As it is rarely possible to study an entire population, data are usually collected from a sample of individuals in order to make inferences about the population of interest. This can be done through a process known as hypothesis testing, the basic principles of which have been outlined in a previous tutorial. This note described how, at the outset, it is important to have a clear research question and know what the outcome variable to be compared is. Once the research question has been stated, the null and alternative hypotheses can be formulated. The null hypothesis (H₀) usually assumes that there is no difference in the outcome of interest between the study groups, whereas the study or alternative hypothesis (H₁) usually states that there is a difference between the study groups. Next, the appropriate statistical test must be selected and conducted to obtain a P-value. This P-value can then be used to make a decision about whether the results are statistically significant and thus whether the null hypothesis can be rejected.

This tutorial will provide examples of how the process of setting and testing a hypothesis is implemented in practice. It will focus on two elementary methods for analysing continuous, normally distributed data: the paired and unpaired t-tests. Continuous data can be measured and can take any value on the scale on which they are measured; examples include height, weight, blood pressure and area coverage.

### Choosing the statistical method

What type of statistical analysis to use depends on the answer to five key questions (table 1), and given answers to these, an appropriate approach to the statistical analysis of the data collected can be decided upon. The type of statistical analysis depends fundamentally on what the main purpose of the study is. In particular, what is the main question to be answered? The data type for the outcome variable will also govern how it is to be analysed, as an analysis appropriate to continuous data would be completely inappropriate for binary data. In addition to what type of datum the outcome variable is, its distribution is also important, as is the summary measure to be used. In this note we will describe methods that assume the data to be plausibly Normally distributed. Highly skewed data require a different analysis from data that are Normally distributed. Statistical tests that make an assumption about the distribution of the data, such as assuming Normality, are referred to as parametric tests. Tests that make no assumption are called non-parametric or distribution-free tests and will be covered in a later note. An important point to note is that it is the test that is parametric or non-parametric, not the data.

Before beginning any analysis, it is important to examine the data, using the techniques described in the first two tutorials in this series; an adequate description of the data should precede and complement the formal statistical analysis. For most studies and for randomised control trials (RCTs) in particular, it is good practice to produce
a table or tables that describe the initial or baseline characteristics of the sample.

The independent samples t-test is used to test for a difference in the mean value of a continuous variable between two independent groups. For example, as part of an RCT of two treatments for venous leg ulcers, one of the main questions of interest was whether there was a difference in the number of ulcer-free weeks between the control and the clinic groups. The number of ulcer-free weeks was continuous and was assumed to be sampled from a Normal distribution. As there were two independent treatment groups assessed in the trial, the most appropriate summary measures for the data were the sample means, with the difference in the mean ulcer-free weeks used to compare the two groups.

When conducting any statistical analysis, it is important to check that the assumptions which underpin the chosen method are valid. The assumptions underlying the two-sample t-test are outlined in table 2. The assumption of Normality can be checked by plotting two histograms, one for each sample; these do not need to be perfect, just roughly symmetrical. The two standard deviations should also be calculated and, as a rule of thumb, one should be no more than twice the other.

The test statistic for the independent samples t-test, t, is calculated as follows:

\[
(1) \quad t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} 
\]

where \(\bar{x}_1\) and \(\bar{x}_2\) are the means of the two groups and \(n_1\) and \(n_2\) are the numbers in the two groups and

\[
(2) \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}
\]

is an estimate of the pooled variance. Once this test statistic has been calculated, it can be compared to values for the \(t\)-distribution on \(n_1 + n_2 - 2\) degrees of freedom in order to obtain the \(P\)-value.

Note that the \(t\)-distribution is used for small sample sizes as an approximation to the Normal distribution. The \(t\)-values can take negative as well as positive values, and as the number of degrees of freedom gets larger, the \(t\)-distribution approaches the Normal distribution. Tables for the \(t\)-distribution can be found in statistical books or through a software package; in EXCEL, for example, probabilities from the \(t\)-distribution can be obtained using the TDIST function. This function requires three arguments, \(X\) = the \(t\)-statistic obtained above (for which the \(P\)-value is required), \(\text{deg}_\text{freedom} = n_1 + n_2 - 2\), and \(\text{tails}\), where 1 indicates a one-sided test and 2 indicates a two-sided test. It is recommended that you always use 2 here to ensure that your test is two-sided.

For the leg ulcer data, there were 120 patients in the clinic group and their mean number of ulcer-free weeks was 20.1. There were 113 patients in the control group and they had a mean number of ulcer-free weeks of 14.2. The pooled estimate of variance, \(s_p^2\), was 5.57. Putting these into equation (1) gives a \(t\)-statistic of 2.49 on 231 degrees of freedom, and this in turn gives a \(P\)-value of 0.014.

The final and most crucial stage of hypothesis testing is to make a decision, based upon this \(P\)-value. A \(P\)-value is the probability of obtaining the study results (or results more extreme) if the null hypothesis is true. It tells you how likely is the result obtained (from the study data), if there truly is no difference in the population. A ‘small’ \(P\)-value, say close to zero, indicates that the results obtained are unlikely when the null hypothesis is true and the null hypothesis is rejected. Alternatively, if the \(P\)
value is ‘large’, then the results obtained are likely when the null hypothesis is true and the null hypothesis is not rejected. Conventionally, the cut-off value or significance level for declaring that a particular result is statistically significant is set at 0.05 (or 5%). Thus if the \( P \)-value is less than this value, the null hypothesis (of no difference) is rejected and the result is said to be statistically significant at the 5% or 0.05 level. For the example above, for the difference in the number of ulcer-free weeks, the \( P \)-value is 0.014. As this is less than the cut-off value of 0.05, there is said to be a statistically significant difference in the number of ulcer-free weeks between the two groups at the 5% level.

### Comparison of paired observations: paired t-test

When there is more than one group of observations, it is vital to distinguish the case where the data are paired from that where the groups are independent. Paired data may arise when the same individuals are studied more than once, usually in different circumstances, or when individuals are paired, as in a case-control study. As part of the same leg ulcer trial described earlier, researchers were interested in assessing whether there was a change in health-related quality of life (HRQoL) between baseline and 3 months for those individuals with a healed leg ulcer (irrespective of study group). HRQoL at baseline and 3 months are both continuous variables, and the data are paired as measurements are made on the same individuals at baseline and 3 months; therefore, interest is in the mean of the differences, not the difference between the two means.

If we assume that the paired differences are Normally distributed, then the best comparative summary measure is the mean of the paired difference in HRQoL between baseline and 3 months for those individuals with a healed leg ulcer (irrespective of study group). HRQoL at baseline and 3 months are both continuous variables, and the data are paired as measurements are made on the same individuals at baseline and 3 months; therefore, interest is in the mean of the differences, not the difference between the two means.

The test statistic for the paired \( t \)-test is again \( t \) and is calculated as

\[
(3) \quad t = \frac{\bar{d}}{se(\bar{d})}
\]

where \( \bar{d} \) is the mean of the paired differences and \( se(\bar{d}) \) is the standard error of \( \bar{d} \) and is estimated as

\[
(4) \quad se(\bar{d}) = \frac{sd(\bar{d})}{\sqrt{n}}
\]

and \( n \) is the number of paired differences. As with the unpaired case, this \( t \)-statistic can then be compared to values for the \( t \)-distribution on \( n – 1 \) degrees of freedom. The mean change (3 months to baseline) in HRQoL for the 36 patients with healed ulcers was \(-7.33\) with an SD of 16.5, and using these in equation (3) gives a \( t \)-value of \(-2.661\), which in turn gives a \( P \)-value of 0.012. As this is less than the nominal level usually set for a statistical significance of 0.05 (or 5%), we can conclude that there is a statistically significant difference in HRQoL between baseline and 3 months. It is worth noting that as the mean change is negative, HRQoL actually declined for these patients over the 3 months!

### Summary

Outlined in this note are simple methods for comparing two groups of continuous data. However, it is important to bear in mind that statistical significance does not necessarily mean that the result obtained is clinically significant or of any practical importance. A \( P \)-value will only indicate how likely the results obtained are when the null hypothesis is true. Much more information, such as whether the result is likely to be of clinical importance, can be gained by calculating a confidence interval, as this a range of plausible values for the estimated quantity.1 In the next tutorial, we will extend the methods outlined above to cover more than two groups.

### References