Clin Orthop Relat Res (2011) 469:619–620
DOI 10.1007/s11999-010-1435-0

In Brief

Statistics in Brief

Statistical Power: What Is It and When Should It Be Used?

Frederick J. Dorey PhD

Published online: 29 June 2010

Background

Although any report formally testing a hypothesis should include an associated p value and confidence interval, another statistical concept that is in some ways more important is the power of a study. Unlike the p value and confidence interval, the issue of power should be considered before even embarking on a clinical study.

Question

What is statistical power, when should it be used, and what information is needed for calculating power?

Discussion

Like the p value, the power is a conditional probability. In a hypothesis test, the alternative hypothesis is the statement that the null hypothesis is false. If the alternative hypothesis is actually true, the power is the probability that one will correctly reject the null hypothesis. The most meaningful application of statistical power is to decide before initiation of a clinical study whether it is worth doing, given the needed effort, cost, and in the case of clinical experiments, patient involvement. A hypothesis test with little power will likely yield large p values and large confidence intervals. Thus when the power of a proposed study is low, even when there are real differences between treatments under investigation, the most likely result of the study will be that there is not enough evidence to reject the H_0 and meaningful clinical differences will remain in question. In that situation a reasonable question to ask would be, was the study worth the needed time and effort to get so little additional information.

The usual question asked involving statistical power is: what sample size will result in a reasonable power (however defined) for the primary hypothesis being investigated. In many cases however, a more realistic question would be: what will the statistical power be for the important hypothesis tests, given the most likely sample size that can be obtained during the duration of the proposed study?

For any given statistical procedure and significance level, there are three statistical concepts closely related to each other. These are the sample size, effect size, and power. If you know any two of them, the third can be determined. To determine the effect size the investigator first must estimate the magnitude of the minimum clinically important difference (MCID) that the experiment is designed to detect. This value then is divided by an estimate of the variability of the data as interpretation of numbers only makes sense relative to the variability of the estimated parameters. Although investigators usually can provide a reasonable estimate of the MCID for a study, they frequently have little idea about the variability of their data. In many cases the standard deviation of the control group will provide a good estimate of that variability. As intuitively it should be easier to determine if two groups differ by a large rather than a small clinically meaningful difference, it follows that a larger effect size usually will result in more

Each author certifies that he or she has no commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.

F. J. Dorey (✉)
Department of Pediatrics, Children's Hospital Los Angeles, 4650 Sunset Blvd, Mailstop 54, Los Angeles, CA 90027, USA
e-mail: fdorey@chla.usc.edu

power. Also, a larger sample size results in more precision of the parameters being estimated thus resulting in more power as the estimates are more likely to be closer to the true values in the target population. (A more-detailed article by Biau et al. [1] discusses the relationships between power and sample size along with examples.)

For power calculations to be meaningful, it first is necessary to decide on the proper effect size. The effect size must be decided first because, for any proposed sample size, an effect size always can be chosen that will result in any desired power. In short, the goals of the experiment alone should determine the effect size. Once a study has been completed and analyzed, the confidence interval reveals how much, or little, has been learned and the power will not contribute any meaningful additional information. In a detailed discussion of post hoc power calculations in general, Hoenig and Heisey [2] showed that if a hypothesis test has been performed with a resulting p value greater than the 5% significance level, then the power for detecting the observed difference will only be approximately 50% or less. However, it can be verified easily with examples that hypothesis tests resulting in very small p values (such as 0.015) could still have a post hoc power even less than 70%; in such a case it is difficult to see how a post hoc power calculation will contribute any more information than what already is known.

There is a very nice relationship between the concepts of hypothesis testing and diagnostic testing. Let the null hypothesis represent the absence of a given disease, the alternative hypothesis represent the presence of the disease, and the rejection of the null hypothesis represent having a positive diagnostic test. With these assumptions, the power is simply equivalent to the sensitivity of the test (the probability the test is positive when the disease is present). In addition, the significance level is equivalent to one minus the specificity of the test, or in other words, the error you are willing to risk of falsely rejecting the null hypothesis simply corresponds to the probability of getting a positive test among patients without the disease.

Myths and Misconceptions

As discussed above the notion of power after the data have been collected does not provide very much additional information about the hypothesis test results. This is illustrated by considering the experiment of flipping a coin 10 times to see if the coin is fair, that is, the probability of heads is 0.5. Suppose you flip the coin 10 times and you get 10 heads. This experiment with only 10 flips has very little power for testing if the coin is fair. However the p value for obtaining 10 heads in 10 flips with a fair coin (the null hypothesis) is very small, so the null hypothesis certainly will be rejected. Thus, even though the experiment has little power, it does not change the fact that an experiment has been conducted and provided convincing evidence that the coin is biased in favor of heads. I do not recommend that you bet on tails.

Another myth is that the power always has to be at least 80% or greater. That might be a reasonable expectation for a clinical study potentially involving great inconvenience or risk to patients. However in a laboratory study or a retrospective correlation study, there is usually no necessity for the power to be that high.

Conclusions

The concept of statistical power should be used before initiating a study to help determine whether it is reasonable and ethical to proceed with a study. Calculation of statistical power also sometimes is useful post hoc when statistically insignificant but potentially clinically important trends are noted, say in the study of two treatments for cancer. Such post hoc tests can inform the reader or future researchers how many patients might be needed to show statistical differences. The power and effect size needed for a study to be reasonable also will depend on the medical question being asked and the information already available in the literature.

References

1. Biau DJ, Kernéis S, Porcher R. Statistics in brief: the importance of sample size in the planning and interpretation of medical research. *Clin Orthop Relat Res*. 2008;466:2282–2288.
2. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat.* 2001;55: 19–24.