Journal of Athletic Training 2015;50(12):1223–1225
doi: 10.4085/1062-6050-51.1.04

www.natajournals.org

technical note

Statistical Primer for Athletic Trainers: The Difference Between Statistical and Clinical Meaningfulness

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

*Department of Health Sciences, Armstrong State University, Savannah, GA; †Department of Physical Therapy and Athletic Training, Northern Arizona University, Flagstaff

Objective: To explain statistical significance and clinical meaningfulness and to provide guidance in evaluating the clinical meaningfulness of a study.

Background: Understanding the results and statistics reported in original research remains a large challenge for many certified athletic trainers, which in turn, may be among the biggest barriers to integrating research into athletic training practice.

Description: Statistical significance reflects the influence of chance on the outcome, whereas clinical meaningfulness reflects the degree to which the differences and relationships reported in a study are relevant to athletic training practice. As consumers of original research, athletic trainers must understand the core factors, most notably sample size, that influence statistical significance.

Recommendations: To assist clinicians in evaluating the clinical meaningfulness of a research study, authors should provide the core elements necessary for interpreting statistical significance and discuss the clinical meaningfulness of statistically significant findings.

Key Words: statistics, *P* values, research design

The expectation that certified athletic trainers should be informed consumers of research has increased because athletic training, like many allied health professions, adheres to principles of evidence-based practice. Despite the increased emphasis on evidence-based practice in our educational programs, journals, and symposia, understanding research remains a challenge for many athletic trainers. A common barrier to integrating research into athletic training practice is comprehending the results and statistics reported in original research. Therefore, the purpose of the forthcoming short series was to examine a few aspects of reporting statistical results to facilitate better understanding between clinicians and researchers. Beginning with this paper, we will provide an explanation of what *statistical significance* means and how to evaluate whether the results of a study reach a threshold of *clinical meaningfulness*. In future papers, we will provide the essentials for understanding statistical power, effect sizes, confidence intervals, and ultimately, how to determine the main purpose of the research (comparison of groups, estimates of treatment effects, or estimates of the probability of a discrete outcome). In meeting the objectives of each paper in this series, a framework of expectations for researchers reporting results in the *Journal of Athletic Training* will be established.

What Does Statistical Significance Mean?

Research relies on the use of samples selected from the target population to infer what could be expected if the entire population had been studied. Given that a sample is used instead of the entire population, our estimates about what exists in the entire population likely differ slightly from the true reality in the population. Using rigorous research design elements, including random sampling and random allocation (assignment), sufficiently sized samples (discussed in the next paper), and reliable outcome measures, helps decrease the discrepancy between the population and the sample being studied. Statistical tests attempt to provide an indication of whether the differences and relationships that the sample data revealed may be considered "true" versus the likelihood that they occurred based on chance alone. The resulting *P* value is the probability of obtaining the results if the hypothesis that no difference or relationship (null hypothesis) exists were true. Thus, the *P* value can be considered an index of the evidence against the null hypothesis. To reduce interpretation of the *P* value to a simple *yes* or *no* regarding whether the differences or relationships are "real," the *P* value is often compared with a threshold point (α level). Statistical significance, or rejection of the null hypothesis, is concluded when the *P* value is less than the α level. Researchers establish the α level early in the research-planning process; most often, .05 is used. The α level provides an estimation of a researcher's willingness to incorrectly conclude (ie, commit a type I statistical error) that a true difference or relationship exists when, in reality, it does not. This approach to interpreting the *P* value is referred to as *null-hypothesis significance testing*. Whereas frequently used, this approach to interpreting *P* values has many challenges, including the arbitrary use of .05 to define the border between concluding *yes* or *no*, that are beyond the purpose of this paper.[1]

To illustrate null-hypothesis significance testing, consider the circumstance in which a researcher is interested in comparing 2 intervention programs for improving ankle dorsiflexion. To conduct the research, a random sample of 40 physically active participants with restricted dorsiflexion

Table 1. Interpreting *P* Values With Respect to Sample Sizes

P Value	Classification	Interpretation
>.15	Clearly not statistically significant	A small sample size may not have sufficient statistical power and may need additional research. A large sample size could show a very small effect size or an inconsistent effect.
.05–.09	Not statistically significant but close to .05 criterion	Result could truly reflect no difference between interventions, an inconsistent effect, or a sample size that is too small; additional research is needed.
.01–.05	Statistically significant but close to .05 criterion	Given a statistical difference, the effect size may be small or inconsistent; additional research is needed.
<.01	Statistically significant	A small sample size could show either a large effect size or consistent effect or both. A large sample size could show a small effect size.

are assigned randomly to 1 of 2 groups: control (standard stretching) or experimental (myofascial release followed by standard stretching). Their ankle-dorsiflexion range of motion is measured with a standard goniometer before and after the intervention program. After the 3-week program, range-of-motion improvements are 5.7° ± 1.7° and 6.8° ± 1.5° for the control and experimental groups, respectively. Based on the results of a test for statistical comparison ($t_{38} = 2.05$, $P = .047$), the researchers claim that the incorporation of myofascial release with stretching resulted in a significantly greater range-of-motion improvement because the computed *P* value was less than .05. Inherent to this interpretation is the premise that the 1.1° difference in range of motion exceeded what would be expected if no difference existed in range-of-motion improvements between the 2 groups.

Interpreting *P* Values

Whereas most researchers frequently use null-hypothesis significance testing with the threshold of .05 defining statistical significance, several considerations to this interpretation are important. For example, in our dorsiflexion-intervention study, little logic exists in accepting the observations when $P = .049$ while rejecting the observations when $P = .051$. The *P* value is a function of several factors, including some of the aforementioned research-design elements. Two potent factors that we will fully examine in a forthcoming paper are sample size and the consistency of the effect (ie, change from pretest to posttest). With smaller sample sizes or inconsistent effects, attaining statistical significance becomes more difficult. For this reason, we advocate considering the *P* value in conjunction with the sample size (Table 1). If the difference between the groups in our dorsiflexion-intervention–study example remained the same (1.1°) but we had included 10 rather than 20 participants per group, we would have concluded a different "answer" about whether the addition of myofascial release before stretching was more beneficial than stretching alone (Table 2). For a simulation of sample-size effects and *P* values, view the "Dance of the *P* Values."[2]

Table 2. Effects of Sample Sizes on Statistical Significance

Sample Size Per Group	Mean Difference Between Groups	*t* Statistic	*P* Value	Statistically Significant at .05?
10	1.1	1.44	.164	No
20	1.1	2.05	.047	Yes
30	1.1	2.51	.015	Yes

Statistical Significance Does Not Mean Clinical Meaningfulness

Statistical significance reflects the influence of chance on the outcome, whereas *clinical meaningfulness* reflects the clinical value of the outcome.[3] In other words, clinical meaningfulness reflects the degree to which the differences and relationships reported in a study are relevant to athletic training practice. In our dorsiflexion-intervention–study example, the results reached statistical significance ($P = .047$); however, the difference in range-of-motion improvement was only 1.1°. Clinical meaningfulness in this example relates to whether an additional 1.1° of range-of-motion improvement is worth the additional time investment to perform the myofascial release before stretching. Whereas minimal risk exists for an adverse event with myofascial release, higher risks and costs (money, time) may need to be considered in addition to the benefit for many other clinical procedures. Furthermore, when determining clinical meaningfulness, we also need to consider the reliability of the measurement tools used and the types of participants included in the study. Unreliable approaches to assessing outcome measures decrease the likelihood of reaching statistical significance. The results of interventions for healthy, college-aged participants, 1 of the most common populations studied in athletic training research, may not be generalizable to other populations and will likely yield smaller effect sizes than in participants with pathologic conditions. Additional tools, such as confidence intervals and effect sizes, can be reported to help readers determine clinical (or applied) meaningfulness. These concepts will be the topic of future papers in this series.

CONCLUSIONS

When the *P* value generated against a hypothesis that no difference or relationship (null hypothesis) exists is smaller than a specified threshold (most often .05), statistical significance is claimed. In a well-designed and executed research study, statistical significance should be interpreted as evidence that the likelihood the results could have occurred based on chance is small. Clinical meaningfulness relates to whether the differences or relationships shown are of sufficient magnitude to influence clinical practice. Whereas no standard rules indicate when results are clinically significant, a few items that may assist in evaluating clinical meaningfulness can be reported. These

items include effect sizes and confidence intervals, which are the topic of a forthcoming paper in this series.

RECOMMENDATIONS

To enhance the abilities of athletic trainers to make decisions about clinical meaningfulness and of researchers to replicate research studies, we recommend that researchers report exact *P* values versus simply indicating whether the *P* value is less than or greater than the α (threshold for statistical significance). Discussion sections should include examinations of the clinical meaningfulness and the statistical significance of the study results. These additions will help clinicians who may be less informed about evaluating clinical meaningfulness. Finally, we recommend including other information, such as confidence intervals and effect sizes, that can assist readers in ascertaining clinical meaningfulness.

REFERENCES

1. Kline RB. *Beyond Significance Testing: Statistical Reform in the Behavioral Sciences*. 2nd ed. Washington, DC: American Psychological Association; 2013.
2. Cumming G. Intro statistics 9: dance of the *P* values. YouTube Web site. https://www.youtube.com/watch?v=5OL1RqHrZQ8. Accessed July 29, 2015.
3. Lang TA, Secic M. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers*. 2nd ed. New York, NY: American College of Physicians; 2006:xvii–xix.

Address correspondence to Bryan L. Riemann, PhD, ATC, FNATA, Department of Health Sciences, Armstrong State University, 154 University Hall, 11935 Abercorn Street, Savannah, GA 31419. Address e-mail to bryan.riemann@armstrong.edu.