In comparing or assessing methods of treatment it is vital that the appropriate number of patients is selected in order to ensure that the conclusions drawn are statistically viable. This annotation describes the relevance of a statistical power analysis in the context of hypothesis testing to the determination of the optimal sample size of a study. The power of the test indicates how likely it is that the test will correctly produce a statistically significant result.

In statistical testing of a hypothesis,\(^1,2\) we use sample data to make a decision about whether we have enough evidence to reject a statistical hypothesis about the population, usually stated as a null hypothesis, in favour of an alternative hypothesis. For example, let us consider a recent randomised controlled trial which compared the outcome of two groups of patients with an intertrochanteric fracture of the hip treated either with a percutaneous compression plate (PCCP) or a conventional compression hip screw (CHS).\(^3\) The null hypothesis is that, in the population, the mean operative blood loss is the same in the two groups. The usual approach is to reject the null hypothesis if the p-value is less than some cut-off value, typically 0.05. The p-value is the probability of obtaining the observed or a more extreme result if the null hypothesis is true, and the cut-off for significance is called the significance level.

When we make a decision, based on the magnitude of the p-value, to reject or not reject a null hypothesis, we have to accept that, because we are dealing with sample data, our decision may be wrong. There are two possible mistakes that we can make. A Type I error occurs when we incorrectly reject the null hypothesis. A Type II error occurs when we incorrectly do not reject the null hypothesis.

Clearly, we want the chances of making these errors to be minimal. The chance or probability of making a Type I error is the p-value that we obtain from the test, and the maximum chance of a Type I error is the significance level. We therefore limit the chance of making a Type I error by fixing the significance level at the outset. Rather than consider the chance of making a Type II error, thereby not rejecting the null hypothesis when it is false, we usually focus on the chance of rejecting the null hypothesis when it is false. This is called the power of the test. It may be expressed as a probability with a value from zero to one or, more usually, as a percentage.

The factors that affect power
A number of factors affect the power of a study. We are more likely to detect a real effect as statistically significant (i.e., the power of the study will be greater) if any one of the following is true; namely if the significance level is higher, e.g. it is 0.05 instead of 0.01, if the data exhibit less variability, if the sample size is larger, or if the important effect size (e.g. the difference in treatment means) is greater. The important effect size is the smallest effect that is clinically or biologically important, and which the investigators would not want to overlook.

The effect of power
We should design our study in the knowledge that we have sufficient power to detect, as statistically significant, a treatment effect of a given size. If the power is too low, usually when the sample size is inadequate, we may fail to detect the treatment effect as statistically significant when there really is an effect. We will have wasted our time, money, labour and patients, and could produce a study which is ethically unacceptable. Generally, the power of a study is not deemed acceptable unless it exceeds about 80%.

However, if the sample size is unduly large, in which case the power is likely to be very high, the study may be unnecessarily time-
The power statement in sample size determination

The factors that affect power affect sample size in a similar fashion since power and sample size are directly related. Therefore, if we are to use a power analysis to determine the optimal sample size of a study in which we intend to test a particular hypothesis, we must specify the required power as well as the values of each of these factors, namely the proposed significance level, the relevant effect size and, if the data are numerical, the expected variability in the data. We can then incorporate these values into special statistical tables, equations, computer programs (e.g., nQuery Advisor v. 7.0, Statistical Solutions, Cork, Ireland) or diagrams to determine the optimal sample size for a given hypothesis test (e.g., the two-sample t-test to compare two means). The justification for the recommended sample size should be contained in a power statement which documents the values of all the factors that were used in the calculation. Such a power statement should always be included in publications, grant applications, ethical approval applications, etc.

For example, consider again the randomised controlled trial to compare the blood loss in patients with an intertrochanteric fracture of the hip treated with either PCCP or CHS. Use of statistical tables for example, would result in the following power statement: using a two-sample t-test with 80% power at the 5% level of significance, 29 patients would be required in each treatment group in order to detect a difference in mean blood loss of 150 ml between them, assuming the SD of blood loss in each group is about 200 ml. A simple approach to evaluating the sample size in these circumstances is to use Lehr's formula, which states that, when the power is 80% and the significance level is 0.05, the number required in each group is approximately equal to 16/(important treatment difference/SD)^2 = 16/(150/200)^2 = 28.4 which, when rounded up, equals 29. Note that the sample size in each group increases to 44 if the significance level is reduced from 0.05 to 0.01, and to 39 if, instead, the power is increased from 80% to 90%.

Retrospective power analysis

Occasionally we want to know whether we can attribute the result of a non-significant hypothesis test to an inadequately powered study. If we are to perform such a post hoc power calculation, we use the same approach as for the determination of sample size but modify it by evaluating the power for the given sample size, instead of vice versa. However, in this calculation, we must not, use the treatment effect (e.g. the difference in means) that we have observed in our sample. Instead, we must be sure to base the calculation on the clinically important treatment effect that was specified at the outset, before the data were collected.

Glossary

**Alternative hypothesis.** In statistical hypothesis testing, if there is evidence to reject the null hypothesis stating that there is no effect in the population, it is rejected in favour of the alternative hypothesis. Typically the test is two-sided, so that the alternative hypothesis assumes an effect in the population but the direction of the effect is unspecified. For example, if the null hypothesis is that two population means are equal, the alternative hypothesis is that they are not equal.

**Chi-squared test.** This is a statistical hypothesis test which tests a null hypothesis that relates to categorical data. For example, it may be used to test the null hypothesis that two or more proportions are equal, or that there is no association between the factors that define a contingency table.

**Confidence interval.** This is loosely interpreted as the range of values that is believed to encompass the true population parameter with a specified degree of certainty, typically 95%. A strict interpretation is that, after repeated sampling, 95% of the confidence intervals so determined will contain the parameter. The confidence interval provides a measure of precision of the estimated value of the parameter. Wide intervals are indicative of poor precision.

**Contingency table.** This is a table of frequencies which shows how individuals are classified into different categories of two or more factors. In a two-way contingency table with two factors of interest, the rows represent the different categories of one factor, the columns represent the different categories of the other factor and the entries in the cells of the table are frequencies.

**Effect of interest.** This is the value of the response or the outcome variable that reflects the comparison of interest (e.g., the difference in two population means).

**Effect size.** This may be the effect of interest but, sometimes, it refers to a standardised measure, such as the difference in two population means divided by the SD of one or other of the populations.

**Hypothesis testing.** This is a formalised statistical method of deciding whether or not there is enough evidence to reject a null hypothesis on the basis of a set of sample observations.

**Important effect size.** This is the smallest treatment effect, such as the difference in means, which is clinically or biologically meaningful.

**Null hypothesis.** In statistical hypothesis testing, this is a statement that relates to the population and assumes no effect, e.g. that two population means are equal.

**Paired t-test.** This is a statistical hypothesis test that tests the null hypothesis that the mean of the differences obtained from a population of paired data is equal to zero.

**Parameter.** This is a summary measure, such as the mean, SD or proportion, which describes a particular feature of a distribution. Its value relates to the population.

**Power.** This is the probability that a treatment effect of a given magnitude will be detected as statistically significant if such an effect actually exists. It is the probability of not making a Type II error which is equal to 1 - the probability of making a Type II error.
Power analysis. This determines the optimal sample size of a study: it requires a specification of the values of the factors that affect the sample size, such as the power, level of significance and the important treatment effect, and an indication of the proposed method of statistical analysis of the data.

Power statement. This provides justification for the proposed sample size of a study by specifying the values of the factors that affect sample size, such as the power, treatment effect and SD of the observations, and indicating the intended method of statistical analysis of the data.

p-value. This is the probability of obtaining the observed results, or more extreme results, if the null hypothesis about the population is true.

Significance level. This is the probability, chosen at the outset of an investigation, which leads to the rejection of the null hypothesis if the p-value obtained from the relevant statistical hypothesis test lies below it. The significance level is often 0.05, in which case the null hypothesis is rejected if $p < 0.05$.

Test statistic. This is a specific formula which is relevant to a particular hypothesis test. When sample data are substituted into the formula, its value is compared with a known probability distribution such as the normal, $t$, chi-squared or $F$-distributions to obtain a p-value. The null hypothesis under test will be rejected if this p-value is less than the significance level, typically 0.05.

Treatment effect. This is the effect of interest that affords a comparison between different treatment groups. For example, it may be the relative risk for binary outcomes, or the difference in means for numerical outcomes. The treatment effect is often referred to as the effect of interest, or sometimes, the effect size, particularly if the comparison groups are not defined by treatments.

Two-sided (-tailed) test. This is a statistical hypothesis test in which the direction of the effect of interest is not specified in the alternative hypothesis.

Two-sample (unpaired) t-test. This is a statistical hypothesis test which tests the null hypothesis that the means from two independent populations are equal.

Type I error. This occurs in a statistical hypothesis test when the null hypothesis is rejected when in fact it is true.

Type II error. This occurs in a statistical hypothesis test when the null hypothesis is not rejected when in fact it is false.

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

References