Northumbria Research Link

1   Title: Distinguishing between statistical significance and practical/clinical meaningfulness using

2   statistical inference.

3   Submission Type: Current opinion

4   Authors: 1. Michael Wilkinson

5   Affiliation: 1. Faculty of Health and Life Sciences

6   Northumbria University

7   Correspondence address: Dr Michael Wilkinson

8   Department of Sport, Exercise and rehabilitation

9   Northumbria University

10   Northumberland Building

11   Newcastle-upon-Tyne

12   NE1 8ST

13   <u>ENGLAND</u>

14   Email: mic.wilkinson@northumbria.ac.uk

15   Phone: 44(0)191-243-7097

16

17   Abstract word count: 232

18   Text only word count: 4505

19   Number of figures = 2; number of tables = 0

20

21

22

23

24

25

26

27

28

29   **Abstract**

30   Decisions about support for predictions of theories in light of data are made using statistical

31   inference. The dominant approach in sport and exercise science is the Neyman-Pearson significance-

32   testing approach. When applied correctly it provides a reliable procedure for making dichotomous

33   decisions for accepting or rejecting zero-effect null hypotheses with known and controlled long-run

34   error rates. Type I and type II error rates must be specified in advance and the latter controlled by

35   conducting an a priori sample size calculation. The Neyman-Pearson approach does not provide the

36   probability of hypotheses or indicate the strength of support for hypotheses in light of data, yet

37   many scientists believe it does. Outcomes of analyses allow conclusions only about the existence of

38   non-zero effects, and provide no information about the likely size of true effects or their practical /

39   clinical value. Bayesian inference can show how much support data provide for different hypotheses,

40   and how personal convictions should be altered in light of data, but the approach is complicated by

41   formulating probability distributions about prior-subjective estimates of population effects. A

42   pragmatic solution is magnitude-based inference, which allows scientists to estimate the true

43   magnitude of population effects and how likely they are to exceed an effect magnitude of practical /

44   clinical importance thereby integrating elements of subjective-Bayesian-style thinking. While this

45   approach is gaining acceptance, progress might be hastened if scientists appreciate the

46   shortcomings of traditional N-P null-hypothesis-significance testing.

47

48

49

50

51

52

53

54

55

56

57

58   **Running head**

59   Distinguishing statistical significance from practical meaningfulness

60

61

62

63 1.0 Introduction

64 Science progresses by the formulation of theories and the testing of specific predictions (or, as has
65 been recommended, the attempted falsification of predictions) derived from those theories via
66 collection of experimental data [1, 2]. Decisions about whether predictions and their parent theories
67 are supported or not by data are made using statistical inference. Thus the examination of theories
68 in light of data and progression of 'knowledge' hinge directly upon how well the inferential
69 procedures are used and understood. The dominant (though not the only) approach to statistical
70 inference in the sport and exercise research is the Neyman-Pearson approach (N-P), though few
71 users of it would recognise the name. N-P inference has a particular underpinning logic that requires
72 strict application if its use is to be of any value at all. In fact, even when this strict application is
73 followed, it has been argued that the underpinning 'black and white' decision logic and value of such
74 'sizeless' outcomes from N-P inference are at best questionable and at worst can hinder scientific
75 progress [3-6] The failure to understand and apply methods of statistical inference correctly can
76 lead to mistakes in the interpretation of results and subsequently to bad research decisions.
77 Misunderstandings have a practical impact on how research is interpreted and what future research
78 is conducted, so impacts not only researchers but any consumer of research. This paper will clarify
79 N-P logic, highlight limitations of this approach and suggest that alternative approaches to statistical
80 inference could provide more useful answers to research questions while simultaneously being more
81 rational and intuitive.

82

83 2.0 The origins of 'classical' statistical inference.

84 The statistical approach ubiquitous in sport and exercise research is often mistakenly attributed to
85 British mathematician and geneticist Sir Ronald Fisher (1890 – 1962). Fisher introduced terms such
86 as 'null hypothesis' (denoted as H_0) and 'significance' and the concept of degrees of freedom,
87 random allocation to experimental conditions and the distinction between populations and samples
88 [7, 8]. He also developed techniques including analysis of variance amongst others. However, he is
89 perhaps better known for suggesting a p of 0.05 as an arbitrary threshold for decisions about H_0 that
90 has now achieved unjustified, sacrosanct status [8]. Fisher's contributions to statistics were
91 immense, but it was Polish mathematician Jerzy Neyman and British statistician Egon Pearson who
92 suggested the strict procedures and logic for null hypothesis testing and statistical inference that
93 predominate today [9].

94

95 3.0 Defining probability.

96 The meaning of probability is still debated among statisticians, but generally speaking, there are two
97 interpretations. The first is subjective and the second objective. Subjective probability is probably
98 the most intuitive and underpins use of statements about probability in everyday life. It is a personal
99 degree of belief that an event will occur e.g. "I think it will definitely rain tomorrow". This is an
100 interpretation of probability generally applied to theories we 'believe' to be accurate accounts of the
101 world around us. In contrast, the objective interpretation of probability is that probabilities are not
102 personal but exist independent of our beliefs. The N-P approach is based on an objective, long-run-

frequency interpretation of probability proposed by Richard von Mises [10]. This interpretation is best and most simply 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 collective / population must be clearly defined. In this example, the collective 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 <u>not</u> to a single event or even a finite number of events. It follows that objective probabilities also do not apply to hypotheses as a hypothesis in the N-P approach is simply retained or rejected in the same way that a single event either happens or does not, and has no associated collective to which an objective probability can be assigned. This might come as a surprise, as 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 N-P statistics says *nothing* about the truth or otherwise 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* [11]. In other words, the truth of H_0 is assumed and is fixed, *p* refers to all data from a 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 $p(D|H)$. I contend that what scientists really want to know (and what most probably think *p* is telling them) is the probability of a hypothesis in light of the data collected, or $p(H|D)$ i.e. 'does my data provide support for, or evidence against the hypothesis under examination?'. The second conditional probability cannot be derived from the first. To illustrate this, Dienes [12] provides a simple and amusing example summarised below:

P(dying within two years|head bitten off by shark) = 1

Everyone that has their head bitten off by a shark will be dead two years later.

P(head bitten off by shark|died in the last two years) ~ 0

Very few people that died in the last two years would be missing their head from a shark bite so the probability would be very close to zero. Knowing $p(D|H)$ does not tell us $p(H|D)$ which is really what we would like to know. Note that the notation '*p*' refers to a probability calculated from continuous data (interval or ratio) whereas '*P*' is the notation for discrete data, as in the example above. Unless the example requires it, the rest of this paper will use '*p*' when discussing associated probabilities and will assume that variables producing continuous data are the topic of discussion.

4.0 Neyman-Pearson logic and decision rules.

N-P statistics are based on the long-run-frequency interpretation of probability so tell us nothing about the probability of hypotheses of interest or how much data support them. Neyman and

143   Pearson were very clear about this and in the introduction of their seminal paper to the Royal

144   Society stated "… as far as a particular hypothesis is concerned, no test based on the (objective)

145   theory of probability can by itself provide any valuable evidence of the truth or falsehood of that

146   hypothesis" [9]. Instead, they set about defining rules to govern decisions about retaining or

147   rejecting hypotheses such that, by following them, in the long run, wrong decisions will not often be

148   made.

149   The starting point of the N-P approach is the formation of a pair of contrasting hypotheses (H_0 and

150   H_1). For example, H_0 could be that μ_s (population mean time to fatigue given supplement x) = μ_p

151   (population mean time to fatigue given placebo), or to put it another way, the difference between μ_s

152   and μ_p is zero. The alternative (H_1) could be $\mu_s > (\mu_p + 20)$ i.e. that the supplement will increase time

153   to fatigue by at least 20 units. Note that H_0 need not be 'no difference' ($\mu_s = \mu_p$) as is usually the case.

154   It could be a hypothesised difference or even range of differences that ought not to be possible

155   given the theory being tested. In fact, under the philosophy of Popper, the latter constitutes a far

156   more severe test of a theory, such that survival of the test (i.e. failure to reject H_0) offers strong

157   corroboration for the theory [1]. By the same token, H_1 ought also to be a specific difference or band

158   of differences because merely specifying that $\mu_s - \mu_p > 0$ is a vague prediction, rules out little and

159   allows for any effect greater than 0. Furthermore, with continuous data, an effect of zero has a

160   probability of precisely zero as does any exact integer so such an H_0 is always false! It would be

161   fruitful to elaborate on this link between philosophy and statistical inference, but it is a digression

162   from the issue at hand, which is how N-P statistics proceed from here.

163   The two hypotheses should be mutually exclusive such that if H_0 is rejected, then by deductive logic

164   H_1 is assumed true and vice versa, if H_0 is not rejected, H_1 is assumed false. However, statistical

165   inference and indeed science does not deal in absolute proofs, truths or falsehoods, there is always a

166   magnitude of uncertainty. If this uncertainty is extended to this example of N-P logic, we have: If H_0

167   then *probably* NOT H_1, data arise consistent with H_1, therefore H_0 is *probably* false.

168   This logic has been challenged. Pollard and Richardson [13] highlight a flaw using the following

169   example: 'if a person is American, they are probably not a member of Congress; person x is a

170   member of Congress therefore person x is probably not American'. Furthermore, Oakes [11] points

171   out that we are concluding the truth of H_1 based on H_0 being unlikely, when H_1 might be even less

172   likely but we shall never know as it has not been tested nor has the likelihood of multiple other

173   possible versions of H_1. This paradox has been called the fallacy of the transposed conditional [3].

174   N-P logic gives rise to two possible errors in decision making, namely wrongly rejecting H_0 when it is

175   actually true (type I error) and wrongly retaining H_0 when it is actually false (type II error). Neyman

176   and Pearson devised procedures whereby the acceptable risk of each type of error were specified in

177   advance of testing (subjectively and according to the type of error the researcher deemed more

178   harmful), and were then fixed and controlled such that, over an infinite number of hypothetical

179   repeats of the experiment, the probability of making each type of error was known [9]. The

180   probability of a type I error is termed α and is conventionally and without reason set at 0.05. The

181   probability of a type II error is termed β. This error rate is less formally agreed and in the majority of

182   research in sport and exercise is never actually specified or controlled, violating N-P decision-rule

183   logic. The few studies that do control β generally specify it at 0.2 giving the study an 80% chance $(1 -$

184    $\beta)$ of correctly rejecting a false H_0 or having 80% statistical power. That researchers class the

185   consequences of a type II error as less harmful than a type I error is interesting and the discussion of
186   this could form a paper in its own right. Nevertheless, for the type II error rate to be fixed, a
187   minimum worthwhile / interesting effect that researchers wish to detect must be specified in
188   advance of data collection, and an appropriate sample size calculated that provides the power (and
189   thus the type II error rate) deemed acceptable. *Exactly* that number of participants should be tested
190   to control the type II error rate at the specified level. Failure to specify β in advance and ensure it is
191   controlled by testing an appropriately-sized sample renders decisions about H_0 impossible in
192   situations where it cannot be rejected. It can also result in effects not large enough to be of practical
193   / clinical importance being deemed 'significant' if a larger-than-necessary sample is collected (i.e. the
194   experiment is overpowered).

195   In the time-to-fatigue example outlined previously, having specified hypotheses and error rates and
196   calculated an appropriately-sized sample, a sample (assumed to be random) is taken from the
197   population(s) of interest. The sample means for the supplement (M_s) and the placebo (M_p) and the
198   difference between them can be calculated. The standard error of the mean difference (SEM_{diff}) can
199   also be calculated. These values are then used to calculate a sample statistic that combines them, in
200   this case a *t* statistic, where $t = (M_s - M_p / SEM_{diff})$. In order to calculate the long-run probability that
201   such a *t* statistic could occur given H_0 is true, the collective that gave rise to this *t* statistic must be
202   defined. The collective in this case is a probability distribution of *t* statistics from an infinite number
203   of hypothetical repeats of the experiment assuming H_0 is true (so having a mean of 0 and an
204   assumed-normal distribution). The distribution represents all values of *t* that are probable given H_0.
205   Now the decision rule is applied by defining a rejection region of the distribution where *t* statistics
206   are deemed so extreme that they would occur infrequently in the long run if H_0 is true. The
207   probability of obtaining a *t* score in that region is equal to the predefined α. Thus, if the observed *t*
208   from the sample data falls into the region of the probability distribution beyond α, in the N-P
209   approach, H_0 is rejected as such a *t* statistic would occur infrequently in the long run if H_0 were true.
210   Note that the interpretation such a finding is that 'an effect exists that should not be likely if there
211   really was no effect'. Little can be concluded about the size of the effect or the practical / clinical
212   value of it, which is arguably much more important [3, 4] (see **Fig 1**)

213

214   **Fig 1**. A distribution of probable *t* scores given H_0 of no mean difference between μ_s and μ_p. Note, the
215   shaded rejection region (representing possible values of *t* as or more extreme than that observed) is
216   in a single tail of the distribution because H_1 in the example above is a directional hypothesis i.e. μ_s >
217   (μ_p + 20). Note μ_s is the population mean time to fatigue after a nutritional supplement, μ_p is the
218   population mean time to fatigue after a placebo, H_0 and H_1 denote the null and experimental
219   hypotheses respectively.

220

221   Note that the *exact* probability of the observed *t* is irrelevant to the decision to reject H_0. It need
222   only be less than α. Furthermore, having set α at 0.05, upon a significant result with *p* of 0.004, an
223   author should not report significance at $p < 0.01$ because this was not the long-run error rate
224   specified before data were collected. This is fairly common though. The requirement for authors to
225   report exact *p* values is also redundant and stems from a mistaken belief that the calculated *p* is in
226   some way a measure of strength of evidence against H_0 such that the lower the *p* the stronger the

227 evidence against H_0 and by extension for H_1. This common misinterpretation of p reveals the
228 researcher's true interpretation of probability i.e. that it is subjective and can be assigned to
229 individual events and hypotheses. This interpretation of probability forms the basis of Bayesian
230 statistical inference that will be introduced shortly. Most researchers probably believe the p value
231 tells them something about the probability of their hypothesis in light of the data i.e. $p(H|D)$, and
232 that the magnitude of p is in some way a continuous measure of the weight of evidence against H_0,
233 when in fact, any given p could simply be a function of random sampling variation [14]. Note also the
234 desire for p to indicate 'magnitude' of evidence in this example. The importance of estimating the
235 likely 'size' of an effect has been recognised as a more important goal of statistical inference [3, 4,
236 15]

237

238 4.1 Other criticisms of Neyman-Pearson statistics

239 N-P statistics are sensitive to the conditions under which a researcher chooses to stop collecting
240 data and perform the analysis, called the stopping rule. For example, a stopping rule could be (and
241 often is) 'test as many participants as is common in the area of interest'. Unless the number of
242 participants happens to match that required to achieve a predefined power to detect a smallest
243 worthwhile effect, this rule is poor. Power is not controlled at any known value and the probability
244 of type II error is unknown. Should a non-significant result arise, the researcher cannot know if the
245 sample statistic arose by chance alone and H_0 should be retained, or the study was not powerful
246 enough to reject H_0 when it was actually false. The only conclusion to draw is one of uncertainty.
247 Another illegitimate stopping rule is to carry on testing participants until a significant result is
248 achieved. The issue here is that, even if H_0 is true, a significant result is guaranteed to occur
249 eventually i.e. both power and α are 1. The legitimate stopping rule under the N-P approach is to
250 calculate the sample size that will yield the required power and β before data are collected, then test
251 that number of participants. An amalgam of the two illegitimate stopping rules described here is
252 setting out to test the number of participants common in the area, and upon analysing the data and
253 finding a non-significant result, adding a few more and testing again to find a significant result (say p
254 = 0.03). The type I error rate for the 'second look' cannot be 0.05, it must be higher because there
255 have been two attempts to reject H_0 (it is actually a little under 0.1). Furthermore, the associated p
256 value of the second attempt is associated with a different collective to the first attempt i.e. a
257 collective defined by the stopping rule 'test the common number of participants, if not significant,
258 add more until significant'. To retain α of 0.05 for the two attempts, each attempt must be carried
259 out at a lower α level. There are many approaches to this, the simplest being the Bonferroni method
260 where each attempt is carried out at an α of 0.05/k and k is the number of attempts to reject H_0. This
261 problem arises any time more than one H_0 is tested and is a particular problem where effects not
262 specified as being of interest before data collection catch the researchers attention after data
263 collection. For example, the research might specify one particular comparison, but the researcher
264 threw in some extra (two) conditions while there was access to the participants, and the additional
265 comparisons show effects that appear interesting. The only effect that can be tested at the 0.05
266 level is the one specified in advance of data collection. The others must be tested at a lower level
267 because they belong to a collective defined by 'perform three t tests: if any of them are significant at
268 α of 0.05, reject that H_0' which actually has an α of just under 0.15 (almost a 15% chance of type I
269 error). The 'family' of tests to perform must be specified before data are collected. This seems

270   illogical as most scientists would agree that if data suggest an interesting effect, why should it
271   matter when you chose to think about the effect. Scientists that think this way are believers in the
272   likelihood law, which put simply, is that all the information relevant to inference is contained in the
273   data [16]. N-P statistics violate the likelihood law because inferential decisions are based on when
274   one chose to think about interesting effects. Given this situation, the value of N-P statistics for
275   making valid inferential judgements about hypotheses has been questioned [3, 4, 11]. Note that
276   while the preceding section has discussed 'significance' testing, the same issues (i.e. multiple testing,
277   unplanned comparisons etc.) also apply to confidence intervals calculated in the frequentist-
278   probability framework, though it must be acknowledged that interval estimation is superior to and
279   more informative than the dichotomous decision procedures of null-hypothesis-significance testing
280   as it offers some estimate of the likely magnitude of an effect though such estimates are still not
281   often framed against pre-determined 'interesting / worthwhile' effects. Many users of frequentist
282   confidence intervals prefer a 95% interval estimate and interpret these in relation to whether the
283   interval spans zero – hence essentially still 'testing' for a null hypothesis of zero effect at a threshold
284   alpha of 0.05 and somewhat missing the point of 'estimating' the likely magnitude of a population
285   effect [4, 6].

286

287   5.0 Bayesian inference – combining prior knowledge with observed data

288   It seems that most scientists wish statistics to provide probabilities of their theories being correct
289   and in fact many believe that a N-P *p* provides this. This is not and cannot be the case with objective
290   probabilities. It can however be the case with a subjective probability. Bayesian inference allows
291   scientists to alter initial degree of belief in a hypothesis in light of experimental data. It is likely that
292   most readers will not have heard of the Bayesian approach as N-P methods are the dominant and
293   unchallenged approach in sport and exercise research and most other sciences. Given that most will
294   scarcely recognise the names of these methods, let alone understand the conceptual differences and
295   issues of their use, unquestioning adoption of N-P statistics is hardly an informed choice.

296   Bayes theorem was developed by fellow of the Royal Society, Reverend Thomas Bayes (1702-1761)
297   while working on the problem of assigning a probability to a hypothesis given observed data. The
298   theorem is directly derived from the axioms of probability theory such that:

299      $$p(H|D) = p(D|H) \times p(H)/p(D)$$

300   *p*(H) is called the prior is a probability distribution of the unknown population effect suggested by
301   the researcher prior to collecting any data. *p*(H|D) is the posterior and is the probability distribution
302   of the unknown population effect (the prior) altered in light of the data that were collected. It
303   represents how prior estimates about an effect should be changed based on observations. *p*(D|H) is
304   the probability of the observed data arising given the prior estimated effect and is called the
305   likelihood of the hypothesis. It is distinct from the *p*(D|H) described in N-P statistics where the
306   hypothesis is held constant and the probability of data that did not occur but might have is
307   considered. Conversely, likelihood is *p*(obtaining *exactly* this sample mean|prior estimated effect)
308   where the likelihood of different effects (e.g. population means) are considered, but the data are
309   fixed. **Fig 2** shows the distinction between the meaning of *p*(D|H) in significance testing versus
310   Bayesian inference. Note the location of the effect of interest (mean difference) on the x axis in each

311   approach. Most researchers "think" like a statistician interested in likelihoods (panel B), yet apply a
312   statistical approach that does not mirror their beliefs (panel A).

313

314   **Fig 2**. Likelihood in Neyman-Pearson and Bayesian inference. (a) – a distribution of probable sample
315   means given H_0 of 'zero' difference; (b) – a distribution of probable population means given the
316   actual observed sample mean. ($M_s - M_p$) in both panels is the location of sample mean difference in
317   time to fatigue after supplementation and placebo respectively. The height of the likelihood curve in
318   panel (b) shows which population mean difference (in this example) is likely given the data. The
319   shaded area in (a) are values for mean difference that are unlikely assuming H_0 of zero difference.

320

321   The outcome of a Bayes analysis is generally expressed as an interval estimate for the magnitude of
322   the true population effect, called a credibility interval. This is similar to a confidence interval except
323   that it can be claimed that *this* interval has a specified probability (say 95%) of including the true
324   population effect.   However, the subjective choice of the components (e.g. mean and SD) of a prior
325   probability distribution for the estimated-unknown population effect can be difficult to defend and,
326   given the same data, two scientists with different prior opinions would obtain different posterior
327   distributions and estimates of the true population effect. Nevertheless, careful consideration of
328   what constitutes a practically / clinically meaningful effect, prior to data collection, is not only a
329   worthwhile venture but a must for meaningful interpretation of data analysis. While it is a
330   requirement of N-P inference to specify a smallest-worthwhile effect to control type II error,
331   'significance' and therefore conclusions relate to rejection of a zero-effect H_0 and is generally
332   irrespective of effect magnitude and therefore of questionable value [3, 4].

333

334   6.0 Magnitude-based inference: a pragmatic solution?

335   The frequentist use of probability dominates sport and exercise sciences, yet Bayesian incorporation
336   of prior beliefs is something that most scientist probably do if not formally at least subconsciously
337   and likelihood-based methods of inference are clearly more intuitive. The days of a clear divide
338   between Bayesian and frequentist philosophies have passed, and pragmatic statisticians [17, 18] and
339   scientists [4, 15] now recommend and practice approaches that combine a frequentist approach to
340   with elements of Bayesian thinking. One such approach, magnitude-based inference [4] focusses on
341   estimating the magnitude of population effects with reference to *a priori* subjective estimates of
342   practically / clinically worthwhile effect magnitudes, without the complication of expressing the
343   latter as a probability distribution. Moreover, the tools and instructions required to perform and
344   interpret such analyses are readily available [19] whereas common statistical-software packages do
345   not offer options for full Bayesian analysis or other hybrid methods such as the calibrated Byes
346   approach [18].

347

348   7.0 Summary and recommendations

349 Significance testing is designed to provide a reliable procedure for making black and white decisions
350 for accepting or rejecting (usually zero-effect) null hypotheses with known and controlled long-run
351 error rates. If that is what a scientist wishes to know, then all is well, but type I and type II error rates
352 must be specified in advance and ought to be based on careful thought about potential costs
353 incurred by each type of error, not dictated simply by convention. It follows that sample size must be
354 determined in advance and that the resulting number of participants are tested to ensure type II
355 error rate is controlled. The outcome of an analysis allows conclusions about the mere existence of
356 non-zero effects but provides no information about the likely size of true effects or their practical /
357 clinical value.

358 If a scientist wishes to estimate the true magnitude of an effect and how likely it is to exceed an
359 effect magnitude of practical / clinical importance, while allowing for elements of subjective
360 Bayesian-style thinking, magnitude-based inference provides a solution. While this approach is
361 gaining acceptance, progress might be hastened if scientists appreciate the shortcomings of
362 traditional N-P null-hypothesis-significance testing. In summary, it is up to the individual scientist to
363 decide what they wish statistics to do for them and be aware of which approach is best suited to this
364 purpose.

365

366 Acknowledgements

369

370 References

371 1.   Popper KR. The Logic of Scientific Discovery. 6th ed. London: Hutchinson & Co Ltd; 1972a.
372 2.   Popper KR. Conjectures and Refutations:The Growth of Scientific Knowledge. 4th ed.
373   London: Routledge and Kegan Paul Ltd; 1972b.
374 3.   Ziliak ST, McClaskey DN. The Cult of Statistical Significance: how the standard error costs us
375   jobs, justice, and lives. Michigan: University of Michigan Press; 2008.
376 4.   Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. Int J Sport
377   Phys Perf. 2006;1:50-7.
378 5.   Krantz DH. The null hypothesis testing controversy in psychology. J American Stat Assoc.
379   1999;94(448):1372-81.
380 6.   Sterne JAC, Smith GD. Sifting the evidence - what's wrong with significance tests? Br Med J.
381   2001;322:226-31.
382 7.   Fisher R. Statistical methods for research workers. London: Oliver and Boyd; 1950.
383 8.   Fisher R. Statisitcal methods and scientific inference. London: Collins Macmillan; 1973.
384 9.   Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses.
385   Philosophical Transactions of the Royal Society of London, Series A. 1933;231:289-337.
386 10.   von Mises R. Probability, Statistics and Truth. 2nd ed. London: Allen and Unwin; 1928.
387 11.   Oakes M. Statistical inference: A commentary for the social and behavioural sciences. New
388   Jersey: Wiley; 1986.
389 12.   Dienes Z. Bayesian versus orthodox statistics: which side are you on? Perspectives on
390   Psychological Science. 2011;6(3):274-90.

391 13. Pollard P, Richardson JTE. On the probability of making type I errors. Psychological Bulletin.
392   1987;102(1):159-63.
393 14. Cumming G. Understanding the new statistics: Effect sizes, confidence intervals and meta
394   analysis. New York: Taylor and Francis Group; 2012.
395 15. Hopkins WG, Marshall SW, Batterham AM, Hanin J. Progressive statistics for studies in sports
396   medicine and exercise science. Medicine and Science in Sports and Exercise. 2009;41(1):3-
397   12.
398 16. Edwards AWF. Likelihood. Cambridge: Cambridge University Press; 1972.
399 17. Kass RE. Statistical inference: the big picture. Statistical Science. 2011;26(1):1-9.
400 18. Little RJ. Calibrated Bayes, for statistics in general, and missing data in particular. Statistical
401   Science. 2011;26(2):162-74.
402 19. Hopkins WG. A new view of statistics. Internet Society for Sport Science. 2000; Available
403   from: http://www.sportsci.org/resource/stats/.

404
405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425   Fig 1
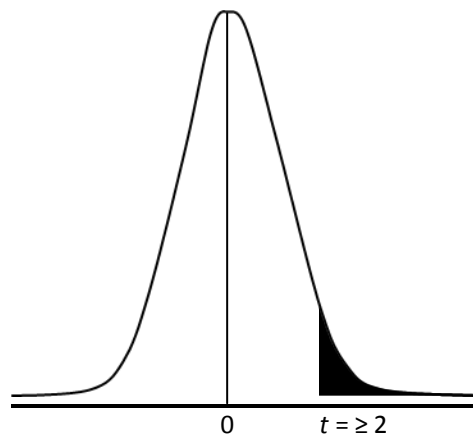
452   Fig 2

453

454

455

456   **(a)**   **(b)**

457    $(M_s — M_p)$ $(M_s - M_p)$
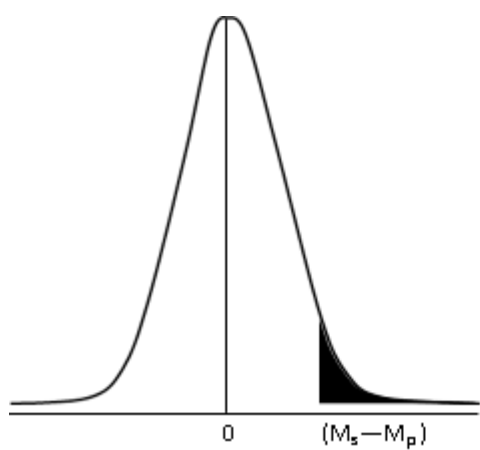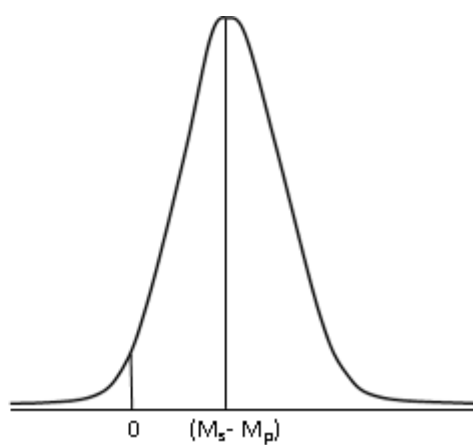
458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473   Fig 3

474