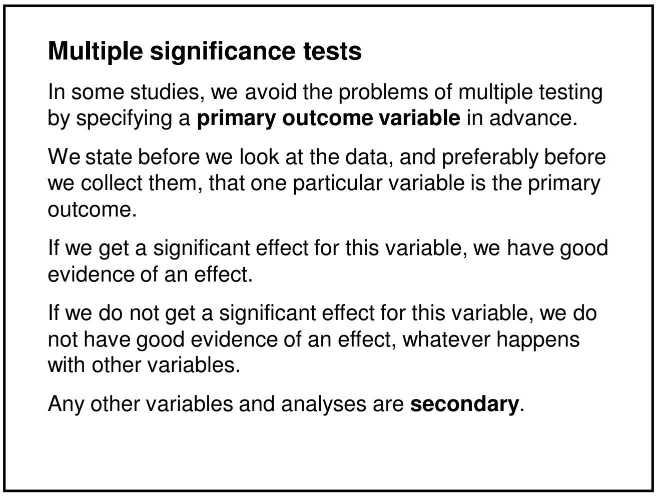
It may be the one in twenty which we should get by chance alone.



Multiple significance tests

This is particularly important when we find that a clinical trial or epidemiological study gives no significant difference overall, but does so in a particular subset of subjects, such as women aged over 60.

If there is no difference between the treatments overall, significant differences in subsets are to be treated with the utmost suspicion.



Multiple significance tests

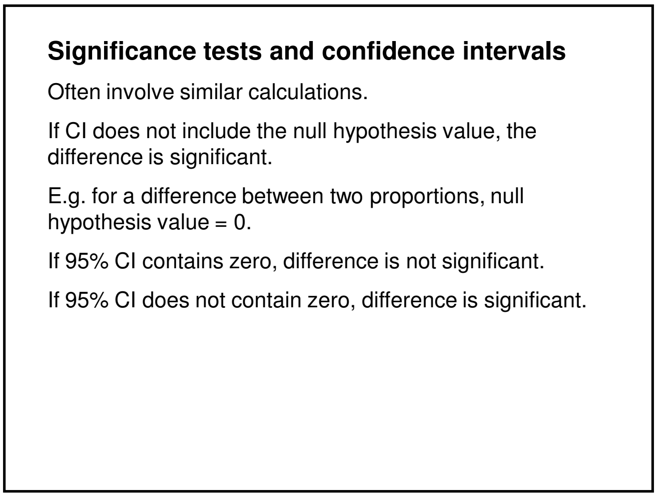
In some studies, we avoid the problems of multiple testing by specifying a **primary outcome variable** in advance.

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.

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



Significance tests and confidence intervals

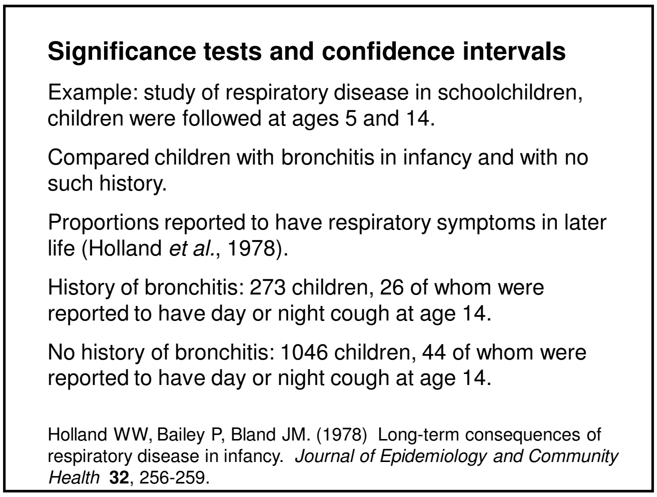
Often involve similar calculations.

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

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

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

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



Significance tests and confidence intervals

Example: study of respiratory disease in schoolchildren, children were followed at ages 5 and 14.

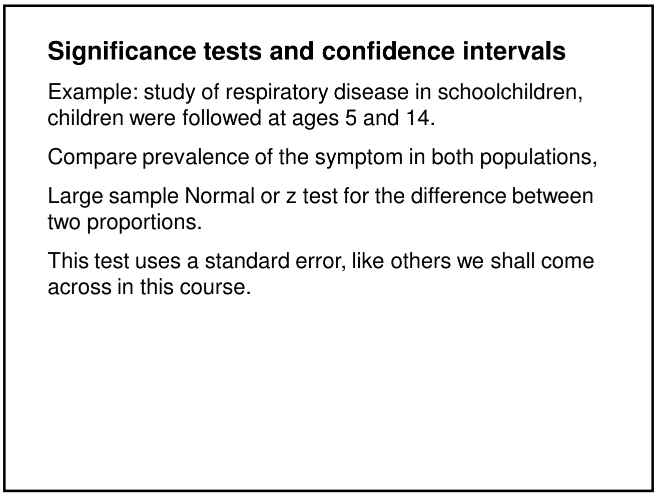
Compared children with bronchitis in infancy and with no such history.

Proportions reported to have respiratory symptoms in later life (Holland *et al.*, 1978).

History of bronchitis: 273 children, 26 of whom were reported to have day or night cough at age 14.

No history of bronchitis: 1046 children, 44 of whom were reported to have day or night cough at age 14.

Holland WW, Bailey P, Bland JM. (1978) Long-term consequences of respiratory disease in infancy. *Journal of Epidemiology and Community Health* **32**, 256-259.



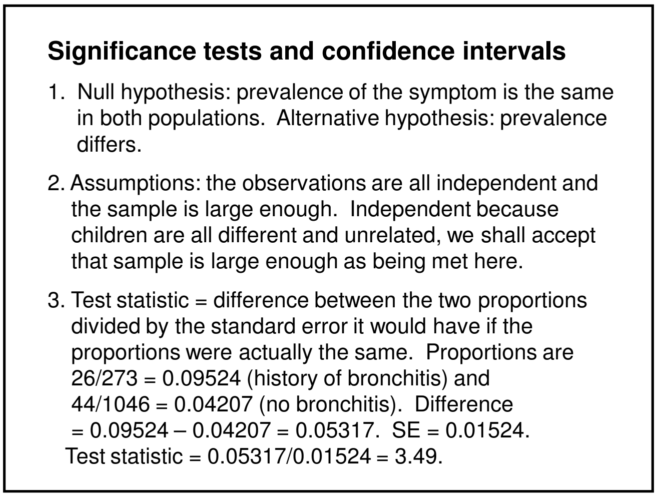
Significance tests and confidence intervals

Example: study of respiratory disease in schoolchildren, children were followed at ages 5 and 14.

Compare prevalence of the symptom in both populations,

Large sample Normal or z test for the difference between two proportions.

This test uses a standard error, like others we shall come across in this course.



Significance tests and confidence intervals

1. Null hypothesis: prevalence of the symptom is the same in both populations. Alternative hypothesis: prevalence differs.
2. Assumptions: the observations are all independent and the sample is large enough. Independent because children are all different and unrelated, we shall accept that sample is large enough as being met here.
3. Test statistic = difference between the two proportions divided by the standard error it would have if the proportions were actually the same. Proportions are 26/273 = 0.09524 (history of bronchitis) and 44/1046 = 0.04207 (no bronchitis). Difference = 0.09524 – 0.04207 = 0.05317. SE = 0.01524.
   Test statistic = 0.05317/0.01524 = 3.49.

Significance tests and confidence intervals

4. If the null hypothesis were true, test statistic would be an observation from the Standard Normal distribution. (Sample is large ➔ both proportions will follow approximately Normal distributions. The distribution of differences should have mean zero if the null hypothesis is true. Dividing by the standard error gives us standard deviation of this distribution = 1.0.)
5. The probability of the test statistic having a value as far from zero as 3.49 is 0.0005.
6. Conclude that the data are not consistent with the null hypothesis. We have strong evidence that children with a history of bronchitis are more likely than other to be reported to have cough during the day or at night at the age of 14.

Significance tests and confidence intervals

Confidence interval: use a different standard error, the standard error when the proportions may not be equal.

SE = 0.0188.

95% confidence interval for the difference is
$0.05317 - 1.96 \times 0.0188$ to $0.05317 + 1.96 \times 0.0188$
= 0.016 to 0.090.

The null hypothesis value of the difference is zero and this is not included in the 95% confidence interval.

We do not include zero as a value for the difference which is consistent with the data.

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 = 0.0188

Significance test: SE = 0.0152

95% CI and 5% significance test sometimes give different answers near the cut-off point.

One- and two-sided tests of significance

In the pronethalol 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 number of attacks on pronethalol is less than the number of attacks on placebo.

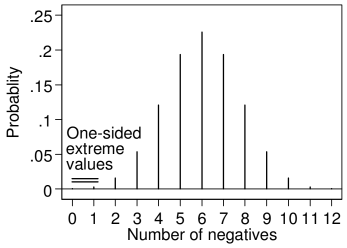
Null hypothesis: in the population, the number of attacks on pronethalol is greater than or equal to the number of attacks on placebo.

P = 0.003, and of course, a higher significance level than the two sided test.

One- and two-sided tests of significance

One sided null hypothesis: the number of attacks on pronethalol is greater than or equal to the number of attacks on placebo.

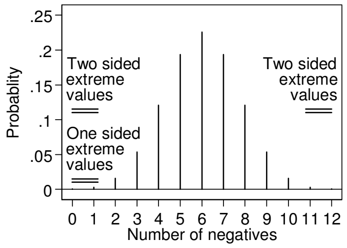
One sided alternative hypothesis: the number of attacks on pronethalol is less than the number of attacks on placebo.

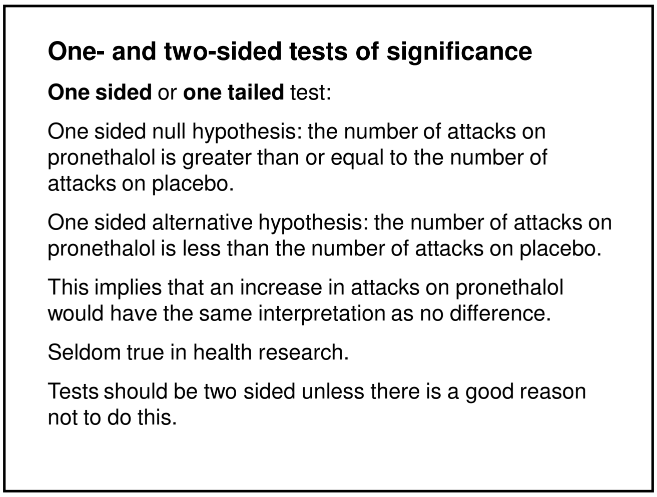


One- and two-sided tests of significance

Two sided null hypothesis: the number of attacks on pronethalol is equal to the number of attacks on placebo.

Two sided alternative hypothesis: the number of attacks on pronethalol is not equal to the number of attacks on placebo.





One- and two-sided tests of significance

One sided or **one tailed** test:

One sided null hypothesis: the number of attacks on pronethalol is greater than or equal to the number of attacks on placebo.

One sided alternative hypothesis: the number of attacks on pronethalol is less than the number of attacks on placebo.

This implies that an increase in attacks on pronethalol would have the same interpretation as no difference.

Seldom true in health research.

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