The *p*-value Case, a Review of the Debate: Issues and Plausible Remedies

Francesco Pauli

Abstract We review the recent debate on the lack of reliability of scientific results and its connections to the statistical methodologies at the core of the discovery paradigm. Null hypotheses statistical testing, in particular, has often been related to, if not blamed for, the present situation. We argue that a loose relation exists: although NHST, if properly used, could not be seen as a cause, some common misuses may mask or even favour bad practices leading to the lack of reliability. We discuss various proposals which have been put forward to deal with these issues.

Keywords Null hypotheses statistical testing · *p*-value · Reproducibility

1 Introduction

A discussion on the role of the hypothesis statistical testing method in jeopardizing the reliability of scientific results is underway in the recent literature across many disciplines [18, 39]. It has been argued that a worrying portion of published scientific results, within various disciplines, are actually false discoveries [25]. This state of things has been related to the widespread use—or abuse—of *p*-values to measure evidence and corroborate new theories [7, 16, 34], to the point that a journal in psychology "banned" *p*-values [48] (although not in a very clear-cut way, for instance, they are allowed in submissions [49]). That of banning *p*-values altogether is not a novel idea nor it is exclusive of psychology [42]. According to a recent survey of 1576 researchers made by Nature [1], more than 90% have heard of a 'crisis of reproducibility'. Most of them think that the crisis is in fact in place and has not been overemphasized. Statistics is seen both as part of the problem and as a mean to improve the situation.

It is worth to note that the false discovery rate (FDR) across science (or a discipline) is not a clear-cut concept: a reference population of findings should be identified and a criterion of falsehood defined. In empirical evaluations a (non random)

F. Pauli (✉)
DEAMS, University of Trieste, Trieste, Italy
e-mail: francesco.pauli@deams.units.it

sample of results is usually considered and falsehood is often equated to lack of replication, which is a different, although related, concept. Notwithstanding how difficult or ambiguous it may be to precisely define the notion, however, the error rate of scientific results is a relevant concept of general interest, as the number of attempts which have been made to quantify it reveals.

Null hypotheses statistical testing (NHST) has a central role in the paradigm which is commonly employed to confirm new scientific theories (Sect. 2), and a long-running controversy on its use is in place. Whilst it may or may not be the culprit of the lack of reliability (Sect. 3), it is relevant to discuss whether alternatives to NHST may lead to a more reliable procedure to confirm scientific results (Sect. 4).

2 Scientific Discoveries and Statistical Testing

Null hypotheses statistical testing (NHST) is a standard topic in academic curricula of various disciplines and a standard tool to analyse data in many scientific fields.

Controversies concerning NHST started since the proposal of significance testing (and p-values) by Fisher and the alternative—and incompatible—procedure for hypotheses testing by Neyman and Pearson. Fisher proposed to measure the strength of evidence of a given observation against a hypothesis on the probabilistic mechanism which generated it with the probability, conditional on that hypothesis, of obtaining a sample at least as extreme as the observed one (p-value) [12]. Neyman and Pearson argued that "no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis" and propose instead a procedure to choose between two alternative hypotheses on the data generating mechanism keeping under control the (conditional) probabilities of making the wrong choice [37].

NHST plays a central role in the procedure—or mindless ritual as some scholars provocatively called it [15]—which is used to corroborate scientific theories. The procedure goes as follows: a theory is posited according to which a relationship is in place between two quantities; in order to corroborate the theory a null hypothesis of absence of relationship is statistically tested using a sample; a confirmation is claimed whenever the null hypothesis is rejected at a specified level, which is usually 5%. Instances of the use of such a procedure abound across disciplines, for practical examples see [4] in medicine, [2] in psychology, [41] in economics, [20] in zoology. The exact p-value is generally taken as a measure of the evidence against the null hypothesis and possibly also as a measure of the evidence in favor of the alternative hypothesis. Also, acceptance of the null is often taken as evidence of the absence of the posited effect.

It is commonly maintained that, of the two procedures, the Fisherian p-value is the more apt to the described task, while the Neyman-Pearson procedure is more apt to problems which are more naturally cast in a decision framework. It is also to be noted, however, that the actual interpretation given to NHST in applications is sometimes a combination of the two. In fact the p-value is used to draw the conclusion but

an alternative hypothesis is also considered (which also helps to clarify what an "extreme result" is in the definition of the p-value) and/or an error probability is attached to the p-value based conclusions [3]. In what follows we refer to NHST having this somewhat imprecise interpretation in mind.

This description does not encompass all uses of NHST in the scientific literature, but it represents the most problematic use and is widespread, and debated, across disciplines. In psychology the debate was already not new in 1994 [7, 28] (it dates back to 1955 according to [44]), and is lively as of today. On the other hand, the use of NHST is increasing, due to academic inertia according to some [44]. In medicine these issues are discussed since the rise of evidence based medicine [16] and still [19]. The practice is also widespread in economics/econometrics [35, 52], although some scholars disagree [23] on the extent of the problem.

While it is sometimes argued that NHST is not used in hard sciences like physics [35, 44], this is not the case: NHST has its place in high energy physics [40], in cosmology [9], in atmospheric sciences [38]. In these contexts, however, it is regarded as "only part of discovering a new phenomenon", the actual degree of belief depending on substantial considerations [8]. A p-value (or, more often, the Z-score, which is the $(1-p)$-quantile of the standard normal distribution) is used as a measure of surprise, which suggests further investigation of the alternatives, in particular on whether they better explain observations. A peculiarity of some hard sciences is that different thresholds for rejection are customary: threshold values commonly used are $Z = 5$ ($p = 2.87 \times 10^{-7}$) and $Z = 1.64$ ($p = 0.05$). The first is used for 'discovery', that is when the alternative hypothesis includes a sought signal and the null is a 'background only' hypothesis; the second is used when the null is a signal.

3 NHST (p-Value), Good, Bad or Neutral?

The debate on the reliability of scientific results is intertwined with the debate on the suitability of the p-value as a measure of evidence. We argue that there is a relation between the use of p-values and the reliability crisis, albeit loose. In fact, some misuses of the p-value are susceptible to exacerbate some issues of the discovery paradigm outlined in Sect. 2.

The concerns which have been raised upon p-values can be categorized in three classes: one related to interpretation; one to the relationship with the size of the effect; the latter related to the role of the alternative.

Misinterpretations of the p-value take different forms, which in some cases are equivalent. The more trivial, yet common, misinterpretation is to relate the p-value to the probability of the null being true. This amounts at wishful thinking, since the probability of the null is what the researcher actually wants. It is barely worth mentioning that such an interpretation is logically wrong (as the p-value is a probability conditional on the null being true) and potentially strongly misleading, as a given p-value is compatible with any value for the probability of the null being true.

Another common misinterpretation (seen even in "serious use of statistics" [3]) is that the p-value is the probability of wrongly rejecting the null (or the probability of the result being due to chance [19]). That is, the (Fisherian) p-value, which is conditional to the sample, is mistaken for the (Neyman-Pearson) significance level, which is a long run error probability. The coexistence of the two approaches, whose logical incompatibility is often under-appreciated by non-statisticians users of statistics, is probably to be blamed for this [3].

A second class of issues arises from the fact that the p-value is a function of both the estimate of the effect size and the size of the sample; a low p-value, indicating a statistically significant effect, is not necessarily associated to a substantially significant effect, and vice versa. In spite of this, in many fields it is common practice to choose models—for example selecting the covariates in a regression analysis—based on the significance of coefficients, that is, only those coefficients which are significantly different from zero at a specified level are reported, implicitly assuming that the others are zero. Within econometrics this practice has been labeled "sign econometrics", interpreting the sign of a significant coefficient regardless of its size, and "star econometrics", ranking importance of variables according to their significance level ignoring their relative sizes [35, 52]. It is contended that the key question in scientific inquiry, establishing "How large is large" (to be substantively relevant), can not be answered by p-values [21, 30, 35, 52]. It has also been said that the p-value alone may be only a measure of how large is the sample, since in many settings the null hypothesis is a nil hypothesis (of absence of any effect) and this is (almost) surely false due to what Meehl [36] calls the *crud factor*—the fact that in many situations the effect is not precisely zero, but the actual scientific hypothesis of interest is the effect being so low to be irrelevant rather than it being exactly zero.

An obvious solution is to complement the information given by the p-value with the estimate of the effect: if the p-value leads to rejection, the estimate is reported. It has been noted, however, that coupling NHST and the size estimate is an issue in certain circumstances: if the true effect size is non zero but such that, given the sample size, the test has low power, then the estimate conditional on the p-value being lower than the significance threshold is upward biased [24].

A number of authors phrase their critics of the p-value saying that it does not convey a valid measure of the evidence against the null or in favor of the alternative. Different meanings may be attached to this, in many cases the same issues outlined above lie at the root of it. For instance, the already mentioned dependence on the sample size reflects a failure of conveying evidence against the null. The fact that the same p-value may correspond to very different probabilities of the null if a Bayesian analysis is performed on the same data is also seen as evidence that it does not convey all information [50]. Finally, from a formal point of view the fact that the p-value does not measure evidence in favor of the alternative is almost obvious since by using the p-value we do not consider an alternative, and the data may be unlikely given the null but even more unlikely given a specific alternative. It is to be noted that this plays well with the fact that the alternative hypothesis is usually phrased vaguely as merely the direction of the effect, if at all.

Rather than being genuine pitfalls of the *p*-value, the above are instances of misuse of it. In fact, in the search for the causes of the alleged low reliability of scientific results, it has been suggested that the *p*-value *per se* is not problematic, rather, it is the use which is made that is questionable, prompting the recent statement by the American Statistical Association on *p*-value use [51].

Roots of the reliability crisis may lie upstream the *p*-value. It has been pointed out that from an epistemological point of view, the procedure outlined in Sect. 2 may not be suitable to corroborate scientific theories [13, 36]. Despite this, its use is widespread, probably due to its simplicity, apparent objectivity and perceived compelling nature as a measure of evidence. In fact, the *p*-value alone is a compelling measure of evidence only if misinterpreted through wishful thinking ("it [NHST] does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" [7]). The diffusion of its misinterpretations may also be seen as a hint that the simplicity of the *p*-value is only illusory.

The objectivity of the *p*-value may be a fallacy as well. In fact, the *p*-value may be seen as an objective measure of a theory (whatever it measures) if the theory and the test to be performed are specified in advance of collecting the data. If this is not the case, the value which is obtained as a result of the testing procedure may be driven not only objectively by the data but by the subjective judgment of the experimenter through the conscious (or even unconscious) processes of *p*-hacking [45] or the "garden of forking paths" (GOFP) [14]. The former, *p*-hacking, refers to the fact that each single researcher or team uses the same data to probe different theories, thus the final *p*-value is the minimum of a set of *p*-values obtained from a number of (related) tests. The latter, GOFP, refers to the fact that a single theory (even pre-specified) may be tested, but the details of the data analysis may be driven by the data through mechanisms such as the selection of the relevant variables or the inclusion/exclusion of observations, thus introducing a bias in the testing procedure. Evidence of such phenomena can be found by analyzing the frequency distribution of samples of *p*-values (*p*-curves, [46]): both phenomena are expected to lead to a relatively high frequency in a (left) neighborhood of the common 5% threshold. This feature has in fact been observed in different disciplines [5, 22, 32]. A further confirmation of the effect of the researcher degrees of freedom on the likelihood of finding significant results comes from a natural experiment in large NHLBI clinical trials, where it has been noted that, upon the introduction of pre-registration, the share of experiments leading to significant results dropped [26].

4 What Then?

Various changes of the procedure have been proposed to make it able to deal with scientific questions. We may broadly distinguish them based on whether the paradigm itself is left unchanged but the *p*-value is substituted by an alternative summary

of compliance of data to theory (Sect. 4.2) or the paradigm is changed altogether (Sect. 4.1).

4.1 Changing the Paradigm: The Hard Way

Gelman [13] advocates a total change of methodology in which the focus is on estimation rather than testing. He argues that the correct interpretation of the p-value makes it (almost) irrelevant to the purpose for which it is used in science (it remains useful in indicating lack of fit of a model for the purpose of deciding how to improve it). Moreover, from a practical point of view, the use of the p-value may mask p-hacking or GOFP. Any conclusion concerning scientific discoveries should rather be derived from the implications of the estimated model. This approach does not, in general, offer a clear-cut (yes/no) answer and requires more expertise in data analysis than what is needed to use NHST.

A different approach is to remedy the limitations of the p-value by complementing it with some other measure related to the reliability of conclusions. The basic idea can be traced back to Meehl [36], who suggested that the strength of an experiment in corroborating a theory can be measured by the precision with which experimental results can be predicted by the theory. More recently, Mayo and Spanos [33] proposed the severity, which is defined as follows. Suppose that $\{t(Y) > t(y_c)\}$ is a Neyman-Pearson rejection region for $H_0 : \theta \leq \theta_0$. If y_0 is observed and H_0 is accepted, this is evidence against $\theta > \theta_1 (> \theta_0)$ and the strength of this evidence is measured by $P(t(Y) > t(y_0)|\theta = \theta_1)$. A similar notion is defined in case of rejection of the null. Loosely speaking, the evidence against a hypothesis is measured by the probability that the test statistic would have shown less agreement with the null had the hypothesis been false. The severity is related to the power (they are equal if the sample is at the boundary of the rejection region) but is a different concept (it is a function of the observed data, thus being a measure of power given the observed sample). It can be said that it "retains aspects of, and also differs from, both Fisherian and Neyman-Pearsonians accounts" [33], in particular it explicitly allows for the alternative hypothesis but also retains the post-data interpretation of the p-value.

A similar approach is used in physics, where a p-value is often complemented by the "median p-value" (the p-value one would get if the observed value of the test statistic is the median of the sample distribution in the alternative hypothesis) or the expected significance level, both measures of the p-value one would get under specific alternatives.

Focusing on avoiding bias phenomena such as p-hacking and GOFP, it has been proposed to apply the principles of blinded analysis [31]. This method was introduced in particle physics [27] and entails adding noise to data and/or masking labels so that the researcher who performs the data analysis can not anticipate the substantive conclusions of his inferences. The main difficulty is to hide enough information to avoid bias but still allowing a meaningful analysis.

Finally, Bayesian tools may be used [17]. Although a well developed technique, Bayesian analysis has never been widely adopted in applications, likely due to the fact that, with respect to NHST, it is less simple to use and not perceived as objective.

4.2 Changing the Paradigm: The Soft Way

The proposals reviewed in Sect. 4.1 imply a major change of paradigm and, most important, they do not share two of the main perceived advantages of NHST: ease of use and objectivity [50]. Although both "advantages" may be fallacious, their explicit absence may render the suggested alternatives less appealing to potential users and prevent their adoption. An alternative approach is to change the least of the paradigm, substituting the p-value with some other synthetic measure which does not share its pitfalls but keeps the (purported) advantages. We review below the main proposed substitutes.

Substituting NHST with confidence intervals [10] is (at least in standard situations) a change in the way in which the results are communicated rather than a change of method. However, it may still be a relevant change since it is plausible that confidence intervals be less prone to misinterpretations (and some empirical evidence confirming this is available [11]).

Scholars from different fields [6, 50] suggest using model selection criteria: the null and alternative hypotheses correspond to two different models, the null hypothesis is then "rejected" if the model corresponding to the alternative is preferred. This is an appealing strategy because of its simplicity of implementation and "objectivity". A number of options is available for the model comparison criterion: AIC and BIC are the ones which are more often put forward. Besides the link with the likelihood ratio, it should be remembered that AIC is related to cross validation, while BIC is the Bayes factor with suitable priors. We note that using BIC may be one way to introduce the Bayes factor as a substitute of NHST without paying the price of the complications of the Bayesian approach. Beside AIC and BIC, other similar criteria may be considered depending on the models under consideration (Mallows C_p, GCV, UBRE score), standard cross validation (leave-one-out, K-fold, fixed samples) may also be used. Also, using the likelihood alone has been suggested [43] (mainly on the grounds that it does not depend on the sample space (that is, on experimenter intention)).

A further model selection method which is suitable for the task is the lasso method, at least whenever the models can be framed in a (generalized) linear model specification and the null hypothesis is that a coefficient is equal to zero. In that case one may accept the null hypothesis if the lasso estimate of the coefficient is null, the penalization weight being chosen somehow, for instance by cross validation.

Finally, the minimal change which has been suggested is to lower the conventional threshold for significance. It has been noted that the 5% threshold was introduced when fewer hypotheses were being tested so it makes sense to change it today.

A lower threshold is usually employed in hard sciences, which appear less affected—although not immune—by the reliability crisis.

One advantage of the above procedures—which admittedly would probably be seen as a disadvantage by many—is that they offer an automatic choice. This may allow to compare their performances by means of a simulation study to assess, under various scenarios, the false discovery rate they would imply if used as a substitute for NHST/*p*-value.

5 Discussion

A number of issues have been raised in the literature concerning the use of NHST and the *p*-value since the introduction of such tools by Neyman-Pearson and Fisher. The debate on whether they are useful or harmful for assessing scientific hypotheses is particularly vivid today and coupled with the debate on the lack of reproducibility and high false discovery rate of scientific results in many disciplines.

In fact, the misuse and misinterpretation of NHST are the reasons why it is often singled out as a major weakness. On the contrary, it can be argued that there are relevant possible reasons for the high FDR/lack of reproducibility which lie upstream the use of NHST.

First, there is a big leap in inferring from the falsification of a null nil hypothesis a confirmation of a specific alternative, particularly when the alternative does not imply a precise prediction of what would have been observed had it been true (i.e., the alternative predicts a positive effect rather than an effect of a given size) [36].

Second, a high number of scientific hypotheses is probed. Each single researcher or team tends to use the same data to probe different theories, thus leading to a multiple testing situation which may be explicit or, more subtly, due to the degrees of freedom in specifying the data processing step and the model. This may be phrased saying that exploratory studies are then treated as confirmatory ones (where by the latter we mean experiments with pre-specified hypotheses and methods) thus creating unrealistic expectations on the reliability of the result (on the probability of it being a false discovery). Moreover, this also happens "science-wide" meaning that, at least in some disciplines, lots of labs and researchers means a high number of hypotheses being tested leading to an uncontrollable multiple testing situation associated to a search for small effects (having the "main ones", the low hanging fruits, already been found) [47].

Based on the above considerations, it is reasonable to think that the "soft" changes to the present paradigm, where basically the *p*-value is substituted by some other measure of concordance/discordance between theory and data would hardly be a solution [29]. Also, it is probably unrealistic to try to devise a synthetic measure of evidence for or against a scientific theory. A "hard" change of paradigm is more promising, however no generally accepted alternative has been identified as of today. Moreover, it is to be noted that most, if not all, promising changes do not give a

clear-cut answer to the posited question (of whether a given theory is true), a circumstance which is likely to make it hard for them to become generally accepted.

Acknowledgements This work was supported by Univesity of Trieste within the FRA project "Politiche strutturali e riforme. Analisi degli indicatori e valutazione degli effetti".

References

1. Baker, M.: Is there a reproducibility crisis? Nature **533**, 452–454 (2016)
2. Beall, A.T., Tracy, J.L.: Women are more likely to wear red or pink at peak fertility. Psychol. Sci. **24**, 1837–1841 (2013)
3. Berger, J.O.: Could Fisher, Jeffreys and Neyman have agreed on testing? Stat. Sci. **18**(1), 1–12 (2003)
4. Boland, M.R., Shahn, Z., Madigan, D., Hripcsak, G., Tatonetti, N.P.: Birth month affects lifetime disease risk: a phenome-wide method. J. Am. Med. Inform. Assoc. ocv046 (2015)
5. Brodeur, A., Lé, M., Sangnier, M., Zylberberg, Y.: Star wars: the empirics strike back. Am. Econ. J. Appl. Econ. **8**(1), 1–32 (2016)
6. Burnham, K., Anderson, D.: P values are only an index to evidence: 20th-vs. 21st-century statistical science. Ecology **95**(3), 627–630 (2014)
7. Cohen, J.: The earth is round ($p < 0.05$). Am. Psychol. **49**, 997–1003 (1994)
8. Cowan, G., Cranmer, K., Gross, E., Vitells, O.: Asymptotic formulae for likelihood-based tests of new physics. Eur. Phys. J. C **71**(2), 1–19 (2011)
9. Cowen, R.: Big bang finding challenged. Nature **510**(7503), 20 (2014)
10. Cumming, G.: The new statistics why and how. Psychol. Sci. **25**, 7–29 (2013)
11. Fidler, F., Loftus, G.R.: Why figures with error bars should replace p values: some conceptual arguments and empirical demonstrations. J. Psychol. **217**(1), 27–37 (2009)
12. Fisher, R.A., et al.: Statistical methods for research workers. In: Statistical Methods for Research Workers, 10th. edn. (1946)
13. Gelman, A.: Commentary: P values and statistical practice. Epidemiology **24**(1), 69–72 (2013)
14. Gelman, A., Loken, E.: The statistical crisis in science. Am. Sci. **102**, 460–465 (2014)
15. Gigerenzer, G.: Mindless statistics. J. Socio-Econ. **33**(5), 587–606 (2004)
16. Goodman, S.N.: Toward evidence-based medical statistics. 1: the p value fallacy. Ann. Intern. Med. **130**(12), 995–1004 (1999)
17. Goodman, S.N.: Toward evidence-based medical statistics. 2: the bayes factor. Ann. Intern. Med. **130**(12), 1005–1013 (1999)
18. Goodman, S.N.: Aligning statistical and scientific reasoning. Science **352**, 1180–1181 (2016)
19. Greenland, S., Poole, C.: Living with p values: resurrecting a bayesian perspective on frequentist statistics. Epidemiology **24**(1), 62–68 (2013)
20. Hart, et al.: Dogs are sensitive to small variations of the Earth's magnetic field. Front. Zool. **10**, 80 (2013)
21. Hauer, E.: The harm done by tests of significance. Accident Analysis & Prevention **36**(3), 495–500 (2004)
22. Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., Jennions, M.D.: The extent and consequences of p-hacking in science. PLoS Biol. **13**(3), e1002,106 (2015)
23. Hoover, K.D., Siegler, M.V.: Sound and fury: Mccloskey and significance testing in economics. J. Econ. Method. **15**(1), 1–37 (2008)
24. Ioannidis, J.P.: Contradicted and initially stronger effects in highly cited clinical research. Jama **294**(2), 218–228 (2005)
25. Ioannidis, J.P.: Why most published research findings are false. PLoS Med. **2**(8), e124 (2005)
26. Kaplan, R.M., Irvin, V.L.: Likelihood of null effects of large nhlbi clinical trials has increased over time. PloS one **10**(8), e0132,382 (2015)

27. Klein, J.R., Roodman, A.: Blind analysis in nuclear and particle physics. Ann. Rev. Nucl. Part. Sci. **55**(1), 141–163 (2005)
28. Krantz, D.H.: The null hypothesis testing controversy in psychology. J. Am. Stat. Assoc. **94**(448), 1372–1381 (1999)
29. Leek, J.T., Peng, R.D.: Statistics: P-values are just the tip of the iceberg. Nature **520**(7549) (2015)
30. Lovell, D.: Biological importance and statistical significance. J. Agric. Food Chem. **61**(35), 8340–8348 (2013)
31. MacCoun, R., Perlmutter, S.: Blind analysis: hide results to seek the truth. Nature **526**(7572), 187–189 (2015)
32. Masicampo, E.J., Lalande, D.R.: A peculiar prevalence of p-values just below.05. Q. J. Exp. Psychol. **65**(11), 2271–2279 (2012)
33. Mayo, D.G., Spanos, A.: Severe testing as a basic concept in a neymanpearson philosophy of induction. Br. J. Philos. Sci. **57**(2), 323–357 (2006)
34. McCloskey, D.: The insignificance of statistical significance. Sci. Am. **272**, 32–33 (1995)
35. McCloskey, D.N., Ziliak, S.T.: The standard error of regressions. J. Econ. Lit. **34**(1), 97–114 (1996)
36. Meehl, P.: The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In: What if there were no significance tests, pp. 393–425. Psychology press (2013)
37. Neyman, J., Pearson, E.S.: On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. R. Soc. Lon. Ser. A **231**, 289–337 (1933)
38. Nicholls, N.: Commentary and analysis: the insignificance of significance testing. Bull. Am. Meteorol. Soc. **82**(5), 981–986 (2001)
39. Nuzzo, R.: Scientific method: statistical errors. Nature **506**(7487), 150–152 (2014)
40. Reich, E.S.: Timing glitches dog neutrino claim. Nature **483**(7387), 17 (2012)
41. Rogoff, K., Reinhart, C.: Growth in a time of debt. Am. Econ. Rev. **100**, 573–578 (2010)
42. Rothman, K.J.: Writing for epidemiology. Epidemiology **9**(3), 333–337 (1998)
43. Royall, R.: Statistical Evidence: A Likelihood Paradigm (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). Chapman and Hall/CRC (1997)
44. Schmidt, F., Hunter, J.: Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In: S.A.S.J. Harlow L.L. (ed.) What if There were no Significance Tests?, pp. 37–64. Psychology Press (1997)
45. Simmons, J.P., Nelson, L.D., Simonsohn, U.: False-Positive psychology-undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. **22**(11), 1359–1366 (2011)
46. Simonsohn, U., Nelson, L.D., Simmons, J.P.: P-curve: a key to the file-drawer. J. Exp. Psychol. Gen. **143**(2), 534–547 (2014)
47. Sterne, J.A.C., Smith, G.D., Cox, D.R.: Sifting the evidence-what's wrong with significance tests? Phys. Ther. **81**(8), 1464–1469 (2001)
48. Trafimow, D.: Editorial. Basic Appl. Soc. Psychol. **36**(1), 1–2 (2014)
49. Trafimow, D., Marks, M.: Editorial. Basic Appl. Soc. Psychol. **37**(1), 1–2 (2015)
50. Wagenmakers, E.J.J.: A practical solution to the pervasive problems of p values. Psychon. Bull. Rev. **14**(5), 779–804 (2007)
51. Wasserstein, R.L., Lazar, N.A.: The ASA's statement on p-values: context, process, and purpose. Am. Stat. **70**(2), 129–133 (2016)
52. Ziliak, S., McCloskey, D.: Size matters: the standard error of regressions in the american economic review. J. Socio-Econ. **33**(5), 527–546 (2004)