Hypothesis Testing and Estimation

Jenny V. Freeman and Steven A. Julious
Medical Statistics Group, School of Health and Related Research, University of Sheffield, Community Sciences Centre, Northern General Hospital, Herries Road, Sheffield, UK.

**Introduction**

In the previous tutorial, we outlined the basic properties of the Normal distribution and discussed the Central Limit Theorem. The Normal distribution is fundamental to many of the tests of statistical significance covered in subsequent tutorials. As a result of the principles of the Central Limit Theorem, the Normal distribution enables us to calculate confidence intervals and make inferences about the population from which the sample is taken. In this tutorial, we explain the basic principles of hypothesis testing (using P-values) and estimation (using confidence intervals). By the end of the tutorial, you will know of the processes involved and have an awareness of what a P-value is and what it is not, and what is meant by the phrase ‘statistical significance’.

**Statistical analysis**

It is rarely possible to obtain information on an entire population, and usually data or information are collected on a sample of individuals from the population of interest. Therefore, one of the main aims of statistical analysis is to use this information from the sample to draw conclusions (‘make inferences’) about the population of interest.

Consider the hypothetical example of a study designed to examine the effectiveness of two treatments for migraine. In the study, patients were randomly allocated to two groups corresponding to either treatment A or treatment B. It may be that the primary objective of the trial is to investigate whether there is a difference between the two groups with respect to migraine outcome; in this case, we could carry out a significance test and calculate a P-value (hypothesis testing). Alternatively, it may be that the primary objective is to quantify the difference between treatments together with a corresponding range of plausible values for the difference; in this case, we would calculate the difference in migraine response for the two treatments and the associated confidence interval for this difference (estimation).

**Hypothesis testing (using P-values)**

Figure 1 shows the steps in the process of hypothesis testing. 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

![Figure 1. Hypothesis testing: the main steps.](image-url)
stated, the null and alternative hypotheses can be formulated. The null hypothesis \( (H_0) \) usually assumes that there is no difference in the outcome of interest between the study groups. The study or alternative hypothesis \( (H_1) \) usually states that there is a difference between the study groups.

In lay terms, the null hypothesis is what we are investigating, while the alternative is what we often wish to show. For example, when comparing a new migraine therapy against control, we are investigating whether there is no difference between treatments. We wish to prove that this null hypothesis is false and demonstrate that there is a difference at a given level of significance.

In general, the direction of the difference (e.g. that treatment A is better than treatment B) is not specified, and this is known as a two-sided (or two-tailed) test. By specifying no direction, we investigate both the possibility that A is better than B and the possibility that B is better than A. If a direction is specified, this is referred to as a one-sided test (one-tailed), and we would be evaluating only whether A is better than B, as the possibility of B being better than A is of no interest. It is rare to do a one-sided test, as they have no power to detect a difference if it is in the opposite direction to the one being evaluated. We will not dwell further on the difference between two-sided and one-sided tests other than to state that the convention for one-sided tests is to use a level of significance of 2.5% – half that for a two-sided test. Usually in studies it is two-sided tests that are done.

A common misunderstanding about the null and alternative hypotheses is that when a statistical test is being carried out, it is the alternative hypothesis (that there is a difference) that is being tested. This is not the case – what is being examined is the null hypothesis, that there is no difference between the study groups; we conduct a hypothesis test in order to establish how likely (in terms of probability) it is that we would have obtained the results that we have obtained, if there truly is no difference in the population.

For the migraine trial, the research question of interest is ‘for patients with chronic migraines, which treatment for migraine is the most effective?’

There may be several outcomes for this study, such as the frequency of migraine attacks, the duration of individual attacks or the total duration of attacks. Assuming we are interested in reducing the frequency of attacks, then the null hypothesis, \( H_0 \), for this research question is ‘there is no difference in the frequency of attacks between treatment A and treatment B groups’ and the alternative hypothesis, \( H_1 \), is ‘there is a difference in the frequency of attacks between the two treatment groups.’

Having set the null and alternative hypotheses, the next stage is to carry out a significance test. This is done by first calculating a test statistic using the study data. This test statistic is then compared to a theoretical value under the null hypothesis in order to obtain a \( P \)-value. The final and most crucial stage of hypothesis testing is to make a decision, based upon the \( P \)-value. In order to do this, it is necessary to understand first what a \( P \)-value is and what it is not, and then understand how to use it to make a decision about whether to reject or not reject the null hypothesis.

So what does a \( P \)-value mean? A \( P \)-value is the probability of obtaining the study results (or results more extreme) if the null hypothesis is true. Common misinterpretations of the \( P \)-value are that it is either the probability of the data having arisen by chance or the probability that the observed effect is not a real one. The distinction between these incorrect definitions and the true definition is the absence of the phrase ‘when the null hypothesis is true’. The omission of this phrase leads to the incorrect belief that it is possible to evaluate the probability of the observed effect being a real one. The observed effect in the sample is genuine, but what is true in the population is not known. All that can be known with a \( P \)-value is, if there truly is no difference in the population, how likely is the result obtained (from the sample). Thus, a small \( P \)-value indicates that difference we have obtained is unlikely if there genuinely was no difference in the population – it gives the probability of obtaining the study results (or results more extreme) (difference between the two study samples) if there actually is no difference in the population.

In practice, what happens in a trial is that the null hypothesis that two treatments are the same is stated, i.e. \( A = B \) or \( A - B = 0 \). The trial is then conducted and a particular difference, \( d \), is observed where \( A - B = d \). Owing to pure randomness, even if the two treatments are the same, you would seldom observe \( A - B = 0 \). Now if \( d \) is small (say a 1% difference in the frequency of attacks), then the probability of seeing this difference under the null hypothesis is very high, say \( P = 0.995 \). If a larger difference is observed, then the probability of seeing this difference by chance is reduced, say \( d = 0.05 \), and then the \( P \)-value could be \( P = 0.562 \). As the difference increases, therefore, so the \( P \)-value falls, such that \( d = 0.20 \) may equate to \( P = 0.021 \). This relationship is illustrated in figure 2: as \( d \) increases, the \( P \)-value (under the null hypothesis) falls.

It is important to remember that a \( P \)-value is a probability, and its value can vary between 0 and 1. 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.
But how small is small? Conventionally, the cut-off value or two-sided 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 (table 1). For the example above, if the P-value associated with the mean difference in the number of attacks was 0.01, we would say, as this is less than the cut-off value of 0.05, that there was a statistically significant difference in the number of attacks between the two groups at the 5% level.

The choice of 5% is somewhat arbitrary, and although it is commonly used as a standard level for statistical significance, its use is not universal. Even where it is used, one study that is statistically significant at the 5% level is not usually enough to change practice; replication is required. For example, to get a licence for a new drug, usually two statistically significant studies are required at the 5% level, which equates to a single study at the 0.00125 significance level. It is for this reason that larger ‘super’ studies are conducted to get significance levels that would change practice, i.e. a lot less than 5%.

Where the setting of a level of statistical significance at 5% comes from is not really known. Much of what we refer to as statistical inference is based on the work of R. A. Fisher (1890–1962), who first used 5% as a level of statistical significance acceptable to reject the null hypothesis. One theory is that 5% was used because Fisher published some statistical tables with different levels of statistical significance and 5% was the middle column. An exercise that we do with students in order to demonstrate empirically that 5% is a reasonable level for statistical significance is to toss a coin and tell the students whether we have observed a head or a tails. We keep saying heads. After about six tosses, we ask the students when they stopped believing we were telling the truth. Usually, about half would say after four tosses and half after five. The probability of getting four heads in a row is 0.063, and the probability of getting five heads in a row is 0.031; hence 5% is a level around which most people would intuitively start to disbelieve a hypothesis!

Although the decision to reject or not reject the null hypothesis may seem clear-cut, it is possible that a mistake may be made, as can be seen from the shaded cells of table 2. For example, a 5% significance level means that we would only expect to see the observed difference (or one greater) 5% of the time under the null hypothesis. Alternatively, we can rephrase this to state that even if the two treatments are the same 5% of the time, we will conclude that they are not and we will make a type I error. Therefore, whatever is decided, this decision may correctly reflect what is true in the population: the null hypothesis is rejected, when in fact it is false, or the null hypothesis is not rejected, when in fact it is true. Alternatively, it may not reflect what is true in the population: the null hypothesis may be rejected, when in fact it is true, which would lead us to a false positive and making a type I error (α); or the null hypothesis may not be rejected, when in fact it is false. This would lead to a false negative and making a type II error (β). Acceptable levels of the type I and type II error rates are set before the study is conducted. As mentioned above, the usual level for declaring a result to be statistically significant is set at a two-sided level of 0.05 prior to an analysis; that is, the type I error rate (α) is set at 0.05 or 5%. In doing this, we are stating that the maximum acceptable probability of
rejecting the null hypothesis when it is in fact true (committing a type I error, \( \alpha \) error rate) is 0.05. The \( P \)-value that is then obtained from our analysis of the data gives us the probability of committing a type I error (making a false-positive error).

The probability that a study will be able to detect a difference, of a given size, if one truly exists is called the power of the study and is the probability of rejecting the null hypothesis when it is actually false (probability of making a type II error, \( \beta \)). It is usually expressed in percentages, so for a study which has 90% power, there is a probability of 0.9 of being able to detect a difference, of a given size, if there genuinely is a difference in the population. An underpowered study is one which lacks the ability, i.e. has very low power, to detect a difference when there truly is a difference. The concepts of power and type I and II errors will be dealt with further in a later tutorial on sample size, as these are important components of sample size calculation.

### Estimation (using confidence intervals)

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. It can only be used to decide whether the results are statistically significant or not; it does not give any information about the likely size of the clinical difference. Much more information, such as whether the result is likely to be of clinical importance, can be gained by calculating a confidence interval. Although in the previous tutorial we talked about the 95% confidence interval for the mean, it is possible to calculate a confidence interval for any estimated quantity (from the sample data), such as the mean, median, proportion, or even a difference. It is a measure of the precision (accuracy) with which the quantity of interest is estimated (in the case of the migraine trial, the quantity of interest is the mean difference in the number of migraine attacks).

Technically, the 95% confidence interval is the range of values within which the true population quantity would fall 95% of the time if the study were to be repeated many times. Crudely speaking, the confidence interval gives a range of plausible values for the quantity estimated; although not strictly correct, it is usually interpreted as the range of values within which there is 95% certainty that the true value in the population lies. For the migraine example, let us assume that the quantity estimated, the mean difference in the number of attacks between the groups, was 3 attacks per month, and the 95% confidence interval for this difference was 1.2 to 4.8 attacks per month. Thus, while the best available estimate of the mean difference was 3 attacks per month, it could be as low as 1.2 or as high as 4.8 attacks per month, with 95% certainty. As the confidence interval excludes 0, we can infer from the observed trial that it is unlikely that there is no difference between treatments. In fact, as we have calculated a 95% confidence interval, we can deduce that the statistical significance is less than 5%. The actual \( P \)-value associated with this difference was 0.01, and given that it is less than 5%, we can conclude that the difference is statistically significant at the 5% level.

As confidence intervals are so informative, and from them we can infer statistical significance as well as quantify plausible values for the population effect, there is a growing consensus that only confidence intervals should be reported for studies. However, it is unlikely that \( P \)-values will ever be eliminated as a way to quantify differences.

### Statistical and clinical significance

So far in this tutorial, we have dealt with hypothesis testing and estimation. However, in addition to statistical significance, it is useful to consider the concept of clinical significance. While a result may be statistically significant, it may not be clinically significant (relevant/important), and conversely, an estimated difference that is clinically important may not be statistically significant. For example, consider a large study comparing two treatments for high blood pressure; the results suggest that there is a statistically significant difference (\( P = 0.023 \)) in the amount by which blood pressure is lowered. This \( P \)-value relates to a difference of 3 mmHg between the two treatments. While the difference is statistically significant, it could be argued that a difference of 3 mmHg is not clinically important. This is supported by the 95% confidence interval of 2.3–3.7 mmHg. Hence, although there is a statistically significant difference, this difference may not be sufficiently large to convince anyone that there is a truly important clinical difference.

This is not simply a trivial point. Often in presentations or papers, \( P \)-values alone are quoted and inferences about differences between groups are made based on this one statistic. Statistically significant \( P \)-values may be masking the fact that the differences have little clinical importance. Conversely, it may be possible to have a \( P \)-value greater than the magic 5% but for there to be a genuine difference between groups: absence of evidence does not equate to evidence of absence.

### Summary

In this tutorial, we have outlined the basic principles of hypothesis testing (using \( P \)-values) and estimation (using confidence intervals). In subsequent tutorials, we will be applying this knowledge when performing statistical significance testing in order to make decisions about the results of analyses.