Applications of Confidence Limits and Effect Sizes in Sport Research

Eric Drinkwater*

School of Human Movement Studies, Charles Sturt University, Panorama Avenue, Bathurst, NSW, Australia, 2795

Abstract: This article describes the origins of the conventional use of null hypothesis significance testing and why this convention has led to difficulties in implementing research results in applied settings. This article continues to explain the the value of expressing research results with confidence limits and effect sizes for sporting application.

As researchers investigating human performance, one of our greatest measures of success is for our research outcome to be implemented by sports coaches for their athletes. While coaches are becoming increasingly receptive to the results of sports scientists, coaches are often frustrated by the inconclusive and numerically cryptic results we report. Conventional null hypothesis significance testing dictates that unless the probability of rejecting the null in error (p-value) is less than 5%, we must accept the null hypothesis that the difference between our groups is zero. But to return to a sports coach after six weeks of a training intervention to report "nothing happened" is frustrating and probably not entirely accurate. It may be possible that the intervention did have an effect, but due to sources of error in human performance testing, the results lacked sufficient consistency to pass the conventional 5% rule. However, is the p-value returned by our results greater than 5% because nothing happened, or is the problem in our use of the arbitrary 5% line in the sand to justify the success or failure of our intervention? After all, "… surely, God loves the .06 nearly as much as the .05." [1, p. 1277].

Origins of the p-values in Null Hypothesis Significance Testing

Initially describing type I and type II error rates was the work of Neyman and Pearson [2]. Neyman and Pearson considered that there was sufficient evidence to reject a null hypothesis if the probability of its rejection in error was less than 5%. The work by Fisher [3] initially described some standard levels (e.g. 1%, 5%, 10%, etc.) of area under the χ^2, t- and f-distributions, thereby making 5% of these distributions widely accessible to researchers. While Fisher only intended percentages of these distributions to add support to inferences drawn from data, Neyman and Pearson argued that in order for research to be used to make decisions, 5% of these distributions was an acceptable α (cut-off point) [4]. Since this time, accepting or rejecting a null hypothesis based on a 5% probability of error has become the norm. The sport science interpretation of Neyman and Pearson's work would be that sports coaches (i.e. research end-users) can only make informed decisions when told if an intervention works or does not work, whereas Fisher would argue that sports coaches should be the ones to decide what probability of error is unacceptably high for their athletes [4].

Confidence Limits

After decades of accepting or rejecting null hypotheses based on p-values of less than or greater than 0.05, there has been a recent criticism of using solely the p-value to accept or reject research findings [5]. Confidence limits express the precision of the mean changes within a sample (or mean differences between samples, hereafter called the *mean estimate*) by expressing upper and lower boundaries within a confidence bandwidth (e.g. 90%, 95%) rather than simply expressing the probability that the mean estimate equals zero. When expressing a mean estimate (e.g. there was 66 W difference between two groups) Fig. (**1**), the true difference between the two groups for the population is unlikely to be exactly 66 W. The 66 W only represents an estimate from the sample; there is certain to be error in the mean estimate when making inferences from the sample to the population. For example, in a recent piece of research [6] we indicated that the estimated difference between two trials was 66 W and, while accepting that 66 W was unlikely to be the exact effect of the intervention, we were 95% certain that the true value of the difference between trials lay between 36 and 96W. In this way, we provided much more useful information about the precision of our estimate, rather than just the 66 W estimate and that the probability of the estimated difference was actually zero was less than 1% [6].

Confidence limits can be derived for any percent level, though most common are 90% and 95%. While some researchers [7] feel that the range of 95% confidence limits is too broad to be useful, others [6] feel that 95% confidence limits are more suitable in the current climate in which many journal reviewers still look for statistical significance of results. A p-value of less than 0.05 (i.e. 'statistically significant') can be derived from 95% confidence limits if both the upper and lower limits are on the same side of the zero (e.g. 36 to 96 W, Figure 1, Series 1); if the upper and lower limits are on different sides of the zero (e.g. -14 to 146 W, Fig. (**1**), Series 2) then the result had a p-value of greater than 0.05.

Effect Sizes

Imagine a new training method that could reduce a person's 100-m sprint time by a statistically significant 0.04 seconds. The primary consideration for a coach in wether to implement the intervention or not is if the change will have a worthwhile, not just a statistically significant, effect on performance. This distinction demonstrates the difference between statistical versus practical significance [8]. Cohen [9] detailed a method of dividing the change score by the stan-

*****Address correspondence to this author at the School of Human Movement Studies, Charles Sturt University, Panorama Avenue, Bathurst, NSW, Australia, 2795, Australia; Tel: +61 2 6338 6116; Fax: +61 2 6338 4065; E-mail: edrinkwater@csu.edu.au

1875-399X/08 **2008 Bentham Science Publishers Ltd.**

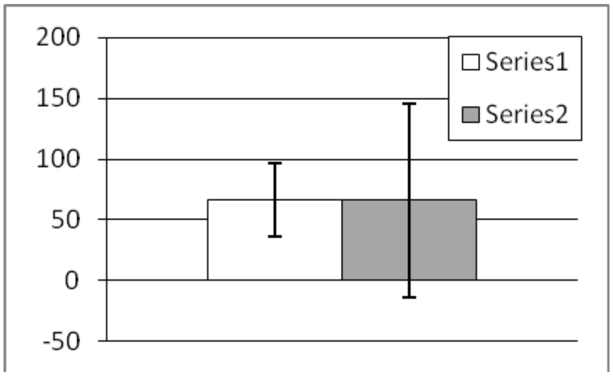

Fig. (1). Power output differences between two trials. Error bars represent 95% confidence limits.

dard deviation (SD) of the raw data to arrive at a standard effect size (*Cohen's d*). While the SD of both the men's and women's 100-m sprint final 2004 Athens Olympic Games was 0.09 s, the SD of a school sports carnival may be 2 seconds. The impact of the 0.04 seconds in the Olympic final would have a meaningful impact (Cohen's d = 0.04 ÷ 0.09 = 0.44), though the impact would be trivial at the school sports carnival (Cohen's d = 0.04 ÷ 2 = 0.02). Hopkins has also conducted extensive work and published resources to calculate likelihoods for effects being clinically beneficial [10].

CONCLUSION

Simply because a result is statistically significant does not necessarily mean it is worthwhile, just as a result that is not statistically significant is not necessarily useless. A coach could be presented with statistically non-significant results indicating that an intervention improved 100-m sprint time by a meaningful 0.04 seconds with 95% confidence limits ranging from -0.02 to 0.10 seconds. The coach may decide that the risk of doing actual harm to an athlete likely to win a bronze medal, thus knocking them out of the medal standing, is too great. A coach may also decide for an athlete likely to place 4th that the risk of doing harm, thus moving the athlete further down the ranking, is worth the chance if it could move the athlete into the medal standings. Furthermore, a statistically significant 0.01 second improvement that has a trivial effect size requiring an extra 10 hours of training per week may not be worth the time and effort required. Regardless, the coach is able to make much more informed decisions about the course of action to be taken based on the research. For these reasons, sport scientists should look to express the mean estimates of their research using standardised effect sizes (i.e. Cohen's d) and confidence limits rather than the making a 'yes' or 'no' decision based on the mean estimate's probability of equalling zero.

REFERENCES

[1] Rosnow RL, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. Am Psychol 1989; 44: 1276-84.

[2] Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika 1928; 20: 170-240.

[3] Fisher RA. Statistical methods for research workers. New York: Hafner 1946.

[4] Lehman EL. The Fisher, Neyman-Pearson theories of testing hypothesis: One theory or two? J Am Stat Assoc 1993; 88(424): 1242-9.

[5] Cohen J. The Earth is Round (p<0.05). Am Psychol 1994 December 1994; 49(12): 997-1003.

[6] Drinkwater EJ, Pritchett EJ, Behm DG. Effect of instability and resistance on unintentional squat lifting kinetics. Int J Sports Physiol Perform 2007; 2(4): 400-13.

[7] Petersen CJ, Wilson BD, Hopkins WG. Effects of modified-implement training on fast bowling in cricket. J Sports Sci 2004; 22(11-12): 1035-9.

[8] Hopkins WG. Probabilities of clinical or practical significance. Sportscience 2002; 6: sportsci.org/jour/0201/wghprob.htm.

[9] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates 1988.

[10] Hopkins WG. A spreadsheet for deriving a confidence interval, mechanistic inference and clinical inference from a p value. Sportscience 2007; 11: 16-20. Available from www.sportsci.org/resource/stats/xcl.xls.

Received: March 12, 2008 Revised: March 28, 2008 Accepted: March 29, 2008