Beyond *p* Values: Interpreting Statistical Results

Geoff Kushnick, School of Archeology and Anthropology, The Australian National University

For many years, Null Hypothesis Significance Testing (NHST) has been the dominant framework for analysing data in biology, psychology, and many other sciences. The familiar *p* value measures how compatible the observed data are with a null hypothesis (typically that there is no effect). By convention, results with $p<0.05$ are often described as "statistically significant."

Despite its widespread use, NHST has attracted considerable criticism. One concern is that *p* values are frequently mis-understood. A statistically significant result does not tell us that a hypothesis is true, nor does it tell us whether an effect is important. Likewise, a non-significant result does not necessarily mean there is no effect. It may simply reflect limited data or insufficient statistical power.

The emphasis on achieving statistical significance has also contributed to questionable research practices (QRPs). These include *p*-hacking, where researchers consciously or unconsciously try multiple analyses, selectively report results, or make analytic decisions that increase the likelihood of obtaining $p<0.05$. Although most researchers do not engage in deliberate misconduct, these practices have been common enough to contribute to concerns about the reproducibility of published research.

There is no universal agreement on how best to address these problems. Some statisticians have argued that significance testing should be abandoned altogether in favour of estimation or Bayesian approaches. Others have suggested retaining NHST but adopting a more stringent significance threshold (for example, $p<0.005$ instead of 0.05). A third perspective is that the problem lies less with the *p* value itself than with how it is used and interpreted. In this view, *p* values remain useful when they are applied appropriately and reported alongside other information.

For this reason, good statistical reporting should rarely rely on a *p* value alone. Effect sizes help us judge the magnitude of an effect, while confidence intervals indicate the precision of our estimates. Together with thoughtful study design, adequate sample sizes, transparency about analytic decisions, and consideration of the broader body of evidence, these approaches provide a much stronger basis for scientific inference than a simple declaration of whether a result is "statistically significant."

Five Ways to Improve Your Statistical Practice:

1. *Report effect sizes, not just p values:* A *p* value tells you whether the data are inconsistent with the null hypothesis, but it does not tell you how large or important an effect is. Always report an appropriate effect size (e.g., Cohen's *d*, Pearson's *r*, odds ratio) alongside significance tests.

2. *Include confidence intervals:* Confidence intervals show the range of values that are reasonably consistent with your data, providing information about the precision and uncertainty of your estimates. They are more informative than a *p* value alone.

3. *Focus on the research question, not just statistical significance:* Ask whether the observed effect is scientifically or practically meaningful. A tiny effect can be statistically significant in a very large sample, while an important effect may fail to reach significance in a small study.

4. *Plan your analyses before looking at the data:* Decide on your hypotheses, variables, and statistical tests in advance whenever possible. This reduces the temptation to try multiple analyses until something becomes significant (often called *p*-hacking) and improves the credibility of your findings.

5. *Interpret results in context:* No single statistic should determine your conclusions. Consider *p* alongside the effect size, confidence interval, sample size, study design, assumptions of the analysis, previous research, and overall evidence quality.

Background Reading List:

Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature*, *567*, 305-307.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management*, *64*(4), 912-923.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z.,…Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6-10.

Betensky, R. A. (2019). The p-Value Requires Context, Not a Threshold. *The American Statistician*, 73(A1), 115-7.

Buyse, M., Pramesh, C., & Ranganathan, P. (2015). Common pitfalls in statistical analysis: Clinical versus statistical significance. *Perspectives in clinical research*, 6(3), 169-170.

Carver, R. P. (1993). The Case Against Statistical Significance Testing, Revisited. *The Journal of Experimental Education*, 61(4), 287-292.

Chavalarias, D., Wallach, J. D., Li, A. H. T., & Ioannidis, J. P. A. (2016). Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. *JAMA*, 315, 1141-8.

Chén, O., Bodelet, J., Saraiva, R. G., Phan, H., Di, J., Nagels, G., Schwantje, T., Cao, H., Gou, J., Reinen, J., Xiong, B., Zhi, B., Wang, X., & de Vos, M. (2023). The roles, challenges, and merits of the p value. *Patterns*, *4*, 100878.

Cohen, J. (1994). The earth is round (p<0.05). *American Psychologist*, 49, 997-1003.

Du Prel, J.-B., Hommel, G., Rohrig, B., & Blettner, M. (2009). Confidence interval or p-value? *Deutsches Arzteblatt International*, *106*(19), 335-339.

Dushoff, J., Kain, M. P., & Bolker, B. M. (2019). I can see clearly now: Reinterpreting statistical significance. *Methods in Ecology and Evolution*, *10*(6), 756-759.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, 1(4), 379–390.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460.

Gelman, A., & Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, *60*(4), 328-331.

Gigerenzer, G. (2004). Mindless statistics. The Journal of Socio-Economics, 33(5), 587-606.

Goodman, S. N. (2016). Aligning statistical and scientific reasoning. *Science*, *352*(6290), 1180-1181.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Europ J Epidem*, *31*(4), 337-350.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, 13(3), e1002106.

Hurlbert, S. H., Levine, R. A., & Utts, J. (2019). Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires. *The American Statistician*, 73(S1), 352-357.

Ilakovac, V. (2009). Statistical hypothesis testing and some pitfalls. *Biochemia Medica*, *19*(1), 10-16.

Imbens, G. W. (2021). Statistical Significance, p-Values, and the Reporting of Uncertainty. *Journal of Economic Perspectives*, 35(3), 157-74.

Ioannidis, J. P. A. (2018). The Proposal to Lower P Value Thresholds to .005. *JAMA*, 319(14), 1429-1430.

Lakens, D. (2021). The Practical Alternative to the p Value Is the Correctly Used p Value. *Persp Psych Sci*, 16(3), 639-648.

Lakens, D. (2022). Why P values are not measures of evidence. *Trends In Ecology & Evolution*, 37(4), 289-290.

Lalongo, C. (2016). Understanding the effect size and its measures. *Biochemia Medica*, 26(2), 150-163.

Leek, J., McShane, B. B., Gelman, A., Colquhoun, D., Nuijten, M. B., & Goodman, S. N. (2017). Five ways to fix statistics. *Nature*, 551, 557-9.

Lovell, D. P. (2013). Biological Importance and Statistical Significance. *Journal of Agricultural and Food Chemistry*, 61(35), 8340-8.

McCloskey, D. N., & Ziliak, S. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press.

McShane, B., Gal, D., Gelman, A., Robert, C., & Tackett, J. (2019). Abandon Statistical Significance. *The American Statistician*, 73(S1), 235-5.

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

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5(2), 241-301.

Nuzzo, R. (2014). Statistical errors: p values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature*, 506, 150-2.

Panagiotakos, D. B. (2008). The value of p-value in biomedical research. *Open Cardiov Med J*, 2, 97-99.

Popovic, G., Mason, T. J., Drobniak, S. M., Marques, T. A., Potts, J., Joo, R., Altwegg, R., Burns, C. C. I., McCarthy, M. A., Johnston, A., Nakagawa, S.,…Pottier, P. (2024). Four principles for improved statistical ecology. *Methods in Ecology and Evolution*, 15(2), 266-281.

Rice, W., & Gaines, S. (1994). 'Heads I win, tails you lose': testing directional alternative hypotheses in ecological and evolutionary research. *Trends Ecol Evol*, 9, 235-7.

Schmidt, M., & Rothman, K. J. (2014). Mistaken inference caused by reliance on and misinterpretation of a significance test. *Int J Cardiol*, 177(3), 1089-1090.

Sedgwick, P. M., Hammer, A., Kesmodel, U. S., & Pedersen, Lars H. (2022). Current controversies: Null hypothesis significance testing. *Acta Obstetricia et Gynecologica Scandinavica*, 101(6), 624-27.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2), 534–547.

Simonsohn, U., Nelson, L., & Simmons, J. (2014). p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. *Perspect Psychol Sci*, 9, 666-81.

Stephens, P., Buskirk, S., & del Rio, C. (2007). Inference in ecology and evolution. *Trends Ecol Evol*, 22, 192-7.

Sterne, J. A. C., Cox, D. R., & Smith, G. D. (2001). Sifting the evidence—what's wrong with significance tests? Another comment on the role of statistical method. *British Medical Journal*, 322(7280), 226-231.

Valentine, J. C., Aloe, A. M., & Lau, T. S. (2015). Life After NHST: How to Describe Your Data Without "p-ing" Everywhere. *Basic Appl Soc Psych*, 37(5), 260-73.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "p < 0.05". *The American Statistician*, 73(S1), 1-19.

Geoff Kushnick is a Senior Lecturer in Biological Anthropology (Human Behavioural Ecology) and a Statistical Person of Knowledge and Experience (SPOKE) in the Statistical Support Network at The Australian National University.

Contact: geoff.kushnick@anu.edu.au