Assessing group-based changes in high-performance sport.
Part 1: Null hypothesis significance testing and the utility of *p* values

Anthony N Turner, Nimai Parmar, Alex Jovanoski, Gary Hearne
London Sport Institute, Middlesex University

Abstract

The role of a strength and conditioning coach (SCC) has evolved over the last 10 years to accommodate the large influx of data now available. As such, today's SCC must extend their skill set to include data analysis, understanding the validity and utility of *p* values, effect sizes, confidence intervals, and terms such as the smallest worthwhile change, and minimal difference. The aim of part one of this two-part review is to define and discuss the utility of null hypothesis significance testing (NHST), *p* values, and error rates. In part two, we introduce effect sizes, measures of variability, and confidence intervals, culminating in recommendations as to which may be the most viable options within the context of performance-based sport, and thus potential methods to report group-based changes.

Introduction

The role of a strength and conditioning coach (SCC) has evolved over the last 10 years, perhaps to the point where the term strength and conditioning scientist may be just as apt. In principle, this is because SCC's are likely to spend as much time behind a computer analysing the streams of data they have just collected, as they are coaching athletes in the gym. While acknowledging that evidence-based practice has always been at the root of this role, perhaps today's shift is due to some of the following: (a) the need to demonstrate an objective approach to appease various stakeholders, (b) an increase in academic scientific-based degrees in this discipline of sport science, and (c) perhaps most causative, an influx of performance-based (affordable) software and hardware that provide vast quantities of performance variables. These advances mean that today's SCC must extend their skill set to include data analysis. While the evolution of this role, of course, requires the development of many other skill sets (Stewart, Comfort, & Turner, 2017), this paper focuses on analysing data, and in particular, assessing group changes (differences) in performance in applied settings. This appears to be an area of interest at the moment given the confusion in the best method to do this, from the implementation of *p*-values (Greenland, et al., 2016; Wasserstein & Lazar, 2016), to reporting effect sizes (Cohen 1988,

1992), confidence intervals (Cumming G. , 2014; Cumming & Finch, 2001), and the utility of analyses incorporating concepts such as the smallest worthwhile change (Hopkins, 2004) and minimal difference (Weir, 2005).

The aim of this two-part review therefore, is to discuss each of these in turn and provide recommendations to the reader as to which may be most viable within the context of performance-based sport, and thus potential methods to report group-based changes over time. Furthermore, such an analysis will also assist practitioners in critiquing relevant research findings when considering adoption of new strategies to practice. Here, in part one, we first define and discuss the utility of null hypothesis significance testing (NHST), *p* values, and error rates, given that NHST is the most commonly taught approach to testing research questions with statistical models, and thus the most well-known and used within the literature (Wasserstein & Lazar, 2016); in essence, this will serve as the platform from which we can develop our statistical approach to analyse our data. In part two then, through a series of worked examples, we will introduce effect sizes, measures of variability, and confidence intervals, culminating in recommendations as to which may be the most viable options for the applied analysis of data relating to group-based changes in strength and conditioning.

Null Hypothesis Significance Testing

We should also start by recognising the work of Sir Ronald Fisher, who is considered a pioneer of statistics and devised the *p* value that we will shortly explain; Fisher also coined the term "*statistically significant*" (Fisher, 1925). Actually, the term "*significant*" is now recognised as a poor choice of word and is consequently considered as one of the seven most misused words in science (Ghose, 2013). What Fisher actually meant, was along the lines of *statistically interesting and requires further scrutiny* (Wasserstein R. , 2019). However, its meaning was taken literally and is one of many reasons around the confusion of what a *p* value is, why statisticians have repeatedly asked us to refrain from using the phrase "*statistically significant*" (Wasserstein, Schirm, & Lazar, 2019; Amrhein, Greenland, & McShane, 2019; Lakens, et al., 2018), and why one journal even felt they had to ban it from use (Traimow & and Marks, 2015).

In actuality, the *p* value (and the NHST framework that commonly relies on it) can be a useful tool, if we appreciate its true meaning and utility (Greenland, 2019), with it offering a first line of defence against us being fooled by random error and any confirmation bias we may have

towards a theory (Lakens, 2019). In understanding the *p* value, we should first note that NHST is part of frequentist statistics, which means that it is concerned with the interpretation of probability, and specifically, long run probability; that is, what the likely results of a study would be, if repeated over and over again (Greenland, et al., 2016; Wasserstein & Lazar, 2016). So, when you appreciate the frequentist statistical framework, you can note that the results of any single study, only tell you what would happen if it were infinitely repeated (Greenland, et al., 2016) and do not actually relate to your single use study. As such, *p* values can never be regarded as evidence of the error or effect in your study (Greenland, et al., 2016). An additional misconception occurs when we don't appreciate the name and literal meaning of the test we are conducting, i.e., NHST. An NHST does exactly that, it typically investigates a test statistic obtained from a parameterized model against a null model, which centres on there being no effect or difference noted in your data, i.e., the result will be zero (however, via chance and random variation, we more than likely observe variability around zero). By way of example, if we introduce a new exercise to an intervention group (to check it works) and compare it to a control group, we are not actually testing the hypothesis that this new exercise will improve performance (which is referred to as the alternate hypothesis), but rather, that there is no difference between the groups (i.e., the null). This can be a confusing concept to grasp, because really, what we want to know, is if our alternate hypothesis (does our new exercise intervention work) is true or false. But as applied sport scientists, using NHST, we must appreciate that this is not the question that we are answering. Instead, we are testing the probability of our data, given our (null) hypothesis, which is written as $P(D|H_0)$. We are not testing our (alternate) hypothesis, given our data, written as $P(H_1|D)$. For the latter, we would need to use Bayesian statistics, but in any case, these are not the same thing and this explains why when reading around this issue, you may see $P(D|H) \neq P(H|D)$. We should also point out that NHST need not always be about null hypothesis (zero effect or difference) testing, and that hypotheses centring on equivalence, non-inferiority, and superiority, can also be tested. For example, Lakens (2017) provides examples where equivalence testing may be more advantageous, especially in scenarios where researchers want to argue for the absence of an effect that is large enough to be worthwhile to examine, and where researchers should also consider the effect size under the alternative hypothesis (we discuss this concept more in part two). Equivalence testing is beyond the scope of this text, so we recommend readers to the work of Lakens (2017); we suspect that in a large number of sports performance based research, this may be more appropriate than null hypothesis testing.

If we continue with the example of comparing the difference between two independent groups against a null model, we would then choose the *t*-test as our statistical model, ensuring we have met the model's underlying assumptions, such as independent groups, normal distribution of the means, and characteristically similar samples for example. By ensuring we have met these assumptions, we can *act* as if the only thing that differs between groups is our training intervention (Lakens, 2019) – all the while acknowledging that our data will always contain random error and thus noise, which we can never fully identify, control or exclude. Neyman and Pearson (1933) suggest the term *act* as a way forward with NHST, given it does not imply we truly believe the results directly relate to our single study, which in any case, did not investigate $P(H_1|D)$. We then run our test and generate our test statistic, which in this case is the *t*-statistic. The *t*-statistic we get, coupled with the sample size, is used to generate a *p* value, which informs us of the probability of obtaining our result (or more extreme), assuming the null hypothesis (and all the statistical model's assumptions) is true. Remember, that if the null hypothesis is true, the difference between groups would be around zero (assuming that was the threshold you decided on), and in this instance, all *p* values are equally likely (Caldwell & Cheuvront, 2019).

Now, if you get a low *p* value, you can then say that your data (not your alternate hypothesis, as you did not test this) is not compatible with the statistical model (and all its underlying assumptions) and thus is *interesting and requires further scrutiny*. The next question that logically arises then, is how low does the *p* value need to be, to spark your interest and for you to be satisfied that you are not merely measuring noise? Well, in answering this, let's first explain what the *p* value actually tells us. Say you obtained a *p* value of 0.03, this would mean that if the null hypothesis were true (and all the assumptions made by the underlying model), the probability of obtaining such a result (or more extreme) is 3%. This is now a good point to also introduce the term alpha (α), which describes the error rate you settled on prior to undertaking the research. By convention, we use an α level of 0.05, which implies that we accept the probability that we will get a false-positive (which is called a Type I error) in 5% of future studies, when using the same model and similar samples. This now brings us full circle back to *statistical significance.* In our example, we got a *p* value of 0.03, which is less than our pre-defined (conventional) α of 0.05. Given this, we historically conclude we have "*statistically significant"* results. Furthermore, when our p value is less than our accepted 5% error rate, if we choose to *act* as if there is in an effect when there really is not, in the long run,

you won't be wrong more than 5 % of the time. To note again, if the null is true, all *p* values are equally likely.

The drive to steer researchers away from using "*statistically significant*" is because it creates a dichotomy of evidence (McShane & Gal, 2017), whereby values on one side are important and meaningful (i.e., statistically significant) and values on the other side are unimportant and unhelpful (i.e., statistically non-significant). For example, if one strength training intervention results in $p = 0.049$, while another $p = 0.051$, we tend to deem the former as being an effective intervention and we would probably plan on implementing it with future athletes. The latter however ($p = 0.051$), would be deemed as non-statistically significant and thus ineffective, and we would therefore not plan to use it any longer. This thinking is of course incorrect and a consequence of categorical thinking. Rather, it has been argued that the *p* value should be treated and reported as a continuous quantity between 0 and 1, e.g., $p = 0.06$ (Greenland, et al., 2016), with us acknowledging that it does not tell us which assumption is incorrect; it could be the null or any of the model's underlying assumptions. Equally, any noted "effect" is subject to the *statistical power* of a test, which is discussed later in this paper. Such dichotomous thinking also drives publication bias (Franco, Malhotra, & Simonovits, 2014) and the "*file drawer effect*" (Rosenthal, 1979), whereby in some cases, we only get to read of studies that were statistically significant; this in turn can motivate *p*-hacking (the practice of flexibly analysing data until the *p* value passes the "significance threshold"). Collectively, these negative consequences have amongst other suggestions including total abandonment of NHST, resulted in calls for the alpha level to be lowered to $p = 0.005$ (Benjamin, et al., 2018). It is suggested that this change in critical threshold would help to reduce the number of published false-positives and the generation of weak evidence.

The appropriate use of *p* values

Firstly, the *p* value is a continuous probability and should be reported as such. While we act as though we have met all the model's underlying assumptions, this is actually very challenging and explains why, when coupled with the random variation that occurs in all data, we have profound study replication issues (Cumming G. , 2014; Caldwell & Cheuvront, 2019). For example, even if the exact same study was repeated with similar samples, it would generate different *p* values most of the time (Cumming G. , 2014; Amrhein, Greenland, & McShane, 2019). As such, it is difficult to profess that the difference between $p = 0.049$ and $p = 0.051$ is anything other than random variation, as opposed to a discrepancy residing only with

incompatibility of the data with the null. Instead then, we should simply say that assuming the null hypothesis were true (and all the assumptions made by the underlying model), the probability of obtaining such a result (or more extreme) is 5.1%. Now given that in sport science, unlike medicine perhaps, the consequence of making a Type 1 error (i.e., a false-positive) will unlikely be fatal or lead to any health complications, let alone lead to injury, it is probably okay to increase the *a priori* α level. Such thinking is in line with Lakens et al., (2018), who state that the α level should be adjusted based on the context at hand and on the cost of false-negatives (Type II errors): a higher α would be used by those for whom false-positives are relatively inconsequential, and lower α would be used by those for whom false-positives could be disastrous. In sport, we argue that in some cases, an α level as high as 0.1 (i.e., a 10% error rate) could be justified. For example, in performance sport, success is based on the smallest of margins (a statement that every Olympic Games final proves testament to) and professional athletes are often butting up against their genetic ceiling. As such, it is more important to reduce false-negatives and thus potential opportunities that may stimulate positive adaptations.

Irrespective of the α level, no decision should be made solely on the strength of a *p* value (Wasserstein & Lazar, 2016), given its indirect link to the alternate hypothesis, which is based on many assumptions that cannot be individually accounted for. Therefore, we need to move away from lazy dichotomous thinking (Gardner & Altman, 1986), and accept the uncertainty of our data and embrace its variation (Wasserstein, Schirm, & Lazar, 2019). One way to do this is to use confidence intervals, as well as acknowledging that one single study can never be taken as conclusive evidence of a new theory, irrespective of how low a *p* value is. With respect to the latter, this is why meta-analyses are so important to any field of study. That said, when considering the aforementioned publication bias, it is also interesting to consider if results derived from meta-analysis are in fact over inflated effect sizes, given that with far less frequency do we read about an intervention that does not work. Finally, we need to identify the magnitude of the effect, which NHST does not do. For example, rather than inferring an effect occurred, it would be far more useful to actually quantify the magnitude of the effect, such that coaches who are looking to adopt this new exercise can base decisions also on whether changes were small, moderate or large. However, we should again be cautious about applying critical thresholds to our data and perhaps consider the smallest effect size of interest (SESOI), which we could calculate if we appreciate the variability in our data or a particular target we are

aiming toward. Effect sizes, including the SESOI, as well as confidence intervals, are discussed in part 2 of this 2-part review.

Statistical Power

The final element to address as part of NHST, is statistical power, which is defined as the ability of a test to detect an effect when one exists. Statistical power should thus be considered by researchers and applied practitioners before they undertake any experiment. This is because studies that are woefully underpowered can be a waste of resources as well as time for all those involved (Caldwell & Cheuvront, 2019). For example, Caldwell and Cheuvront (2019) illustrated the results of a data simulation test, involving 100,000 repetitions, to demonstrate the distribution of *p*-values when the null hypothesis was false (i.e., there was an effect). When they ran the simulation with 80% statistical power, 80,000 simulations (80%) correctly identified a "statistically significant" effect, meaning that in 20,000 simulations (20%), a type II error (false-negative) occurred. They then repeated the simulations, but this time with 50% power, and unsurprisingly, 50,000 simulations (50%) correctly identified a "statistically significant" effect and the remaining 50,000 simulations (50%) missed it, generating a type II error. The natural conclusion to be reached here is, are studies worth doing (even from an ethical perspective) when the chance of finding an effect, if one exists, is 50-50? Probably not given you increase the risk of making erroneous conclusions if your decisions are based solely on *p*-values. So, let's now look at how we determine statistical power but first noting that again, as with *p*-values and confidence intervals, this probability is defined over repetitions of the same study design and so is a frequency probability (Greenland, et al., 2016).

Statistical power can be calculated using a host of statistical software, some are free such as G* Power, whereby you need only enter your pre-determined α level, sample size, and the SESOI. Given the requirement of these data, statistical power is considered a *conditional probability* (Caldwell & Cheuvront, 2019). For example, using G*Power for the calculation of statistical power, if we wanted to compare two independent groups (to see who could jump highest for example), and we used the conventional α of 0.05 and the conventional power of 80%, as well as aiming to detect an effect size (or magnitude of difference between groups) of half a standard deviation, we would need 64 participants per group. Increasing the α to 0.1, reduces the number of participants to 51 per group. It is not hard to appreciate therefore, that many studies undertaken in sport are likely underpowered and some true effects are missed.

This understanding should serve to justify the need for additional analysis such as effect sizes and confidence intervals to make inferences of whether an effect or difference was observed or not.

Conclusion

The roles and responsibilities of today's SCC means they must extend their skill set to include data analysis. NHST, along with its derived *p* value, can be a useful tool for this, if we appreciate its true meaning and utility, with it offering a first line of defence against us being fooled by random error and any confirmation bias we may have towards a theory. Importantly, we must note that NHST is part of the frequentist framework of statistics and thus refers to long run probability, with results from any single study used to infer what would happen if the study was repeated over and over again under identical conditions with different but identically distributed samples. Also, we must appreciate that our ability to find an effect, when one exists, is affected by statistical power – if this is too low, the utility of NHST is questionable when the expected effect (or difference between groups) is hypothesised to be small. Finally, if we do choose to use thresholds to limit the error within which we are happy to operate, we should choose an alpha level that represents the context at hand and the risks associated with Type I and II errors. Either way, we must recognise that the *p* value is a continuous variable, and thus should be reported as such. Therefore, practitioners using *p* values should conclude with a statement along the following lines (in this example let's say we got $p = 0.083$): *Assuming the null hypothesis were true and all the assumptions made by the underlying model, the probability of obtaining such a result or more extreme, is 8.3%. Furthermore, given our alpha level of 0.1, if we choose to act as if there is in an effect when in fact there is not, in the long run, we won't be wrong more than 10 % of the time.*

Winter et al., (2014) nicely summarise the essence of NHST via Karl Popper's principle of falsifiability, that is, before something can be accepted the opposite has to be shown to be untenable. So, in closing, it is prudent to again reinforce that we are analysing the probability of our data, given our (null) hypothesis, which is written as $P(D|H_0)$. In performance-based sport, however, we may determine that NHST isn't necessary or appropriate, going straight to methods that determine the *practical significance* of our data using effect sizes, and embrace the uncertainty of our data through confidence intervals. Through a series of worked examples, these will be explored in Part 2.

References

1. Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature, 567*, 305.
2. Benjamin, D., Berger, J., Johannesson, M., Nosek, B., Wagenmakers, E., Berk, R., & Cesarini, D. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6.
3. Caldwell, A., & Cheuvront, S. (2019). Basic statistical considerations for physiology: The journal Temperature toolbox. *Temperature, 6*, 181-210.
4. Cumming, G. (2014). The new statistics: Why and how. *Psychological science, 25*(1), 7-29.
5. Cumming, G., & Finch, S. (2001). A primer on the understand- ing, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532-574.
6. Fisher, R. (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver & Boyd .
7. Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345(6203), ., 345*, 1502-1505.
8. Gardner, M., & Altman, D. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J, 292*, 746-750.
9. Ghose, T. (2013, April 2). *"Just a Theory": 7 Misused Science Words*. Retrieved from Scientific American: https://www.scientificamerican.com/article/just-a-theory-7-misused-science-words/
10. Greenland, S. (2019). Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician, 73*, 106-114.
11. Greenland, S., Senn, S., Rothman, K., Carlin, J., Poole, C., Goodman, S., & Altman, D. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology, 31*(4), 337-350.
12. Hopkins, W. (2004). How to interpret changes in an athletic performance test. *Sportscience , 8*, 1-7.
13. Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science, 8*(4), 355-362.
14. Lakens, D. (2019). *What is a p-value?* Retrieved from YouTube: https://www.youtube.com/watch?v=RVxHlsIw_Do
15. Lakens, D., Adolfi, F., Albers, C., Anvari, F., Apps, M., Argamon, S., & Buchanan, E. (2018). Justify your alpha. . *Nature Human Behaviour, 2*, 168-171.

16. McShane, B., & Gal, D. (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association, 112*, 885-895.
17. Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, 231*, 289-337.
18. Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638-641.
19. Stewart, P., Comfort, P., & Turner, A. (2017). Strength and conditioning: Coach or scientist? . In A. Turner, & P. Comfort, *Advanced Strength and Conditioning: An Evidence-based Approach.* (pp. 1 - 11). London: Routledge.
20. Traimow, D., & and Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*, 1-2.
21. Wasserstein, R. (2019). *Misunderstandings of 'statistical significance'*. Retrieved from You Tube: https://www.youtube.com/watch?v=Wu4-OK_91EM&t=271s
22. Wasserstein, R., & Lazar, N. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician, 70*(2), 129-133.
23. Wasserstein, R., Schirm, A., & Lazar, N. (2019). Moving to a world beyond "p< 0.05". *The American Statistician, 73*, 1-19.
24. Weir, J. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J. Strength Cond. Res, 19*(1), 231–240.
25. Winter, E., Abt, G., & Nevill, A. (2014). Metrics of meaningfulness as opposed to sleights of significance. *Journal of Sport Sciences, 31*(10), 901-902.