STATISTICAL BRIEFING: TYPE 1 AND TYPE 2 ERRORS

CHRISTOPHER R. LAMB

WHEN CONSIDERING THE possibility that there is a difference between two or more groups, the *alternative hypothesis* states that there is a difference and the *null hypothesis* states that there is no difference. Statistical tests generally produce results stating the likelihood that the null hypothesis is true. Depending on the results of our tests, we may decide to retain or reject the null hypothesis, and any result that enables us to reject it represents evidence for the alternative hypothesis.[1]

Just as there are four possible outcomes (true positive, false positive, true negative, false negative) when using a diagnostic test, there are four possible outcomes of hypothesis testing:

	Decision	
Null Hypothesis	Reject Null Hypothesis	Retain Null Hypothesis
False	Correct	Type 2 error
True	Type 1 error	Correct

Rejecting the null hypothesis when it is true is known as a Type 1 error. An example of this would be concluding that there is a significant difference between two groups when the difference observed was in fact due to random variation. Retaining the null hypothesis when it is false is known as a Type 2 error. An example of this would be concluding that there is no difference between groups when too small a sample has been collected to reveal a difference that truly exists.

If we collect data and analyze it as a means of looking for evidence to support our alternative hypothesis, we can generally obtain a result in the form of a test statistic. Attached to each value of a test statistic is a probability, called the *P*-value, which describes the chance of getting this result if the null hypothesis was true (Type 1 error).

Generally, an arbitrary probability of 0.05 (5%) is chosen as the cut-off between "significant" and "not significant" results: results associated with *P*-values < 0.05 are unlikely to have been obtained as a result of random variation, hence the result is considered significant and the null hypothesis may be rejected; whereas results with *P*-values > 0.05 are deemed insufficient evidence (not significant) to reject the possibility that the null hypothesis is true.

Use of the word significant when describing the interpretation of test results does not necessarily imply that the result is real or that it is important. It is possible to find significant results associated with a difference that is real, but too small to be clinically important. Conversely, it is possible for a difference to be real and important but statistically insignificant because of some deficiency in study methodology. "Not significant" does not necessarily mean "no difference".[2]

When there really is a difference between groups, we should be able to detect it. For this to occur, the likelihood of a Type 2 error must be below a certain level, usually taken to be 0.2 (20%). This makes the probability of detecting a significant difference equal to $1–0.2 = 0.8$, which means there is an 80% chance of detecting a difference if one exists. This probability is known as the statistical power.

Knowing the statistical power becomes particularly important when the result of a study is negative because there is a need to distinguish a true negative result (i.e., there really is no statistically significant difference) from a false negative result (i.e., there is an underlying difference, but the study was not powerful enough to find it).[3] It is normal practice to determine the power of a study before any data are collected in order to recognize when a lack of power could be an issue, and, if possible, to take steps to increase the power of the study.

REFERENCES

1. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 1. Hypothesis testing. Can Med Assoc J 1995;152:27–32.

2. Altman DG, Bland JM. Statistics notes: absence of evidence is not evidence of absence. BMJ 1995;311:485.

3. Kuzon WM Jr., Urbanchek MG, McCabe S. The seven deadly sins of statistical analysis. Ann Plast Surg 1996;37:265–272.