Received: 18 June 2024 | Accepted: 14 July 2024

DOI: 10.1111/opo.13369

OPO THE COLLEGE OF OPTOMETRISTS

LETTER TO THE EDITOR

Recommended improvements to the statistical guidelines

To the Editor,

The website for this journal currently provides its official 'Statistical Guidelines' (https://onlinelibrary.wiley.com/page/journal/14751313/homepage/statistical_guidelines.htm) in the form of several editorial articles. Much of the material in those articles is valuable and correct. However, some of the articles contain statements that require clarification. In the present letter, two such articles are critically evaluated: 'Is There a Large Sample Size Problem?'[1] and 'When to Use the Bonferroni Correction'[2] (both by Armstrong). Updates to the statistical guidelines are then recommended.

CRITIQUE OF 'IS THERE A LARGE SAMPLE SIZE PROBLEM?'

The larger a sample is, the more closely it will tend to resemble the population that it came from, and thus the more reliable the inferences about that population will tend to be. Larger samples not only tend to provide more statistical power for hypothesis testing but also tend to provide a more accurate and more precise estimation of the population parameter (e.g., effect or correlation) that is being investigated.[3,4] Thus, although there can of course be logistical and financial (and in some cases, ethical) constraints on what sample size is feasible, there is generally no such thing as a sample being 'too large' for purposes of statistical inference. It is therefore surprising that Armstrong[1] proposed that researchers should limit their sample sizes to avoid 'both a small and large sample size problem'. Armstrong has made the same claim in other articles as well, stating for example that 'having too large a sample size may create as many problems of interpretation as having an inadequate sample size',[5] and that 'both a too small and a too large a sample size can cause problems in testing the hypothesis and in interpretation'.[6]

Armstrong's[1] concern was based on the fact that large sample sizes increase the likelihood of statistical significance when the effect or correlation being investigated is non-zero, even in cases where that effect or correlation is too small to be clinically relevant. The resulting 'large sample size problem', according to Armstrong, is that large samples make statistical significance a poor indicator of clinical relevance, since 'there is the increased chance of demonstrating a statistically significant effect which is too small to be clinically meaningful'.[1] However, the problem in that scenario is not that the sample size is too large. Rather, the problem is that statistical significance is being interpreted as clinical relevance—which is inappropriate regardless of the sample size.[3,4,7,8] The American Statistical Association's official statement on *p*-values was unequivocal on that point: 'A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result'.[7]

When the goal is to evaluate the size of an effect or correlation (or of some other parameter), the most appropriate inferential tool is typically a *confidence interval*—not a *p*-value.[3,4,7,8] A confidence interval proposes a lower and upper bound for the parameter being estimated. That is, it provides margins of error below and above the point estimate. Thus, for example, if the confidence interval's lower bound exceeds a predesignated threshold of clinical relevance, that could be taken as evidence in favour of a clinically relevant effect.

What about the situation that Armstrong[1] was concerned about, where a large sample reveals 'a statistically significant effect which is too small to be clinically meaningful'? In that case, appropriately concluding that there is evidence for a non-zero effect (due to statistical significance) would not prevent the researcher from also appropriately concluding that there is evidence against the effect being clinically meaningful on average (due to the confidence interval being entirely within some designated clinically irrelevant range).[9] Thus, although Armstrong[1] blamed large sample size for an inferential 'problem' that allegedly 'can affect all statistical procedures to some extent', there is not actually a problem as long as each statistical procedure is used for its proper inferential purpose: The *p*-value can be used to evaluate evidence against the effect being zero, while the confidence interval can be used to make a more specific inference about how large the effect is.[3,4,7–9] In fact, the larger the sample, the smaller the standard error will be for the estimation,[3] and thus the more likely that the confidence interval will be narrow enough to provide conclusive evidence about whether the effect is indeed 'too small to be clinically meaningful'.

Thus, when a researcher is interested in whether an effect or correlation is large enough to be clinically relevant, the preferable approach is to use a large sample and compute a confidence interval—yet Armstrong[1] cautioned against large samples and did not mention confidence intervals anywhere in the article. Armstrong did rightly call for researchers to consider the correlation's magnitude in addition to its statistical significance. But Armstrong depicted that evaluation of magnitude as being based on a point estimate, such as the sample correlation or the r^2.

Importantly, even when the point estimate is examined in conjunction with the *p*-value, that is not an adequate substitute for using a confidence interval, since neither the *p*-value nor the point estimate provides margins of error for the estimation.[3,4]

To be clear, confidence intervals are imperfect tools that, like any other statistic, are subject to misuse and misinterpretation.[4] Moreover, confidence intervals alone are not sufficient for thoroughly evaluating clinical relevance. Indeed, evaluating clinical relevance is often a complex task that requires considering many types of information (some statistical, some not) besides just the estimated magnitude of the effect on average. Nonetheless, confidence intervals are typically essential tools when clinical relevance is of interest.[3,4,7,8]

In short, there is not a 'large sample size problem' in statistical inference. Larger samples tend to provide better inferences, not worse. That principle holds true regardless of whether the focus is on statistical significance or on confidence intervals, as long as those tools are used properly. Although Armstrong may be correct that some misguided researchers use statistical significance improperly (e.g., 'if a result is statistically significant it is assumed to be clinically relevant'),[1] those researchers should be encouraged to stop doing that and to use confidence intervals,[3] not encouraged to use smaller samples.

CRITIQUE OF 'WHEN TO USE THE BONFERRONI CORRECTION'

Armstrong's[2] article on Bonferroni adjustment made some true and important statements. For example, it rightly noted that many studies do not adequately address the problem of multiple testing. It also correctly noted that Bonferroni adjustment is not always the best method for a given multiple-testing situation. And perhaps most importantly, the article rightly stated that researchers should be transparent about which aspects of their analysis were planned a priori. However, the article also contains errors and misunderstandings.

The first error is in the unsourced formula $1-(1-\alpha)/T$, where α is the 'critical *p* level' (more commonly called the alpha level) for each test, and T is the number of tests.[2] Armstrong made two conflicting claims—neither of which is correct—about what that formula represents. First, Armstrong claimed that the formula gives the probability of at least one Type I error (a probability known as the *familywise Type I error rate* or FWER) when all null hypotheses are true. But Armstrong was presumably thinking of $1-(1-\alpha)^T$, which is the correct formula when the tests are statistically independent[10] (though Armstrong did not mention assuming independence). Indeed, that correct formula produces the value of ~0.64 that Armstrong gave as the FWER for $\alpha=0.05$ and $T=20$ (see also Armstrong et al.[11]), whereas inputting those same values of α and T to Armstrong's erroneous formula produces a nonsensical value of 0.9525.

Adding to the confusion, after incorrectly claiming that the formula $1-(1-\alpha)/T$ gives the FWER, Armstrong then claimed in the next sentence that the same formula was for something else: an adjusted alpha level for which α/T is 'an approximation'.[2] But that is not correct either, as Armstrong's formula does not yield values anywhere near α/T (except when $T=1$). Instead of the erroneous formula $1-(1-\alpha)/T$, presumably in this case Armstrong was thinking of $1-(1-\alpha)^{1/T}$, which is the adjusted alpha level when using the Šidák procedure to control the FWER.[12] Indeed, the Šidák-adjusted alpha level does tend to be near the classical α/T Bonferroni-adjusted alpha level. For example, when $\alpha=0.05$ and $T=2$, the Bonferroni-adjusted alpha level for each test is $0.05/2=0.025$, and the Šidák-adjusted alpha level for each test is $1-(1-0.05)^{1/2}\approx 0.0253$.

Armstrong also claimed that Bonferroni-adjusted testing 'is often concerned with the wrong hypothesis and is in actuality a test of the "universal" H_O, that is, if 20 different comparisons were made on two groups, that the two groups were identical in *all* comparisons'[2] (see also Armstrong et al.[11]). The same claim has been made by many other articles that, like Armstrong, cited an opinion piece by Perneger[13] (examples of such citations were compiled in Frane[14]). However, the claim is a known myth, and despite its appearance in many non-statistical sources, it has been unequivocally debunked in statistical journals.[14,15] Indeed, it is a proven mathematical fact—not a matter of opinion—that Bonferroni-adjusted testing provides decisions about all the individual null hypotheses, not just a single global decision about the omnibus ('universal') null hypothesis. Simple proofs of that fact, based on Boole's inequality, have been provided.[10,15,16] Note also that in the presumably rare cases that the omnibus null hypothesis is actually the hypothesis of interest, there are more efficient ways to test that hypothesis than to conduct Bonferroni-adjusted tests of all the individual hypotheses.[15] Thus, there is no apparent justification for Armstrong's[2] recommendation that Bonferroni-adjusted testing be used when 'a single test of the "universal null hypothesis" (H_O) that *all* tests are not significant is required'.

Some of Armstrong's[2] other recommendations are confusing as well, and some appear to be inconsistent. For example, Armstrong recommended not using any adjustment when 'a study is exploratory involving *post-hoc* testing of unplanned comparisons which are regarded as hypotheses for further investigation'. Yet, Armstrong recommended considering Bonferroni adjustment when 'a large number of tests are carried out without preplanned hypotheses in an attempt to establish any results that may be significant'. It is not clear what the defining difference is between those two scenarios. Presumably, Armstrong meant to emphasise the 'exploratory' nature of the study in the first scenario, since formal adjustment is often not

considered necessary when the tests themselves are essentially informal anyway.[17,18] But both scenarios involve fishing through the data without preplanned hypotheses to see what turns up. Thus, the second scenario is effectively just as exploratory as the first. And in both scenarios, any statistically significant 'results' that the researcher might 'establish' would be better considered as 'hypotheses for further investigation', rather than as discoveries or confirmatory demonstrations.

Indeed, it is important for readers to understand that if hypotheses were not planned before seeing the data, then the hypotheses typically cannot be meaningfully 'tested'—except in an informal sense—on those same data, due to the effects of selection bias on Type I error rates.[7,10,14,17,19] Even Bonferroni adjustment typically cannot remedy that problem.[10,17] That is because to overcome selection bias, the adjustment would theoretically need to account not only for the finite number of actually conducted tests, but also for the potentially indeterminate number of tests that would have been conducted in alternative scenarios where the data had come out differently and different hypotheses had looked promising.[10] Thus, when doing an analysis of unplanned hypotheses, deciding whether to adjust or not is typically far less important than acknowledging that the 'tests' should not be interpreted (or presented) as formal hypothesis tests at all. Moreover, researchers should not have a false sense of security that they can transform their unplanned testing into a rigorous, confirmatory analysis merely by applying Bonferroni adjustment.

Armstrong also recommended not using any multiple-testing adjustment 'if the study is restricted to a small number of planned comparisons'.[2] It is not clear exactly what was meant by 'a small number'. But unless that small number is 1, the FWER can be substantially inflated if no adjustment is applied. For example, even for just two planned comparisons, the FWER is nearly doubled if no adjustment is applied (assuming true null hypotheses and roughly independent tests). And for just three planned comparisons, the FWER may be nearly tripled if no adjustment is applied—meaning a ~1 in 7 chance of making at least one Type I error when $\alpha = 0.05$. Thus, it is not clear what the mathematical or theoretical basis could be for the recommendation. Moreover, multiple-testing adjustments are already inherently less stringent for smaller numbers of tests, so there is no apparent reason for special leniency to be required in such cases.

The idea that 'planned' tests should be inherently exempt from multiple-testing adjustment, either in general or when the tests are few in number, is a common assertion by researchers. But that assertion appears to lack adequate foundation.[14,20] To be clear, there are situations where a small number of planned tests do not need adjustment. However, in such cases, the reason the tests are exempt from adjustment is not simply *because* they are a small number of planned tests. For instance, in some situations, the FWER is inherently controlled by the testing structure (e.g., when using *serial gatekeeping* and the test sequence is defined a priori in the analysis plan).[10,18]

It is worth explaining why Bonferroni adjustment differs from other multiple-testing procedures, and why a researcher might choose one procedure over another. First of all, the more positive the inter-test dependence is (i.e., the more that statistical significance in one test would make statistical significance in another test more likely), the less the FWER is inflated by unadjusted tests, and thus the less stringent an adjustment can be while still controlling the FWER.[15,21] Bonferroni adjustment makes no assumptions about inter-test dependence,[15] whereas Šidák adjustment achieves its marginally lower stringency by assuming the inter-test dependence is nonnegative[12] (which is typically a reasonable assumption, at least for two-sided tests). Other FWER-control procedures can reduce stringency more substantially by exploiting positive inter-test dependence. In some cases, that positive inter-test dependence is inherent to the study design (e.g., when using a procedure that is specially optimised for comparing 'all possible pairs' of groups).[21] In other cases, the procedure itself creates positive inter-test dependence (e.g., when using a multi-step procedure like Holm[22] or Hommel,[23] though such methods often cannot be applied to two-sided confidence intervals). To minimise the increased risk of Type II errors, researchers often want to choose the most 'powerful' (i.e., least stringent) adjustment procedure that reliably controls the FWER in the given context.[15,21]

That said, although the FWER is the most commonly referenced overall Type I error rate, it is not the only overall Type I error rate that can be considered. For example, the classical α/T Bonferroni adjustment controls not only the FWER, but also the *per-family Type I error rate* (the expected number of Type I errors), which is a stricter standard than the FWER.[24,25] There are also procedures that are designed to control only the *false discovery rate* (a more lenient standard than the FWER), though their use is largely confined to contexts where the number of tests is very large (e.g., in the hundreds or thousands) and strong conclusions about individual hypotheses are not required.[15,25]

Heuristically speaking, multiple-testing adjustment should be considered when an analysis involves multiple unique opportunities for a false finding. That leaves room for subjectivity in some cases, regarding what constitutes a single 'analysis' (i.e., which tests should be included in a given *family*)[10,21] and what constitutes a 'finding'. In the case of clinical trials for new drugs, industry guidances often resolves those issues.[18] But in some contexts, especially when the potential costs of Type I and Type II errors are not as clear-cut, researchers may reasonably disagree about whether or how adjustment should be applied, depending on how the test results are expected to be used. In any case, decisions about how to adjust should be incorporated into the analysis plan a priori.[18] And such decisions should be made thoughtfully, based on correct understanding of the principles involved, not based on arbitrary criteria (e.g., exempting

tests from adjustment simply because they are planned and few in number).

CONCLUSIONS

The misunderstandings that have been discussed here are important because they involve fundamental statistical principles that have implications for how research should be conducted and interpreted. Moreover, when misunderstandings in the scientific literature are left uncorrected, they can spread and have real consequences on research practice. For example, Perneger's[13] false claim that Bonferroni-adjusted testing only addresses the 'universal' null hypothesis was repeated by Armstrong[2] (and by many other authors, e.g., Schulz and Grimes[26]) and in turn has been repeated by dozens of articles that cite Armstrong.[2] In fact, a Google Scholar search for citations of Armstrong[2] that contain the exact phrase 'universal null hypothesis' yields 13 documents from the year 2023 alone, mostly by researchers attempting to justify their unadjusted tests. The number of researchers who have not directly cited Armstrong, but nonetheless have been directly or indirectly influenced by the false claim, is indeterminate.

In the interest of promoting readers' understanding and reducing the further spread of misconceptions, I recommend that updates be made to this journal's statistical guidelines. For example, in place of the two articles[1,2] that have been critiqued here, readers could be directed to other resources, such as the American Statistical Association's statement on *p*-values.[7] I also recommend Calin-Jageman and Cumming[4] as well as Hawkins and Samuels,[9] which give practical, nontechnical guidance on how to use confidence intervals properly. Readers may also be interested in Frane,[14] which discusses some common misconceptions about multiple-testing adjustment. Lastly, readers should know about websites such as clinicaltrials.gov and osf.io, which allow researchers to register their study protocols and analysis plans a priori in a verifiable way and thus bolster the credibility of their planned tests.

In any case, researchers should be wary of claims about statistical principles when those claims are not supported by mathematical logic or by citations of authoritative statistical sources, especially if those claims tell the researchers what they want to hear (e.g., that their multiple tests are inherently exempt from adjustment for some reason). Statistical misconceptions are abundant in the nonstatistical literature (as noted in previous articles[14,15,17,20]), and the medical literature is no exception.

AUTHOR CONTRIBUTIONS

Andrew V. Frane: Conceptualization (equal); investigation (equal); writing – original draft (equal); writing – review and editing (equal).

FUNDING INFORMATION

The author has no funding to declare.

CONFLICT OF INTEREST STATEMENT

The author has no conflict of interest to declare.

Andrew V. Frane

Occidental College, Los Angeles, California, USA

Correspondence
Andrew V. Frane, Occidental College, Los Angeles, CA, USA.
Email: avfrane@ucla.edu

ORCID

Andrew V. Frane https://orcid.org/0000-0002-5057-7567

REFERENCES

1. Armstrong RA. Is there a large sample size problem? *Ophthalmic Physiol Opt*. 2019;39:129–30.
2. Armstrong RA. When to use the Bonferroni correction. *Ophthalmic Physiol Opt*. 2014;34:502–8.
3. Gardner MJ, Altman DG. Confidence intervals rather than *p* values: estimation rather than hypothesis testing. *Stat Med*. 1986;292:746–50.
4. Calin-Jageman RJ, Cumming G. The new statistics for better science: ask how much, how uncertain, and what else is known. *Am Stat*. 2019;73:271–80.
5. Armstrong RA, Eperjesi F, Gilmartin B. The use of correlation and regression methods in optometry. *Clin Exp Optom*. 2005;88:81–8.
6. Armstrong RA. Should Pearson's correlation coefficient be avoided? *Ophthalmic Physiol Opt*. 2019;39:316–27.
7. Wasserstein RL, Lazar NA. The ASA's statement on *p*-values: context, process, and purpose. *Am Stat*. 2016;70:129–33.
8. Livingston EH. Study design and statistics. In: Christiansen SL, Iverson C, Flanagin A, Livingston EH, Fischer L, Manno C, et al., editors. AMA manual of style: a guide for authors and editors. 11th ed. New York: Oxford University Press; 2020.
9. Hawkins AT, Samuels LR. Use of confidence intervals in interpreting nonstatistically significant results. *JAMA*. 2021;326:2068–9.
10. Hochberg Y, Tamhane AC. Multiple comparison procedures. New York: Wiley; 1987.
11. Armstrong RA, Davies LN, Dunne MCM, Gilmartin B. Statistical guidelines for clinical studies of human vision. *Ophthalmic Physiol Opt*. 2011;31:123–36.
12. Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc*. 1967;62:626–33.
13. Perneger TV. What's wrong with Bonferroni adjustments. *BMJ*. 1998;316:1236–8.
14. Frane AV. Misguided opposition to multiplicity adjustment remains a problem. *J Mod Appl Stat Meth*. 2019;18:eP2836. https://doi.org/10.22237/jmasm/1556669400
15. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. *Stat Med*. 2014;33:1946–78.
16. Dunn OJ. Estimation of the means of dependent variables. *Ann Math Stat*. 1958;29:1095–111.
17. Bender R, Lange S. Adjusting for multiple testing—when and how? *J Clin Epidemiol*. 2001;54:343–9.
18. Committee for Proprietary Medicinal Products. Points to consider on multiplicity issues in clinical trials. London: European Agency for the Evaluation of Medicinal Products; 2002.

19. Selvin HC, Stuart A. Data-dredging procedures in survey analysis. *Am Stat*. 1966;20:20–3.
20. Frane AV. Planned hypothesis tests are not necessarily exempt from multiplicity adjustment. *J Res Pract*. 2015;11:P2. Retrived from: https://files.eric.ed.gov/fulltext/EJ1083896.pdf. Accessed 20 July, 2024.
21. Frane AV. Experiment-wise Type I error control: a focus on 2×2 designs. *Adv Meth Pract Psychol Sci*. 2021;4:1–20.
22. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6:65–70.
23. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*. 1988;75:383–6.
24. Frane AV. Are per-family Type I error rates relevant in social and behavioral science? *J Mod Appl Stat Meth*. 2015;14:12–23.
25. Lawrence J. Familywise and per-family error rates of multiple comparison procedures. *Stat Med*. 2019;38:3586–98.
26. Schulz KF, Grimes DA. Multiplicity in randomised trials I: endpoints and treatments. *Lancet*. 2005;365:1591–5.