Hypothesis tests: the *t*-tests

Introduction

Invariably investigators wish to ask whether their data answer certain questions that are germane to the purpose of the investigation. It is often the case that these questions can be framed in terms of population parameters.

As an example, take a trial of various methods of pain relief following thoracotomy (*British Journal of Anaesthesia* 1995, 75, 19-22). As part of this trial one group received intrathecal Fentanyl for post-operative pain relief and another received intrathecal saline. The outcome variable we will consider is peak expiratory flow rate (PEFR) in litres/min at one hour after admission to the High Dependency Unit. This will be larger if pain relief is better. Not only is this a useful surrogate for pain relief, it has direct relevance as a good PEFR will reduce post-operative pneumostasis and associated chest infections. For the present it is assumed that PEFR has a Normal distribution in the two groups. If the mean of the Fentanyl group is μ_F and the mean of the saline group is μ_S, then interest may focus on the difference between these quantities. The challenge is how to make inferences about these essentially unknown quantities.

How Hypothesis Tests Work: The Null Hypothesis.

It is often the case that interest will focus on $\tau = \mu_F - \mu_S$, the difference in the mean PEFR between the two treatments. Although we do not know this difference, we do know two things that are relevant.

i) We know the difference between the treatment means in the groups treated in the trial, say m_F-m_S.
ii) We know (or at least can estimate) the standard error of this quantity.

The information in ii) allows the investigator to know how far the estimate in i) is likely to stray from the quantity of primary interest, namely τ. Judging the likely values of τ is something that could be done using a confidence interval, as described in the document 'Standard Errors and Confidence Intervals'. In this document we will consider an alternative approach, widely known as 'Significance testing' but more properly and more helpfully called 'Hypothesis testing'. However, there are close links between confidence intervals and hypothesis tests, as described in Appendix I.

If an investigator observes a difference between the saline and Fentanyl groups, perhaps noting that the mean PEFRs in the two groups are unequal, then the natural response is to explain this difference, ideally by concluding that there is a difference in the effectiveness of the treatments. However, the first possibility to try to exclude is that the difference might simply be due to chance. This is the question that hypothesis testing addresses.

A general observation can be made at this early stage, is that even if the hypothesis test concludes that a difference might well be due to chance, it does not follow that it

is due to chance. This point is important because it is often overlooked and serious misinterpretations arise as a consequence.

The first stage in the formulation of a hypothesis test is to be precise about what is meant by a difference 'being due to chance'. In the context of a comparison between two groups, it is that the two populations from which the two samples have been drawn are, in fact identical. The approach is to *assume* that the populations are identical and see what follows. This assumption is known as the *null hypothesis*. What *ought* to follow if the assumption is justified is that the two samples should be similar – in fact that they should only differ by sampling error. The technicalities of hypothesis testing centre on how to measure the difference between the samples and how to decide when a difference is surprisingly large. The procedure yields a probability, the *P-value*, which measures how likely it is that a difference as large as that observed could occur *if the null hypothesis were true*. The logic of the hypothesis test is that we then conclude that

Either:

i) we have seen an event that occurs with probability given by the P-value
or
ii) the null hypothesis is false.

If the P-value is small, then we may be disinclined to believe that i) has occurred and we opt for ii), i.e. we hold that the hypothesis test has provided evidence against the null hypothesis. The strength of this evidence is often measured by the size of the P-value: a value of $P<0.05$ is conventionally held to be reasonable evidence against the null hypothesis, with $P<0.01$ being strong evidence and $P<0.001$ being very strong evidence. Of course, these are rather arbitrary conventions. Another piece of terminology is that the *P*-value is the Type I error rate, being the probability of making the error of stating that there is a difference between the groups when there is no difference.

The mechanics of producing the *P*-value differ according to the type of outcome variable but the interpretation of the *P*-value itself is the same in all cases. In this document we will consider only the case when the variable has a Normal distribution.

Comparing the Fentanyl and saline groups: the unpaired t-test

The application of the above general discussion to this case requires the following.

i) Identification of the values of m_F-m_S that are likely *if the null hypothesis is true*.
ii) Use of this information to quantify how likely is the observed value of m_F-m_S.

If the null hypothesis is true then m_F-m_S has a Normal distribution with *zero mean*. The standard deviation of this distribution is simply the standard error of m_F-m_S, which is written SE for the moment. As was described in the document 'Standard Errors and Confidence Intervals', the values of m_F-m_S will tend to fall within a couple of standard errors of the mean. Thus, under the null hypothesis we would expect m_F-m_S, to be in the interval ±2SE: in other words we would expect the ratio (m_F-m_S)/SE to be, roughly speaking, between ±2.

This is, in essence, Student's *t*-test. It is based on the *t*-statistic which is the ratio

$$\frac{m_1 - m_2}{SE},$$

and large values of this are unlikely if the null hypothesis is true. The preceding discussion illustrates that if the null hypothesis is true then this ratio would generally have values of between ±2. There are, of course, many details which this heuristic introduction ignores. The main ones are:

a) precisely how likely are particular values of the above ratio;
b) how is the SE calculated;
c) are large positive and large negative values handled in the same way.

Issue c) will be left to Appendix II. The way in which the SE is computed depends on details of the structure of the data. There are two types of *t*-test, paired and unpaired *t*-tests and they differ by the way they compute the SE. The present example compares two groups of patients and this requires the unpaired version of the test. Further discussion of the distinction between the types of test and how to compute standard errors is in the next section.

A summary of the data from the trial is given below.

Group	Size	Mean PEFR (l/min)	SD (l/min)
Fentanyl	10	235	47
Saline	10	137	58

Table 1: summary of Fentanyl vs saline trial

As will be shown later, the value for the SE in this application is 23.76 l/min, and the observed difference in means is 235 - 137 = 98 l/min, so the *t*-statistic is 98/23.76 = 4.12. On the basis of the previous discussion we should expect this to indicate that an unusual event has occurred, or that the null hypothesis is false. Can we be more precise?

Figure 1 shows the plot of the distribution of the *t*-statistic if the null hypothesis is true. The observed value of the *t*-statistic is shown by the arrow. The shaded area shows the proportion of all *t*-statistics that are more extreme than the observed value and this is only 0.001 of all such statistics. This proportion is the *P*-value, in other words P=0.001. Notice that this proportion includes values larger than the observed value of 4.12 and smaller than –4.12. The reason for including the possible values less than –4.12 is linked to the distinction between one-sided and two-sided tests and is explained in Appendix II.

What does this *P*-value mean? It means that *if the null hypothesis is true* then a difference in mean PEFR as large or larger than 98 l/min would occur with probability 0.001. As this is a very small probability then it is more tenable to assume that the null hypothesis is false.

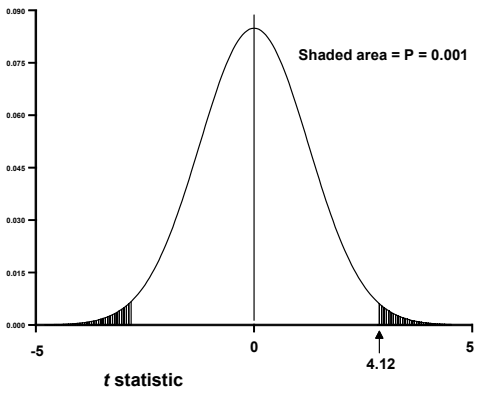

Figure 1: distribution of *t*-statistics if the null hypothesis is true.

What if a much larger *P*-value had been obtained, say *P*=0.45? In this case a difference in means as large or larger than that obtained would be quite likely to occur if the null hypothesis were true. In other words, if the null hypothesis were true, then this sort of value for the difference in means would be the sort of value you would expect to see quite often (45% of the time, in fact). Consequently, it is now not sensible to conclude that there is evidence against the null hypothesis. However, it is important to realise that this is not the same as asserting that the null hypothesis is true. The data are compatible with $\tau = \mu_F - \mu_S = 0$, but they are also compatible with a range of values around 0: this range is given by the confidence interval – see Appendix I.

Types of t*-test and their assumptions*

There are two types of *t*-test, known by a variety of names, such as paired and unpaired, one-sample and two-sample, dependent and independent. Both sorts test a null hypothesis that the mean of a population has a pre-specified value (usually 0). The difference between them depends on the structure of the data, which in turn is reflected in the way the standard error is calculated. However, in order to understand some of the manoeuvres involved it is useful to consider at the same time the assumptions made by the *t*-test.

Unpaired *t*-test

In this version of the test two quite unrelated samples are taken as the basis of the comparison. In particular this means that there is no basis on which a value in one sample can be associated with a corresponding value in the other sample. The above example comparing groups of patients given saline or Fentanyl uses an unpaired version of the test. This is because the study uses two groups, each comprising ten patients that are quite separate.

In these circumstances the *t*-test assumes:

I) that each sample is drawn from a Normal population;

II) these populations have a common standard deviation, σ

The null hypothesis that is most usually tested is that the means of these populations are the same. Several remarks are appropriate at this point.

a) If the populations are non-Normal, but the departure from Normality is slight then this violation of the assumptions is usually of little consequence.
b) The assumption of equal standard deviations may seem to constitute a rather severe restriction. In practice it is often reasonably close to true and, where it is not, it is often related to the data being skewed: this is discussed in a later lecture.
c) It should be remembered that a Normal distribution is characterised by its mean and standard deviation: a test that assessed the equality of means and paid no heed to the standard deviations would not be all that useful. It is often of interest to assess whether or not the samples are from the same population and assuming that the means are the same does not specify that the populations are the same unless the standard deviations are also the same.

If these assumptions seem justified the standard error SE referred to above should be computed. If the two the samples have sizes N and M then the standard error of the difference in means can be calculated by statistical theory to be

$$\sigma\sqrt{\frac{1}{M}+\frac{1}{N}},$$

where σ is the common standard deviation of the two Normal populations. The best estimate of σ draws on information from both samples and is a pooled estimate. This is described in most introductory tests but a detailed understanding of this step is not necessary, as most software will automatically compute the correct value. However, in some programs a questionable version of the *t*-test which does not assume equal standard deviations is sometimes used and it is important to ensure that the version using a pooled estimate is used. In the above example the pooled estimate of σ was 53.2 l/min, which is seen to be between the standard deviations of the separate samples.

Paired Test

In this version of the *t*-test two samples are compared but now there is a link between the samples. As the name suggests, there is a sense in which a value in one sample can be meaningfully associated with a corresponding value in the other sample. The following example will illustrate this.

Another aspect of the study comparing Fentanyl and saline was carried out on ten patients receiving conventional pain relief (PCA using morphine). The PEFR was measured on each of these patients on admission to the High Dependency Unit and an hour later. These figures are shown below (all values in l/min: data made available by kind permission of Dr I.D. Conacher, Freeman Hospital).

PEFR on admission to HDU	PEFR one hour post admission	Change
100	110	10
80	60	-20
180	160	-20
60	80	20
210	200	-10
130	80	-50
80	90	10
80	60	-20
120	80	-40
250	280	30
Table 2: data on PCA patients		

The important distinction between this set of data and the one in which two groups of patients were compared is that each observation in the 'post-admission' column is linked with an observation in the 'admission' column in a natural way. To be specific, the value of 110 l/min at the top of the second column is observed on the same individual as the value of 100 l/min seen at the top of the first column. It seems sensible that any appropriate analysis would acknowledge this structure in the analysis of these data. This structure would not be acknowledged if an unpaired *t*-test were used. Consider the following table, which reproduces the first two columns from the table above. The third column is a random re-ordering of the second column. If an unpaired test was used to compare the first two columns then the results would be identical to the results from comparing the first and third columns. It seems unreasonable that losing such an important aspect of the data in this way should have no effect.

PEFR on admission to HDU (I)	PEFR one hour post admission (II)	(II) randomly re-ordered
100	110	90
80	60	80
180	160	160
60	80	110
210	200	280
130	80	60
80	90	200
80	60	60
120	80	80
250	280	80
Table 3: re-ordered data from table 2		

The paired *t*-test takes account of the pairing by forming the differences within each pair, as shown in the third column of table 2. Once these differences have been formed the original observations can be discarded for the purposes of the test, as only the differences are analysed. If the mean PEFR is the same one hour after admission as on admission to HDU, then the differences will come from a population with zero mean. The paired *t*-test is a way of assessing whether a single sample of differences comes from a population with mean zero.

The paired *t*-statistic can be found from the formula

$$\frac{m_1 - m_2}{SE},$$

as before, although SE is computed differently. However, the usual way to compute the statistic as

$$\frac{\bar{d}}{SE}$$

where $\bar{d}$ is the mean of the sample of differences. The *SE* is simply the standard error of these difference computed as the standard deviation divided by the square root of the sample size, as explained in 'Standard Errors and Confidence Intervals'.

The value of the numerator, $\bar{d}$, is the same as that in the numerator of the unpaired *t*-test. The means of the three columns in table 2 are, respectively, 129, 120, -9 and the last of these is clearly the difference of the first two. However, the denominators of the paired and unpaired *t*-tests are not the same. The SE in the unpaired case is based on the standard deviation which measures the variation *between* the patients. The act of taking the difference between the two values of PEFR on the same patient would generally be expected to remove this source of variation from the data. This is well illustrated by the standard deviations of the three columns in table 2, which are, respectively, 64, 72 and 26. The standard deviation of the differences, being unaffected by inter-patient variation, is much lower than the other two values. The SE of the differences is $26/\sqrt{10} = 8.2$, so the paired *t*-test is $-9/8.2 = -1.09$ and this gives a *P*-value of 0.30.

This gives another reason for using the paired version of the test when it is appropriate. It is usually the case that the pairing in the data will lead to a more sensitive experiment but this advantage will be lost if it is not reflected in the arithmetic used for the analysis of the data. Only if the SE appropriate for the paired data is used will the correct precision be ascribed to the difference in means.

The assumptions underlying the paired *t*-test are simple. In fact there is only one: that the differences come from a Normal distribution. Note that it is not necessary to specify the distribution of the individual observations, just their difference. As with the unpaired case, modest departures from this assumption are not usually troublesome.

Appendix I: hypothesis testing and confidence intervals

One way of imprecisely describing hypothesis tests is to claim that they assess whether it is plausible that the sample to hand could have been drawn from a population with parameters satisfying the null hypothesis. On the other hand, in an equally imprecise way, confidence intervals provide a range of plausible values for the parameter of interest. These aims seem so similar that it is natural to ask if there is a more formal link between hypothesis tests and confidence intervals. The answer is yes – there are quite strong links between these two

entities. The link is perhaps best introduced by use of an example and the above application of the paired *t*-test is suitable.

The test of the null hypothesis that the mean PEFR was the same on admission and one hour post-admission gives P=0.3. A 95% confidence interval for the difference in mean PEFRs is (–27.6, 9.6) l/min. The hypothesis test yields a P value that indicates that an observed mean change in PEFR more extreme than that observed would occur in 30% of samples if the population mean change were 0. In other words the data are entirely compatible with a population mean change of 0. On the other hand, the confidence interval spans 0, again indicating that a population mean change of 0 is compatible with the observed data. In general, if 0 is included in a 95% confidence interval for the *t*-test then the associate hypothesis test will give $P>0.05$. Conversely, if $P<0.05$, then the 95% confidence interval will not include 0.

Of course, while the data may be compatible a population mean change of 0, they may also be compatible with other population mean changes. Although the details have not been covered above, it is entirely feasible to test the null hypothesis that the population mean change is x for any specified number x, not just 0. If you performed such a test and deemed that the data were compatible with a population mean change of x if you obtained $P>0.05$, then which values of x would be compatible with your sample? The answer is that the values within the 95% confidence interval are the values of x that are compatible with your data in this sense. So, for example, if you tested the null hypothesis that the population mean change in PEFR over the first hour after admission was x, for any x between –27.6 and 9.6 l/min, then every hypothesis test would yield $P>0.05$. Indeed this equivalence is true whether or not the confidence interval includes 0. The same equivalence applies between 99% confidence intervals and tests yielding $P>0.01$ and so on.

Appendix II: one and two sided tests. (Not Assessed)

In the calculation of the P-value illustrated by figure 1, not only was the probability of *t*-statistics larger than the observed value computed, the probability of values below the negative of the *t*-statistic was also taken into account. This is because the test being performed is a *two-sided* test. If the P-value only comprised the probability in one tail of the distribution in figure 1, the test would be known as a one-sided test. For a given value of the *t*-statistic the one-sided P-value is half of the two-sided P-value.

One-sided tests are not widely used. This unpopularity might seem surprising for a device which halves P-values. A one-sided test is appropriate if it is thought that the only way in which the null hypothesis could be discredited is if $\mu_F - \mu_S > 0$. In this case large negative values of the *t*-statistic could only occur by chance. While this is tempting, in practice it would seldom be prudent. An extreme difference between two treatments cannot be ignored simply because the direction of the difference is contrary to prior expectations. However, this is a necessary consequence of adopting a one-sided test.