

Statistical Power—So What?!

AS CLINICIANS WE read a variety of research articles related to our clinical practice. If you are anything like me, you often come across articles that mention the words *statistical power*. What is statistical power and why is it important to me as a clinician? The purpose of this issue's column is to shed some light on this concept in a practical, user-friendly format.

Most research studies conducted in our profession are designed to assess differences between or among groups. For example, a study designed to test the effects of various ankle braces on athletic performance between football and soccer players would compare performance measurements (e.g., 40-yd-dash times, vertical-jump height) between the braces and different athletes. After collecting and analyzing the data, the researchers would attempt to draw some conclusions from the statistical findings. Some inference is then made based on these conclusions with respect to the general population or intended target group (i.e., athletes, patients).

Readers of medical literature should be careful in interpreting research findings, especially for studies that fail to report statistical power. Greenfield et al. state that it is imperative to calculate statistical power and sample size during the design phase of a study.¹ Statistical power is the probability that a significant difference between or among groups will be detected with a statistical test. In statistics, power is the ability to avoid a Type II error,¹ which occurs when researchers accept the null hypothesis when they should not have. An example of a Type II error would be concluding that strengthening the peroneal muscles makes no difference in minimizing ankle sprains, when in fact it is indeed beneficial. Table 1 shows a diagram of a truth table for the null hypothesis and should serve as a nice review of hypothesis testing. A well-designed study calculates power (sample size) a priori (before data collection), not after the study concludes. Post hoc power calculations that use differences observed

in the study are useless, because they tell one nothing more than what the *p* value has already indicated regarding the significance of the statistical test.

Another way of thinking about statistical power is to use legal jargon. Think of researchers formulating a study design by first developing a "trial strategy." They decide on design, dependent measures, the number of measures per participant, and how many participants to test. Each factor will have an impact on how likely the evidence (results of the study) presented to the jury (reader) is to be convincing beyond a reasonable doubt (the *power* of the study). The statistical tests applied to the data are used to determine whether the null hypothesis (Group A = Group B: No differences between the groups exist) can be rejected. In other words, the evidence was powerful enough to convince beyond a reasonable doubt. An interesting Web page created by Rob Becker called the "OJ Page" uses this concept to introduce the idea of statistical power: <http://trochim.human.cornell.edu/OJtrial/ojhome.htm>

There are four interrelated components that influence the conclusions reached from a statistical test: sample size, effect size, alpha level, and power.² Sample size usually refers to the number of participants studied. This is perhaps the one variable that is easiest to manipulate. Effect size can be thought of as the "potency" of the treatment. It provides an estimate of the degree to which the treatment influenced the outcome. It is generally thought of as the standardized difference between the means. Sometimes this is referred to as the meaningfulness of the difference. Alpha level (level of significance) is defined as the

TABLE 1. NULL-HYPOTHESIS (H_0) TRUTH TABLE :

	H_0 True	H_0 False
Accept	accurate decision	Type II error (β)
Reject	Type I error (α)	accurate decision

TABLE 2. THE STATISTICAL INFERENCE DECISION MATRIX

What We Conclude	In Reality	
	H ₀ (null hypothesis) true, H ₁ (alternative hypothesis) false. There is <i>no</i> relationship; there is <i>no</i> difference, no gain; our theory is <i>wrong</i> .	H ₀ (null hypothesis) false, H ₁ (alternative hypothesis) true. There is a relationship; there is a difference or gain; our theory is <i>correct</i> .
<p>We accept the null hypothesis (H₀) and reject the alternative hypothesis (H₁).</p> <p>We say "There is no relationship; there is no difference, no gain; our theory is wrong."</p>	<p>1 - α (e.g., .95): Confidence Level</p> <p>The odds of saying there is no relationship, difference, gain, when in fact there is none; the odds of correctly not confirming our theory. <i>95 times out of 100, when there is no effect, we'll say there is none.</i></p>	<p>β (e.g., .20): Type II Error</p> <p>The odds of saying there is no relationship, difference, gain, when in fact there is one; the odds of not confirming our theory when it is true. <i>20 times out of 100, when there is an effect, we'll say there isn't.</i></p>
<p>We reject the null hypothesis (H₀) and accept the alternative hypothesis (H₁).</p> <p>We say "There is a relationship; there is a difference or gain; our theory is correct."</p>	<p>α (e.g., .05): Type I Error (significance level)</p> <p>The odds of saying there is a relationship, difference, gain, when in fact there is not; the odds of confirming our theory incorrectly. <i>5 times out of 100, when there is no effect, we'll say there is.</i> We should keep this small when we cannot risk wrongly concluding that our program works.</p>	<p>1 - β (e.g., .80): Power</p> <p>The odds of saying there is a relationship, difference, gain, when in fact there is one; the odds of confirming our theory correctly. <i>80 times out of 100, when there is an effect, we'll say there is.</i> We generally want this to be as large as possible.</p>

Note. Adapted with permission from W.M. Trochim.²

odds that the results might have occurred by chance. For example, most studies use an alpha level of $p < .05$, which means that the researchers are willing to accept a 1 in 20 chance that they would be wrong in their statistical conclusion based on the treatment studied. Power, as previously defined, indicates that you will observe the treatment effect when in fact it does occur. It is generally accepted that the power of a study should be at least .80, or 80%.¹ This would mean that 80% of the time the researcher will be correct when accepting the conclusion that there is no difference between treatment groups. The power of the statistical test will also depend on the critical nature of committing a Type II error.¹ This is especially important in studies in which statistical conclusions could affect human life (i.e., pharmaceutical studies). In these studies, a priori power estimates are set between .95 and .99. Table 2 illustrates the interrelationships among these four components.

Readers should look for evidence of power calculation (sample-size determination) in the methods sec-

tion of research articles, especially in studies reporting no statistically significant differences between groups, treatments, or participants. Researchers should elaborate on their use of a priori power calculations in the discussion section. Readers should be cautious about statistically significant results that occur with acceptable statistical power but have very little "clinical" significance. This is one area in which clinicians can provide insight to researchers and help bridge the gap that can develop between research significance and clinical significance. ■

References

1. Greenfield MLV, Kuhn JE, Wojtys EM. A statistics primer: power analysis and sample size determination. *Am J Sports Med.* 1997;25: 138-140.
2. Trochim WM. *Statistical power*. Research Methods Knowledge Base Web site. Available at: <http://trochim.human.cornell.edu/kb/power.htm>. Accessed May 27, 2003.

Tom Kaminski is an associate professor and director of athletic training education at the University of Delaware.