Journal of Athletic Training 2018;53(7):716–719
doi: 10.4085/1062-6050-284-17
© by the National Athletic Trainers' Association, Inc
www.natajournals.org

Statistical Primer for Athletic Trainers: Understanding the Role of Statistical Power in Comparative Athletic Training Research

Monica R. Lininger, PhD, LAT, ATC*; **Bryan L. Riemann, PhD, ATC, FNATA†**

*Department of Physical Therapy and Athletic Training, Northern Arizona University, Flagstaff; †Department of Health Sciences, Georgia Southern University, Savannah

Objective: To describe the concept of statistical power as related to comparative interventions and how various factors, including sample size, affect statistical power.

Background: Having a sufficiently sized sample for a study is necessary for an investigation to demonstrate that an effective treatment is statistically superior. Many researchers fail to conduct and report a priori sample-size estimates, which then makes it difficult to interpret nonsignificant results and causes the clinician to question the planning of the research design.

Description: *Statistical power* is the probability of statistically detecting a treatment effect when one truly exists. The α level, a measure of differences between groups, the variability of the data, and the sample size all affect statistical power.

Recommendations: Authors should conduct and provide the results of a priori sample-size estimations in the literature. This will assist clinicians in determining whether the lack of a statistically significant treatment effect is due to an underpowered study or to a treatment's actually having no effect.

Key Words: statistics, reporting statistical findings, beta

A certified athletic trainer (AT) is treating a student-athlete who has sustained a tibial plateau fracture and reads a study about the effects of using an ultrasound bone stimulator to decrease healing time. After reading that the study showed no difference between the bone stimulator and a placebo with regard to bone healing, the AT thinks, "The rationale for the treatment studied in this published paper is valid, and anecdotally, I have seen great responses with the exact protocol in similar patients. But are there things that could explain the lack of statistical significance?"

Whereas there might be numerous answers to the clinician's question, including subtle differences in the protocol and setting, we will focus this article on 2 possible statistical explanations: sample-size calculations and type II statistical error. A common question asked by researchers during the design of an intervention study is "How many participants do I need to see a difference, if one actually exists, between those in the treatment group and those in the control group?" This question may appear simple, but unfortunately, the answer is not that straightforward. It is important for clinicians to understand study methods and appraise whether the statistical power was adequate when nonsignificant results are reported.

In the first paper of the "Statistical Primer for Athletic Trainers" series, we presented the difference between statistical significance and clinical meaningfulness.[1] We encouraged authors to provide additional statistical evidence, such as effect sizes and confidence intervals, along with the traditional *P* value, in the second paper of the series.[2] In this installment of the series, we will focus on the determinants used to estimate sample size. It is beyond the scope of this article to supply an all-encompassing explanation of sample-size calculations, but interested readers are referred to several resources for greater detail.[3–8]

What Is Statistical Power?

Statistical power is the probability of statistically detecting a treatment effect when one truly exists. In other words, statistical power is the likelihood of rejecting a null hypothesis that states there is no difference between groups.[1] In research, we study samples to make inferences about a population of interest. A certain amount of uncertainty or error is associated with this approach, especially if the sample was not randomly selected. In research decisions, the 2 types of possible sources of error are typically referred to as type I and type II statistical errors (Table 1). A *type I error* (false-positive) occurs when the authors conclude that a treatment did work (ie, a statistically significant treatment effect was detected) although there was actually no difference in the effectiveness of the interventions. As discussed in the first paper of this series,[1] the α level is associated with the likelihood of making a type I error, and it usually is specified as .05. The second type of error is a *type II error* (false-negative), which occurs when investigators fail to reject the null hypothesis (ie, a statistically significant treatment effect was not detected) when a treatment effect was indeed present. This situation is problematic because the study conclusion was that the groups did not differ when, in fact, they did. The probability of making a type II error is associated with β, which represents the likelihood of concluding that the 2 groups were equal when, in fact, they

Table 1. Type I and Type II Errors

	Result of Statistical Test	
Truth	Fail to Reject Null Hypothesis	Reject Null Hypothesis
No difference between groups	No error (probability $= 1 - \alpha$)	Type I error—false-positive (probability $= \alpha$)
Difference between groups	Type II error—false-negative (probability $= \beta$)	No error (probability $= 1 - \beta$)

were not. Typically, more emphasis is placed on minimizing the risk of type I errors than of type II errors; however, not statistically finding an effect that truly exists can be as important as statistically finding an effect that does not really exist. For example, if an AT suspects an injury but there is no underlying condition, the student-athlete will unnecessarily miss practice or playing time (type I error or a false-positive). However, an AT could conclude that no injury is present and allow the student-athlete to compete when, in reality, an injury is present (type II error or a false-negative).

Statistical power is quantitatively represented as $1 - \beta$, where β is the probability of making a type II error. The complement of β ($1 - \beta$) is *statistical power*, the probability of correctly rejecting the null hypothesis when it is false. Another way to understand statistical power is to say that a research study with a β equal to .10 would have a statistical power of 0.90, or a 90% chance of detecting a treatment effect that is truly present. A study with a statistical power of 0.90 has a much better chance of rejecting the null hypothesis when it is indeed false than does a study with a statistical power of 0.70.

Factors Influencing Statistical Power

Statistical power is influenced by the α level, a measure of the expected difference between the groups being studied, and the sample size. As a review, the α *level* is a quantifiable measure of the researcher's willingness to commit a type I error.[1] The most commonly used α level is .05, which means that 5% of the time, a researcher is willing to incorrectly say the groups differ when, in reality, they do not.

It is important to have a clear understanding of the expected difference between the 2 groups when estimating statistical power. Not only do we need to examine the expected difference between the groups but also the predicted similarity (*homogeneity*) or variation (*heterogeneity*) in how participants respond to the treatment as well as the variability of the outcome measure among participants in the control or placebo group. It is easier to achieve statistical significance when less variability (ie, more consistency in the treatment effect) is seen in the outcome measure than when a great deal of variability is present. A frequent method used to combine the differences and variability into a single value for estimating power is the effect size. The *effect size* is the magnitude of the difference between 2 groups relative to the variability and was examined in the second paper of this series.[2] If the treatment effect in a study is small, it may not be possible to reject the null hypothesis (stating there is no difference). When calculating statistical power, it might not be clear what effect size to use. An effect size can be derived from pilot data or the previous literature or it can be an arbitrary value (ie, small, medium, large) that corresponds to what the researcher thinks is most appropriate.[9]

Finally, sample size is the last component that we can objectify and is traditionally used to calculate statistical power. As discussed and illustrated earlier in the series,[1] it is generally easier to achieve statistical significance with large sample sizes than with small sample sizes because the former provide better population estimates, more precise confidence intervals, and smaller standard errors. As a result, one might think the easy solution is to incorporate as many participants as possible to ensure high statistical power. Unfortunately, this approach may result in wasting time and money, yet more important are the ethical concerns about research conducted with human participants. For example, if 50 participants (25 in each group) are sufficient to demonstrate a treatment is effective, involving 100 participants exposes an additional 50 participants to the risks of the study. The converse is also true: If the sample size required for a study is 100 but only 50 participants can be enrolled due to, for example, time and money, the study should not be undertaken because the participants would be exposed to the risks of the study for potentially no societal benefit.

In research, it is most common to set statistical power at 0.80 or higher and the α level at .05, which then leaves the effect size and sample size as unknown factors in the calculation. Typically, an investigator will estimate the sample size before starting a study (a priori) in order to determine how many participants are needed to minimize the risk of a type II error. The ideal choice for the anticipated effect size would be based on a clinically relevant change. Further guidance regarding the potency of an intervention to promote change typically comes from the previous literature or pilot studies. Using the minimal detectable change or minimum important difference as the expected difference may also help to address the clinical relevance of the difference between groups as well as the sample size needed for appropriate statistical testing. With a larger effect size, a smaller sample size can achieve the same statistical power at the traditional .05 α level. For instance, with an effect size of 0.50 and α level of .05, 128 participants would be needed to achieve a statistical power of 0.80 (Table 2). However, if the effect size was larger, such as 0.80, then only 52 participants would be needed to achieve the same statistical power of 0.80. Occasionally, an a priori power analysis is conducted by selecting effect sizes based on the standard interpretation conventions described previously.[2] Whereas this approach will establish the sample size needed to reach statistical significance with the chosen effect size, we advocate avoiding this approach because it fails to account for the clinical relevance of the change.

We, along with other authors,[10,11] caution against performing statistical power analyses only after data collection has been completed (post hoc). Sometimes this is referred to as *retrospective power*. By definition, if a statistically significant result is attained, post hoc power computation will reveal an adequately powered study; however, this does not rule out the possibility of a type I error (ie, small sample size with a few outliers). In contrast,

Table 2. Factors Influencing Statistical Power

Sample Size (N)	Effect Size	α Level	Statistical Power (1 − β)
128	0.50	.05	0.80
52	0.80	.05	0.80
102	0.50	.10	0.80
42	0.80	.10	0.80

if a study does not reach statistical significance, a type II error could have occurred or there may truly be no difference between the 2 groups. Reporting retrospective power does not offer any additional information to explain nonsignificant results. Rather, it is more beneficial to explain unexpected nonsignificant findings using effect sizes, confidence intervals, and the variability of participant responses to the intervention and to examine whether the null hypothesis could, in fact, be true.[12] However, one may report post hoc power calculations with the proviso that they be used for planning research studies and conducting meta-analyses but not for interpreting the results of the current study.

Methods to Improve and Estimate Statistical Power

Now that we have described statistical power and its influencing factors, it is important to review methods for improving the statistical power of a study. By studying a sample that is more homogeneous (ie, participants are all similar), such as limiting the focus to collegiate female soccer players instead of all collegiate athletes (ie, both sexes, all sports), the overall between-subjects variability will likely decrease. This will in turn increase the effect size. However, a more homogeneous sample also carries disadvantages, such as the inability to generalize (ie, external validity) to a larger group (eg, all collegiate athletes versus collegiate female soccer players). Another approach to decreasing variability with the intent of increasing statistical power is to use a repeated-measures design. By measuring the dependent variable on multiple occasions, the intersubject variability will decrease, ultimately increasing the precision of the study and its overall statistical power. Statistical power can also be increased by using a reliable measurement. When researchers use an unreliable tool for measuring the outcome variable, the variability within the scores increases, thereby decreasing the statistical power.

A second option is to use an intervention that has a large effect size. As previously mentioned, a larger effect size will be easier to detect statistically than a smaller effect size, which increases the probability of rejecting the null hypothesis when it is false (greater statistical power). Whereas some aspects of an intervention can be adjusted to increase an effect size, there are limits to how much an intervention can be manipulated. For example, although one might produce a greater intervention effect by having patients complete 1 hour of rehabilitation 5 days per week instead of 1 hour of rehabilitation 3 days per week, the trade-off might be that the target population cannot commit to and comply with such a time demand.

It is beyond the scope of this article to provide a detailed list of all the computer applications available for calculating statistical power. However, we would like to mention a few for the reader who is interested in learning more. Some of the common applications include Power Analysis and Sample Size (pricing varies, available at https://www.ncss.com/software/pass/), G*Power (free, available at http://www.gpower.hhu.de/), SAS (pricing varies, available at https://www.sas.com/order/product.jsp?code=PERSANLBNDL), and SPSS (pricing varies, available at http://www-01.ibm.com/software/analytics/spss/products/statistics/base).

Challenges In Estimating Statistical Power

Even with the best-planned methods that include an a priori power analysis and use of a promising intervention based on a solid rationale and previous research, attaining statistical significance is not guaranteed. The most obvious explanation is that the intervention does not work as hypothesized, regardless of how sound the rationale or previous research was. Despite the heavy bias in the literature toward publishing only studies with statistical significance, knowing that an intervention is effective is as important as knowing that it is not effective. Another alternative explanation, as discussed earlier, is the chance of a type II error; specifically, the intervention had an effect, but our statistical results do not lead us to conclude there was an effect.

One may ask, "How can we have a type II error when we included an a priori power analysis?" The oversimplified answer is that we are estimating. Similar to relying on a sample to estimate the population, when computing the sample size needed using an a priori power analysis, we are really just estimating. Whereas α and β are most often established by convention, as described previously, the other essential elements used to establish the effect size (mean differences and variance) are estimated from samples (research or pilot work). Based on random sampling differences, it is possible to obtain a study sample that has a different variance than the values used for the power analysis. Thus, the estimated sample size from a power analysis is best treated as a conservative minimal estimate, and therefore, prudent practice would include slight increases to the sample size to account for the uncertainty.

CONCLUSIONS

In summary, statistical power (1 − β) is the probability of finding a statistically significant difference between treatment and control groups when one truly exists. Statistical power is influenced by the α level, the expected difference between groups (effect size), and the sample size. Sample size is often the only factor that is readily under the control of the investigator. Other ways to improve statistical power include assembling a more homogeneous sample, using an intervention that has a large effect size as long as it maintains clinical relevance, or applying a repeated-measures design.

RECOMMENDATIONS

The sample size of a research project needs to be large enough to reach statistical significance if there is indeed a difference between the means of the treatment and control groups. We encourage authors to perform and report a priori sample-size estimations in manuscripts submitted to the *Journal of Athletic Training*. Performing these calculations will help to ensure that statistical power can be achieved and that extra participants are not being exposed to an experimental treatment that may be potentially harmful. The publication of these sample-size estimations, as well as the rationale for the effect size used (difference and variance estimates), will allow the reader to determine whether the study was possibly underpowered or it is more likely that the treatment did not work as hypothesized.

REFERENCES

1. Riemann BL, Lininger M. Statistical primer for athletic trainers: the difference between statistical and clinical meaningfulness. *J Athl Train*. 2015;50(12):1223–1225.
2. Lininger MR, Riemann BL. Statistical primer for athletic trainers: using confidence intervals and effect sizes to evaluate clinical meaningfulness. *J Athl Train*. 2016;51(12):1045–1048.
3. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988:274–288.
4. Cohen J. A power primer. *Psychol Bull*. 1992;112(1):155–159.
5. Ellis PD. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge, UK: Cambridge University Press; 2010.
6. Kraemer HC, Thiemann S. *How Many Subjects? Statistical Power Analysis in Research*. Newbury, CA: Sage Publishers; 1987.
7. Lipsey MW. *Design Sensitivity: Statistical Power for Experimental Research*. Newbury, CA: Sage Publishers; 1990.
8. Murphy KR, Myors B, Wolach A. *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*. New York, NY: Routledge; 2009.
9. Beck TW. The importance of a priori sample size estimation in strength and conditioning research. *J Strength Cond Res*. 2013;27(8):2323–2337.
10. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med*. 1994;121(3):200–206.
11. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat*. 2001;55(1):19–24.
12. Greenland S. On sample-size and power calculations for studies using confidence intervals. *Am J Epidemiol*. 1988;128(1):231–237.

Address correspondence to Monica R. Lininger, PhD, LAT, ATC, Department of Physical Therapy and Athletic Training, Northern Arizona University, PO Box 15094, Flagstaff, AZ 86011. Address e-mail to monica.lininger@nau.edu.