Type I and Type II errors: what are they and why do they matter?

Introduction

Hypothesis tests are commonly used when we wish to make a decision. For example, we may wish to decide whether a new intervention results in an improved patient outcome compared with the gold standard treatment. In this setting, Type I and Type II errors are fundamental concepts to help us interpret the results of the hypothesis test.[1] They are also vital components when calculating a study sample size.[2,3] We have already briefly met these concepts in previous *Research Design and Statistics* articles[2,4] and here we shall consider them in more detail.

Hypothesis tests

Type I and Type II errors can be defined once we understand the basic concept of a hypothesis test. As we have seen previously,[4,5] here we construct a null hypothesis and an alternative hypothesis. The null hypothesis is our study 'starting point'; the hypothesis against which we wish to find sufficient evidence to be able to *reject* or *disprove* it. Typically the null hypothesis is the hypothesis of no effect, or no difference between the arms of the study. The alternative hypothesis is the alternative scenario that we believe will be true should we find sufficient evidence to reject the null hypothesis.

So, for example, we may be investigating the efficacy of a new drug compared with a standard of care treatment. Our null hypothesis may therefore be 'the true percentage responding to the new drug is the same as the percentage responding to the standard of care drug'. The alternative hypothesis reflects the alternative scenario that we wish to consider. For example, our alternative hypothesis perhaps could be 'there is a true difference in the percentage responding to the new drug compared with the percentage responding to the standard of care'.

We then collect data and conduct a hypothesis test. We use the results of this test to determine whether we have sufficient evidence to reject the null hypothesis, and therefore conclude that our new drug leads to an improved patient outcome. We usually do so by calculating a P value based on our data and determining whether it is below a predetermined probability cut-off (e.g. <0.05).[5]

Statistical errors in hypothesis testing

However, underlying this result, there will be a 'truth'; either the new drug truly has equivalent efficacy to the standard of care or it does not. We can therefore find ourselves in four different situations:

(1) The new drug truly has a benefit, and we correctly come to this conclusion, based on the results of our hypothesis test;
(2) The new drug truly has a benefit, but we incorrectly conclude that there is no evidence of a benefit, based on the results of our hypothesis test;
(3) A new drug truly has no benefit, and we correctly come to this conclusion based on the results of our hypothesis test;
(4) A new drug truly has no benefit, but we incorrectly conclude that the new drug is beneficial, based on the results of our hypothesis test.

Thus, there are two situations in which we have come to the correct conclusion (scenarios 1 and 3) and two situations in which we have made an error (scenarios 2 and 4). Scenario 4 is known as a *Type I* error and scenario 2 is known as a *Type II* error.

Type I errors

A Type I error is defined as incorrectly concluding that there is a difference between groups when none truly exists.[1] This error is also frequently known as α (alpha). The size of this error that we are willing to accept is typically fixed prior to conducting the hypothesis test. It becomes our predetermined probability cut-off to decide whether to reject the null hypothesis or not. The most common level taken for a Type I error is at 5% (i.e. we will accept a $P < 0.05$ as sufficient evidence to reject the null hypothesis). However, it can of course be fixed at any level, such as at 1%, 0.1% or 0.01%, depending on the exact situation, and to which extent we wish to minimize the chances of making this error. In fact, the only way that we can minimize the chance of making a Type I error is by requiring a smaller P value to be observed before making the decision to reject the null hypothesis.

Type II errors

A Type II error is defined as incorrectly concluding that there is no difference between groups when in fact a difference truly exists.[1] This error is also frequently known as β (beta). The probability of making a Type II error is usually strongly influenced by the study sample size and, for continuous variables, by the amount of variability present in the data.[3] Therefore, we can usually minimize the chances of making a Type II error by increasing the number of individuals included in our study, or by reducing the variability in our data if at all possible (for example, by reducing assay variability). An article by Qualls *et al.*[6] also demonstrates that using a parametric rather than non-parametric statistical test when data are not normally distributed can also increase the chance of making a Type II error.

Conclusions

I hope that this review has given a brief explanation of hypothesis tests, and the types of errors that can be made. These hypothesis tests are particularly useful when we wish to make a decision, such as investigating the efficacy of a new drug compared with a gold standard. This approach requires examining the strength of evidence in our study to determine whether to reject the null hypothesis or not (usually by determining whether the *P* value is above or below a predefined value). In practice, however, we usually complement this approach by also considering the strength of the evidence against the null hypothesis. We do so by calculating the *P* value rather than just considering whether it is above or below a cut-off.[7] For example, *P* values of 0.049 and 0.0001 may be interpreted differently in practice as there is much stronger evidence to be able to reject the null hypothesis in the latter case, despite both being below a 0.05 cut-off.

C J Smith PhD
Research Department of Infection and Population Health, UCL, Royal Free Campus, Rowland Hill Street, London NW3 2PF, UK
Email: c.smith@ucl.ac.uk

DOI: 10.1258/phleb.2012.012j04

References

1 Swinscow TDV, revised by MJ Campbell. *Statistics at Square One*. Chapter 5 Differences Between Means: Type I and Type II Errors and Power… Chichester, UK: BMJ Publishing Group, 1997
2 Smith CJ. How many patients do I need? Sample size and power calculations. *Phlebology* 2011;**26**:44–5
3 Machin D, Campbell MJ, Fayers PM, Pinol APY. *Sample Size Tables for Clinical Studies*. 2nd edn. Oxford, UK: Blackwell, 1997
4 Smith CJ, Fox ZV. The use and abuse of hypothesis tests: how to present *P* values. *Phlebology* 2010;**25**:107–12
5 Bland JM. *An Introduction to Medical Statistics*. 3rd edn. Oxford, UK: Oxford Medical Publications
6 Qualls M, Pallin DJ, Schuur JD. Parametric versus non-parametric statistical tests: the length of stay example. *Acad Emerg Med* 2010;**17**:1113–21
7 Fisher RA. The arrangement of field experiments. *J Minist Agric GB* 1926;**33**:503–13