Title: Clinical and practical importance versus statistical significance: limitations of conventional statistical inference.

Submission Type: Research methodology

Authors: Michael Wilkinson

Affiliation: Faculty of Health and Life Sciences

Northumbria University

Correspondence address: Dr Michael Wilkinson

Department of Sport, Exercise and Rehabilitation

Northumbria University

Northumberland Building

Newcastle-upon-Tyne

NE1 8ST

ENGLAND

Email: mic.wilkinson@northumbria.ac.uk

Phone: 44(0)191-243-7097

Abstract word count: 193

Text only word count: 4727

Abstract

Decisions about support for therapies in light of data are made using statistical inference. The dominant approach is null-hypothesis-significance-testing. Applied correctly it provides a procedure for making dichotomous decisions about zero-effect null hypotheses with known and controlled error rates. Type I and type II error rates must be specified in advance and the latter controlled by a priori sample size calculation. This approach does not provide the probability of hypotheses or the strength of support for hypotheses in light of data. Outcomes allow conclusions only about the existence of non-zero effects, and provide no information about the likely size of true effects or their practical / clinical value. Magnitude-based inference, allows scientists to estimate the 'true' / large sample magnitude of effects with a specified likelihood, and how likely they are to exceed an effect magnitude of practical / clinical importance. Magnitude-based inference integrates elements of subjective judgement central to clinical practice into formal analysis of data. This allows enlightened interpretation of data and avoids rejection of possibly highly-beneficial therapies that might be 'not significant'. This approach is gaining acceptance, but progress will be hastened if the shortcomings of null-hypothesis-significance testing are understood.

Introduction

The scientific method is characterised by the formulation of theories and the evaluation of specific predictions derived from those theories against experimental data. Decisions about whether predictions and their parent theories are supported or not by data are made using statistical inference. Thus the examination of theories and the evaluation of therapies in light of data and progression of 'knowledge' hinge directly upon how well the inferential procedures are used and understood. The dominant approach to statistical inference is null-hypothesis-significance testing (NHST). NHST has a particular underpinning logic that requires strict application if its use is to be of any value at all. Even when this strict application is followed, it has been argued that the underpinning 'yes or no' decision logic and the value of the 'sizeless' outcomes produced from NHST are at best questionable and at worst can hinder scientific progress (Ziliak and McCloskey, 2008, Batterham and Hopkins, 2006, Krantz, 1999, Sterne and Smith, 2001). The failure to understand and apply methods of statistical inference correctly can lead to mistakes in the interpretation of results and subsequently to bad research decisions. Misunderstandings have a practical impact on how research is interpreted and what future research is conducted, so impacts not only researchers but any consumer of research. This paper will clarify NHST logic, highlight limitations of this approach and suggest an alternative approach to statistical inference that can provide more useful answers to research questions while simultaneously being more rational and intuitive.

Scientific inference

With a clear picture of scientific inference, it is easier to understand the 'fit' of different statistical approaches to what we wish the scientific method to achieve. Science is a way of working involving formulation of theories or guesses about how the world works, calculation of the specific consequences of those guesses (i.e. hypotheses about what should be observed if the theory is correct), and the comparison of actual observations to those predictions (Chalmers, 1999). This description of the scientific method is not contentious. How observations are used to establish the 'truth' of theories is more contentious. The nature of philosophy is such that there will never be wholesale agreement, but a philosophy of science proposed by Sir Karl Popper is generally considered an ideal to strive towards. Popper wrote that theories must make specific predictions and importantly, those predictions should be potentially falsifiable through experiment (Popper, 1972b). It was Popper's falsifiability criteria that differentiated his philosophy from the consensus approach of 'truth by verification' that predominated previously, while simultaneously overcoming the problem of inductive reasoning highlighted by Scottish philosopher David Hume (Hume, 1963). In short, Popper showed that it was impossible to 'prove' a theory no matter how many observations verified it, but that a single contrary observation could 'disprove' or falsify a theory. This thesis is often explained using the 'white swan' example. Imagine a hypothesis that all swans are white. No amount of observations of white swans could prove the hypothesis true as this assumes all other swans yet to be observed will also be white and uses inductive reasoning. A single observation of a black (or other non-white) swan could however, by deductive reasoning, disprove the hypothesis (Ladyman, 2008). In Popper's philosophy, scientists should derive *specific* hypotheses from general theories and design experiments to attempt to falsify those hypotheses. If a hypothesis withstands attempted falsification, it and the parent theory are not proven, but have survived to

face further falsification attempts. Theories that generate more falsifiable predictions and more specific predictions are to be preferred to theories whose falsifiable predictions are fewer in number and vague. This latter point is particularly important in relation to NHST and will be expanded upon later.

Truth, variability and probability

Critics of Popper argue that, in reality, scientists would never reject a theory on the basis of a single falsifying observation and that there is no absolute truth that more successful theories move towards (Kuhn, 1996). Popper agreed and acknowledged that it would be an accumulation of falsifying evidence that 'on balance of probability' would lead to the conclusion that a theory had been disproven (Popper, 1972b). Herein lie two important links between statistical and scientific inference, namely that *probability* must be the basis for conclusions about theories because of *variability* in the results of different experiments on the same theory. Uncertainty is inescapable, but statistics can allow quantification of uncertainty in the light of variability. British polymath Sir Ronald Fisher first suggested a method of using probability to assess strength of evidence in relation to hypotheses (Fisher, 1950, Fisher, 1973). Fisher's contributions to statistics include the introduction of terms such as 'null hypothesis' (denoted as H_0), 'significance' and the concept of degrees of freedom, random allocation to experimental conditions and the distinction between populations and samples (Fisher, 1950, Fisher, 1973). He also developed techniques including analysis of variance amongst others. He is perhaps better known for suggesting a p (probability) of 0.05 as an arbitrary threshold for decisions about H_0 that has now achieved unjustified, sacrosanct status (Fisher, 1973).

Fisher's null

Fisher's definition of the null hypothesis was very different from what we currently understand it to mean and is possibly the root cause of philosophical and practical problems with NHST that will be discussed in this paper. In Fisher's work, the null was simply the hypothesis we attempt to 'nullify' or in other words 'falsify' (Fisher, 1973). With this understanding, he was actually referring to what we now call the 'experimental' hypothesis (denoted as H_1) and his procedures were well aligned with Popper's falsification approach. The conventional zero-point null hypothesis and the procedures for establishing a decision-making procedure about H_0 (i.e. retain or reject) that predominate today were created by Polish mathematician Jerzy Neyman and British statistician Egon Pearson (Neyman and Pearson, 1933). Despite the $p < 0.05$ being attributed to Fisher as a threshold for making a decision about (his version of) H_0, he was opposed to the idea of using threshold probabilities and argued vigorously in the literature with Neyman and Pearson about this (Ziliak and McCloskey, 2008). Instead, Fisher argued that probability could be used as a continuous measure of strength of evidence against the null hypothesis (Fisher, 1973), a point that, despite his genius, he was gravely mistaken about.

Defining probability

Generally speaking, there are two interpretations of probability in statistics. The first is subjective and the second objective. Subjective probability is the most intuitive and describes a personal degree of belief that an event will occur. It also forms the basis of the Bayesian method of inference. In contrast, the objective interpretation of probability is that probabilities are not personal but exist independent of our beliefs. The NHST approach and Fisher's ideas are based on an objective interpretation of probability proposed by Richard von Mises (von Mises, 1928). This interpretation is best illustrated using a coin-toss example. In a fair coin, the probability of heads is 0.5 and reflects the proportion of times we *expect* the coin to land on heads. However, it cannot be the proportion of times it lands on heads in any *finite* number of tosses (e.g. if in 10 tosses we see 7 heads, the probability of heads is not 0.7). Instead, the probability refers to an *infinite number of hypothetical* coin tosses referred to as a 'collective' or in more common terms a 'population' of scores of which the real data are assumed to be a sample. The population must be clearly defined. In this example, it could be all hypothetical sets of 10 tosses of a fair coin using a precise method under standard conditions. Clearly, 7 heads from 10 tosses is perfectly possible even with a fair coin, but the more times we toss the coin, the more we would expect the proportion of heads to approach 0.5. The important point is that the probability applies to the hypothetical-infinite collective and *not* to a single toss or even a finite number of tosses. It follows that objective probabilities also do not apply to hypotheses as a hypothesis in the NHST approach is simply retained or rejected in the same way that a single event either happens or does not, and has no associated population to which an objective probability can be assigned. Most scientists believe a *p* value from a significance test reveals something about the probability of the hypothesis being tested (generally the null). Actually a *p* value in NHST says *nothing* about the likelihood of H_0 or H_1 or the strength of evidence for or against either one. It is the probability of data as extreme or more extreme than that collected occurring in a hypothetical-infinite series of repeats of an experiment *if H_0 were true* (Oakes, 1986). In other words, the truth of H_0 is assumed and is fixed, *p* refers to all data from a hypothetical distribution probable under or consistent with H_0. It is the conditional probability of the observed data assuming the null hypothesis is true, written as $p(D|H)$.

Null-Hypothesis-Significance Testing logic

Based on the objective interpretation of probability, the NHST approach was designed to provide a dichotomous decision-making procedure with known and controlled long-run error rates. Neyman and Pearson were clear about this and in the introduction of their classic paper to the Royal Society stated "… as far as a particular hypothesis is concerned, no test based on the (objective) theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis" (Neyman and Pearson, 1933). Instead, they set about defining rules to govern decisions about retaining or rejecting hypotheses such that wrong decisions would not often be made, but the probability of making them in the long run would be known.

The starting point of the N-P approach is a pair of contrasting hypotheses (H_0 and H_1). For example, H_0 could be that μ_I (population mean ankle dorsi-flexion angle given therapy x) = μ_C (population mean ankle dorsi-flexion angle given no therapy – i.e. control group), or to put it another way, the difference between μ_I and μ_C is zero. The alternative (H_1) is then generally of the form $\mu_I \neq \mu_C$ i.e. the

population mean of the therapy and control groups will not be equal / will differ. Here we have the first philosophical issue with the conventional use of NHST. Under the philosophy of Popper, a hypothesis should be a specific prediction such that it is highly falsifiable. Popper argued a theory that allows everything explains nothing (Popper, 1972a) i.e. falsifying a null of 'no difference' simply allows for any magnitude of difference in any direction – hardly a severe test of a theory! Furthermore, the hypothesis under consideration (i.e. a zero-effect null) is not actually the hypothesis of interest, but is simply a straw man that the researcher does not believe or they would not be performing the experiment. Not surprisingly, Popper was not a supporter of NHST (Dienes, 2008). A practical issue is also raised here. Ignoring the philosophical problem, if a null of 'no difference' is rejected, a question that remains is how big is the effect? It is generally the size of effect of a therapy versus a control condition / group that is of real interest, not simply that the effect is different from zero in some unspecified amount and direction.

The illogic of NHST

Because H_0 and H_1 are mutually exclusive, if H_0 is rejected, by deduction H_1 is assumed true and vice versa, if H_0 is not rejected, H_1 is assumed false. However, statistical inference and indeed science does not deal in absolute proofs, truths or falsehoods, there is always uncertainty. If this uncertainty is extended to this example, we have: If H_0 then *probably* NOT H_1, data arise consistent with H_1, therefore H_0 is *probably* false. This logic has been challenged. Pollard and Richardson (1987) highlight a flaw using the following example: 'if a person is American, they are probably not a member of Congress; person x is a member of Congress therefore person x is probably not American'. Furthermore, Oakes (1986) points out that we are concluding the truth of H_1 based on H_0 being unlikely, when H_1 might be even less likely but we shall never know as it has not been tested, nor has the likelihood of multiple other possible versions of H_1. This paradox has been called the fallacy of the transposed conditional (Ziliak and McCloskey, 2008)

Errors in decision making with NHST

NHST logic gives rise to two possible errors in decision making, namely, wrongly rejecting H_0 when it is actually true (type I error) and wrongly retaining H_0 when it is actually false (type II error). To correctly use NHST, the researcher chooses the acceptable risk of each error type in advance of testing (subjectively and according to the type of error the researcher deems more harmful). The NHST procedure ensures these error rates are fixed and controlled such that, over an infinite number of hypothetical repeats of the experiment, the probability of making each type of error is known (Neyman and Pearson, 1933). The probability of a type I error is termed α and is conventionally and without reason set at 0.05. The probability of a type II error is termed β. This error rate is less formally agreed and in the majority of research is never actually specified or controlled, violating NHST decision rules. The few studies that do control β generally specify it at 0.2 giving the study an 80% chance (1 – β) of correctly rejecting a false H_0, or 80% statistical power. For the type II error rate to be fixed, a minimum worthwhile / interesting effect that researchers wish to detect must be specified in advance of data collection, and an appropriate sample size calculated that provides the specified power (and thus type II error rate). *Exactly* that number of participants

must be tested to control the type II error rate at the specified level. Failure to specify β in advance and to control it by testing an appropriately-sized sample renders decisions about H_0 impossible when it cannot be rejected i.e. was the effect really likely to be zero or was it different from zero but undetectable due to low power? The only conclusion that can be drawn is one of uncertainty. Failure to test the 'correct' number of participants can also result in effects not large enough to be of practical / clinical importance being deemed 'significant' if a larger-than-necessary sample is used (i.e. the experiment is overpowered). It should be acknowledged that, in reality, the sample sizes in published studies are rarely based on an appropriate calculation. In fact, few studies would actually take place if the estimated sample size had to be obtained, as they are often prohibitively large. This is unlikely to change, but researchers simply need to be aware that, strictly, the decision logic of NHST only applies when both error rates are actually controlled.

An example of NHST in practice

In the 'therapy-versus-control-group' example outlined in the previous section, having specified hypotheses and error rates and calculated an appropriately-sized sample, samples (assumed to be random) are taken from the hypothetical collectives of interest. The sample means for the therapy (M_t) and the control (M_c) groups and the difference between them can be calculated. The standard error of the mean difference (SEM_{diff}) can also be calculated. These values are used to calculate a sample statistic that combines them, in this case a *t* statistic, where $t = (M_i - M_c / SEM_{diff})$. In order to calculate the long-run probability that such a *t* statistic could occur given H_0, the hypothetical collective that gave rise to this *t* statistic must be defined. The collective in this case is a probability distribution of *t* statistics from an infinite number of hypothetical repeats of the experiment assuming H_0 is true (so having a mean *difference* between therapy and control groups of 0 and an assumed-normal distribution). This theoretical distribution represents all values of *t* that are probable if H_0 were true. Now the decision rule is applied by defining a rejection region of the distribution where *t* statistics are deemed so extreme that they would occur infrequently in the long run if H_0 were true. The probability of obtaining a *t* score in that region is equal to the predefined α. Thus, if the observed *t* from the sample data falls into the region of the probability distribution beyond α, H_0 is rejected as such a *t* statistic would occur infrequently in the long run if H_0 were true. Note that the interpretation of such a finding is that 'an effect exists in the sample that should not be likely if there really was no effect in the collective sampled – therefore, there is likely to be an effect larger than zero in the collective sampled. If you find this confusing, you are not alone. Little can be concluded about the size of the effect of the therapy versus the control or the practical / clinical value of it, which is arguably much more important (Ziliak and McCloskey, 2008, Batterham and Hopkins, 2006).

Note that the *exact* probability of the observed *t* is irrelevant to the decision to reject H_0. It need only be less than α. The requirement for authors to report exact *p* values is actually redundant and stems from a mistaken belief that *p* is in some way a measure of strength of evidence against H_0 such that the lower the *p* the stronger the evidence against H_0 and by extension for H_1. It has already been discussed that *p* is a conditional probability of the observed *data* occurring assuming a fixed H_0, i.e. $p(D|H_0)$, as such *p* is <u>not</u> an indicator about either hypothesis. Most researchers believe the *p* value tells them something about the probability of their hypothesis in light of the data i.e. $p(H|D)$,

and that the magnitude of *p* is in some way a continuous measure of the weight of evidence against H_0, when in fact, any given *p* could simply be a function of random sampling variation (Cumming, 2012). It is worth expanding on this point to highlight how trust in *p* as a form of evidence is misplaced.

In the example provided, the *t* statistic for which *p* is calculated is derived from two random samples taken from hypothetical-infinite collectives. Different samples would produce different means and therefore a different *t* statistic and a different *p* value. The *p* value is thus a randomly-fluctuating variable that can and does jump in and out of the rejection region when an experiment is repeated exactly as before. Being so unreliable, how can a researcher possibly have trust in a *p* value as a source of evidence on which to base a decision about their hypotheses? (Cumming, 2012). Note also the desire for *p* to indicate 'magnitude' of evidence and effect. The importance of estimating the likely 'size' of an effect has been recognised as the most important goal of statistical inference but a *p* value is not it (Ziliak and McCloskey, 2008, Hopkins et al., 2009, Batterham and Hopkins, 2006).

Size matters, but so does uncertainty

The goal of statistical inference is to estimate likely 'true / large-sample' effects based on random samples from the collective(s) of interest. However, because different samples always produce different estimates, we must express the uncertainty of our estimates. This is achieved using confidence intervals. The exact definition of a confidence interval is debated, but it is generally accepted to be a plausible range in which the true population effect would be likely to fall with repeats of the experiment. The number of repeats is infinite and 'likely to fall' refers to the percentage of times a calculated interval would contain the 'true' effect (conventionally 95%). In the context of the therapy-control example already used, if we repeated the experiment an infinite number of times and calculated a confidence interval each time, 95% of them would contain the 'true' effect and 5% would not. We can never know whether the interval calculated in *this* study is one of those that does or does not contain the 'true' effect. Taking a pragmatic view however, as 95% of all intervals *will* contain the 'true' effect then our interval is more likely to be one of those that does than one of those that does not. The use of interval estimation instead of NHST is becoming a necessity for publication in many journals. In fact, the International Committee of Medical Journal Editors (2010) now make the following statement in their guidelines; "quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as P values, which fail to convey important information about effect size and precision of estimates"

Suppose a 95% confidence interval for the mean difference between the control and therapy group ankle dorsi-flexion angle was calculated as 1° to 10° in favour of the therapy group. A pragmatic interpretation is that the intervention is likely to result in an improvement in ankle dorsi-flexion range of between 1° and 10° *more* than no therapy. Assume the mean difference between the sample groups was 5.5° and that the *p* value of a NHST (*t* test in this case) was < 0.05. Using the latter, the therapy would be deeded successful, with the best estimate of the 'true / large sample' effect of the therapy being an improvement in ankle dorsi flexion of 5.5°. Using the confidence

interval, we are still confident that there is a benefit of the therapy over the control, but the size of the improvement might be as little as 1° or as large as 10°. The confidence interval factors sampling variability into the estimate and thus expresses the uncertainty about what the true mean difference between therapy and control might be. Imagine how the discussion section of a paper might differ with the NHST and confidence-interval results. The key to the discussion section in both situations should be the *context* of what size of improvement in dorsi flexion is beneficial (for function, quality of life, as a return for time / cost invested in the therapy etc.) (Batterham and Hopkins, 2006, Ziliak and McCloskey, 2008). Given these factors, the conclusion might be that the benefit of the therapy is uncertain as it could be hardly beneficial at all or extremely beneficial. If a worthwhile improvement was say 1°, then the therapy is more than likely to be worthwhile, but if an improvement is not worthwhile or beneficial unless is exceeds 5°, then the conclusion will be less favourable. Note that for a therapy to be deemed 'successful' with *reasonable likelihood* using this approach, two conditions must be satisfied: 1) the confidence interval must exclude zero and; 2) the lower limit of the confidence interval must be equal to or greater a smallest clinically worthwhile / beneficial effect. Both conditions can lead to very conservative conclusions (Atkinson and Nevill, 2001). Suppose that in the above example, a smallest-beneficial improvement was deemed to be 3°, the conclusion would have to be that the therapy is possibly of no use even though it could be very useful indeed. If the interval ran from -1° to 10°, again the conclusion would be unfavourable because not only might the true effect be no effect at all, it might also result in a worsening of ankle range, though the true effect could still be a gain in dorsi-flexion of up to a 10°, and more of the interval lies above the smallest-beneficial effect than below it! There is a real danger of throwing the baby out with the bath water. Clearly, what is required is a method that allows calculation of the likelihood of the true effect in relation to a smallest effect of clinical / practical importance. Researchers and in particular clinical practitioners make these judgements subjectively anyway, so why not incorporate them into statistical inference. There is a method for doing just this (Batterham and Hopkins, 2006, Hopkins et al., 2009).

Magnitude-based inference

Magnitude-based inference (MBI) involves calculating the chances (probability) that the 'true' effect exceeds or is less than an *a priori* determined smallest-worthwhile / clinical or practically-important effect. Proponents of this approach argue that the criteria of reasonable certainty adopted in the confidence interval approach is too stringent and can result in conclusions of therapies / interventions being non beneficial when in fact the likelihood of them being practically / clinically worthwhile are extremely high (Batterham and Hopkins, 2006). Many researchers feel uncomfortable making subjective decisions about the value of the smallest worthwhile effect and argue that hypothesis testing is more scientific and objective. However, it was pointed out earlier that estimating the sample size to control type II error in a NHST-based study, requires an estimate of the smallest-worthwhile effect that is no less subjective. The choice of the smallest beneficial / clinically worthwhile effect size provides the crucial context against which the results of the study should be interpreted.

Making inferences about 'true' effect magnitudes based on smallest-worthwhile effects facilitates more enlightened interpretations of data, though the process and the interpretation of the outcome

can be more challenging. It is often easier to "inspect the *p* value…..and then declare that either there is or there is not an effect" (Battherham and Hopkins, 2006, p.56), but to do so might prevent new knowledge about practically beneficial therapies and restrain progress of research. Lack of knowledge of MBI need not be a barrier as the tools and instructions required to perform, interpret and present such analyses are readily available and simple to apply (Hopkins, 2000). The MBI approach is also congruent with Popper's falsification approach whereby the expected effect of a therapy is also the smallest worthwhile / important effect for the therapy to be supported. Note that the smallest-worthwhile effect is precise and is the effect of interest (rather than a zero-effect null), consistent with the falsification approach (Popper, 1972a).

The MBI approach is gaining acceptance in some of the most popular periodicals for sport, exercise and medicine-related research (Hopkins et al., 2009, May et al., 2007, Hopkins et al., 2011). For a therapy-related example of the application of MBI in a published study, readers are referred to May *et al.* (2007) which examined the effectiveness of anti-inflammatory gel application in the treatment of wrist tenosynovitis in kayakers. In addition to the fictional ankle dorsi flexion example described above, the study helps to further demonstrate the practical application of the MBI approach.

Conclusions

Significance testing is a procedure for making black and white decisions about zero-effect null hypotheses with known and controlled long-run error rates. Type I and type II error rates must be specified in advance and a required sample size calculated and tested to ensure type II error rate is controlled at the specified level. The outcome allows conclusions about the likely existence of non-zero effects but provides no information about the likely size of true effects or their practical / clinical value. The approach is also at odds with accepted philosophies of science.

To estimate the true size of an effect and its likelihood in relation an effect magnitude of practical / clinical importance, magnitude-based inference provides the solution. The approach is gaining acceptance and progress will be hastened if researchers appreciate the shortcomings of traditional NHST. It is recommended that researchers begin to incorporate the subjective-clinical judgements commonly made in light of experimental data, and expressions of uncertainty, into their inferential statistical analysis. This will ensure more considered and enlightened interpretations of data and avoid discounting possibly highly practically / clinically beneficial treatments because they are not statistically significant.

Key points

Even used properly, NHST only gives yes / no decisions about zero-effect null hypotheses which are always false and of no interest anyway.

Estimates of the size of true / large sample effects in NHST do not encompass uncertainty due to sampling variation.

Outcomes of NHST are without context of what is clinically / practically important.

Statistical inference should estimate the likely size of true effects using confidence intervals.

Confidence intervals should be interpreted relative to *a priori* specified clinically / practically worthwhile effect magnitudes.

Probabilities of true effects in relation to clinically / practically worthwhile effects should form the basis for interpretation of experimental data.

References

Atkinson G & Nevill AM (2001). Selected issues in the design and analysis of sport performance research. *J Sports Sci,* **19,** 811-827.

Batterham AM & Hopkins WG (2006). Making meaningful inferences about magnitudes. *Int J Sports Physiol Perf,* **1,** 50-57.

Chalmers AF (1999). *What is this thing called science?,* Buckingham, Open University Press.

Cumming G (2012). *Understanding the new statistics: Effect sizes, confidence intervals and meta analysis,* New York, Taylor and Francis Group.

Dienes Z (2008). *Understanding Psychology as a Science: an introduction to scientific and statistical inference,* Basingstoke, Palgrave Macmillan.

Fisher R (1950). *Statistical methods for research workers.,* London, Oliver and Boyd.

Fisher R (1973). *Statisitcal methods and scientific inference,* London, Collins Macmillan.

Hopkins WG. 2000. *A new view of statistics. Internet Society for Sport Science* [Online]. Available: http://www.sportsci.org/resource/stats/.

Hopkins WG, Batterham AM, Impellizeri FM, Pyne DB & Rowlands DS (2011). Statistical perspectives: all together NOT. *J Physiol,* **589,** 5327-5329.

Hopkins WG, Marshall SW, Batterham AM & Hanin J (2009). Progressive statistics for studies in sports medicine and exercise science. *Med Sci Sports Exerc,* **41,** 3-12.

Hume D (1963). *An enquiry concerning human understanding,* Oxford, Oxford University Press.

Icmje. 2010. *Uniform requirements for manuscripts submitted to biomedical journals* [Online]. Available: http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html#e [Accessed 16/09/14 2014].

Krantz DH (1999). The null hypothesis testing controversy in psychology. *J Am Stat Assoc,* **94,** 1372-1381.

Kuhn TS (1996). *The Stucture of Scientific Revolutions,* Chicago, The University of Chicago Press.

Ladyman J (2008). *Understanding philosophy of science,* Oxford, Routledge.

May JJ, Lovell G & Hopkins WG (2007). Effectiveness of 1% diclofenac gel in the treatment of wrist extensor tenosynovitis in long distance kayakers. *J Sci Med Sport,* **10,** 59-65.

Neyman J & Pearson ES (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos Trans R Soc Lond A,* **231,** 289-337.

Oakes M (1986). *Statistical inference: A commentary for the social and behavioural sciences,* New Jersey, Wiley.

Pollard P & Richardson JTE (1987). On the probability of making type I errors. *Psychol Bull,* **102,** 159-163.

Popper KR (1972a). *The Logic of Scientific Discovery,* London, Hutchinson & Co Ltd.

Popper KR (1972b). *Conjectures and Refutations:The Growth of Scientific Knowledge,* London, Routledge and Kegan Paul Ltd.

Sterne JaC & Smith GD (2001). Sifting the evidence - what's wrong with significance tests? *BMJ,* **322,** 226-231.

Von Mises R (1928). *Probability, Statistics and Truth,* London, Allen and Unwin.

Ziliak ST & Mccloskey DN (2008). *The Cult of Statistical Significance: how the standard error costs us jobs, justice, and lives,* Michigan, University of Michigan Press.