Health Sciences M.Sc. Programme

Applied Biostatistics

Week 5: Significance tests

Testing a hypothesis

A significance test enables us to measure the strength of evidence which the data supply for or against some proposition of interest. For example, Table 1 shows the results of a crossover trial of pronethalol for the treatment of angina, the number of attacks over four weeks on each treatment. 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? We can see that the number of attacks is highly variable from one patient to another, and it is quite possible that this is true from one occasion to another as well. So it could be that some patients would have fewer attacks while on pronethalol than while on placebo quite by chance. 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.

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 do this by finding 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

We shall now find a way of testing this null hypothesis, using a method called the **sign test**. An obvious start is to consider the differences between the number of attacks on the two treatments for each patient, as in Table 1. 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. Another way of saying this is that the probability of a change being negative would be equal to the probability of it becoming positive. These would both be 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. This is quite easy to investigate mathematically. We can work out the probability that 12 tosses of a coin would show any given number of heads. This is also the proportion of occasions on which that 12 tosses of a coin would show the given number of heads. These probabilities are shown in Table 2. We call this the Binomial distribution with $n = 12$ and $p = 0.05$.

Table 1. Trial of pronethalol for the prevention of angina pectoris (data of Pritchard *et al.*, 1963)

Patient number	Number of attacks while on:		Difference, placebo minus pronethalol	Sign of difference
	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	+

Table 2. Probability distribution for the number of heads out 12 flips of a coin, Binomial distribution with n = 12 and p = 0.5

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

We can show these probabilities graphically, as in Figure 1. This shows each probability as a vertical line. It is done this way because only the integer values have any probability.

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, the number of subjects, n, is the number of subjects for whom there is a difference, one way or the other. Those for whom the difference is zero contribute no information. If they were coins which fell on their edge, we would flip them again. In the clinical trial all we can do is exclude them.

Figure 1. The Binomial distribution for the number of heads in 12 flips of a coin

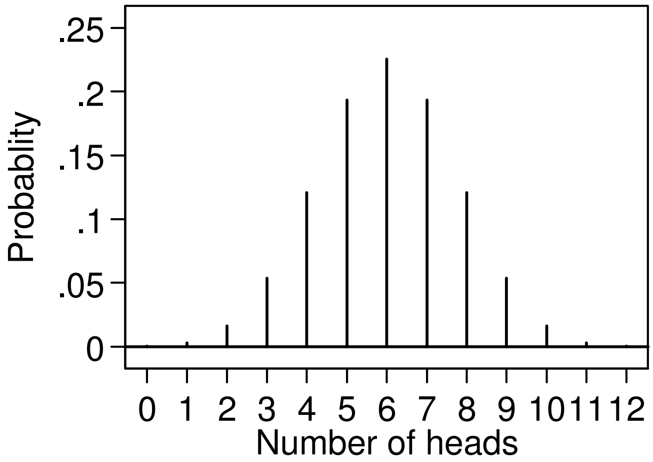
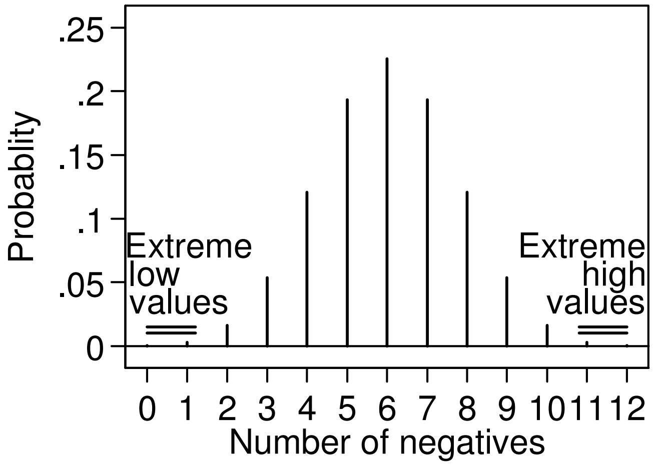


Figure 2. Extremes of the Binomial Distribution for the sign test



If the null hypothesis were true, what would be the probability of getting an observation from this distribution as extreme as the value we have actually observed? The number of negative differences is 1. The probability of getting 1 negative change = 0.00293. This is not a likely event in itself. However, we are interested in the probability of getting a value as far or further from the expected value, 6, as is 1, and clearly 0 is further and must be included. The probability of no negative changes = 0.00024. So the probability of one or fewer negative changes is 0.00293 + 0.00024 = 0.00317. We said that the alternative hypothesis was that there was a difference in either direction. We must, therefore, consider the probability of getting a value as extreme on the other side of the mean, that is 11 or 12 negatives (Figure 2). The probability of 11 or 12 negatives = 0.00293 + 0.00024 = 0.00317. Hence, the probability of getting as extreme a value as that observed, in either direction, is 0.00317 + 0.00317 = 0.00634. This means that 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.

Table 3. Types of error 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

Thus, we would have observed a very unlikely event if the null hypothesis were true. This means that 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.)

Principles of significance tests

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. 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.

For the sign test we have just done, we have

1. The null hypothesis is ‘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’, the alternative hypothesis is ‘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’.
2. The only assumption required is that the patients are independent, which is true here as they are all different people.
3. The test statistic is the number of negatives (= 1).
4. If the null hypothesis were true, this would be an observation from the Binomial distribution with $n = 12$, $p = 0.5$.
5. The probability of a value of the test statistics as far from what we would expect as 1 is P = 0.006.
6. Our conclusion is that the data are inconsistent with the null hypothesis.

There are many different significance tests, all of which follow this pattern. If the data are not consistent with the null hypothesis, the difference is said to be

statistically significant. If the data do not support the null hypothesis, it is sometimes said that we reject the null hypothesis, and if the data are consistent with the null hypothesis it is said that we accept it. Such an ‘all or nothing’ decision making approach is seldom appropriate in medical research. It is preferable to 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. This is a common misconception. The null hypothesis is either true or it is not; it is not random and has no probability.

Significance levels and types of error

We must still consider the question of 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 **alpha error**. We get an **error of the second kind**, **type II error**, or **beta** error if we decide in favour of a null hypothesis which is in fact false. These errors are set out in Table 3.

Now 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. There is not a great difference between probabilities of 0.06 and 0.04, and they surely indicate similar strength of evidence. It is better to regard probabilities around 0.05 as providing some evidence against the null hypothesis, which increases in strength as the probability falls. 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

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.

On the other hand, 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 P=0.0072 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. 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. Whatever we do, we should never quote it as P<0.000, as I have seen. This is impossible.

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 getting a 'not significant' (i.e. correct) decision. 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 $0.95\times 0.95 \times 0.95 \ldots \times 0.95 = 0.36$. 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 expect to get one spurious significant result.

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. However, it is clear that 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.

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.

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. Other significant effects are only an indication that another study may be justified.

Significance tests and confidence intervals

Significance tests and confidence intervals often involve similar calculations. For example, we can test the null hypothesis that two groups have the same mean and we can find a confidence interval for the difference between the means. If the 95% confidence interval for the difference does not include the null hypothesis value, the difference is significant at the 5% level. If the 95% confidence interval for the difference includes the null hypothesis value, the difference is not significant at the 5% level.

For example, in a study of respiratory disease in schoolchildren, children were followed at ages 5 and 14. We looked at the proportions of children with bronchitis in infancy and with no such history who were reported to have respiratory symptoms in later life (Holland *et al.*, 1978). We had 273 children with a history of bronchitis before age 5 years, 26 of whom were reported to have day or night cough at age 14. We had 1046 children with no bronchitis before age 5 years, 44 of whom were reported to have day or night cough at age 14. We shall test the null hypothesis that the prevalence of the symptom is the same in both populations, against the alternative that it is not. We shall use a test called the 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. It follows the structure described above and for this lecture we shall not go into the details of the method. It works like this.

1. The null hypothesis is that the prevalence of the symptom is the same in both populations. The alternative that it is not.
2. The assumptions of the test are that the observations are all independent, which they are because these are all different, unrelated children, and that the sample is large enough, we shall accept as being met here.
3. The test statistic is the difference between the two proportions divided by the standard error it would have if the proportions were actually the same. The two proportions of children reported to have cough are 26/273 = 0.09524 for children with a history of bronchitis and 44/1046 = 0.04207 for those with no bronchitis. The difference between these proportions is = 0.09524 – 0.04207 = 0.05317. The standard error for this difference if the two proportions are actually the same is estimated to be = 0.01524. The test statistic is therefore 0.05317/0.01524 = 3.49.
4. If the null hypothesis were true, this would be an observation from the Standard Normal distribution. This is because sample is large and 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 quite small, 0.0005.
6. We therefore conclude that the data are not consistent with the null hypothesis and 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.

What about the confidence interval? For this, we use a different standard error, the standard error when the proportions may not be equal. This is SE = 0.0188. The 95% confidence interval for the difference is 0.05317 – 1.96 × 0.0188 to 0.05317 + 1.96 × 0.0188 = 0.016 to 0.090. Although the difference is not very well estimated, it is well away from zero and gives us clear evidence that children with bronchitis reported in infancy are more likely than others to be reported to have respiratory symptoms in later life.

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.

Note that the standard error used here is different when the null hypothesis is true from that which is used for the confidence interval, when of course we do not say that there is no difference. The null hypothesis may contain information about the standard error and in the 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 and this alters the standard error for the difference. As a result, 95% confidence intervals and 5% significance tests sometimes give different answers near the cut-off point.

One-sided tests

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. A **one-sided** or **one-tailed** test considers the possibility of differences in one direction only. For the pronethalol example, we would have:

- 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.

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

The one-sided null hypothesis implies that an increase in attacks on pronethalol would have the same interpretation as no difference. This kind of interpretation is seldom true in health research. Biological interventions rarely produce only one kind of effect. If our treatment produces a disadvantage, we want to know about it and we want our statistical methods to detect it. Tests should be two sided unless there is a good reason not to do this.

Figure 3. One-sided test for the pronethalol study

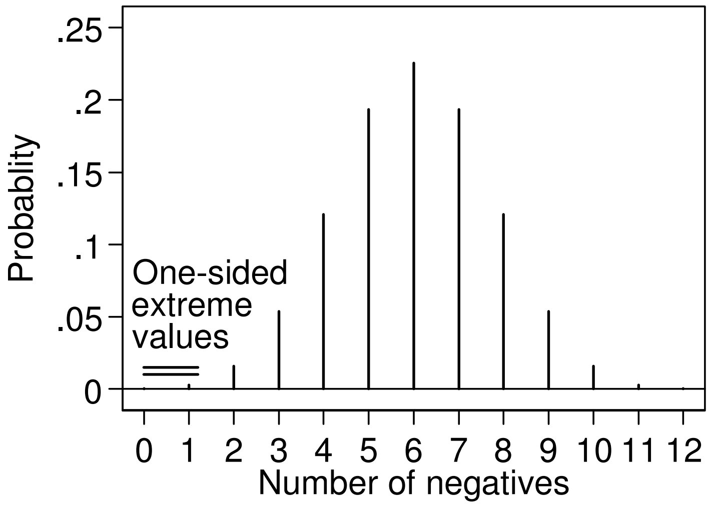
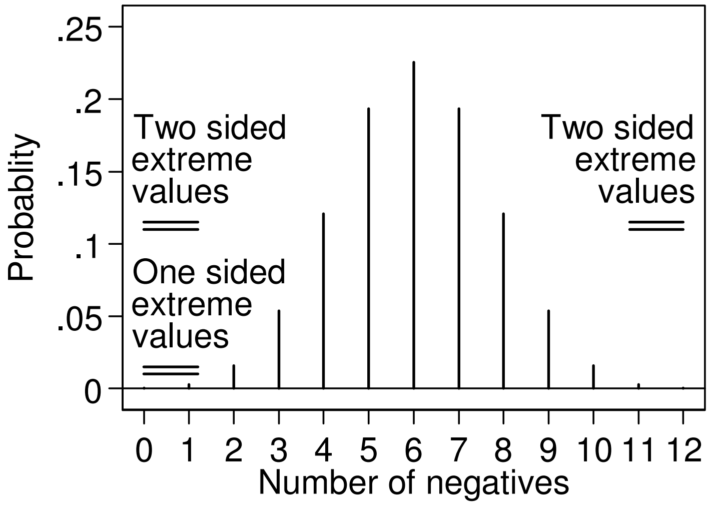


Figure 4. Two-sided test for the pronethalol study



J. M. Bland
15 August 2006

References

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

Pritchard BNC, Dickinson CJ, Alleyne GAO, Hurst P, Hill ID, Rosenheim ML, and Laurence DR. (1963) Report of a clinical trial from Medical Unit and MRC Statistical Unit, University College Hospital Medical School, London. *British Medical Journal* **2**, 1226-1227.