***PeerJ preprint* Commentary**

Key steps to avoiding artistry with significance tests

C. Patrick Doncaster[1*] and Thomas H. G. Ezard[2]

[1]Biological Sciences, University of Southampton, SO17 1BJ, UK. cpd@soton.ac.uk

[2]Ocean and Earth Science, University of Southampton, SO14 3ZH, UK.
T.Ezard@soton.ac.uk

*Correspondence author. E-mail: cpd@soton.ac.uk

Abstract

Statistical significance provides evidence for or against an explanation of a population of interest, not a description of data sampled from the population. This simple distinction gets ignored in hundreds of thousands of research publications yearly, which confuse statistical with biological significance by referring to hypothesis-testing analyses as demonstrating significant results. Here we identify three key steps to objective reporting of evidence-based analyses. Firstly, by interpreting *P*-values correctly as explanation not description, authors set their inferences in the context of the design of the study and its purpose to test for effects of biologically relevant size; nowhere in this process is it informative to use the word 'significant'. Secondly, empirical effect sizes demand interpretation with respect to a size of relevance to the test hypothesis. Thirdly, even without an *a priori* expectation of biological relevance, authors can and should interpret significance tests with respect to effects of reliably detectable size.

Key-words: frequentist statistics, model fitting, null hypothesis, p-values, significance testing

Statistical analysis provides one of the most powerful tools for generalizing from sampled data. All too often, it results in some of the most artful descriptions of significant results. Recent commentaries have provided clear guidance on the meaning of *P* values (see Glossary) and the limitations of significance testing (Wasserstein & Lazar, 2016, and references cited therein). Critiques draw attention to the misapplication of frequentist statistics to the significance of rare events (e.g., Ioannidis, 2005), the increasing prioritization of *P* values over effect sizes (e.g., Burnham & Anderson, 2002; Murtaugh, 2014; Chavalarias et al., 2016), selection bias from *P*-value hacking (Ziliak, 2017), the unpredictability of *P* values and need for their empirical calibration (Halsey et al., 2015; Lazzeroni, Lu & Belitskaya-Lévy, 2016; Claridge-Chang & Assam, 2016; Bruns & Ioannidis, 2016), inflated significance from model selection (Forstmeier & Schielzeth, 2011), and a reducing explanatory power of significance testing in ecology (Low-Décarie, Chivers & Granados, 2014). When appropriately applied to well-powered studies, frequentist statistics nevertheless retain broad applicability as a mechanism for estimating the compatibility of data to a refutable hypothesis.

In this article we address a different issue with significance testing that arises only from confusion about the interpretation of tests. The issue is that authors routinely misrepresent *P* values as evidence for or against significant pattern in the data. The foundational logic of statistical analysis determines that *P* values apply only to inferences about the population sampled by the data, not to descriptions of the sample itself. Statistical significance provides evidence for or against an explanation of the population of interest; statistical significance says nothing about patterns in the sample and does not provide evidence of biological significance that may or may not have been described by parameter estimates. Although science has long recognized the non-equivalence of statistical and biological significance (Berkson, 1938), the problem of statistical explanation masquerading as description evidently still awaits effective articulation.

Here we describe three key steps to avoiding artistry with significance tests, by objective reporting of evidence-based analyses. Firstly, we demonstrate the benefits in exposing the study design to critical appraisal that obtain from separating the explanation provided by a significance test from the description of effect size that follows after the test. Secondly, we review the deceptive attractions of the confidence interval, and warn against using it to bridge across explanation and description. Confidence intervals cannot circumvent the need to interpret empirical effect sizes with respect to a size of relevance to the test hypothesis. Finally, we propose an interim solution to the difficulty that many empirical studies lack *a priori* knowledge of an effect size of biological relevance against which to calibrate the power of significance tests. They can still run useful significance tests with respect to effects of reliably detectable size.

The enduring appeal of significant results

The words ‘significant’ or ‘significantly’ appear in the abstract or title of over 400,000 articles listed in the Web of Science Core Collection for the year 2016, amounting to 18% of all records for that year. Usage has risen 3.8-fold since 1996, outstripping a 2.5-fold rise in the annual production of all articles. In environmental sciences and ecology, usage has risen 4.2-fold over the same period compared to a 3.0-fold rise in annual production. The words appear in over 19,000 environmental and ecology articles published in 2016, amounting to 27% of all records in the discipline and more than in any other research area except Oncology.

These trends belie a renaissance over the last 20 years in Bayesian analysis and information-theoretic modelling that de-emphasizes statistical significance. The articles that cite significant entities include several of the most influential of all research outputs, with the top ten having Impact Factors exceeding 400 (citations within 24 months of publication). The reports of main findings nevertheless involve ambiguous claims in all cases except a minority that make reference to clearly non-statistical significance (e.g., “significant advances in the field” – see Box 1). To appreciate the reason why this is such a universal problem requires a close inspection of the meaning of statistical significance.

[Box 1 here]

Statistical significance explains the population not the data

A refutable null hypothesis H_0 and its test alternative H_1 always make propositions about pattern in a population of interest, from which the study takes data samples for analysis. The hypotheses concern a specified explanation of the domain of inference set by the population, not the significance of an effect on the samples that represent it. When using frequentist statistics, each P value describes the probability of data at least as deviant given H_0, and thus the probability of making an error by rejecting H_0. The inference it permits therefore concerns an explanation of the population, not a description of the data sampled from it. An example will illustrate this distinction and its consequences.

Consider a field experimental test for the effectiveness of a pesticide treatment on crop yield. Replicate independent plots, representatively sampling a population of crop plants of interest, were randomly assigned to a low or a high dosage of the pesticide, or to a water control. The study authors might correctly report a one-way analysis of variance with Helmert contrasts as: "Crop yield depended on treatment ($F_{2,33} = 4.39$, $P = 0.02$), with no evidence of a difference between low and high dosages of pesticide (pesticide vs control contrast: $t_{33} = 2.92$, $P = 0.006$; low vs high dosage contrast: $t_{33} = 0.51$, $P = 0.61$)."

To claim that "crop yield depended significantly on treatment" would misinterpret P, which finds the data incompatible with the null hypothesis, as a description of the data, which finds different sample means. The analysis never tests for, let alone finds, a significant difference between sample means. The correct inference, that "yield depended on treatment" within the population of interest, is evidenced by the low probability of a false positive "($F_{2,33} = 4.39$, $P = 0.02$)" using valid assumptions about the design of sampling from the population. Having established the presence of a treatment effect, a description or illustration of its size can inform the biological significance of the effect within the domain of inference (set by the population) and thus the interpretation of the test.

To claim that "there was no significant difference between the dosages (low vs high dosage contrast: $t_{33} = 0.51$, $P = 0.61$)" would mislead, in implying that they differed albeit not significantly. Worse yet, it would be wrong, because the P value relates to a hypothesized absence of difference in the population, not in samples from the population. One can draw a subtly different inference, however, that "low and high dosages did not differ detectably in their effect on yield." This statement reports an explanation as far as we can ascertain it from the test. Now also it becomes clear that we would want to have calculated *a priori* the power of the design to detect an effect of biologically relevant size, to provide the reader with a level of confidence in the apparent equality. Indeed, regardless of the significance of an effect, we would do well to evaluate it against an *a priori* size threshold (see Box 2).

[Box 2 here]

Expunging the word solves the problem

These details of wording may appear fussy. Perhaps use of 'significant' seems an acceptable shorthand for interpreting the data. If a two-sample difference test has $P < 0.05$, then surely the samples differ significantly? No, they do not. The sampled populations probably differ, given a well-powered test and valid assumptions about sampling, by a small or a large amount that is estimable from some parametric measure of the difference between the two samples (see Faul et al., 2007 for power calculation, and Lakens, 2013 for effect-size estimation). If a regression has $P < 0.05$, then surely the data show a significant trend? No, they do not. The distribution of sample data provides convincing evidence of a trend in the population, given a well-powered test and valid assumptions about sampling. The regression slope quantifies the estimated size of effect, and its confidence interval illustrates the strength of evidence against the H_0 of no trend. In short, both significance tests provide evidence of pattern in the population of interest; neither test provides evidence of significant pattern in the data.

The wording used to report results betrays the authors' motivations in designing the study. Reference to results being significant restricts the domain of inference to the sample data, which sets authors and audience on the path of treating hypothesis testing and explanation as different enterprises. Yet if the sample is the population, then statistical significance has no meaning; all that is left to do is describe biological significance in the magnitudes of parameters calculated from the data. Authors wishing to fit their data to statistical significance find ample opportunity with descriptions of effects that "approached significance" (used in 95 abstracts across all subject areas in 2016) or samples that differed "albeit/although/but/however not significantly" (476 uses), where authors may have wished to see difference, and differences that "were not significant" (1,334 uses), where authors may have wished not to see them (see a full compilation in Hankins, 2013). The data remain resolutely immutable; they cannot be fitted up to anything (Hilborn & Mangel, 1997). Removing the reference to significance removes the opportunity to fit the data to an explanation, by coercing the statement into a conventional report on the detection of effects in alternative models fitted to data.

Confidence intervals alone tell an unreliable story

A shift in focus from significance to detection of effects reinforces the reality that *P* values relate fundamentally to replication, treatment levels and the different responses among them. It sets inferences in the context of the scope and power of the study, and the validity of assumptions underpinning the statistical models. It thereby opens the way to scrutiny of every stage in the data pipeline of evidence-based analysis (Leek & Peng, 2015a; Leek & Peng, 2015b).

The core principles of explanation and description in significance testing apply to statistical analysis using confidence intervals (CIs). Several influential papers have recommended CIs as more informative than the all-or-nothing approach of significance testing (Halsey et al., 2015, Johnson, 1999, Nakagawa & Cuthill, 2007). The 95% CI encompasses the range of plausible values of the null hypothesis, given only the sample data and the assumption of normality. It thus appears to provide more information than the *P* value for a specified H_0, because it encompasses all plausible H_0. We should exercise great care, however, in using the CI for *post hoc* rejection of alternative H_0. This is because the power of a hypothesis-testing study is quantified with respect to an effect size of relevance to the test hypothesis which itself pertains to the refutable null hypothesis. Each H_0 therefore demands a separate power calculation. In consequence, the CI generally provides no more useful information than that given by the *P* value, because it derives from the same data and assumptions (Murtaugh, 2014; van Helden, 2016). As a visual representation of the significance test, moreover, it conceals the pattern of data distribution, which will underpin the assumptions of the test. Although it illustrates the margin of error around the effect size estimate (Halsey et al., 2016), it requires the same interpretation as the *P* value with respect to the power of the study to detect an effect of relevant size (see Box 2). The following example will illustrate this point.

Consider three alternative sampling strategies for measuring change in crop yield due to a pesticide application (figure 2). Study A obtains an average gain in yield of 41.0 kg/ha across a sample of 10 fields. Its 95% CI does not include H_0: $\mu = 0$. It thus finds that a population with a normal distribution of equally variable gains around $\mu = 0$ will yield sample means at least as deviant as the observed one in less than 5% of equally-replicated samples. In contrast, $\mu = 10$ or 70 kg/ha, both lying within the CI, will yield sample means at least as deviant in more than 5% of samples. Study A can report a detectable change in yield (rejection of H_0: $= 0$, $t_9 = 2.758$, $P = 0.022$). An alternative study B, however, with twice the replication and consequently a smaller CI, obtains a lower sample mean from the same population and fails to reject the null hypothesis of no change ($t_{19} = 1.331$, $P = 0.199$, figure 2). Does this more powerful study provide a more robust explanation? We can't tell without evaluating outcomes against an effect size of relevance to the test hypothesis.

Small-sample studies have little reliability in testing for small effects. Suppose the breakeven gain in yield for a cost-effective pesticide is $\delta = 10$ kg/ha, in a population of fields with a standard deviation of $\sigma = 45$ kg/ha (consistent with the observed variability around sample means). We would therefore wish to detect a positive effect for any true standardized gain above $\delta/\sigma = 0.222$. In this case, studies A and B have respectively only 16% and 25% power to detect a positive effect at $\alpha = 0.05$, in a population with $\delta/\sigma = 0.222$ (R commands in Doncaster and Davey, 2017; see also Faul et al., 2007). Study A thus has an 84% probability of Type-II error: failure to reject H_0 of no positive effect, given a true standardized mean at this threshold δ/σ. The few occasions on which this design correctly rejects H_0 will, moreover, almost certainly arise by virtue of its sample mean overestimating such a small true mean (Halsey et al., 2015; Lemoine et al., 2016). With its lower confidence limit lying below the sample mean by $t_{[0.025]}\cdot\sigma/\sqrt{N} = 32$ kg/ha on average, given $\sigma = 45$, the sample mean is more likely than not to overestimate a true means of anything up to 32 kg/ha when $P < 0.05$. From the observed results, we can only conclude that the pesticide effect in study A may grossly overestimate its cost-effectiveness; moreover, the absence of detectable effect in study B has up to 75% chance of undervaluing small but cost-effective gains due to the pesticide.

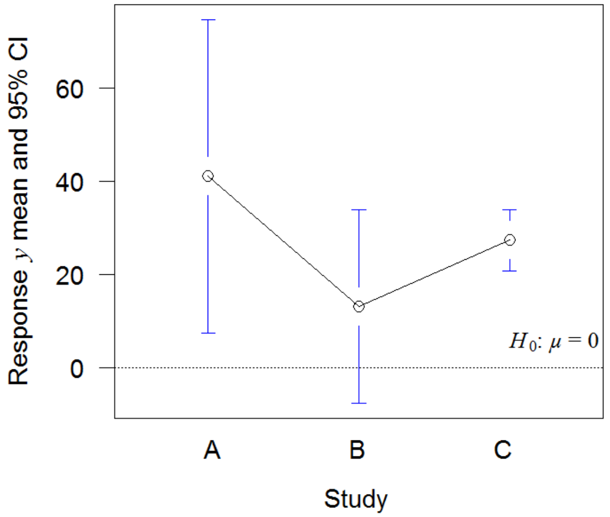


Figure 2. Three one-sample studies of the same population, with true mean estimated by each to lie within the confidence interval given by the blue vertical line above and below its sample mean (plotted with R package 'gplots' by Warnes et al., 2016, R Core Team, 2017). $N = 10$, 20, 170 for A, B, and C respectively.

We need a much larger sample to draw robust conclusions of biological relevance. Study C has 170 observations, which provide at least 90% power to reject H_0 of no gain, given a cost-effective true gain. We can reject H_0: = 0 ($t_{169} = 8.289$, $P < 0.001$, figure 2). Its CI is smaller, indicating higher precision in estimating the population mean μ. The lower confidence limit lies so far above = 0, moreover, that our rejection of this H_0 is very unlikely to be caused by a haphazardly overestimated size of the true mean. Because the CI lies well above even the 10-kg/ha threshold of relevance, we might wish also to reject H_0: = 10 ($t_{169} = 5.254$, $P < 0.001$, figure 2). This *post hoc* test comes with an often neglected caveat, however, that the rejection of H_0: = 10 may well be caused by a haphazardly overestimated size of the true mean. Because the power calculation applied only to = 0, and not to = 10, the interpretation: "yield change exceeds 10 kg/ha ($t_{169} = 5.254$, $P < 0.001$)." cannot reliably say by how much it exceeds this threshold.

Small-sample studies can have useful predictive power, if they test for the presence of large effects. For example, study A has 90% power to detect a positive effect given a true standardized effect $\delta/\sigma = 1.0$, and study B has 90% power given $\delta/\sigma = 0.67$. Suppose that the breakeven gain for the pesticide is $\delta = 45$ kg/ha for $\sigma = 45$ kg/ha. Then study A has 90% power to detect a yield increase given $\delta = 45$, or in the other direction it has 90% power to detect a less-than cost-effective increase given $\delta = 0$. Study B likewise has >90% power to detect these categories of effect size. From the CI of study A we conclude that yield changes (rejection of H_0: = 0, $P < 0.05$), but the estimated amount is less than cost-effective. From the CI of study B we conclude that if there is any yield change, it is less than cost effective (rejection of H_0: = 45, $P < 0.05$). These conclusions reflect the reality that the datasets for figure 2 were generated in R by random sampling from a normal distribution with specified parameters $\pm \sigma = 26 \pm 45$ kg/ha ($\delta/\sigma = 0.58$).

Computer-generated data allow us the privilege of repeating each study multiple times, to play out the advantages of study replication predicted by the statistics (figure 3). In accordance with the threshold α = 0.05 for significance, all three designs reject the true = 26 in ~5% of repeats, showing in figure 3 by ~5 bars in each study being ether red lying above = 26 or blue lying below it. Design C nevertheless produces vastly more consistent estimates than designs A and B. Design A fails to reject a null hypothesis of no effect in ~63% of repeats (~63 of its red bars lying below = 0), reflecting its 37% power to detect an effect at the true δ/σ = 0.58. If we meta-analyzed 17 studies of design A, however, we would match the replication of one study C, and therefore also its power (Borenstein et al., 2010; Koricheva, Gurevitch & Mengersen, 2013).

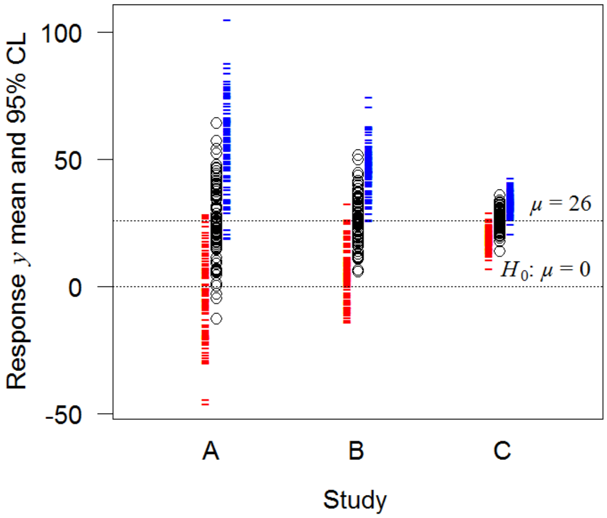


Figure 3. Sample means (black), and lower (red) and upper (blue) 95% confidence limits, from 100 repeats of each of the three studies in figure 2.

Often researchers have the opportunity only for one test of a treatment effect in a single study, without prior knowledge of an effect size relevant to the test hypothesis. Then there is no value in reporting a *post hoc* power analysis (Lenth, 2001). This would lead only to a nonsensical conclusion, of the sort that Study A in figure 2 had 82% power to detect its observed effect size of 41 kg/ha with the observed standard deviation of 47 kg/ha. Such statements ignore the high risk of the study estimate having inflated a much smaller true effect. It does make sense, however, to include in the description of study design the lower threshold of true standardized effect that gives the study 90% power to reject the null hypothesis (e.g., $\delta/\sigma = 1.0$ for a study of design A). Any lower power than this risks substantial imprecision and inaccuracy (Halsey et al., 2015). The threshold provides a caveat for robustness that future-proofs the study inferences against some eventual alignment of effect size with biological relevance.

Does information-theoretic modelling with likelihood tests circumvent the issue of underpowered tests giving unreliable estimates of effect size? Unfortunately not, because these methods use the same statistical information. Differences in Akaike’s Information Criterion (AIC), for example, may have direct equivalents in *P* values (Murtaugh, 2014). They can distinguish the more parsimonious of alternative models, but all candidate models will have poor explanatory precision and descriptive accuracy in an underpowered design.

Concluding remarks

In seeking to generalize from individual samples, the scientific pursuit of knowledge opposes the artistic quest for significant examples of universal truths (Kundera, 1986). Scientists take artistic license by making claims for significant pattern in their samples. They can easily excise the suspicion of fitting their data to a desired model by refraining from any reference to the significance of the data when reporting statistical analyses. A greater difficulty arises in evaluating the precision of significance tests and the accuracy of effect-size estimates, which are done with respect to an effect size of biological relevance. Studies frequently lack such prior knowledge, in which case authors can still usefully report the size of true effect for which the study has 90% power to detect its presence. Or why not instead give Bayesian statistics a try? The Bayesian requirement for a prior probability distribution often deters researchers, and yet it is no more arduous than the frequentist requirement for a standardized size δ/σ of relevance to the test hypothesis (McCarthy, 2007; Beaumont, 2010; Love et al., 2017; Rouder et al., 2017). Clearly there is a need for well-informed training of quantitative methods in graduate schools (Barraquand et al., 2014), which have a key position of influence in promoting logical analysis, and in curbing inappropriate manipulations of terminology and imprecise or inaccurate reporting of inferences.

Acknowledgments

We thank Bjorn Robroek for comments on an earlier draft.

Funding statement

THGE is funded by NERC Fellowship NE/J018163/1.

References

Barraquand F, Ezard THG, Jørgensen PS, Zimmerman N, Chamberlain S, Salguero-Gómez R, Curran TJ, Poisot T. 2014. Lack of quantitative training among early-career ecologists: a survey of the problem and potential solutions. *PeerJ* 2:e285.

Beaumont MA. 2010. Approximate Bayesian computation in evolution and ecology. Annual. *Review of Ecology, Evolution and Systematics* 41:379-406.

Berkson J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association* 33:526-336.

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. 2010. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods* 1:97-111.

Bruns SB, Ioannidis JPA. 2016. *P*-curve and *p*-hacking in observational research. *PLoS ONE* 11:e0149144.

Burnham KP, Anderson DR. 2002. *Model Selection and Multi-Model Inference: a Practical Information-Theoretic Approach*. Tokyo: Springer.

Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. 2016. Evolution of reporting *P* values in the biomedical literature, 1990-2015. *Journal of the American Medical Association* 315:1141-1148.

Claridge-Chang A, Assam PN. 2016. Estimation statistics should replace significance testing. *Nature Methods* 13:108-109.

Doncaster CP, Davey AJH. 2017. *Examples of Analysis of Variance and Covariance. Available at www.southampton.ac.uk/~cpd/anovas/datasets/* (accessed 8 August 2017).

Doncaster CP, Davey AJH, Dixon PM. 2014. Prospective evaluation of designs for analysis of variance without knowledge of effect sizes. *Environmental and Ecological Statistics* 21:239-261.

Faul F, Erdfelder E, Lang A-G, Buchner A. 2007. G*Power 3: a flexible statistical power analysis programme for the social, behavioural, and biomedical sciences. *Behavior Research Methods* 39:175-191.

Forstmeier W, Schielzeth H. 2011. Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology* 65:47-55.

Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. 2015. The fickle *P* value generates irreproducible results. *Nature Methods* 3:179-185.

Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. 2016. Response to "Confidence intervals are no salvation from the alleged fickleness of the *P* value". *Nature Methods* 13:606-606.

Hankins M. 2013. Still not significant. *Available at* https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/ (accessed 8 August 2017).

Hilborn R, Mangel M. 1997. *The Ecological Detective: Confronting Models with Data*. Princeton: Princeton University Press

Ioannidis JPA. 2005. Why most published research findings are false. *PLoS Medicine* 2:696–701.

Johnson D. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63:763-772.

Kassambara A. 2017. STHDA: Statistical Tools for High-Throughput Data Analysis. *Available at* http://www.sthda.com (accessed 8 August 2017).

Koricheva J, Gurevitch J, Mengersen K. 2013. *Handbook of Meta-analysis in Ecology and Evolution*. Princeton: Princeton University Press.

Kundera M. 1986. *L'Art du Roman*. Mesnil-sur-l'Estrée: Éditions Gallimard.

Lakens D. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology* 4:Article 863.

Lazzeroni L, Lu Y, Belitskaya-Lévy I. 2016. Solutions for quantifying *P*-value uncertainty and replication power. *Nature Methods* 13:107-108.

Leek JT, Peng RD. 2015a. Opinion: Reproducible research can still be wrong: adopting a prevention approach. *Proceedings of the National Academy of Sciences USA* 112:1645-1646.

Leek JT, Peng RD. 2015b. *P* values are just the tip of the iceberg. *Nature* 520:612.

Lemoine NP, Hoffman A, Felton AJ, Baur L, Chaves F, Gray J, Yu Q, Smith MD. 2016. Underappreciated problems of low replication in ecological field studies. *Ecology* 97:2554-2561.

Lenth RV. 2001. Some practical guidelines for effective sample size determination. *American Statistician* 55:187-193.

Love J et al. 2017. JASP release 0.8.1.2. *Available at www.jasp-stats.org* (accessed 16 August 2017).

Low-Décarie E, Chivers C, Granados M. 2014. Rising complexity and falling explanatory power in ecology. *Frontiers in Ecology and the Environment* 12:412-418.

McCarthy MA. 2007. Bayesian Methods for Ecology. Cambridge University Press.

Murtaugh PA. 2014. In defense of P values. *Ecology* 95:611-617.

Nakagawa S, Cuthill IC. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews* 82:591-605.

R Core Team 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. *Available at https://www.R-project.org* (accessed 14 August 2017).

Rouder JN, Morey RD, Verhagen J, Swagman AR, Wagenmakers E-J. 2017. Bayesian analysis of factorial designs. *Psychological Methods* 22:304-321.

van Helden J. 2016. Confidence intervals are no salvation from the alleged fickleness of the *P* value. *Nature Methods* 13:605-606.

Warnes GR et al. 2016. Package ‘gplots’. CRAN repository. *Available at https://cran.r-project.org/package=gplots* (accessed 14 August 2017).

Wasserstein RL, Lazar NA. 2016. The ASA's statement on *p*-values: context, process, and purpose. *American Statistician* 70:129-133.

Ziliak S. 2017. *P* values and the search for significance. *Nature Methods* 14:3-4.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3394v1 | CC BY 4.0 Open Access | rec: 6 Nov 2017, publ: 6 Nov 2017

Text Boxes

Glossary

Assumptions: The necessary preconditions for fitting any statistical model to data. No form of generalization from the data is possible without assumptions. They provide the context for, and the means of evaluating, explanations of the population sampled by the data.

Confidence interval (CI): The range of plausible values of a refutable null hypothesis given only the data and assumptions about their distribution. The CI of a mean sampled from a normal distribution lies between 95% limits at $t_{[0.025]}$·SE above and below the mean. The 95% CI thus encompasses the range of values of a true mean with ≤ 95% chance of not producing as deviant a sample mean. Bootstrapping provides a generic means of calculating CI.

Design: Data collection requires designing to meet the specifications of the statistical model that will test a hypothesis of interest. The test hypothesis drives the design of evidence-based data analysis for reproducible inferences from a replicable study; different designs addressing the same broad hypothesis are liable to produce mixed results and different effect sizes.

Effect size: The size of treatment effect on a response (e.g., a difference between means or a regression slope), sometimes standardized against error variation. Effect size is estimated from data independently of significance, and is only sensible to report for a detectable effect.

Hypothesis: A proposition about a population of interest. A test hypothesis, H_1, is a proposition of biologically informative pattern; it is calibrated against a refutable null hypothesis, H_0, of no such pattern. Hypothesis-testing distinguishes alternative explanations of the population, and can be applied to predicting future trends.

Model: A statistical model defines the test and null hypotheses in the form of an equation. The model is tested against data in order to find the best fitting structure, always with respect to its underpinning assumptions. For example, a test hypothesis of biodiversity varying with forest age could take the additive model: Biodiversity = Age + ε, with variation due to Age calibrated against error variation ε. The refutable null hypothesis is: biodiversity = ε.

Population: The entire set of measurable units encompassed by a test hypothesis (e.g., avian biodiversity across all tropical secondary forests in Central America). Study design requires a clear definition of the population, in order to sample representatively from it. The population then defines the scope of inference of the study. Hypothesis-testing statistics are run on samples from a population, not on observations of the entire population.

Power: The probability of a given sampling strategy detecting an effect if it is present in the sampled population at a specified size. Statistical power = $1-\beta$, where β is the probability of making a Type-II error: failure to reject a false null hypothesis, given a true standardized effect size, sampling strategy, test statistic and threshold α of Type-I error. Power analysis provides the means to design studies for precise detection of effects and accurate estimation of their sizes.

***P* value**: The proportion calculated by a frequentist statistic equal to the probability of data at least as deviant as the observed set, given the null hypothesis H_0, and thus the probability of making an error by rejecting H_0. The reliability of the *P* value depends on the power of the study to detect an effect of specified size.

Replication: The number of independent observations randomly sampled from a population of interest that together provide evidence for pattern in the population. No statistics are possible without replication within samples. Small samples will have low power to detect all but large effects. In field studies, large samples may risk violating the assumption of independent observations due to spatial autocorrelation.

Significance: (*i*) The statistical probability of falsely rejecting a null hypothesis (the '*P* value'), in relation to the upper threshold α of acceptable probability in making this Type-I error (often set at 0.05). The relative size of *P* informs an explanation of the population in terms of test and null hypotheses, given the design of data collection. (*ii*) Where *P* is sufficiently small to reject the null hypothesis, parameter estimates from the data inform a description of the impact of effects: their biological significance. For example, forest age influenced species richness ($F_{1,10} = 4.98$; $P < 0.05$), on average adding one additional species with every seven additional years of age.

Statistic: The quantitative measure used to distinguish between competing models. Frequentist statistics make the distinction on the basis of a *P* value; inferences depend on specifying an *a priori* threshold of biological relevance in the size of effect, which determines the detection power of the study. Bayesian statistics quantify relative evidence for the test and null hypotheses, for example in terms of the odds of the data under each; inferences depend on specifying a prior probability distribution of the effect size, for calibrating the posterior distribution given the data.

Treatment: A test factor or variable that is hypothesized to influence a response variable.

Type-I and –II errors: see ‘Power’, ‘*P* value’, and ‘Significance’.

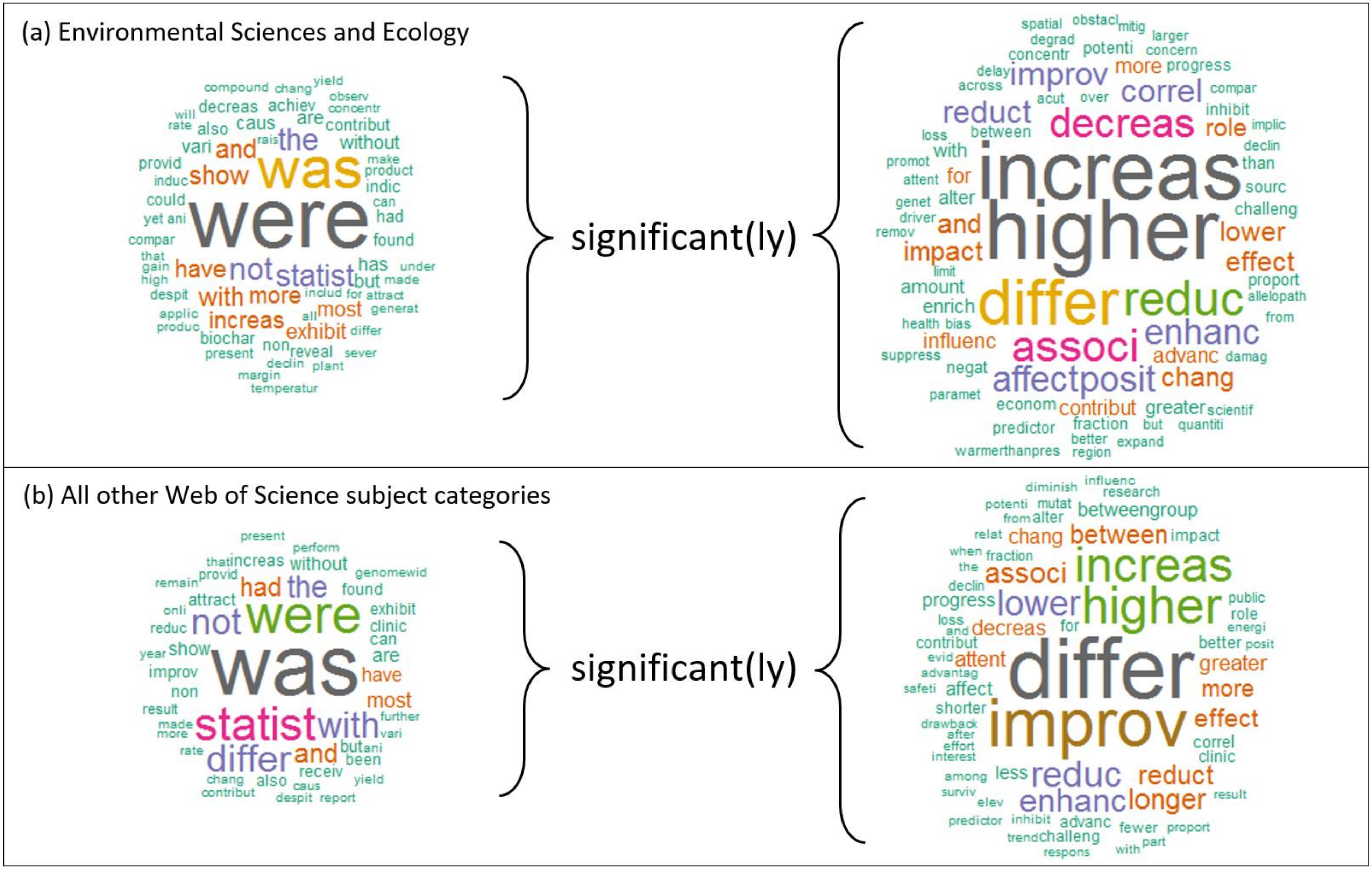
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3394v1 | CC BY 4.0 Open Access | rec: 6 Nov 2017, publ: 6 Nov 2017

Box 1. Did you really mean that?

Abuses of significance abound in the research literature. Figure 1 depicts the most frequent links to the word stem 'significant' in abstracts that used it, from 1000 research papers published in 2016 and listed in the Web of Science Core Collection. Papers are partitioned into the 500 most-cited in each of two thematic areas: (a) Environmental Sciences combined with Ecology, and (b) all other Web of Science categories combined.

The four word clouds illustrate in font size and color the frequencies of all word stems occurring at least twice. Amongst environmental sciences and ecology papers, for example, preceding phrases ('… significant/ ly') took the form 'were significant/ly' 58 times (ranked 1st), and 'achiev/e/ed/es/ing significant/ly' 5 times (20th); while following phrases ('significant/ly …') took the form 'significant/ly higher' 42 times (1st), and 'significant/ly advance/e/ed/es/ing' 6 times (20th).

The word clouds show few conjunctions that refer unambiguously to non-statistical significance (e.g., eight occurrences across all subjects of 'significant/ly challeng/e/ed/es/ing'). Most posit statistically significant results, with a marked preference for bigger or better outcomes.



(a) Environmental Sciences and Ecology

(b) All other Web of Science subject categories

Figure 1. (Ab)uses of 'significant(ly)'. Word stems immediately preceding (left) and following (right) the word stem 'significant' (R scripts by Kassambara, 2017).

Box 2. Benefits of testing for a reliably detectable effect

Ecological studies frequently aim to detect the presence of test effects without anticipating a threshold effect size of biological relevance. The absence of any such *a priori* threshold greatly limits interpretation. A statistically non-significant outcome cannot distinguish between a true absence of effect, and a truly present effect of too-small magnitude for likely detection with the given sampling strategy (Doncaster, Davey & Dixon, 2014). Conversely, a significant outcome cannot distinguish whether the effect size is accurately estimable, or overestimated by an underpowered study (Halsey et al., 2015; Lemoine et al., 2016). We recommend reporting at least the true effect size for which the study has 90% power to reject a false null hypothesis.

Significance testing functions best when effect-size matters. For example, farmers may wish to find out whether a pesticide is cost effective, in terms of raising yield sufficiently to remunerate the cost of its application. Their interest is in the presence of a useful effect. A good experimental study would choose a design with high power, say 90%, to detect a gain in yield if it has breakeven magnitude. Does such a well-powered study then align statistical significance with biological significance? No, it only aligns statistical with biological non-significance. If the test statistic reports $P > 0.05$, the farmers can conclude that the pesticide has no useful effect, within an accepted 10% threshold of error in failing to detect a breakeven effect, for an accepted 5% upper threshold of error in rejecting a true null hypothesis of no effect. If alternatively the test reports $P < 0.05$, the farmers can conclude that the pesticide influences yield, within the accepted 5% threshold of chance of the data being compatible with no effect. The small P value provides no assurance of a breakeven gain in yield. All that the statistics tell us is that any true gain of at least breakeven magnitude will have $P < 0.05$ in at least 90% of tests with this study design, given valid assumptions. Smaller true gains also have a good chance of detection with this design, albeit less than 90%. The effect size needs estimating from the data, to find out whether it indeed exceeds the breakeven gain.

Bibliographical narrative

CPD is a Professor in Ecology at the University of Southampton, whose research covers evolutionary ecology, population and community dynamics, and conservation. He is co-author of ‘Analysis of Variance and Covariance: How to Choose and Construct Models for the Life Sciences’ published by Cambridge University Press in 2007.

THGE is an Associate Professor in Evolutionary Ecology at the University of Southampton, whose research covers evolutionary ecology, particularly the ways in which population and community structure interacts with environmental change to shape ecological and evolutionary dynamics. He addresses these issues by developing the interface of mathematical and statistical models.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3394v1 | CC BY 4.0 Open Access | rec: 6 Nov 2017, publ: 6 Nov 2017