EDITORIAL

Evolution in statistics: \( P \) values, statistical significance, kayaks, and walking trees

Douglas Curran-Everett\(^1\,2\)

\(^1\)Division of Biostatistics and Bioinformatics, National Jewish Health, Denver, Colorado; and \(^2\)Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, Denver, Colorado

Submitted 19 March 2020; accepted in final form 7 April 2020

The scientific literature is littered with the adjective significant and the descriptive phrase statistically significant. Established members of the statistical community now recommend that a scientific paper simply report an actual \( P \) value divorced from any word or phrase that reflects statistical significance (Hurlbert SH, Levine RA, Utts J. Am Stat 73, Suppl 1: 352–357, 2019). In this EDITORIAL, I illustrate why this deceptively simple change is important.

A Brief History of Hypothesis Tests, \( P \) Values, and Significance

Early hypothesis tests, from the Trial of the Pyx in 1279 through the assessment of a discrepant celestial measurement in the 1700s, mandated a binary outcome: the measurement of weight or position either was or was not within some allowable deviation (see Ref. 6).

Between the 1800s and the early 1900s, the focus of a hypothesis test shifted from whether the measurement of a coin or a star was within an allowable deviation to whether some event in mathematics or science could be attributed to chance alone, but the outcome of that hypothesis test remained a binary one (see Ref. 6). If some event—if some difference—was unlikely to have resulted from chance alone, then Edge-worth described that difference as significant, very significant, or as significant and not accidental (15). Thirty-five years later, Boring (2) warned about interpreting a significant difference in the absence of scientific context.

In his pivotal Statistical Methods for Research Workers (16), Sir Ronald Fisher also used significant when he discussed the magnitude of a deviation he regarded as beyond chance, and he defined 0.05 as the benchmark for when he considered some deviation as significant:

The value for which \( P = .05 \), or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant.

In the intervening 100 yr, Fisher’s significance level of 0.05 assumed mystic proportions despite subsequent but less visible elaborations by Fisher himself:

If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent. point), or one in a hundred (the 1 per cent. point). Personally, the writer prefers to set a low standard of significance at the 5 per cent. point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.

[Ref. 17 (1926)]

The attempts that have been made to explain the cogency of tests of significance in scientific research . . . seem to miss the essential nature of such tests. A [person] who “rejects” a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. For when the hypothesis is correct he will be mistaken in just 1% of these cases, and when it is incorrect he will never be mistaken in rejection. . . . However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. Further, the calculation is based solely on a hypothesis, which, in the light of the evidence, is often not believed to be true at all, so that the actual probability of erroneous decision, supposing such a phrase to have any meaning, may be much less than the frequency specifying the level of significance.

[Ref. 18 (1956)]

Recent suggestions that the critical significance level \( \alpha \)—the benchmark for how statistically unusual a result must be before we reject the corresponding null hypothesis—be set at 0.005 or 0.001 (23, 25) minimize the chance that we get a false positive and improve reproducibility (see Ref. 10), but a lower, more stringent significance level \( \alpha \), by itself, fails to address the binary nature of the benchmark (1).

Suppose we define beforehand \( \alpha = 0.05 \) and then obtain, using actual data, \( P = 0.051 \). Is our scientific conclusion going to differ from our conclusion had \( P = 0.049 \)? I hope not. The same logic applies had we defined \( \alpha = 0.001 \): does 0.0011 truly differ from 0.0009? No.

One strategy that circumvents this problem with a binary benchmark\(^1\) is to report the actual \( P \) value rather than simply \( P < 0.05 \) or \( P > 0.05 \) (see Refs. 12 and 14). This strategy, however, has met with limited success in journals published by the American Physiological Society (APS): guidelines for

---

\(^1\) In contrast to the limitations of a binary benchmark in statistics, a binary decision in medicine is typically essential: a surgeon will either operate or not, or an oncologist will either prescribe chemotherapy or not. There may be uncertainty associated with the decision—will there be a good result?—but a clinical decision must be made.
The Vagaries of P Values

A thought experiment I have posed before (see Refs. 9 and 10), suppose we want to learn if some intervention affects the biological thing we care about. If we use two groups—a control group and an experimental group—we might ask if our samples came from the same or different populations. Therefore, the null and alternative hypotheses, \( H_0 \) and \( H_1 \), are:

\[
\begin{align*}
H_0 &: \text{The samples come from the same population.} \\
H_1 &: \text{The samples come from different populations.}
\end{align*}
\]

If we want to know whether the populations have the same mean, we can write these as

\[
H_0 : \Delta \mu = 0 \\
H_1 : \Delta \mu \neq 0,
\]

where \( \Delta \mu \), the difference in population means, is the difference between the means of the experimental and control populations.

We know from previous simulations in which the null hypothesis \( H_0 : \Delta \mu = 0 \) is true (see Refs. 9 and 10) that the observed \( P \) values that result from the test of this null hypothesis are distributed over a wide range of values (Fig. 1): only 100α% of them will be smaller than the critical significance level \( \alpha \).

What may be less obvious is that the observed \( P \) values from the simulated test of a false null hypothesis (Fig. 2) are distributed also over a wide range of values (9, 10, 21); see Fig. 3 and Table 2. In this situation, it is the power of the statistical test (see Ref. 7) that determines the proportion of observed \( P \) values that will be smaller—and larger—than the critical significance level \( \alpha \).

The dichotomy between statistical significance and scientific importance can be illustrated by taking progressively larger numbers of observations from each population in Fig. 2 (Table 3): statistical significance increases (\( P \) decreases), but the scientific importance as estimated by \( \Delta \bar{y} \), the difference between sample means, remains constant.

Evolving Beyond Statistical Significance

In 2016 the American Statistical Association (ASA) issued a position statement that discussed \( P \) values and the notion of...
Fig. 3. The distributions of observed $P$ values from 100,000 replications of simulations in which the null hypothesis, $H_0: \Delta \mu = 0$, is false. We drew at random two samples, each with 10 (top) or 23 (bottom) observations, from the normal distributions in Fig. 2, did a two-sample $t$ test (using a critical significance level $\alpha = 0.05$), and then repeated this process to generate a total of 100,000 replications. The proportion of replications in which $P$ was less than 0.001, 0.01, 0.05, or 0.05 is listed in the upper portion of each graph.

Table 2. Percentiles of observed $P$ values when the null hypothesis is false

| $n$ | Power | Percentile |
|-----|-------|------------|
| 10  | 0.56  | 0.0002, 0.0007, 0.04, 0.05, 0.13, 0.37, 0.72 |
| 23  | 0.91  | <0.0001, 0.0001, 0.001, 0.002, 0.01, 0.05, 0.16 |

Values are $n$, the no. of observations drawn from each population in Fig. 2; theoretical power of the 2-sample $t$ test used to evaluate the null hypothesis $H_0: \Delta \mu = 0$; and percentiles of the distributions of observed $P$ values depicted in Fig. 3. When power is 0.91, 9% of the observed $P$ values are greater than $\alpha = 0.05$.

Table 3. Limitations of statistical significance

| $n$ | $\Delta \mu$ | $SE \{\Delta \mu\}$ | $df$ | $t$ | $P$ | 95% CI | CI Width |
|-----|-------------|----------------|-----|-----|-----|--------|----------|
| 2   | 1           | 0.531          | 2   | 1.883 | 0.20 | -1.28 to +3.28 | 4.56 |
| 4   | 1           | 0.294          | 6   | 3.399 | 0.01 | 0.28 to 1.72   | 1.44 |
| 8   | 1           | 0.528          | 14  | 1.894 | 0.08 | -0.13 to +2.13 | 2.26 |
| 10  | 1           | 0.364          | 18  | 2.745 | 0.01 | 0.23 to 1.77   | 1.54 |
| 15  | 1           | 0.384          | 28  | 2.604 | 0.01 | 0.21 to 1.79   | 1.58 |
| 20  | 1           | 0.396          | 38  | 2.523 | 0.02 | 0.20 to 1.80   | 1.60 |
| 23  | 1           | 0.282          | 44  | 3.550 | 0.0009 | 0.43 to 1.57 | 1.14 |
| 25  | 1           | 0.239          | 48  | 4.180 | 0.0001 | 0.52 to 1.48 | 0.96 |
| 32  | 1           | 0.220          | 62  | 4.543 | 0.00003 | 0.56 to 1.44 | 0.88 |

Values are $n$, the no. of observations drawn from each population in Fig. 2; $\Delta \mu$, the difference between sample means; $SE \{\Delta \mu\}$, the standard error of the difference between sample means; $df$, degrees of freedom; $t$, the $t$ test statistic used to evaluate the null hypothesis, $H_0: \Delta \mu = 0$; $P$, the probability associated with $t$ and the corresponding $df$; 95% CI, 95% confidence interval for $\Delta \mu$, the difference between population means; CI Width, width of the 95% CI for $\Delta \mu$. Values from 100,000 replications of $P$. The difference $\Delta \mu$ and the corresponding 95% CI for $\Delta \mu$ remains constant, and precision of the scientific impact as estimated by the width of the confidence interval increases. See Refs. 8 and 14 for additional detail about the 2-sample $t$ test used in this simulation. [Adapted from Ref. 14; the sampling process is depicted in Fig. 5 of Ref. 14.]

A label of statistical significance adds nothing to what is already conveyed by the value of $p$.

This deceptively simple change affords benefits beyond what you might imagine. Dropping the word significant or the phrase statistically significant prevents a reflexive association of scientific importance with a mere statistical result. This is helpful for two reasons. First, it is quite possible to have a statistically convincing change that is of little or no scientific relevance (see Ref. 5). And second, a binary distinction between not significant and statistically significant can be associated with a trivial difference in the magnitude of the underlying estimate of some effect (20).

To appreciate how straightforward this change is to implement, consider this portion of the abstract from a March 2020 APS Select paper (28):

Male, but not female, collecting duct Bmal1 knockout (CDBmal1KO) mice had significantly lower 24-h mean arterial pressure (MAP) than flox controls (105 ± 2 vs. 112 ± 3 mmHg for male mice and 106 ± 1 vs. 108 ± 1 mmHg for female mice, by telemetry). After 6 days on a high-salt (4% NaCl) diet, MAP remained significantly lower in male CDBmal1KO mice than in male flox control mice (107 ± 2 vs. 113 ± 1 mmHg), with no significant differences between genotypes in female mice (108 ± 2 vs. 109 ± 1 mmHg). . . . However, MAP remained lower in male CDBmal1KO mice than in male flox control mice (124 ± 2 vs. 130 ± 2 mmHg).

This is how that portion of the abstract could have been written to align with the recommendation of Refs. 22 and 26:

---

In this revision I have deleted the standard error associated with each mean pressure (11, 12). Because Ref. 28 did not report actual $P$ values, I have used $P = 0.00$ as a placeholder.

---

In this revision, I have deleted the standard error associated with each mean pressure (11, 12). Because Ref. 28 did not report actual $P$ values, I have used $P = 0.00$ as a placeholder.
Male, but not female, collecting duct 
Bmal1 knockout (CDBmal1KO) mice had lower 24-h mean arterial pressure (MAP) than flox controls [105 vs. 112 mmHg for male mice ($P = 0.00$) and 106 vs. 108 mmHg for female mice ($P = 0.00$), by telemetry]. After 6 days on a high-salt (4% NaCl) diet, MAP remained lower in male CDBmal1KO mice than in male flox control mice [107 vs. 113 mmHg ($P = 0.00$)], with no differences between genotypes in female mice [108 vs. 109 mmHg ($P = 0.00$)]. . . However, MAP remained lower in male CDBmal1KO mice than in male flox control mice [124 vs. 130 mmHg ($P = 0.00$)].

Reference 22 gives more examples.

In April 2019, Hurlbert, Levine, and Utts, the authors of Ref. 22, contacted the Editors-in-Chief of the journals published by the APS. In their e-mail they encouraged the Editors to purge the phrase statistically significant from the APS journals’ future papers, and they argued that APS could spearhead this reform. The APS Publications Committee tabled discussion of this recommendation.

In 1991 Ralph Fletcher wrote Walking Trees, a memoir of his experiences helping teachers in New York City schools learn how to teach writing (19). The title stems from a story Heath, a first-grader, wrote about a family trip to Florida:

See, me and my mommy and daddy went to Florida, and we saw the walking trees down there. They walk with their roots.

For Fletcher, Heather’s trees became a metaphor for the terribly glacial rate of change in education and the Herculean effort required to make even the smallest progress. In my 2017 EDITORIAL (11), I wrote that trying to change the reporting practices of statistics was like trying to change the direction of an ocean liner with a kayak. Whether the metaphor is kayaks or walking trees, so it is with the use of statistics within science.

In my 2017 EDITORIAL, I also announced that I had asked the Associate Editors and Editorial Board of Advances to actively promote two of the 2004 guidelines for reporting statistics (12). I wrote that I understood it was difficult to change entrenched practices and that I understood change was slow, but that did not mean we should not try. When the mainstream statistical community—virtually en masse—recommends a simple, specific course of action—report an actual $P$ value without phrases that reflect statistical significance—the scientific community would do well to heed that recommendation.

ACKNOWLEDGMENTS

I thank Ronald Wasserstein (Executive Director, American Statistical Association) for his time and for his helpful comments and suggestions.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author.

AUTHOR CONTRIBUTIONS

D.C.-E. analyzed data; prepared figures; drafted manuscript; edited and revised manuscript; approved final version of manuscript.

REFERENCES

1. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 567: 305–307, 2019. doi:10.1038/d41586-019-00857-9.

2. Boring EG. Mathematical vs. scientific significance. Psychol Bull 16: 335–338, 1919. doi:10.1037/h0074554.

3. Cox DR. Statistical significance tests. Br J Clin Pharmacol 14: 325–331, 1982. doi:10.1111/j.1365-2125.1982.tb01987.x.

4. Curran-Everett D. Explorations in statistics: standard deviations and standard errors. Adv Physiol Educ 32: 203–208, 2008. doi:10.1152/advan.90123.2008.

5. Curran-Everett D. Explorations in statistics: confidence intervals. Adv Physiol Educ 33: 87–90, 2009. doi:10.1152/advan.00006.2009.

6. Curran-Everett D. Explorations in statistics: hypothesis tests and $P$ values. Adv Physiol Educ 33: 81–86, 2009. doi:10.1152/advan.90218.2008.

7. Curran-Everett D. Explorations in statistics: power. Adv Physiol Educ 34: 41–43, 2010. doi:10.1152/advan.00010.2010.

8. Curran-Everett D. Explorations in statistics: permutation methods. Physiol Educ 56: 181–187, 2012. doi:10.1152/advan.00072.2012.

9. Curran-Everett D. Explorations in statistics: statistical facets of reproducibility. Adv Physiol Educ 40: 248–252, 2016. doi:10.1152/advan.00042.2016.

10. Curran-Everett D. CORP: Minimizing the chances of false positives and false negatives. J Appl Physiol (1985) 122: 91–95, 2017. doi:10.1152/japplphysiol.00397.2016.

11. Curran-Everett D. Small steps to help improve the caliber of the reporting of statistics. Adv Physiol Educ 41: 321–323, 2017. doi:10.1152/advan.00049.2017.

12. Curran-Everett D, Benos DJ. Guidelines for reporting statistics in journals published by the American Physiological Society. Adv Physiol Educ 28: 85–87, 2004. doi:10.1152/advan.2004.19.2004.

13. Curran-Everett D, Benos DJ. Guidelines for reporting statistics in journals published by the American Physiological Society; the sequel. Adv Physiol Educ 31: 295–298, 2007. doi:10.1152/advan.2007.2007.

14. Curran-Everett D, Taylor S, Kafadar K. Fundamental concepts in statistics: elucidation and illustration. J Appl Physiol (1985) 85: 775–786, 1998. doi:10.1152/jappl.1998.85.3.775.

15. Edgeworth FY. Methods of statistics. J Stat Soc London Jubilee: 181–217, 1895.

16. Fisher RA. Statistical Methods for Research Workers. London: Oliver and Boyd, 1925.

17. Fisher RA. The arrangement of field experiments. J Minist Agric GB 33: 503–513, 1926.

18. Fisher RA. Statistical Methods and Scientific Inference. London: Oliver and Boyd/Longman Group, 1966.

19. Fletcher R. Walking Trees. Portsmouth, NH: Heinemann, 1991, p. 202.

20. Gelman A, Stern H. The difference between “significant” and “not significant” is not itself statistically significant. Am Stat 60: 328–331, 2006. doi:10.1198/000313060X152649.

21. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle $P$ value generates irreproducible results. Nat Methods 12: 179–185, 2015. doi:10.1038/nmeth.3288.

22. Hurlbert SH, Levine RA, Utts J. Coup de grâce for a tough old bull: “statistically significant” expires. Am Stat 73, Suppl 1: 352–357, 2019. doi:10.1080/00031305.2018.1543616.

23. Johnson VE. Revised standards for statistical evidence. Proc Natl Acad Sci USA 110: 19313–19317, 2013. doi:10.1073/pnas.1313476110.

24. Snedecor GW, Cochran WG. Statistical Methods (6th ed.). Ames, IA: Iowa State Univ. Press, 1967.

25. Sterne JAC, Davey Smith G. Sifting the evidence—what’s wrong with significance tests? BMJ 322: 226–231, 2001. doi:10.1136/bmj.322.7290.226.

26. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond “$p < 0.05$”. Am Stat 73, Suppl 1: 1–19, 2019. doi:10.1080/00031305.2019.1583913.

27. Wasserstein RL, Lazar NA. The ASA’s statement on $p$-values: context, process, and purpose. Am Stat 70: 129–133, 2016. doi:10.1080/00031305.2016.1154108.

28. Zhang D, Jin C, Obi E, Rhoads MK, Soliman RH, Sedaka RS, Allan JM, Tao B, Speed JS, Pollock JS, Pollock DM. Loss of circadian gene Bmal1 in the collecting duct lowers blood pressure in male, but not female, mice. Am J Physiol Renal Physiol 318: F710–F719, 2020. doi:10.1152/ajprenal.00364.2019.