J. For. Res. (2022) 33:731–739
https://doi.org/10.1007/s11676-022-01471-8

EDITORIAL

Six statistical issues in scientific writing that might lead to rejection of a manuscript

Evgenios Agathokleous[1] · Lei Yu[2]

Received: 21 February 2022 / Accepted: 17 March 2022 / Published online: 9 April 2022

Abstract Communication plays an important role in advancing scientific fields and disciplines, defining what knowledge is made accessible to the public, and guiding policymaking and regulation of public authorities for the benefit of the environment and society. Hence, what is finally published is of great importance for scientific advancement, social development, environmental and public health, and economic agendas. In recognition of these, the goal of a researcher is to communicate research findings to the scientific community and ultimately, to the public. However, this may often be challenging due to competition for publication space, although to a lesser extent nowadays that online-only publications have expanded. This editorial introduces six statistics-related issues in scientific writing that you should be aware of. These issues can lead to desk rejection or rejection following a peer review, but even if papers containing such issues are published, they may prevent cumulative science, undermine scientific advancement, mislead the public, and result in incorrect or weak policies and regulations. Therefore, addressing these issues from the early research stages can facilitate scientific advancement and prevent rejection of your paper.

Keywords Journal editor · Peer review · Rejection · Science communication · Scientific writing

Introduction

Owing to the dedicated work of its editorial office, the diligent work of its academic editors and peer reviewers, and contributions of authors from around the world, the *Journal of Forestry Research* (JFR) has been transformed into a prominent forestry journal. With a 2020 CiteScore[12] of 2.8, JFR ranks 40th among 142 journals of forestry, agricultural and biological sciences, while the updated 2021 tracker value increased to 3.8 (www.scopus.com; last updated 6 March 2022; accessed 17 March 2022). As the journal increases its profile of the world's forestry journals, more submissions are expected, resulting in a decreasing percentage of manuscripts that can be accepted and published.

JFR is published by non-profit, China-based academic societies and institutions and is not subject to policies of publishing that aim at maximizing economic profit

Project funding: This work is co-supported by the Startup Foundation for Introducing Talent of Nanjing University of Information Science & Technology (NUIST), Nanjing, China (Grant No. 003080), the Jiangsu Distinguished Professor program of the People's Government of Jiangsu Province, and the Outstanding Action Plan of Chinese Sci-tech Journals (Grant No. OAP–C–077).

The online version is available at http://www.springerlink.com.

Corresponding editor: Yanbo Hu.

✉ Evgenios Agathokleous
evgenios@nuist.edu.cn

✉ Lei Yu
relayjfr@163.com

[1] School of Applied Meteorology, Nanjing University of Information Science and Technology (NUIST), Nanjing 210044, People's Republic of China

[2] The Editorial Board of Journal of Forestry Research, Northeast Forestry University, Harbin 150040, People's Republic of China

[1] https://www.elsevier.com/connect/what-is-citescore-and-why-should-you-care-about-it

[2] Note: CiteScore is used as an index of journal ranking for illustrative purposes and its use does not imply it is considered the best or more appropriate than other of the many indexes existing.

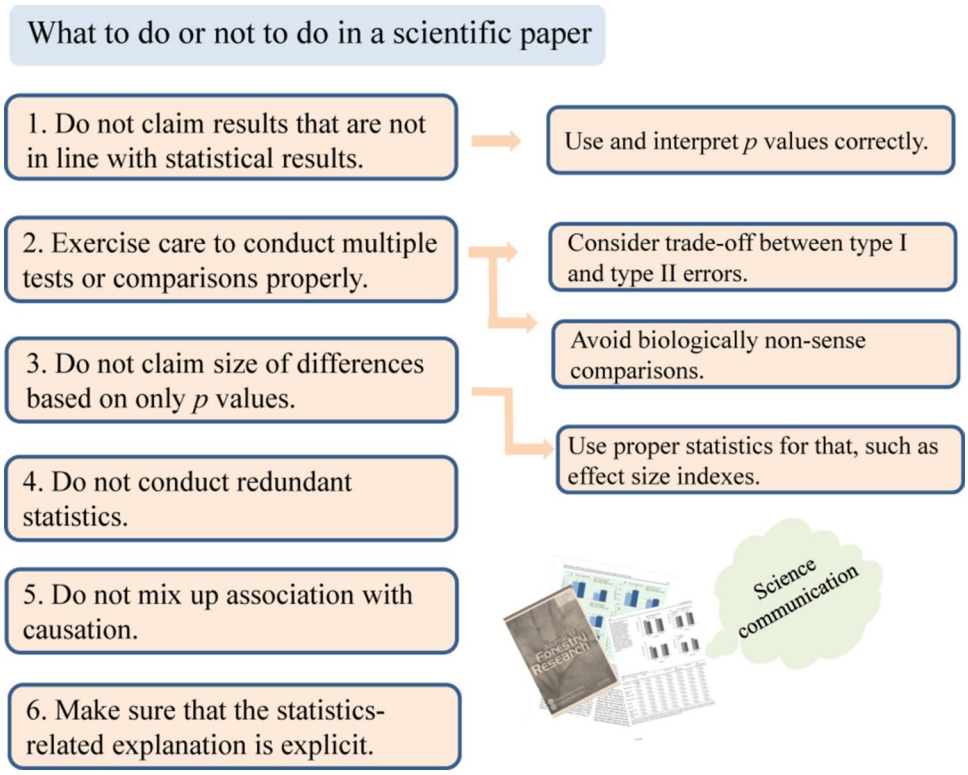

Fig. 1 Statistics-related issues in scientific writing

(Agathokleous 2022). Hence, the journal publishes a specific maximum number of manuscripts annually which means that no additional papers may be published, even if all are excellent, report cutting-edge scientific findings, or are game changing. For example, there were over 1600 submissions in 2021, of which only 7% were accepted for publication. With limited space, the number of papers that may be desk rejected (rejected by editors without assigning it to peer reviewers) is increasing. A desk rejection decision does not always have to do with the science itself or the manuscript quality, but it may simply be that the paper is not considered to be competitive enough among other submissions or because the journal has different publishing priorities at a given time. However, in a journal with competition for space, there are always reasons that can lead to a desk rejection, and statistics-related issues in scientific writing are among the top ones. The following is a list of issues based on ones I[3] have encountered frequently as an associate editor and then as an associate editor-in-chief of JFR, as well as in the framework of my editorial[4] and review[5] works in other scientific journals (Fig. 1). When such issues exist in an original manuscript, a set of them are commonly observed. However, as mentioned, these are based on our own experience (L.Y. is Deputy Editor-in-Chief of JFR) and fields of expertise in which we actively engage peer review and do not cover all the statistical areas such as mathematical modeling, computing systems like artificial neural networks, and machine learning. Moreover, we focus on statistics-related

[3] When 'I' and 'my' appear hereafter they indicate a personal view of the first author (E.A.). The subjective personal pronoun 'we' and the possessive 'our' refer to both authors hereafter.

[4] Associate Editor of Forestry Research. Editorial Board Member of Science of The Total Environment (STOTEN); Current Opinion in Environmental Science & Health (COESH); Journal of Environmental Science and Health, Part A: Toxic/Hazardous Substances and Environmental Engineering; Journal of Environmental Science and Health, Part B: Pesticides, Food Contaminants, and Agricultural Wastes; Plant Stress; Climate; Sci; Frontiers in Forests and Global Change; and Water Emerging Contaminants & Nanoplastics. Guest Managing Editor and Guest Editor in several journals, including STOTEN; Agriculture, Ecosystems and Environment; Current Opinion in Toxicology; COESH; Atmosphere; and Agronomy.

Footnote 4 (continued)

[5] Reviewed approximately 600 papers for 85 journals (https://publons.com/researcher/1194915/evgenios-agathokleous).

issues in scientific writing but not on technical aspects of statistical procedures themselves, such as the nature of dataset and data distribution, and code availability, i.e. the issue of making data and programing codes of data analyses publicly available.

Results claimed are not in line with statistical results[6] (the issue of *p* values)

Inference is made regarding the differences between experimental conditions, whereas either there is no statistical support for the comparison or the statistical result is not in agreement with the conclusion. The latter is more prevalent. In this case, authors often claim 'marginal' differences when the *p* value approaches or exceeds 0.1; the worst I have observed was for a *p* value higher than 0.2 and considered to be significant. Conversely, other authors have asserted that there was no difference if the *p* value was approximately 0.05. The former case is more severe. Regarding the latter, "*surely God loves the 0.06 as much as the 0.05*" (Rosnow and Rosenthal 1989). However, it is my view that if *p* values are to be used, there should be some acceptable range as a reference point. For example, numbers are commonly rounded up if the previous decimal is ≥ 5. I do not see any reason justifying that a *p* value in the range of 0.051 − 0.054 is statistically different from a *p* value of 0.045 − 0.050. A *p* value of approximately 0.05 suggests that the findings warrant further investigation and is enlightening. But these are my views and journals rarely have specific guidelines regarding the use of *p* values. Therefore, it remains highly subjective, resting with the editor's understanding, knowledge and ultimately opinion regarding what he/she finds acceptable. Nevertheless, I believe most, if not all editors, would find unacceptable the claim of significance when *p* values are approaching or exceeding 0.1. If it means to say what we like, no matter the statistics, why are we producing statistics? As an independent editor, I cannot force authors to replace *p* values with other measures or use them with complementary more informative metrics, but I expect authors to reach a conclusion based on *p* values in a logical and justified manner. Above all, we should remember that how *p* values are used defines what results are published and thus directs science and the progress of social and environmental development. Considering the widespread, subjective and highly personalized interpretation and use of *p* values, and how this can affect scientific progress (Dorey 2011; Masicampo and Lalande 2012), all biology journals should set precise guidelines for the interpretation and use of *p* values in consultation with editorial board members and statisticians. This includes JFR as well.

It should be added that the use of *p* values in biology has long been criticized by numerous statisticians. There is a famous quote: "*scientists the world over use them, but scarcely a statistician can be found to defend them. Bayesians in particular find them ridiculous, but even the modern frequentist has little time for them.*" (Senn 2001). Some scientists believe that the bar for statistical significance should be raised to 0.005 or 0.001 (Johnson 2013), while others call for the retirement of the statistical significance and the use of confidence intervals instead (Amrhein et al. 2019). In fact, *p* values can be replaced by or used together with other, more integrated indexes such as effect size estimates and their intervals (e.g. Agathokleous et al. 2016), which can lead to better informed decisions (Connor 2004; Nakagawa 2004; Muff et al. 2022a, b), while Bayesian counterparts (e.g. Bayes' factor) perform better (Goodman 2008; Johnson 2013; Wiens and Nilsson 2017). *p* values were not meant to be the sole criterion to attribute differences and compare magnitudes (Lew 2012; Nuzzo 2014; Agathokleous 2022; Alexander and Davis 2022). However, I do not believe that the replacement of *p* values will be occurring soon and, since they have become the backbone of biology, the so called 'gold standard' of validity (Nuzzo 2014), they should be used correctly. More details about statistical inference and bad practices, including the problematic hybrid interpretation of statistical results between Fisher's *p* values and the strict Neyman–Pearson approach, can be found in the literature (Connor 2004; Goodman 2008; Lew 2012; Nuzzo 2014; Muff et al. 2022b).

Concluding this section, we would draw attention to final points regarding the reporting of *p* values, should *p* values be used. First, no *p* value should be reported as being equal to zero. It could be < 0.001 or < 0.0001 but never = 0.000. Second, reporting only *p* values without other information is impractical. A minimum requirement would be the simultaneous reporting of the value of the statistic (e.g., *F*, *t*, or *U*). For *p* values over 0.05, the exact value should be stated instead of writing $p > 0.05$. As a point of reference, reading a widely used publication would be enlightening and helpful, such as the Publication Manual of the American Psychological Association (APA 2019).

Issues with multiple tests or comparisons

Scientific research has become more demanding in the twenty-first century due to the increased need for multi-factorial experimental designs in some disciplines (Rillig et al. 2019, 2021). This reflects a considerable increase in statistical testing within a study. For example, ecological research is often multidimensional, including numerous variables. If

[6] Note: this does not imply that biologically important results are statistically significant results. Statistically non-significant results can be biologically or practically important results and vice versa.

one examines the association of 15 soil quality parameters with the alpha diversity of communities of microorganisms, the probability of detecting one or more p values smaller than 0.05 increases from 5 to approximately 54%! And, if more than one index of alpha diversity is considered, this probability increases further. This leads to the question of how much uncertainty lies behind the results and conclusions of an array of studies. As the number of statistical tests and comparisons increases, the uncertainty and probability of rejection or acceptance of null hypotheses can increase, depending on how it was accounted for. But this is another issue that remains largely subjective and personalized, as journals rarely have specific guidelines on this.

Modifications of traditional statistical testing procedures are widely applied in a range of research fields to decrease Type I errors (i.e., rejection of a null hypothesis that is true). Perhaps the widespread and most widely used modification is the Bonferroni correction, a modification of alpha (α) by dividing it with k number of statistical tests or comparisons. That is, for a study with 10 tests, the corrected α would be 0.005 ($\alpha = 0.05/10$), if α was set at 0.05. The application of Bonferroni corrections reduces statistical power, highly increasing Type II errors, (i.e., acceptance of a null hypothesis that is false), and potentially contributing to a publication bias which eventually can thwart scientific advancement (Nakagawa 2004). For example, researchers that find many variables to be insignificant might simply choose to omit them from their paper and thus never covered by future meta-analyses, thereby contributing to a 'file-drawer effect'[7] and publication bias (Nakagawa 2004; Fanelli 2010). If the accumulation of knowledge is thwarted, an entire scientific field may be suppressed (Nakagawa 2004). A further type of correction is the sequential Holms-Bonferroni method (Holm 1979) which controls the family-wise error rate while reducing statistical power to a lesser extent compared to the standard Bonferroni correction; however, the probability of Type II errors remains considerably high (Nakagawa 2004). Including less relevant or biologically irrelevant variables in a study leads to unnecessarily increased probability of Type I errors, which often results in reviewers pointing to the need of corrections such as Bonferroni (Nakagawa 2004). Based on these issues, Nakagawa proposed that "*the practice of reviewers demanding Bonferroni procedures should be discouraged, (and also, researchers should play their part in carefully selecting relevant variables in their study)*" (Nakagawa 2004). These are not new issues and have long been known. For example, ending the use of Bonferroni procedures and starting to report effect sizes and/or confidence intervals for effect size or alternatives was proposed two decades ago in animal behavior and behavioral ecology research (Nakagawa 2004).

Some journals have specific guidelines about multiple testing or comparisons. An example is that of the *Annals of Applied Biology*, the journal of the Association of Applied Biologists, where more specific author guidelines regarding statistics have been developed and put into effect, while statistics editors also evaluate relevant submissions (Kozak and Powers 2017; Powers and Kozak 2019; Butler 2021). This is an example that can serve as a reference point for further development in the JFR as well as in other journals. Author guidelines of the *Annals of Applied Biology* discourage using comparisons not based on biological hypothesis, stating "*In particular, the use of multiple comparison adjustments such as Duncan's or Tukey's is not acceptable, nor is the use of letters to denote treatments which are 'not significantly different from each other'*." (https://onlinelibrary.wiley.com/page/journal/17447348/homepage/forauthors.html; Accessed 19 February 2022). Instead, it has been suggested to conduct post hoc comparisons that are of most interest, using the value of least significant difference (LSD) based on the relevant standard error of the difference (SED) from the analysis of variance (ANOVA) (Kozak and Powers 2017). Similarly, in unbalanced studies with an unequal number of experimental units (replicates) among experimental conditions or treatments, SED values may differ among comparisons; however, only post hoc comparisons of most interest should be made, but LSDs and SEDs should be reported for each comparison (Kozak and Powers 2017). Another suggestion is that, where a large number of variables exists, controlling the 'false discovery rate', the fraction of rejecting true null hypothesis, may be more appropriate than controlling the probability of even one false rejection of null hypothesis (Nakagawa 2004).

There are further options that can help with the trade-off between Type I and Type II errors. For example, the use of orthogonal or non-orthogonal linear contrasts is a good alternative, albeit their use is often complicated, difficult, or even impractical in terms of application, interpretation, and presentation, especially in light of current publishing policies of many journals. In fact, based on my experience as a reviewer, editor, referee, and author of literature reviews of numerous scientific papers, the use of post hoc comparisons in most cases is incorrect and problematic, while often planned (a priori) comparisons should be made. In highly multi-factorial studies, the number of biologically irrelevant comparisons is also high, many of which provide little or no useful information. This may be illustrated by a hypothetical example. A researcher studies the effect of various doses of the antibiotic tetracycline (0, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10,000 $\mu M\ L^{-1}$) on saplings of a poplar clone grown in either charcoal-filtered air (i.e., air pollutants are

[7] Meaning that negative or non-significant results are permanently stored in the researchers' drawer instead of being published, thus favoring the publication of positive and easier to publish results.

eliminated; clean atmosphere) or ozone-enriched air (polluted atmosphere). However, many of the possible comparisons are biologically irrelevant. For example, it is irrational to compare the effects of 0.001 μM tetracycline L^{-1} on plants raised in charcoal-filtered air with the effects of concentrations of $0.01 - 10{,}000$ μM tetracycline L^{-1} on plants in ozone-enriched air. Researchers should strive for planned comparisons wherever possible (Ruxton and Beauchamp 2008; see also Wiens and Nilsson 2017). If reviewers criticize the use of correctly applied a priori comparisons, it is important to address their comments and justify why a priori comparisons are correct and should be retained. In a paper my colleagues and I published six years ago, contrasts were used to examine the most biologically relevant questions/comparisons (Agathokleous et al. 2016). There were three reviewers and while endorsing the work, all had some comment(s) on the statistics and/or the way the results were presented; if post hoc comparisons among all means were done, reviewers would be satisfied. In fact, one of the issues raised was that the use of different specific questions, and thus contrasts, made the interpretation of figures and results more difficult, and dictated the repeated return to the questions/contrasts. The reviewers' comments were helpful to thoroughly revise the manuscript by completely changing the presentation of the results, including display elements. However, this is an example where a major revision would be a minor one if post hoc comparisons were used. It could also be a rejection if there were other critical deficiencies in the paper or if one or more reviewers had recommended rejection and the handling editor were unqualified.

Incorrect claims of sizes of differences

As noted previously, *p* values alone do not indicate variations in the size of differences among experimental conditions (Agathokleous 2022). For example, if the *p* values of the effect of treatments A and B compared to control C were 0.011 and 0.002, no inference should be made that treatment B had a larger effect than treatment A, yet such phenomena frequently occur in manuscripts submitted to journals. An inference that may be made in this case is that if treatments A and B had no real effect, a difference from the controls of equal or larger magnitude would be observed in 1.1% and 0.2% of study repetitions, respectively, due to random error.[8] In another example, the null hypothesis is rejected for the effect of liquid chemical treatments D and E on the mycorrhizae colonization of roots of pine seedlings grown in a cambisol soil, and the arithmetic means of treatments D and E were 50% and 10% greater than the arithmetic means of the water-treated control. Speculation that "chemical treatments D and E significantly increased mycorrhizae colonization, and chemical D had a more pronounced[9] effect" is inappropriate and misleading. The point is that *p* values say nothing about the magnitude of the effects or differences among experimental conditions. They only indicate the probability of a similar or more extreme finding than the one obtained in the study, given that the null hypothesis is true and the assumptions underlying the analysis are true to some extent (Butler 2021). A practice I often observe in manuscript submissions is drawing conclusions about the effect of size based only on *p* values or even differences in arithmetic means, such as denoting differences in treatment effects or ranking susceptibility/tolerance of different organisms or groups of organisms (Agathokleous and Saitanis 2020). Such a practice not only is harmful for the progress of science but is also misleading and thus, have societal implications (Agathokleous and Saitanis 2020). Whenever it is needed to make inference about the size of differences between experimental conditions, *p* values are insufficient. In fact, statistical significance or insignificance does not translate to biological importance (Ziliak and McCloskey 2008; Butler 2021), but effect sizes and their improving indexes can be used for biological or for practical importance (Agathokleous et al. 2016). There is a variety of effect size indexes that can be used, each with its own characteristics (Sullivan and Feinn 2012; Solla et al. 2018). Analysis of these indexes is beyond the scope of this paper, but there are various user-friendly software packages operating online or offline for the estimation of effect sizes as well as their improving indexes (e.g., Lenhard and Lenhard 2016; Agathokleous and Saitanis 2020; https://lbecker.uccs.edu/; https://goodcalculators.com/effect-size-calculator/; https://effect-size-calculator.herokuapp.com/). The availability of such computational tools makes calculation easier, even to those who might dislike making such calculations. The only task the user must do is to input the required data.

Redundant statistics

A problematic practice is to conduct redundant statistics. Although it might seem surprising, this problem exists in manuscript submissions even today. For example, in one study with single and combined effects of two factors each with two levels, the researchers carried out a two-way analysis of variance, but they also conducted independent *t* tests between experimental conditions within each factor. As a researcher, you may want to ensure that your manuscript contains no redundant statistics. Ask yourself whether some other statistical test that you have already conducted can provide answers to the questions your new statistical test

[8] Note: the error rate is tightly linked with *p* value (Sellke et al. 2001).

[9] Any synonym may be used.

is going to answer. Consider it for a while and think. If the answer is yes, then you should not conduct this further statistical test.

A situation I have observed many times is one of reviewers asking authors to conduct different tests to trace more significant results (and editors passively transferring reviewer comments onto authors). Such a practice reflects fishing for significant results. The more statistical tests/comparisons one runs, the more significant results will likely be found. As a basic principle, no changes to the statistics (by adding additional statistics) should be done without a clear purpose of doing so, such as due to problematic or incorrect methodology. Conducting different statistical tests also reflects redundancy, even if someone does not report all the results. As mentioned before, as long as you can justify why you did what you did, the chances that you will be asked to change your statistics are lower. Even if you are asked, it does not mean you should make changes, but doing so might enhance the chances of having your paper accepted.

Mixing up association with causation

It might be difficult to believe but mixing up association with causation occurs frequently. Association is a relationship between two variables. An X variation in the values of one variable is associated with a Y variation in the values of another. Association can represent causation, but in many cases it does not. If your study does not account for causation, no inference should be made to claim or imply causation. For example, you could state that "factor A was negatively associated with factor B" but you should not state that "factor B decreased due to factor A". If you want to claim causation based on association, you only need to distinguish between causal and non-causal associations (Stovitz et al. 2019; Kukull 2020). Otherwise, if your study does not support causation, be careful not to state or imply causation.

Lack of sufficient information

Insufficient statistical information is among the most important aspects that may determine the fate of a manuscript submission. Yet insufficient information about statistics appears widely in the literature (Kramer et al. 2016). As noted, often it is about justifying what one did in the scientific process. If what you did is correct, it cannot be rejected. Even if there are cases where alternatives might be advantageous, the question for an editor would be whether a potential change in the statistical processes would be beneficial (beneficial does not mean more 'statistical significances'). What would such a change add to the scientific content of the paper? Is such a change really needed? Would such a change be rather harmful, such as violating basic principles of statistics like fishing for significance and favoring type I errors over type II errors and vice versa? These are some questions an editor must answer when performing evaluations or re-assessments following peer review. These are some examples amongst many. The point is that if you detail adequately why you acted as you did and perhaps why you did not do something else,[10] you facilitate the work of editors and can prevent possibly unfair or incorrect criticisms by reviewers, thus enhancing the chances for a smooth peer review process. However, if the information about the experimental design and/or data analysis is insufficient to evaluate the robustness and validity of the study and does not permit its replication, a desk rejection is very likely. Here, I draw attention to some issues I encounter frequently, but those who have a keen interest in more detailed explanations can refer to the guidelines of the *Annals of Applied Biology* (Kozak and Powers 2017; Powers and Kozak 2019; Butler 2021) or *Science* (https://www.science.org/content/page/science-journals-editorial-policies#statistical-analysis).

The first issues that immediately come to mind are the lack of clarification of sample sizes, experimental and statistical units, and measures of dispersion around the mean, which should be done for each type of analyses. Without this information, the validity of the study cannot be assessed and replicated, which are the minimum requirements of scientific research. The meaning of replicate is often unclear or what is claimed to be a replicate is not valid. Author guidelines of the *Annals of Applied Biology* state that "*Particular care should be taken to explain what is meant by a replicate; only biological replication from independent units can be used to assess variation within and between treatments. Authors should consult a statistician if they require assistance in making inferences from designed experiments*" (https://onlinelibrary.wiley.com/page/journal/17447348/homepage/forauthors.html; Accessed 19 February 2022). Special attention should be given to the correct experimental unit, and thus the real replicates. Real replicates and the issue of pseudoreplication have been discussed extensively in the literature (Hurlbert 1984, 2004, 2013; Hawkins 1986; Potvin and Tardif 1988; Heffner et al. 1996; Oksanen 2001; Cottenie and De Meester 2003). Numerous reviewers recommend that a paper be rejected because the study was based on

[10] Excessive information is discouraged. There is no need to explain why you did not do each of the candidate alternative tests/procedures. This should be justified only in cases where what you did may be disputed such as when it comes to trade-offs between type I and type II errors.

pseudoreplicates and not real ones. In some cases, authors do not identify what the replicates were. In other instances, however, a study may be acceptable and equally important even if there were no real replicates, assuming there was still statistical support. For these reasons, the experimental and statistical units should be properly identified, and, where real replicates did not exist in a study, clarification should be made as to why the study is still valid and important. Finally, reporting arithmetic means without any measure of dispersion around the mean is unscientific. Arithmetic means by themselves are of little –if any- value either biologically or statistically. Hence, these are the first issues we suggest exercising care to explain explicitly.

A frequently occurring issue is lack of clarification of whether data transformation was applied. This is important information and should be made clear, especially when it comes to statistical tools which lead to false conclusions if data were not transformed, as is often the case for multivariate statistics.

No specification of the type of statistical model applied and/or the type of effects/factors is another occurring issue, and care should be made to specify these. Failure to conduct a dependent-samples analysis also occurs, whereas the experimental design would require such an analysis. It might also be the case that it is unclear if a study was based upon a dependent-samples design. Hence, it is important to clarify whether it is a dependent sample design.

The failure to clarify what post hoc test had been made is another observation that is well known (Ruxton and Beauchamp 2008). Therefore, if a post hoc test is applied, it is important to identify the test. [As noted in Sect. 2, specific guidelines regarding p values, α values, and multiple testing and comparisons are difficult to find.] In the absence of specific guidelines, the peer review process and acceptance of a manuscript for publication depends on academic editors. Independent academic editors however, should remain objective and unaffected by their opinion with what is correct or appropriate. But to help the editor and enhance your publication chances, it is important to justify why you applied an α correction or not. Especially because the selection and use of α correction is multi-dimensional and depends on a series of factors (Armstrong 2014).

Repeated measures can be applied to give more biological information in several cases (Powers and Kozak 2019), which is often the case for many research papers submitted to JFR. However, I frequently encounter (across journals) papers where repeated measures (or dependent-samples analysis) could be applied to provide comprehensive biological information but was not applied, and/or it is unclear if it was applied.

Finally, there is no harm in clarifying whether your hypothesis testing was one- or two-tailed. Although most journals rarely request clarification, there are some that do (e.g., *Science*; https://www.science.org/content/page/science-journals-editorial-policies#statistical-analysis). Commonly a two-tailed hypothesis test is the case; however, if it was one-tailed, it is important to ensure that the p values reported, if you used p values, are the correct ones. That is, in many cases the p values should be divided by two because most traditional data analysis software provide results for two-tailed hypothesis testing.

Conclusion

The purpose of this paper is not to create more questions than answers. However, as academic editors, we can raise authors' awareness about these issues, thus helping make proper selection of statistical tools from the earliest research stages. Authors cannot be forced to follow specific protocols but we can provide a basis for authors to consider and follow to make a correct selection of statistical procedures. No editors would reject a paper with justification of the procedure used simply because their opinions differ. But justifying the procedure that authors follow shows awareness of the issues and permits a proper evaluation of the study and the paper itself. We believe that any editor would appreciate a careful selection of tests or comparisons considering how type I and type II errors are affected.

It should be mentioned that this editorial should not be interpreted as suggesting that authors should simply satisfy the requirements of editors and journals –although it is often about compromise. Authors obtain funding, conduct research, and write up their results. This effort is often an outcome from government support (i.e., taxpayer funded), and authors should always bear in mind that the best choice is the one that can contribute to cumulative knowledge and society overall and not one that will facilitate the profit agenda of a publisher. You are free to follow or not to follow any editor's or journals' guidelines. The ultimate decision should be based on what would be ethically correct and fairer with respect to cumulative science and society, and not what would give a pass to a specific journal. If reviewers require the exclusion of specific data because they are not 'statistically significant' or for any other reason, you should ask yourself whether this is honest and ethically correct and what the implications to cumulative science and overall society might be. If you disagree with a particular guideline and you can provide robust scientific justification for it, you can always try a rebuttal, even if rarely successful. We hope you find this information useful. Editors look for good reasons to accept papers (Binkley et al. 2020), instead of searching for reasons to reject papers, and the methodology behind the statistics, beginning from the experimental design, is often a core determinant. Therefore, provide editors reasons for

a peer review to eventually accept rather than reject your paper.

Acknowledgements The authors are grateful to Dr. Ricardo Antunes Azevedo, Editor-in-Chief of the *Annals of Applied Biology* and Professor at the Departamento de Genética, Escola Superior de Agricultura "Luiz de Queiroz"/Universidade de São Paulo (ESALQ/USP), Brazil, for sharing information about editorial policies of the *Annals of Applied Biology*. Information about the Chinese *Journal of Forestry Research* was provided by the journal's office.

Declarations

Conflict of interest Any commercial name cited in this manuscript, e.g., of a journal or software, is not for advertisement, and the authors do not encourage or discourage the use of their services. Readers should make their own search and select products or services that suit them. The views presented herein are those of the authors and do not represent views of the editorial board or the editorial office of the journal, the publisher, the author's institution, or funding bodies that supported the authors. The authors declare that there are no conflicts of interest.

References

Agathokleous E (2022) Mastering the scientific peer review process: tips for young authors from a young senior editor. J Res 33:1–20

Agathokleous E, Saitanis CJ (2020) Plant susceptibility to ozone: a tower of Babel? Sci Total Environ 703:134962. https://doi.org/10.1016/j.scitotenv.2019.134962

Agathokleous E, Paoletti E, Saitanis CJ, Manning WJ, Sugai T, Koike T (2016) Impacts of ethylene diurea (EDU) soil drench and foliar spray in *Salix sachalinensis* protection against O3-induced injury. Sci Total Environ 573:1053–1062

Alexander BCS, Davis AS (2022) Perspective: Scientific rigor or ritual? statistical significance in pest management science. Pest Manag Sci 78:847–854

Amrhein V, Greenland S, McShane B (2019) Retire statistical significance. Nature 567:305–307

Armstrong RARA (2014) When to use the Bonferroni correction. Ophthal Physiol Opt 34:502–508

Binkley D, Fernandez ME, Fredricksen T, Mäkinen H, Prescott C, Tomé M (2020) How to avoid having your manuscript rejected: Perspectives from the Editors of Forest Ecology and Management. For Ecol Manage 473:118321

Butler RC (2021) Popularity leads to bad habits: Alternatives to "the statistics" routine of significance, "alphabet soup" and dynamite plots. Ann Appl Biol 180:182–195

Connor JT (2004) The value of a *p*-valueless paper. Am J Gastroenterol 99:1638–1640

Cottenie K, De Meester L (2003) Comment to Oksanen (2001): reconciling Oksanen (2001) and Hurlbert (1984). Oikos 100:394–396

Dorey F (2011) Statistics in brief: Interpretation and use of *p* values: all *p* values are not equal. Clin Orthop Relat Res 469:3259–3261

Fanelli D (2010) Do pressures to publish increase scientists' bias? An empirical support from US States data. PLoS One 5:e10271. https://doi.org/10.1371/journal.pone.0010271

Goodman S (2008) A dirty dozen: twelve *p*-value misconceptions. Semin Hematol 45:135–140

Hawkins CP (1986) Pseudo-understanding of pseudoreplication: a cautionary note. Bull Ecol Soc Am 67:184–185

Heffner RA, Butler MJ IV, Reilly CK (1996) Pseudoreplication revisited. Ecology 77:2558–2562

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70

Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments. Ecol Monogr 54:187–211

Hurlbert SH (2004) On misinterpretations of pseudoreplication and related matters: a reply to Oksanen. Oikos 104:591–597

Hurlbert SH (2013) Pseudofactorialism, response structures and collective responsibility. Austral Ecol 38:646–663

Johnson VE (2013) Revised standards for statistical evidence. Proc Natl Acad Sci USA 110:19313–19317

Kozak M, Powers SJ (2017) If not multiple comparisons, then what? Ann Appl Biol 171:277–280

Kramer MH, Paparozzi ET, Stroup WW (2016) Statistics in a horticultural journal: problems and solutions. Hort Sci 51:1073–1078

Kukull WA (2020) Association, cause, and causal association, revised: reasoning and methods. Rosenberg's Mol Genet Basis Neurol Psychiatr Dis 65:121–128

Lew MJ (2012) Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know P. Br J Pharmacol 166:1559–1567

Masicampo EJ, Lalande DR (2012) A peculiar prevalence of p values just below.05. Q J Exp Psychol 65:2271–2279

Muff S, Nilsen EB, O'Hara RB, Nater CR (2022a) Response to 'Why P-values are not measures of evidence' by D Lakens. Trends Ecol Evol 37:291–292

Muff S, Nilsen EB, O'Hara RB, Nater CR (2022b) Rewriting results sections in the language of evidence. Trends Ecol Evol 37:203–210

Nakagawa S (2004) A farewell to Bonferroni: the problems of low statistical power and publication bias. Behav Ecol 15:1044–1045

Nuzzo R (2014) Scientific method: statistical errors. Nature 506:150–152

Oksanen L (2001) Logic of experiments in ecology: is pseudoreplication a pseudoissue? Oikos 94:27–38

Potvin C, Tardif S (1988) Sources of variability and experimental designs in growth chambers. Funct Ecol 2:123

Powers SJ, Kozak M (2019) Repeated measures: There's added value in modelling over time. Ann Appl Biol 175:129–135

Rillig MC, Ryo M, Lehmann A, Aguilar-Trigueros C, Buchert S, Wulf A, Iwasaki A, Roy J, Yang G (2019) The role of multiple global change factors in driving soil functions and microbial biodiversity. Science 366:886–890

Rillig MC, Ryo M, Lehmann A (2021) Classifying human influences on terrestrial ecosystems. Glob Chang Biol 27:2273–2278

Rosnow RL, Rosenthal R (1989) Statistical procedures and the justification of knowledge in psychological science. Am Psychol 44:1276–1284

Ruxton GD, Beauchamp G (2008) Time for some a priori thinking about post hoc testing. Behav Ecol 19:690–693

Sellke T, Bayarri MJ, Berger JO (2001) Calibration of p values for testing precise null hypotheses. Am Stat 55:62–71

Senn S (2001) Two cheers for *P*-values? J Epidemiol Biostat 6:193–204
Solla F, Tran A, Bertoncelli D, Musoff C, Bertoncelli CM (2018) Why a p-value is not enough. Clin Spine Surg 31:385–388
Stovitz SD, Verhagen E, Shrier I (2019) Distinguishing between causal and non-causal associations: implications for sports medicine clinicians. Br J Sports Med 53:398–399
Sullivan GM, Feinn R (2012) Using effect size—or why the *P* value is not enough. J Grad Med Educ 4:279–282
Wiens S, Nilsson ME (2017) Performing contrast analysis in factorial designs: from nhst to confidence intervals and beyond. Educ Psychol Meas 77:690–715
Ziliak ST, McCloskey DN (2008) The cult of statistical significance : how the standard error costs us jobs, justice, and lives, 1st edn. University of Michigan Press
APA (2019) Publication Manual of The American Psychological Association, 7th Edition. The American Psychological Association.
Lenhard W, Lenhard A (2016) Computation of effect sizes. Retrieved from: https://www.psychometrica.de/effect_size.html. Psychometrica.