Reviewing scientific manuscripts: how much statistical knowledge should a reviewer really know?

James P. Morton

Advan in Physiol Edu 33:7-9, 2009. ;
doi: 10.1152/advan.90207.2008

Advances in Physiology Education is dedicated to the improvement of teaching and learning physiology, both in specialized courses and in the broader context of general biology education. It is published four times a year in March, June, September and December by the American Physiological Society, 9650 Rockville Pike, Bethesda MD 20814-3991. ESSN: 1522-1229. Visit our website at http://www.the-aps.org/.

Adv Physiol Educ 33: 7–9, 2009;
doi:10.1152/advan.90207.2008.

Reviewing scientific manuscripts: how much statistical knowledge should a reviewer really know?

James P. Morton

Research Institute for Sport and Exercise Sciences, Liverpool John Moores University, Liverpool, United Kingdom

Submitted 18 November 2008; accepted in final form 19 December 2008

Morton JP. Reviewing scientific manuscripts: how much statistical knowledge should a reviewer really know? *Adv Physiol Educ* 33: 7–9, 2009; doi:10.1152/advan.90207.2008.—In the sequel to their guidelines for reporting statistics in American Physiological Society journals, Curran-Everett and Benos highlighted that the initial guidelines of 2004 have had little effect on the statistical reporting practices of authors. In the present article, I suggest that the guidelines have also had little impact on both journal reviewers and editors. I present three cases of statistical reporting practices in which there appears to be considerable discrepancies between the author and reviewer and, moreover, inconsistencies between reviewers. I argue that for authors to comply with these guidelines, the initial challenge is to have a team of reviewers who are also willing to accept the unfamiliar. Indeed, the opinions of reviewers who are ill informed about relatively novel statistical methods and recommended reporting practices may have implications for the final editorial decision on the suitability of submitted manuscripts for publication.

standard deviation; statistical significance; sample size; confidence interval

READERS of *Advances in Physiology Education* are likely to be familiar with the article by Benos et al. (4) entitled "How to review a paper." The authors of this article essentially provided an introductory guide to the ethics underlying the peer reviewing of scientific manuscripts. Within their article, Benos and colleagues (4) suggested that one of the checklist criteria for journal reviewers should be to "delineate the strengths and weakness of methodology/experimental/statistical approach/interpretation of results." To arrive at this professional judgement, the question that is subsequently asked in the present article is "how much *statistics* should a reviewer really know?"

My motivation for writing this article not only stems from the numerous reviews of submitted manuscripts that my colleagues and I have received over the last few years of research but also from the recent article from Curran-Everett and Benos (7), in which the authors provided a sequel to their initial guidelines for reporting statistics published in Americal Physiological Society (APS) journals (6). In their second offering, the authors surmised that the guidelines of 2004 have had little impact on statistical reporting practices of authors within APS journals. In the present article, I adopt an alternative approach by suggesting that the initial guidelines may also not have had the desired effect on journal reviewers (and editors) as had initially been hoped. I provide three cases of statistical procedures (that are likely to be common to many authors) in which there seems to be considerable discrepancies between the author and reviewers and, moreover, inconsistencies between reviewers. This article is therefore intended to provide a thought-provoking and provocative article highlighting the need for all of those involved in the research process to adhere to the same principles and guidelines when it comes to the reporting of statistics in physiology-related journals.

SD Versus SE Versus Confidence Intervals

As an undergraduate and postgraduate student, I believe I was well schooled on the underlying differences in statistical meaning between the above three terms. Having been educated on statistics in my early research method training and in accordance with the initial APS guidelines (6), I therefore chose to report data variability in the text, figures, and tables as means (SD) when beginning my research career. Nevertheless, I soon received reviewers' comments along the following lines: my "error bars look big," my "data appear highly variable," and I should "change SD to SE." As noted by Curran-Everett and Benos (7) in the sequel to their guidelines and also by Atkinson (1), I too am therefore of the opinion that many researchers report SE (as opposed to SD) merely for cosmetic reasons, despite the fact that they provide no valid estimate of data variability. Indeed, if it is considered that SE is calculated as $SD/\sqrt{n}$ (where n is the sample size), the size of the error bar is considerably reduced and the data suddenly appear as much cleaner with less variability present. While there are distinct statistical differences between SD and SE (in both definition and size), it should be noted that there are scientists who also advocate the use of SE because it provides an inferential statistic (as opposed to a descriptive statistic) of true population values (5). Clearly, the issue of SD versus SE is therefore also dependent on what the researcher(s) is trying to convey in their report.

In one of the most recent reviews I received, the reviewer also commented "The error bars on the graphs and the confi-

Address for reprint requests and other correspondence: J. P. Morton, Research Institute for Sport and Exercise Sciences, Liverpool John Moores Univ., 15-21 Webster St., Liverpool L3 2ET, UK (e-mail: J.P.Morton@ljmu.ac.uk).

dence intervals in the tables should be changed from SD to SE." However, in this particular article, there were no confidence intervals present anywhere in the figures or tables, and the term "confidence intervals" was not even present anywhere in the text! Regular comments such as these suggest to me that many physiologists do not appear to understand what are considered as the fundamental statistical components of research (6, 7). I find this slightly worrying for authors in that if it cannot be guaranteed that reviewers will understand such fundamentals, then what hope is there for researchers who are required to employ a more complex or (dare I say) novel and unfamiliar statistical procedure? One may reason that if authors simply justify their statistical choices in the METHODS section of their report with appropriate references and terminology, then reviewers are therefore unlikely to question this aspect of their report. Nevertheless, a colleague of mine undertook such an approach, only for the reviewer to respond with "You are a physiologist, not a statistician!" Similarly, another colleague was greeted with "The use of statistical gobbledygook is very distracting to the paper, do you need all this statistical mumbo jumbo?!" In contrast, one member of my department appeared to have a statistically astute reviewer who actually directed the author to the first APS guidelines of 2004. When taken together, I can't help but feel that such readily apparent inconsistencies between how much statistics reviewers appear to know may have implications for the final editorial decision on the suitability of submitted manuscripts for publication.

If the P Value Is Not <0.05, It's Not Significant!

In a recent article, I decided to also comply with the APS guidelines by reporting precise *P* values. Furthermore, in those instances where $P = 0.07$, I argued that these findings approached statistical significance, and that, as a team of authors, we considered these findings of physiological significance. Nevertheless, the reviewer replied by stating 'If *P* is not <0.05 it's not significant" and subsequently requested that this aspect of the report be revised. Rather than run the risk of upsetting the reviewer and not having the report accepted for publication, I addressed the reviewer's request, and all mentions of physiological significance were removed from the revised manuscript. I therefore reverted to the conventional method of reporting whether *P* was less than or greater than 0.05, and, as a result, I felt that the published article lost some of the physiological message that we were trying to convey.

According to *guideline 6*, Curran-Everett and Benos (6) also suggested that we should report uncertainty about significance by reporting confidence intervals. In my own teaching and research domain of exercise physiology, there has also been a call for the regular reporting of confidence intervals (3). The latter authors suggested that simply reporting a *P* value by itself is not enough, as it does not provide information on the direction or size of the effect. Furthermore, a *P* value of >0.05 (e.g., 0.07, as in the above example) does not necessarily imply there is no worthwhile effect given that a combination of small sample size and individual variability (as is frequently the case in human studies involving invasive measurements) can mask important effects. Reporting a confidence interval, however, provides an interval estimate of the true population value of the statistic and provides a more meaningful based inference of the magnitude of effect. Such an approach therefore allows researchers (and readers) to formulate a more informed opinion on the physiological significance (as opposed to statistical significance per se) of the findings. In a recent manuscript where I attempted to justify my statistical rationale with appropriate references and limited use of jargon, I did just that and based my interpretation of data around both *P* values and confidence intervals. Surprisingly, the review passed off with relatively little comments on the use of confidence intervals, and the report was accepted with minor revision. Thankfully, in this instance, the reviewer was willing to accept the unfamiliar.

Your Sample Size Is Too Low!

An increasingly occurring theme of reviewers' comments that I and my colleagues receive appear to relate to the issue of sample size. Typically, the comments allude to sample sizes that are too low, even when a significant effect has been found! Nevertheless, in those instances where a low sample size was employed and no significant treatment effect was observed, authors can expect reviewers to question the statistical power of the findings. It should be noted, however, that the calculation of post hoc (i.e., retrospective) power is effectively meaningless as, indeed, a study in which $P = 0.05$ will have a retrospective power of essentially 50% and the power will further decrease as $P > 0.05$ (8). In this regard, authors should therefore include justification for their chosen sample size in the METHODS section of their manuscript. Where the research and treatment are novel, this process requires a degree of intuitive guess work on behalf of the authors as the research team will have to collectively decide on the minimal effect size deemed as physiologically relevant as well as the likely variability of the magnitude of the effect (authors should also refer to previous publications if anticipated effect sizes are relatively well known). Finally, the researchers must also decide on their α-level of significance (e.g., $P < 0.05$ or 0.01) and the required statistical power (e.g., 80% or 90%). Once the research team has decided on these vital elements, there are numerous statistical software packages in which inputting these data will readily calculate the number of subjects (or animals) required to achieved the desired statistical power. Interestingly, this is one element of statistical reporting practices that was not included in the original or second APS guidelines and that (to my knowledge of exercise physiology) is very rarely included in APS publications. Nevertheless, if such information is included in your manuscript, it is one less thing that the reviewer can question. I have provided a brief example of this below (adapted from Ref. 14), and interested authors are also directed to Batterham and Atkinson (2) for a further primer.

> Using the NQuery statistical power software (Statistical Solutions, Cork, Ireland) and appropriate statistical guidelines (2), it was estimated that a sample size of six would enable detection of a mean difference of 50% in basal muscle heat shock protein levels between trained and untrained humans, assuming a SD of differences equal to 25% and statistical power of 80%. This effect size and SD are based on those values cited in previous studies that have observed training-induced increases in muscle heat shock protein content in humans (9, 10).

While the calculation of retrospective power provides no meaningful information, researchers may wish to consider undertaking other related posthoc analysis with the view to informing future research design. For example, if the variability observed in the actual data set was greater than that

estimated a priori, authors can subsequently determine if the sample size was sufficient to achieve statistical significance of the predicted physiological effect, that is, of course, assuming a real physiological effect of the chosen intervention even exists!

Concluding Remarks

As an author, I welcomed the APS guidelines of 2004 and 2007. However, more often than not I have found myself at loggerheads with reviewers who are perhaps not aware of the guidelines or who have chosen to ignore them. As a scientist, I find this somewhat contradictory. Indeed, in searching for the unknown, scientists are constantly striving for precision and clarity in both research design and research methods. For example, a researcher would never enter the laboratory without the most appropriate measurement tool(s) required to test the hypothesis in question. Why then do some authors appear to have difficulty in adopting the correct statistical analysis and reporting procedures, given that statistics are also a vital part of the research process when it comes to the testing of a hypothesis? A likely answer to this question is, of course, related to what researchers *perceive* to be the correct or most appropriate statistical methods (7). An overriding question, however, is how can those researchers who choose to adhere to the recommended guidelines or employ relatively novel statistical procedures do so without running the risk of entering a battle with reviewers and editors?

Perhaps one solution is to simply reference and justify our statistical approach [e.g., in accordance with APS guidelines (6), all data are presented as means (SD), etc.] in the METHODS section of our manuscripts and trust that the reviewers are ready to comply. Maybe even a note to the editor in your cover letter alluding to any unfamiliar approach may also do the trick. Alternatively, an online tutorial and/or a list of explicit checklist criteria (with subsequent verification of approval) at both the author submission and reviewer acceptance stage may also remind both parties of the recommended practices. As a side note, I would like to offer some hope by suggesting that, given time, the APS guidelines *can* and *will* make a difference if journal editors are willing to enforce their adoption. One only has to look at the application of the Consolidated Standards for Reporting Trials (CONSORT) statement (checklist criteria and a flow diagram for what should be included in reporting randomized control trials) for a source of inspiration (12). Indeed, since its first appearance in the mid-1990s, there are now over 300 scientific journals that directly endorse the CONSORT statement (11). Finally, what is now readily apparent to me is that precisely *what* and *how* we teach statistics to physiology students also needs to be revised so that the next generation of decision makers are not having the same debates on what are perceived to be the correct and incorrect way to do and report statistics. Whatever your strategy, I hope that the present article has served its purpose in reminding those of us with reviewing responsibilities that we, too, should also be statistically astute when undertaking this important role. I welcome your thoughts.

ACKNOWLEDGMENTS

The author thanks Dr. Mark Scott and Prof. Greg Atkinson for useful scientific discussions and helpful comments on early drafts of this manuscript.

REFERENCES

1. **Atkinson G.** Analysis of repeated measurements in physical therapy research. *Phys Ther Sport* 2: 194–208, 2001.
2. **Batterham AM, Atkinson G.** How big does my sample size need to be? A primer on the murky world of sample size estimation. *Phys Ther Sport* 6: 153–163, 2005.
3. **Batterham AM, Hopkins G.** Invited commentary: making meaningful inferences about magnitudes. *Int J Sports Physiol Perf* 1: 50–57, 2006.
4. **Benos DJ, Kirk KL, Hall JE.** How to review a paper. *Adv Physiol Educ* 27: 47–52, 2003.
5. **Cumming G, Fidler F, Vaux DL.** Error bars in experimental biology. *J Cell Biol* 177: 7–11, 2007.
6. **Curran-Everett D, Benos DJ.** Guidelines for reporting statistics in journals published by the American Physiological Society. *J Appl Physiol* 97: 457–459, 2004.
7. **Curran-Everett D, Benos DJ.** Guidelines for reporting statistics in journals published by the American Physiological Society: the sequel. *Adv Physiol Educ* 31: 295–298, 2007.
8. **Howell DC.** *Statistical Methods in Psychology.* Belmont, CA: Wadsworth, 2007.
9. **Liu YL, Baur S, Optiz-Gress A, Altenburg D, Lehmann M, Steinacker JM.** Human skeletal muscle HSP70 response to physical training depends on exercise intensity. *Int J Sports Med* 21: 351–355, 2000.
10. **Liu YL, Mayer S, Optiz-Gress A, Zeller C, Lormes W, Baur S, Lehmann M, Steinacker JM.** Human skeletal muscle HSP70 response to training in highly trained rowers. *J Appl Physiol* 86: 101–104, 1999.
11. **Moher D, Simera I, Schulz KF, Hoey J, Altman DG.** Helping editors, peer reviewers and authors improve the clarity, completeness and transparency of health related research. *BMC Med* 6: 13, 2008.
12. **Moher D, Schulz KF, Altman DG.** The CONSORT statement: revised recommendations for improving the quality of reporting of parallel group randomised control trials. *Ann Intern Med* 134: 657–662, 2001.
13. **Morton JP, MacLaren DPM, Cable NT, Campbell IT, Evans L, Kayani A, McArdle A, Drust B.** Trained men display increased basal levels of heat shock proteins. *Med Sci Sports Exer* 40: 1255–1262, 2008.