ESSAY

Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

Sander Greenland[1] · Stephen J. Senn[2] · Kenneth J. Rothman[3] · John B. Carlin[4] ·

Keywords Confidence intervals · Hypothesis testing · Null testing · P value · Power · Significance tests · Statistical testing

Introduction

effect sizes. We may also test hypotheses that the effect does or does not fall within a specific range; for example, we may test the hypothesis that the effect is no greater than a particular amount, in which case the hypothesis is said to be a one-sided or dividing hypothesis [7, 8].

Much statistical teaching and practice has developed a strong (and unhealthy) focus on the idea that the main aim of a study should be to test null hypotheses. In fact most descriptions of statistical testing focus only on testing null hypotheses, and the entire topic has been called "Null Hypothesis Significance Testing" (NHST). This exclusive focus on null hypotheses contributes to misunderstanding of tests. Adding to the misunderstanding is that many authors (including R.A. Fisher) use "null hypothesis" to refer to any test hypothesis, even though this usage is at odds with other authors and with ordinary English definitions of "null"—as are statistical usages of "significance" and "confidence."

Uncertainty, probability, and statistical significance

A more refined goal of statistical analysis is to provide an evaluation of certainty or uncertainty regarding the size of an effect. It is natural to express such certainty in terms of "probabilities" of hypotheses. In conventional statistical methods, however, "probability" refers not to hypotheses, but to quantities that are hypothetical frequencies of data patterns under an assumed statistical model. These methods are thus called frequentist methods, and the hypothetical frequencies they predict are called "frequency probabilities." Despite considerable training to the contrary, many statistically educated scientists revert to the habit of misinterpreting these frequency probabilities as hypothesis probabilities. (Even more confusingly, the term "likelihood of a parameter value" is reserved by statisticians to refer to the probability of the observed data given the parameter value; it does not refer to a probability of the parameter taking on the given value.)

Nowhere are these problems more rampant than in applications of a hypothetical frequency called the P value, also known as the "observed significance level" for the test hypothesis. Statistical "significance tests" based on this concept have been a central part of statistical analyses for centuries [75]. The focus of traditional definitions of P values and statistical significance has been on null hypotheses, treating all other assumptions used to compute the P value as if they were known to be correct. Recognizing that these other assumptions are often questionable if not unwarranted, we will adopt a more general view of the P value as a statistical summary of the compatibility between the observed data 847('Nuwee)-10.89999961(o1.2235hs)-32owoul-10.89provideifted stor tes29(wfor)-250.se00122(h32)371(o

P value is a number computed from the data and thus an analysis result, unknown until it is computed.

Moving from tests to estimates

We can vary the test hypothesis while leaving other assumptions unchanged, to see how the P value differs across competing test hypotheses. Usually, these test hypotheses specify different sizes for a targeted effect; for example, we may test the hypothesis that the average difference between two treatment groups is zero (the null hypothesis), or that it is 20 or -10 or any size of interest. The effect size whose test produced $P = 1$ is the size most compatible with the data (in the sense of predicting what was in fact observed) if all the other assumptions used in the test (the statistical model) were correct, and provides a point estimate of the effect under those assumptions. The effect sizes whose test produced $P > 0.05$ will typically define a range of sizes (e.g., from 11.0 to 19.5) that would be considered more compatible with the data (in the sense of the observations being closer to what the model predicted) than sizes outside the range—again, if the statistical model were correct. This range corresponds to a $1 - 0.05 = 0.95$ or 95 % confidence interval, and provides a convenient way of summarizing the results of hypothesis tests for many effect sizes. Confidence intervals are examples of interval estimates.

Neyman [76] proposed the construction of confidence intervals in this way because they have the following property: If one calculates, say, 95 % confidence intervals repeatedly in valid applications, 95 % of them, on average, will contain (i.e., include or cover) the true effect size. Hence, the specified confidence level is called the coverage

3. A significant test result ($P \leq 0.05$) means that the test hypothesis is false or should be rejected. No! A small P value simply flags the data as being unusual if all the assumptions used to compute it (including the test hypothesis) were correct; it may be small because there was a large random error or because some assumption other than the test hypothesis was violated (for example, the assumption that this P value was not selected for presentation because it was below 0.05). P 0.05 only means that a discrepancy from the hypothesis prediction (e.g., no difference between treatment groups) would be as large or larger than that observed no more than 5 % of the time if only chance were creating the discrepancy (as opposed to a violation of the test hypothesis or a mistaken assumption).
4. A nonsignificant test result ($P > 0.05$) means that the test hypothesis is true or should be accepted. No! A large P value only suggests that the data are not unusual if all the assumptions used to compute the P value (including the test hypothesis) were correct. The same data would also not be unusual under many other hypotheses. Furthermore, even if the test hypothesis is wrong, the P value may be large because it was inflated by a large random error or because of some other erroneous assumption (for example, the assumption that this P value was not

hypothesis, these assumptions include randomness in sampling, treatment assignment, loss, and missingness, as well as an assumption that the P value was not selected for presentation based on its size or some other aspect of the results.

10. If you reject the test hypothesis because $P \leq 0.05$, the chance you are in error (the chance your "significant finding" is a false positive) is 5 %. No! To see why this description is false, suppose the test hypothesis is in fact true. Then, if you reject it, the chance you are in error is 100 %, not 5 %. The 5 % refers only to how often you would reject it, and therefore be in error, over very many uses of the test across different studies when the test hypothesis and all other assumptions used for the test are true. It does not refer to your single use of the test, which may have been thrown off by assumption violations as well as random errors. This is yet another version of misinterpretation #1.
11. $P = 0.05$ and $P \leq 0.05$ mean the same thing. No! This is like saying reported height $= 2$ m and reported height 2 m are the same thing: "height $= 2$ m" would include few people and those people would be considered tall, whereas "height 2 m" would include most people including small children. Similarly, $P = 0.05$ would be considered a borderline result in terms of statistical significance, whereas $P \quad 0.05$ lumps borderline results together with results very incompatible with the model (e.g., $P = 0.0001$) thus rendering its meaning vague, for no good purpose.
12. *P* values are properly reported as inequalities (e.g., report "$P < 0.02$" when $P = 0.015$ or report "$P > 0.05$" when $P = 0.06$ or $P = 0.70$). No! This is bad practice because it makes it difficult or impossible for the reader to accurately interpret the statistical result. Only when the P

95 % probability of containing the true value; nonetheless, such computations require not only the assumptions used to compute the confidence interval, but also further assumptions about the size of effects in the model. These further assumptions are summarized in what is called a prior distribution, and the resulting intervals are usually called Bayesian posterior (or credible) intervals to distinguish them from confidence intervals [18].

Symmetrically, the misinterpretation of a small P value as disproving the test hypothesis could be translated into:

20. An effect size outside the 95 % confidence interval has been refuted (or excluded) by the data. No! As with the P value, the confidence interval is computed from many assumptions, the violation of which may have led to the results. Thus it is the combination of the data with the assumptions, along with the arbitrary 95 % criterion, that are needed to declare an effect size outside the interval is in some way incompatible with the observations. Even then, judgements as extreme as saying the effect size has been refuted or excluded will require even stronger conditions.

As with P values, naïve comparison of confidence intervals can be highly misleading:

21. If two confidence intervals overlap, the difference between two estimates or studies is not significant. No! The 95 % confidence intervals from two subgroups or studies may overlap substantially and yet the test for difference between them may still produce $P < 0.05$. Suppose for example, two 95 % confidence intervals for means from normal populations with known variances are (1.04, 4.96) and (4.16, 19.84); these intervals overlap, yet the test of the hypothesis of no difference in effect across studies gives $P = 0.03$. As with P values, comparison between groups requires statistics that directly test and estimate the differences across groups. It can, however, be noted that if the two 95 % confidence intervals fail to overlap, then when using the same assumptions used to compute the confidence intervals we will find $P < 0.05$ for the difference; and if one of the 95 % intervals contains the point estimate from the other group or study, we will find $P > 0.05$ for the difference.

Finally, as with P values, the replication properties of confidence intervals are usually misunderstood:

22. An observed 95 % confidence interval predicts that 95 % of the estimates from future studies will fall inside the observed interval. No! This statement is wrong in several ways. Most importantly, under the model, 95 % is the frequency with which other unobserved intervals will contain the true effect, not how frequently the one interval being presented will contain future estimates. In fact, even under ideal conditions the chance that a future estimate will fall within the current interval will usually be much less than 95 %. For example, if two independent studies of the same quantity provide unbiased normal point estimates with the same standard errors, the chance that the 95 % confidence interval for the first study contains the point estimate from the second is 83 % (which is the chance that the difference between the two estimates is less than 1.96 standard errors). Again, an observed interval either does or does not contain the true effect; the 95 % refers only to how often 95 % confidence intervals computed from very many studies would contain the true effect if all the assumptions used to compute the intervals were correct.
23. If one 95 % confidence interval includes the null value and another excludes that value, the interval excluding the null is the more precise one. No! When the model is correct, precision of statistical estimation is measured directly by confidence interval width (measured on the appropriate scale). It is not a matter of inclusion or exclusion of the null or any other value. Consider two 95 % confidence intervals for a difference in means, one with limits of 5 and 40, t5(co)19996(n9

hypotheses with data and wish to compare hypotheses with this measure, we need to examine their P values directly, not simply ask whether the hypotheses are inside or outside the interval. This need is particularly acute when (as usual) one of the hypotheses under scrutiny is a null hypothesis.

Common misinterpretations of power

The power of a test to detect a correct alternative hypothesis is the pre-study probability that the test will reject the test hypothesis (e.g., the probability that P will

represented by parameters denoted by Greek letters. “Model

hypothesis probabilities. For example, under common statistical models, one-sided P values can provide lower bounds on probabilities for hypotheses about effect directions [45, 46, 112, 113]. Whether such reinterpretations can eventually replace common misinterpretations to good effect remains to be seen.

A shift in emphasis from hypothesis testing to estimation has been promoted as a simple and relatively safe way to improve practice [5, 61, 63, 114, 115] resulting in increasing use of confidence intervals and editorial demands for them; nonetheless, this shift has brought to the fore misinterpretations of intervals such as 19–23 above [116]. Other approaches combine tests of the null with further calculations involving both null and alternative hypotheses [117, 118]; such calculations may, however, may bring with them further misinterpretations similar to those described above for power, as well as greater complexity.

Meanwhile, in the hopes of minimizing harms of current practice, we can offer several guidelines for users and readers of statistics, and re-emphasize some key warnings from our list of misinterpretations:

(a) Correct and careful interpretation of statistical tests demands examining the sizes of effect estimates and confidence limits, as well as precise P values (not just whether P values are above or below 0.05 or some other threshold).

(b) Careful interpretation also demands critical examination of the assumptions and conventions used for the statistical analysis—not just the usual statistical assumptions, but also the hidden assumptions about how results were generated and chosen for presentation.

(c) It is simply false to claim that statistically nonsignificant results support a test hypothesis, because the same results may be even more compatible with alternative hypotheses—even if the power of the test is high for those alternatives.

(d) Interval estimates aid in evaluating whether the data are capable of discriminating among various hypotheses about effect sizes, or whether statistical results have been misrepresented as supporting one hypothesis when those results are better explained by

43. Greenland S. Transparency and disclosure, neutrality and balance: shared values or just shared words? J Epidemiol Community Health. 2012;66:967–70.
44. Greenland S, Poole C. Problems in common interpretations of statistics in scientific articles, expert reports, and testimony. Jurimetrics. 2011;51:113–29.
45. Greenland S, Poole C. Living with P-values: resurrecting a Bayesian perspective on frequentist statistics. Epidemiology. 2013;24:62–8.
46. Greenland S, Poole C. Living with statistics in observational research. Epidemiology. 2013;24:73–8.
47. Grieve AP. How to test hypotheses if you must. Pharm Stat. 2015;14:139–50.
48. Hoekstra R, Finch S, Kiers HAL, Johnson A. Probability as certainty: dichotomous thinking and the misuse of p-values. Psychon Bull Rev. 2006;13:1033–7.
49. Hurlbert Lombardi CM. Final collapse of the Neyman–Pearson decision theoretic framework and rise of the neoFisherian. Ann Zool Fenn. 2009;46:311–49.
50. Kaye DH. Is proof of statistical significance relevant? Wash Law Rev. 1986;61:1333–66.
51. Lambdin C. Significance tests as sorcery: science is empirical—significance tests are not. Theory Psychol. 2012;22(1):67–90.
52. Langman MJS. Towards estimation and confidence intervals. BMJ. 1986;292:716.
53. LeCoutre M-P, Poitevineau J, Lecoutre B. Even statisticians are not immune to misinterpretations of null hypothesis tests. Int J Psychol. 2003;38:37–45.
54. Lew MJ. Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know P. Br J Pharmacol. 2012;166:1559–67.
55. Loftus GR. Psychology will be a much better science when we change the way we analyze data. Curr Dir Psychol. 1996;5:161–71.
56. Matthews JNS, Altman DG. Interaction 2: Compare effect sizes not P values. Br Med J. 1996;313:808.
57. Pocock SJ, Ware JH. Translating statistical findings into plain English. Lancet. 2009;373:1926–8.
58. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. N Eng J Med. 1987;317:426–32.
59. Poole C. Beyond the confidence interval. Am J Public Health. 1987;77:195–9.
60. Poole C. Confidence intervals exclude nothing. Am J Public Health. 1987;77:492–3.
61. Poole C. Low P-values or narrow confidence intervals: which are more durable? Epidemiology. 2001;12:291–4.
62. Rosnow RL, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. Am Psychol. 1989;44:1276–84.
63. Rothman KJ. A show of confidence. NEJM. 1978;299:1362–3.
64. Rothman KJ. Significance questing. Ann Intern Med. 1986;105:445–7.
65. Rozeboom WM. The fallacy of null-hypothesis significance test. Psychol Bull. 1960;57:416–28.
66. Salsburg DS. The religion of statistics as practiced in medical journals. Am Stat. 1985;39:220–3.
67. Schmidt FL. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychol Methods. 1996;1:115–29.
68. Schmidt FL, Hunter JE. Methods of meta-analysis: correcting error and bias in research findings. 3rd ed. Thousand Oaks: Sage; 2014.
69. Sterne JAC, Davey Smith G. Sifting the evidence—what's wrong with significance tests? Br Med J. 2001;322:226–31.
70. Thompson WD. Statistical criteria in the interpretation of epidemiologic data. Am J Public Health. 1987;77:191–4.
71. Thompson B. The "significance" crisis in psychology and education. J Soc Econ. 2004;33:607–13.
72. Wagenmakers E-J. A practical solution to the pervasive problem of p values. Psychon Bull Rev. 2007;14:779–804.
73. Walker AM. Reporting the results of epidemiologic studies. Am J Public Health. 1986;76:556–8.
74. Wood J, Freemantle N, King M, Nazareth I. Trap of trends to statistical significance: likelihood of near significant P value becoming more significant with extra data. BMJ. 2014;348:g2215. doi:10.1136/bmj.g2215.
75. Stigler SM. The history of statistics. Cambridge, MA: Belknap Press; 1986.
76. Neyman J. Outline of a theory of statistical estimation based on the classical theory of probability. Philos Trans R Soc Lond A. 1937;236:333–80.
77. Edwards W, Lindman H, Savage LJ. Bayesian statistical inference for psychological research. Psychol Rev. 1963;70:193–242.
78. Berger JO, Sellke TM. Testing a point null hypothesis: the irreconcilability of P-values and evidence. J Am Stat Assoc. 1987;82:112–39.
79. Edwards AWF. Likelihood. 2nd ed. Baltimore: Johns Hopkins University Press; 1992.
80. Goodman SN, Royall R. Evidence and scientific research. Am J Public Health. 1988;78:1568–74.
81. Royall R. Statistical evidence. New York: Chapman and Hall; 1997.
82. Sellke TM, Bayarri MJ, Berger JO. Calibration of p values for testing precise null hypotheses. Am Stat. 2001;55:62–71.
83. Goodman SN. Introduction to Bayesian methods I: measuring the strength of evidence. Clin Trials. 2005;2:282–90.
84. Lehmann EL. Testing statistical hypotheses. 2nd ed. Wiley: New York; 1986.
85. Senn SJ. Two cheers for P-values. J Epidemiol Biostat. 2001;6(2):193–204.
86. Senn SJ. Letter to the Editor re: Goodman 1992. Stat Med. 2002;21:2437–44.
87. Mayo DG, Cox DR. Frequentist statistics as a theory of inductive inference. In: J Rojo, editor. Optimality: the second Erich L. Lehmann symposium, Lecture notes-monograph series, Institute of Mathematical Statistics (IMS). 2006;49: 77–97.
88. Murtaugh PA. In defense of P-values (with discussion). Ecology. 2014;95(3):611–53.
89. Hedges LV, Olkin I. Vote-counting methods in research synthesis. Psychol Bull. 1980;88:359–69.
90. Chalmers TC, Lau J. Changes in clinical trials mandated by the advent of meta-analysis. Stat Med. 1996;15:1263–8.
91. Maheshwari S, Sarraj A, Kramer J, El-Serag HB. Oral contraception and the risk of hepatocellular carcinoma. J Hepatol. 2007;47:506–13.
92. Cox DR. The planning of experiments. New York: Wiley; 1958. p. 161.
93. Smith AH, Bates M. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. Epidemiology. 1992;3:449–52.
94. Goodman SN. Letter to the editor re Smith and Bates. Epidemiology. 1994;5:266–8.
95. Goodman SN, Berlin J. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med. 1994;121:200–6.
96. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat. 2001;55:19–24.
97. Senn SJ. Power is indeed irrelevant in interpreting completed studies. BMJ. 2002;325:1304.
98. Lash TL, Fox MP, Maclehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol. 2014;43:1969–85.

99. Dwan K, Gamble C, Williamson PR, Kirkham JJ, Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. PLoS One. 2013;8:e66844.
100. Page MJ, McKenzie JE, Kirkham J, Dwan K, Kramer S, Green S, Forbes A. Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of healthcare interventions. Cochrane Database Syst Rev. 2014;10:MR000035.
101. You B, Gan HK, Pond G, Chen EX. Consistency in the analysis and reporting of primary end points in oncology randomized