Rafi and Greenland *BMC Medical Research Methodology* (2020) 20:244
https://doi.org/10.1186/s12874-020-01105-9

BMC Medical Research Methodology

TECHNICAL ADVANCE Open Access

Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise

Zad Rafi[1*] and Sander Greenland[2]

Abstract

Background: Researchers often misinterpret and misrepresent statistical outputs. This abuse has led to a large literature on modification or replacement of testing thresholds and *P*-values with confidence intervals, Bayes factors, and other devices. Because the core problems appear cognitive rather than statistical, we review some simple methods to aid researchers in interpreting statistical outputs. These methods emphasize logical and information concepts over probability, and thus may be more robust to common misinterpretations than are traditional descriptions.

Methods: We use the Shannon transform of the *P*-value *p*, also known as the binary surprisal or *S*-value $s = -\log_2(p)$, to provide a measure of the information supplied by the testing procedure, and to help calibrate intuitions against simple physical experiments like coin tossing. We also use tables or graphs of test statistics for alternative hypotheses, and interval estimates for different percentile levels, to thwart fallacies arising from arbitrary dichotomies. Finally, we reinterpret *P*-values and interval estimates in unconditional terms, which describe compatibility of data with the entire set of analysis assumptions. We illustrate these methods with a reanalysis of data from an existing record-based cohort study.

Conclusions: In line with other recent recommendations, we advise that teaching materials and research reports discuss *P*-values as measures of compatibility rather than significance, compute *P*-values for alternative hypotheses whenever they are computed for null hypotheses, and interpret interval estimates as showing values of high compatibility with data, rather than regions of confidence.
Our recommendations emphasize cognitive devices for displaying the compatibility of the observed data with various hypotheses of interest, rather than focusing on single hypothesis tests or interval estimates. We believe these simple reforms are well worth the minor effort they require.

Keywords: Confidence intervals, Cognitive science, Bias, Data interpretation, Evidence, Hypothesis tests, Information, *P*-values, Statistical significance, Models, statistical

* Correspondence: zad@lesslikely.com
Available Code: The R scripts to reproduce the graphs in the text can be obtained at https://osf.io/6w8g9/
[1]Department of Population Health, NYU Langone Medical Center, 227 East 30th Street, New York, NY 10016, USA
Full list of author information is available at the end of the article

Background

Statistical science is fraught with psychological as well as technical difficulties, yet far less attention has been given to cognitive problems than to technical minutiae and computational devices [1, 2]. If the issues that plague science could be resolved by mechanical algorithms, statisticians and computer scientists would have disposed of them long ago. But the core problems are of human psychology and social environment, one in which researchers apply traditional frameworks based on fallacious rationales and poor understanding [1, 3]. These problems have no mathematical or philosophical solution, and instead require attention to the unglamorous task of developing tools, interpretations and terminology more resistant to misstatement and abuse than what tradition has handed down.

We believe that neglect of these problems is a major contributor to the current crisis of statistics in science [4–9]. Several informal descriptions of statistical formulas may be reasonable when strictly adhered to, but nevertheless lead to severe misinterpretations in practice. Users tend to take extra leaps and shortcuts, hence we need to anticipate implications of terminology and interpretations to improve practice. In doing so, we find it remarkable that the *P*-value is once again at the center of the controversy [10], despite the fact that some journals strongly discouraged reporting *P*-values decades ago [11], and complaints about misinterpretation of statistical significance date back over a century [12–14]. Equally remarkable is the diversity of proposed solutions, ranging from modifications of conventional fixed-cutoff testing [15–18] to complete abandonment of traditional tests in favor of interval estimates [19–21] or testing based on Bayesian arguments [22–26]; no consensus appears in sight.

While few doubt that some sort of reform is needed, the following crucial points are often overlooked:

1) There is no universally valid way to analyze data and thus no single solution to the problems at hand.
2) Careful integration of contextual information and technical considerations will always be essential.
3) Most researchers are under pressure to produce definitive conclusions, and so will resort to familiar automated approaches and questionable defaults [27], **with** or **without** *P*-values or "statistical significance" [28].
4) Most researchers lack the time or skills for reeducation, so we need methods that are simple to acquire quickly based on what is commonly taught, yet are also less vulnerable to common misinterpretation than are traditional approaches (or at least have not yet become as widely misunderstood as those approaches).

Thus, rather than propose abandoning old methods in favor of entirely new ones, we will review simple cognitive devices, terminological reforms, and conceptual shifts that encourage more realistic, accurate interpretations of conventional statistical summaries. Specifically, we will advise that:

a) We should replace decisive-sounding, overconfident terms like "significance," "nonsignificance" and "confidence interval," as well as proposed replacements like "uncertainty interval," with more modest descriptors such as "low compatibility," "high compatibility" and "compatibility interval" [29–31].
b) We should teach alternative ways to view *P*-values and interval estimates via information measures such as *S*-values (surprisals), which are the negative logarithms of the *P*-values; these measures facilitate translation of statistical test results into results from simple physical experiments [31, 32].
c) For quantities targeted for study, we should replace single *P*-values, *S*-values, and interval estimates by tables or graphs of *P*-values or *S*-values showing results for relevant alternative hypotheses as well as for null hypotheses.
d) We should from the start teach that the usual interpretations of statistical outputs are often misleading even when they are technically accurate. This is because they *condition* on background assumptions (i.e., they treat them as given), and thus they ignore what may be serious uncertainty about those assumptions. This deficiency can be most directly and nontechnically addressed by treating them *unconditionally*, shifting their logical status from assumptions to components of the tested framework.

We have found that the last recommendation (to decondition inferences [31]) is the most difficult for readers to comprehend, and is even resisted and misrepresented by some with extensive credentials in statistics. Thus, to keep the present paper of manageable length we have written a companion piece [33], which explains in depth the rationale for de-emphasizing traditional conditional interpretations in favor of unconditional interpretations.

An example

We will display some of these problems and recommendations with published results from a record-based cohort study of serotonergic antidepressant prescriptions during pregnancy and subsequent autism spectrum disorder (ASD) of the child (Brown et al. [34]). Out of 2837 pregnancies that had filled prescriptions, approximately 2% of the children were diagnosed with ASD. The paper

first reported an adjusted ratio of ASD rates (hazard ratio or HR) of 1.59 when comparing mothers with and without the prescriptions, and 95% confidence limits (CI) of 1.17 and 2.17. This estimate was derived from a proportional-hazards model which included maternal age, parity, calendar year of delivery, neighborhood income quintile, resource use, psychotic disorder, mood disorder, anxiety disorder, alcohol or substance use disorder, use of other serotonergic medications, psychiatric hospitalization during pregnancy, and psychiatric emergency department visit during pregnancy.

The paper then presented an analysis with adjustment based on a high-dimensional propensity score (HDPS), in which the estimated hazard ratio became 1.61 with a 95% CI spanning from 0.997 to 2.59. Despite the estimated *61% increase* in the hazard rate in the exposed children and an interval estimate including ratios as large as 2.59 and no lower than 0.997, the authors still declared that there was no association between in utero serotonergic antidepressant exposure and ASD because it was not "statistically significant." This was a misinterpretation of their own results because an association was not only present, but also quite close to the 70% increase they reported from previous studies [35]. Yet the media simply repeated Brown et al.'s misstatement that there was no association after adjustment [36].

This type of misreporting remains common, despite the increasing awareness that such dichotomous thinking is detrimental to sound science and the ongoing efforts to retire statistical significance [23, 29, 37–42]. To aid these efforts, we will explain the importance of showing results for a range of hypotheses, which may help readers see why conclusions such as in Brown et al. [34, 36] represent dramatic misinterpretations of statistics – even though the reported *numeric* summaries are correct. We will also explain why it would be correct to instead have reported that "After HDPS adjustment for confounding, a 61% hazard elevation remained; however, under the same model, every hypothesis from no elevation up to a 160% hazard increase had $p > 0.05$; Thus, while quite imprecise, these results are consistent with previous observations of a positive association between serotonergic antidepressant prescriptions and subsequent ASD. Because the association may be partially or wholly due to uncontrolled biases, further evidence will be needed for evaluating what, if any, proportion of it can be attributed to causal effects of prenatal serotonergic antidepressant use on ASD incidence." We believe this type of language is careful and nuanced, and that such cautious attention to detail is essential for accurate scientific reporting. For simplicity and consistency with common practice we have used the 0.05 cutoff in the description, but recognize that researchers may be better served by choosing their descriptive approach as well as decision cutoffs based on background literature and error costs, rather than using traditional conventions [16].

Methods

Making sense of tests, I: the *P*-value as a compatibility measure

The infamous *observed P*-value p (originally called the observed or attained "level of significance" or "value of *P*" [43–45]) is a measure of compatibility between the observed data and a targeted test hypothesis **H**, given a set of background assumptions (the background model) which are used along with the hypothesis to compute the *P*-value from the data. By far the most common example of a test hypothesis **H** is a traditional null hypothesis, such as "there is no association" or (more ambitiously) "there is no treatment effect." In some books this null hypothesis is the only test hypothesis ever mentioned. Nonetheless, the test hypothesis **H** could just as well be "the treatment doubles the risk" or "the treatment halves the risk" or any other hypothesis of practical interest [46]; we will argue such alternatives to the null *should* also be tested whenever the traditional null hypothesis is tested. Our discussion will also apply when **H** concerns multiple parameters and thus the test involves multiple degrees of freedom, for example a general test of linearity of trend (dose-response) when a treatment has 5 levels (which has 3 degrees of freedom).

With this general background about the test hypothesis, the other key ingredient in traditional statistical testing is a test statistic, such as a Z-score or χ^2, which measures the discrepancy between the observed data and what would have been expected under the test hypothesis, given the background assumptions. We can now define an observed *P*-value p as the probability of the test statistic being *at least* as extreme as observed *if* the hypothesis **H** targeted for testing *and* every assumption used to compute the *P*-value (the test hypothesis **H** *and* the background statistical model) were correct [46]. Those background assumptions typically include a host of conditions such as linearity of responses and additivity of effects on a given scale; appropriateness of included variables (e.g., no intermediates for the effect under study); unimportance of omitted variables (e.g., all important confounding is controlled), random errors in a given family, no selection bias, and full accounting for measurement error and model selection.

This accurate and technical description does not accord well with human psychology, however: It is often said by Bayesians that researchers want a probability for the targeted test hypothesis (posterior probability of **H**), not a probability of observations. This imperative is indicated by the many "intuitive" – and incorrect – verbal definitions and descriptions of the *P*-value that amount to calling it the probability of the test hypothesis, which

is quite misleading [46]. Such errors are often called *inversion fallacies* because they invert the role of the observations and the hypothesis in defining the *P*-value (which is a probability for the observed test statistic, not the test hypothesis).

A standard frequentist criterion for judging whether a *P*-value is valid for statistical testing is that all possible values for it from zero to one are equally likely (uniform in probability) if the test hypothesis and background assumptions are correct. We discuss this criterion in more detail in the Supplement. With this validity criterion met, we can also correctly describe the *P*-value without explicit reference to repeated sampling, as the *percentile* or proportion at which the observed test statistic falls in the distribution for the test statistic under the test hypothesis and the background assumptions [47, 48]. The purpose of this description is to connect the *P*-value to a familiar concept, the percentile at which someone's score fell on a standard test (e.g., a college or graduate admissions examination), as opposed to the remote abstraction of infinitely repeated sampling.

Making sense of tests, II: the *S*-value

Even when *P*-values are correctly defined and valid, their scaling can be deceptive due to their compression into the interval from 0 to 1, with vastly different meanings for absolute differences in *P*-values near 1 and the same differences for *P*-values near 0 [31], as we will describe below. One way to reduce test misinterpretations and provide more intuitive numerical results is to translate the *P*-values into probabilities of outcomes in familiar games of chance.

Consider a game in which one coin will be tossed and we will bet on tails. Before playing however we want evidence that the tossing is acceptable for our bet, by which we mean not biased toward heads, because such loading would make our losing more probable than not. To check acceptability, suppose we first do s independent test tosses and they *all* come up heads. If the tossing is acceptable, the chance of this happening is at most $½^s$, the chance of all heads in s unbiased (fair) tosses. The smaller this chance, the less we would trust that the game is acceptable. In fact we could take s as measuring our evidence against acceptability: If we only did one toss and it came up heads ($s = 1$) that would be unsurprising if the tossing were unbiased for then it would have chance ½, and so would provide barely any evidence against acceptability. But if we did 10 tosses and all came up heads ($s = 10$) that would be surprising if the tossing were unbiased, for the chance of that is then $½^{10} \approx 0.001$, and so would provide considerably more evidence against acceptability.

With this setting in mind, we can now gauge the evidence supplied by a *P*-value p by seeing what number s of heads in a row would come closest to p, which we can find by solving the equation $p = ½^s$ for s. The solution is the negative base-2 logarithm of the *P*-value, $s = \log_2(1/p) = -\log_2(p)$, known as the binary Shannon information, surprisal, logworth, or *S-value* from the test [31, 49, 50]. The *S*-value is designed to reduce incorrect probabilistic interpretations of statistics by providing a nonprobability measure of information supplied by the test statistic against the test hypothesis **H** [31].

The *S*-value provides an absolute scale on which to view the information provided by a valid *P*-value, as measured by calibrating the observed p against a physical mechanism that produces data with known probabilities. A single coin toss produces a binary outcome which can be coded as 1 = heads, 0 = tails, and thus requires only two symbols or states to record or store; hence the information in a single toss is called *bit*, short for binary digit, or a *shannon*. The information describing a sequence of s tosses requires s bits to record or store; thus, extending this measurement to a hypothesis **H** with *P*-value p, we say the test supplied $s = -\log_2(p)$ bits of information against **H**.

We emphasize that, without further restrictions, our calibration of the *P*-value against coin-tossing is only measuring information *against* the test hypothesis, not in support of it. This limitation is for the purely logical reason that there is no way to distinguish among the infinitude of background assumptions that lead to a test with the same or larger *P*-value and hence the same or smaller *S*-value. There is no way the data can *support* a test hypothesis except relative to a fixed set of background assumptions. Rather than taking the background assumptions for granted, we prefer instead to adopt a refutational view, which emphasizes that any claim of support will be undermined by assumption uncertainty, and is thus best avoided. This caution applies regardless of the test statistic used, whether *P*-value, *S*-value, Bayes factor, or posterior probability.

As with the *P*-value, the *S*-value refers only to a particular test with particular background assumptions. A different test based on different background assumptions will usually produce a different *P*-value and thus a different *S*-value; thus it would be a mistake to simply call the *S*-value "the information against the hypothesis supplied **by the data**", for it is always a test of the hypothesis conjoined with (or conditioned on) the assumptions. As a basic example, we may contrast the *P*-value for the strict null hypothesis (of no effect on any experimental unit) comparing two experimental groups using a *t*-test (which, along with randomization, assumes normally distributed responses under the null hypothesis), to the *P*-value from a permutation test (which assumes only randomization).

Finally, as explained in the Supplement, the *S*-value can also be expressed using other logarithmic units such as natural (base-e) logs, $-\ln(p)$, which is mathematically more convenient but not as easy to represent physically.

Evaluating *P*-values and fixed-cutoff tests with *S*-values

With the *S*-value in hand, a cognitive difficulty of the *P*-value scale for evidence can be seen by first noting that the difference in the evidence provided by *P*-values of 0.9999 and 0.90 is trivial: Both represent almost no information against the test hypothesis, in that the corresponding *S*-values are $-\log_2(0.9999) = 0.00014$ bits and $-\log_2(0.90) = 0.15$ bits. Both are far less than 1 bit of information against the hypothesis – they are just a fraction of a coin toss different. In contrast, the information against the test hypothesis in *P*-values of 0.10 and 0.0001 is profoundly different, in that the corresponding *S*-values are $-\log_2(0.10) = 3.32$ and $-\log_2(0.0001) = 13.3$ bits; thus $p = 0.001$ provides 10 bits more information against the test hypothesis than does $p = 0.10$, corresponding to the information provided by 10 additional heads in a row. The contrast is illustrated in Fig. 1, along with other examples of the scaling difference between *P* and *S* values.

As an example of this perspective on reported results, from the point and interval estimate from the HDPS analysis reported by Brown et al. [34], we calculated that the *P*-value for the "null" test hypothesis **H** that the hazard ratio is 1 (no association) is 0.0505. Using the *S*-value to measure the information supplied by the HDPS analysis against this hypothesis, we get $s = -\log_2(0.0505) = 4.31$ bits; this is hardly more than 4 coin tosses worth of information against no association. For comparison, when setting the test hypothesis **H** to be that the hazard ratio is 2 (doubling of the hazard among the treated), we calculated a *P*-value of about 0.373. The information supplied by the HDPS analysis against this test hypothesis is then measured by the *S*-value as $s = -\log_2(0.373) = 1.42$ bits, hardly more than a coin-toss worth of information against doubling of the hazard among the treated. In these terms, then, the HDPS results supply roughly 3 bits more information against no association than against doubling of the hazard, so that doubling (HR = 2) is more compatible with the analysis results than is no association (HR = 1).

S-values help clarify objections to comparing *P*-values to sharp dichotomies. Consider that a *P*-value of 0.06 yields about 4 bits of information against the test hypothesis **H**, while a *P*-value of 0.03 yields about 5 bits of information against **H**. Thus, $p = 0.06$ is about as surprising as getting all heads on four fair coin tosses while $p = 0.03$ is one toss (one bit) more surprising. Even if one is committed to making a decision based on a sharp cutoff, *S*-values illustrate what range around that cutoff corresponds to a trivial information difference (e.g., any *P*-value between 0.025 and 0.10 is less than a coin-toss difference in evidence from $p = 0.05$).

S-values also help researchers understand more subtle problems with traditional testing. Consider for example the import of the magical 0.05 threshold (α-level) that is wrongly used to declare associations present or absent. It has often been claimed that this threshold is too high to be regarded as representing much evidence against **H** [15, 26], but the arguments for that are usually couched in Bayesian terms of which many remain skeptical. We

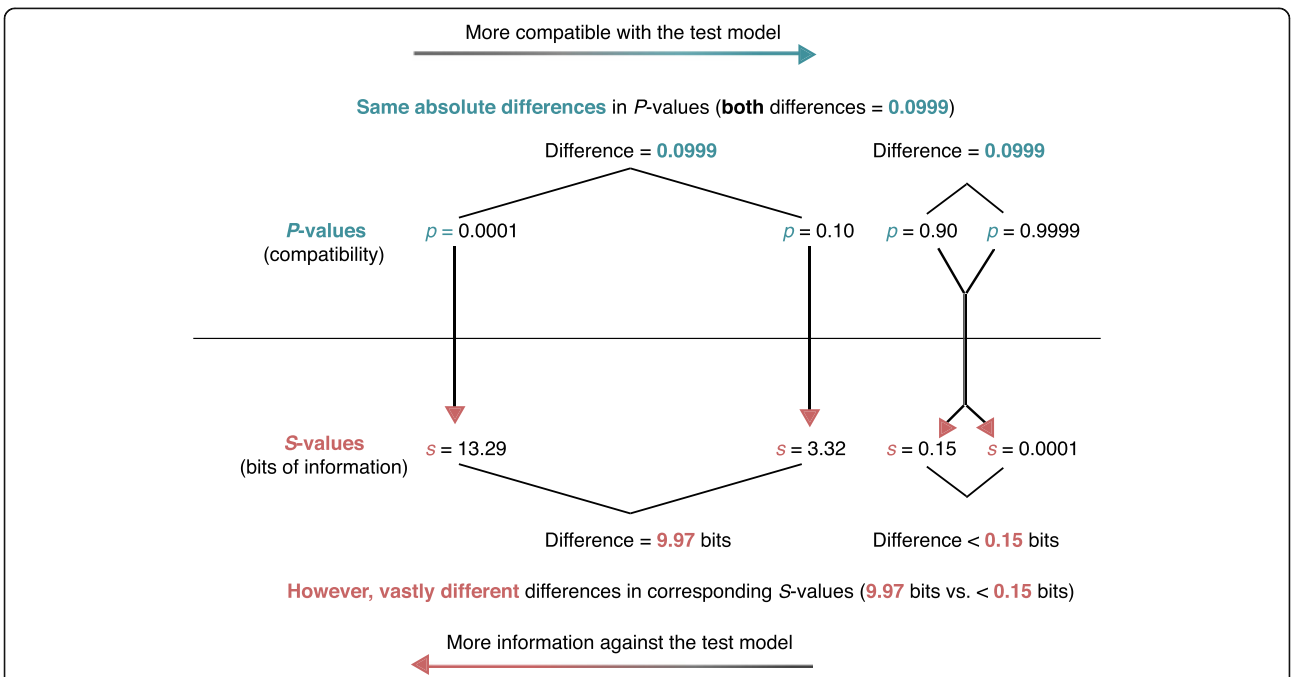

Fig. 1 Comparison of *P*-value and *S*-value scales. Top labels: Data compatibility with test model as measured by *P*-values. Bottom labels: Information against test model as measured by *S*-values

can however see those objections to 0.05 straightforwardly by noting that the threshold translates into requiring an *S*-value of only $-\log_2(0.05) = 4.32$ bits of information against the null; that means $p = 0.05$ is barely more surprising than getting all heads on 4 fair coin tosses.

While 4 heads in a row may seem surprising to some intuitions, it does in fact correspond to doing only 4 tosses to study the coin; a sample size of $N = 4$ binary outcomes would rarely qualify as a basis for (say) recommending a new treatment even if all 4 patients recovered but the recovery rate without the treatment was known to be 50%. Thus, like other proposals, the *S*-value calls into question the traditional $\alpha = 0.05$ standard, and may help users realize how little information is contained in most *P*-values when compared to the thousands of bits of information in a typical cell-phone directory. Further crucial information will be given by *P*-values and *S*-values tabulated for several *alternative* hypotheses, interval estimates over varying percentiles, and graphs of data and information summaries such as those illustrated below.

Further advantages of *S*-values

Unlike probabilities, *S*-values are unbounded above and can be added over independent information sources to create simple summary tests [55 p. 80; see also our supplement]. They thus provide a scale for comparing and combining test results across studies that is aligned with information rather than probability measurement [31]. Another advantage of *S*-values is that they help thwart inversion fallacies, in which a *P*-value is misinterpreted as a probability of a hypothesis being correct (or equivalently, as the probability that a statement about the hypothesis is in error). Hypothesis probabilities computed using the data are called *posterior probabilities* (because they come after the data). It is difficult to confuse an *S*-value with a posterior probability because the *S*-value is unbounded above, and in fact will be above 1 whenever the *P*-value is below 0.50.

Probabilities of data summaries (test statistics) given hypotheses and probabilities of hypotheses given data are identically scaled, leading users to inevitably conflate *P*-values with posterior probabilities. This confusion dominates observed misinterpretations [46] and is invited with open arms by "significance" and "confidence" terminology. Such mistakes could potentially be avoided by giving actual posterior probabilities along with *P*-values. Bayesian methods provide such probabilities but require prior distributions as input; in turn, those priors require justifications based on often contentious background information. While the task of creating such distributions can be instructive, this extra input burden has greatly deterred adoption of Bayesian methods; in contrast, *S*-values provide a direct quantification of information without this input.

Table 1 provides a translation of the *P*-value to the binary *S*-value $s = -\log_2(p)$. It also gives the corresponding maximum-likelihood ratio (MLR), and the deviance-difference or likelihood-ratio test statistic 2ln(MLR), assuming that **H** is a simple hypothesis about one parameter (e.g., that a mean difference or regression coefficient is zero) and that the statistic has a 1 degree of freedom χ^2 distribution; see the Appendix and Supplement for further details. The MLR and deviance statistic are themselves often treated as measures of information against **H** under the background assumptions (fortuitously, when rounding to the nearest integer, the binary *S*-value and deviance statistic coincide in the often-contentious *P*-value range of 0.005 to 0.10). The table also shows the different alpha levels used in various fields and the stark contrast in information associated with these cutoffs. The alpha levels used in particle physics and genome-wide association studies (GWAS) are extremely small because in those areas false positives are considered far more likely and costly than false negatives: Discovery declarations in particle physics require "5 sigmas", nearly 22 bits of information against **H** (corresponding to all heads in 22 fair coin tosses), while GWAS requires nearly 27 bits; for discussions of these choices see [51, 52].

Further details of the relations among these and other measures are given in the Supplement. Table 2 presents these measures as computed from the Brown et al. report [34]; it can again be seen that by any measure there is more information against the null (equal hazards across treatment, $S = 4.31$) than against doubling of the hazard (HR = 2, $S = 1.42$), so the claim that these results demonstrate or support no association is simply wrong.

In summary, the *S*-value provides a gauge of the information supplied by a statistical test in familiar terms of coin tosses. It thus complements the probability interpretation of a *P*-value by supplying a mechanism that can be visualized with simple physical experiments. Given amply documented human tendencies to underestimate the frequency of seemingly "unusual" events [53], these experiments can guide intuitions about what evidence strength a given *P*-value actually represents.

Replace unrealistic "confidence" claims with compatibility measures

Confidence intervals (commonly abbreviated as CI) have been widely promoted as a solution to the problems of statistical misinterpretation [19, 21]. While we support their presentation, such intervals have difficulties of their own. The major problem with "confidence" is that it encourages the common confusion of the CI percentile level (typically 95%) with the probability that the true value of the parameter is in the interval (mistaking the CI for a Bayesian posterior interval) [46], as in

Table 1 *P*-values and binary *S*-values, with corresponding maximum-likelihood ratios (MLR) and deviance (likelihood-ratio) statistics for a simple test hypothesis **H** under background assumptions **A**

P-value p (compatibility of **H** with data given **A**)	*S*-value $s = -\log_2(p)$ (information against **H** given **A** in bits)	Maximum-likelihood ratio against **H** given **A**	Deviance statistic 2ln(MLR)
0.99	0.014	1.00	0.00016
0.90	0.15	1.01	0.016
0.50	1.00	1.26	0.45
0.25	2.00	1.94	1.32
0.10	3.32	3.87	2.71
0.05	4.32	6.83	3.84
0.025	5.32	12.3	5.02
0.01	6.64	27.6	6.63
0.005	7.64	51.4	7.88
0.0001	13.3	1935	15.1
5 sigma[a] (~ 2.9 in 10 million)	21.7	5.2×10^5	26.3
1 in 100 million (GWAS)	26.6	1.4×10^7	32.8
6 sigma[a] (~ 1 in a billion)	29.9	1.3×10^8	37.4

[a]5 and 6 sigma cutoffs are the upper standard-normal tail probabilities at 5 and 6 standard deviations above the mean [51]

statements such as "we are 95% *confident* that the true value is within the interval."

The fact that "confidence" refers to the procedure behavior, *not* the reported interval, seems to be lost on most researchers. Remarking on this subtlety, when Jerzy Neyman discussed his confidence concept in 1934 at a meeting of the Royal Statistical Society, Arthur Bowley replied, "I am not at all sure that the 'confidence' is not a confidence trick." [54]. And indeed, 40 years later, Cox and Hinkley warned, "interval estimates cannot be taken as probability statements about parameters, and foremost is the interpretation 'such and such parameter values are consistent with the data.'" [55]. Unfortunately, the word "consistency" is used for several other concepts in statistics, while in logic it refers to an absolute condition (of noncontradiction); thus, its use in place of "confidence" would risk further confusion.

To address the problems above, we exploit the fact that a 95% CI summarizes the results of varying the test hypothesis **H** over a range of parameter values, displaying all values for which $p > 0.05$ [56] and hence $s < 4.32$ bits [31, 57]. Thus the CI contains a range of parameter values that are more compatible with the data than are values outside the interval, under the background assumptions [31, 46]. Unconditionally (and thus even if the background assumptions are uncertain), the interval shows the values of the parameter which, when combined with the background assumptions, produce a test model that is "highly compatible" with the data in the sense of having less than 4.32 bits of information against it. We thus refer to CI as *compatibility* intervals rather than *confidence* intervals [30, 31, 57]; their abbreviation remains "CI." We reject calling these intervals "uncertainty intervals," because they do not capture uncertainty about background assumptions [30].

Another problem is that a frequentist CI is often used as nothing more than a null-hypothesis significance test (NHST), by declaring that the null parameter value (e.g., HR = 1) is supported if it is inside the interval, or refuted if it is outside the interval. These declarations defeat the

Table 2 Reanalysis of the Brown et al. HDPS results [34][a]

Test Hypothesis (H)	*P*-value (compatibility)	*S*-value (bits of information)	Maximum-likelihood ratio	Likelihood-ratio statistic
Halving of hazard, HR = 0.5	1.6×10^{-6}	19.3	1.0×10^5	23.1
No association (null), HR = 1	0.0505	4.31	6.77	3.82
Point estimate, HR = 1.61	1.00	0.00	1.00	0.00
Doubling of hazard, HR = 2	0.373	1.42	1.49	0.79
Tripling of hazard, HR = 3	0.01	6.56	26.2	6.53
Quintupling of hazard, HR = 5	3.3×10^{-6}	18.2	5.0×10^4	21.7

[a]Computed from the normal approximations given in the Appendix

P-values, *S*-values, maximum-likelihood ratios, and likelihood-ratio statistics for several test hypotheses about the hazard ratio (HR) computed from Brown et al. HDPS results [34].

use of interval estimates to summarize information about the parameter, and perpetuate the fallacy that information changes abruptly across decision boundaries [40, 46, 57, 58]. In particular, the usual 95% default forces the user's focus onto parameter values that yield $p > 0.05$, without regard to the trivial difference between (say) $p = 0.06$ and $p = 0.04$ (an information difference far smaller than a coin toss). Even differences conventionally seen as "large" are often minor in information terms, e.g., $p = 0.02$ and $p = 0.16$ represent a difference of only $\log_2(0.16/0.02) = 3$ coin tosses, underscoring the caution that the difference between "significance" and "nonsignificance" is not significant [59].

To address this problem, we first note that a 95% interval estimate is only one of a number of arbitrary dichotomization of possibilities of parameter values (into either inside or outside of an interval). A more accurate picture of information is then obtained by examining intervals using other percentiles, e.g., proportionally-spaced compatibility levels such as $p > 0.25$, 0.05, 0.01, which correspond to 75, 95, 99% CIs and equally-spaced *S*-values of $s < 2$, 4.32, 6.64 bits. When a detailed picture is desired, a table or graph of *P*-values and *S*-values across a broad range of parameter values seems the clearest way to see how compatibility varies smoothly across the values.

Gradations, not dichotomies

Graphs of *P*-values or their equivalent have been promoted for decades [40, 60–62], yet their adoption has been slight. Nonetheless, *P*-value and *S*-value graphing software is now available freely through several statistical packages [63, 64]. A graph of the *P*-values p against possible parameter values allows one to see at a glance which parameter values are most compatible with the data under the background assumptions. This graph is known as the *P*-value function, or compatibility, consonance, or confidence curve [40, 60–62, 65–69]; the "severity curve" ([18], fig. 5.5) is a special case (see Supplement). Transforming the corresponding *P*-values in the graph to *S*-values produces an *S*-value (surprisal) function.

Most studies not only examine but also present results for multiple associations and models, and examining or presenting graphs for each of the results may be impractical. Nonetheless, as in the Brown et al. example, there is often a "final" analysis or set of results that is used to generate the highlighted conclusions of the study. We strongly advise inspecting graphs for those analyses before writing conclusions, and presenting the graphs in the paper or at least in a supplementary file. As mentioned above, it is quite easy to now construct these curves using various statistical packages [63].

Example, continued

Figures 2 and 3 give the *P*-value and *S*-value graphs produced from the Brown et al. [34] data, displaying an estimated hazard ratio of 1.61 and 95% limits of 0.997, 2.59 (see Appendix for computational details). Following the common (and important) warning that *P*-values are *not* hypothesis probabilities, we caution that the *P*-value graph is *not* a probability distribution: It shows *compatibility* of parameter values with the data, rather than plausibility or probability of those values given the data. This is not a subtle difference: compatibility is a much weaker condition than plausibility. Consider for example that complete fabrication of the data is always an explanation *compatible* with the data (and indeed has happened in some influential medical studies [70]), but in studies with many participants and authors involved in all aspects of data collection it becomes so implausible as to not even merit mention. We emphasize then that all the *P*-value ever addresses in a direct logical sense is compatibility; for hypothesis probabilities one must turn to Bayesian methods [31].

The *P*-value graph rises past HR = 1 (no association, a parameter value which we have only plotted for demonstration purposes) until it peaks at the point estimate of 1.61, which coincides with the smallest *S*-value. The graphs show how rapidly the *P*-values fall and the *S*-values rise as we move away from the point estimate. CIs at the 75, 95, and 99% levels can be read off the graph as the range between the parameter values where the graph is above $P = 0.25$, 0.05, and 0.01. Both Figs. 2 and 3 illustrate how misleading it is to frame discussion in terms of whether *P* is above or below 0.05, or whether the null value is included in the 95% CI: Every hazard ratio from 1 to 2.58 is more compatible with the Brown et al. data according to the HDPS analysis, and has less information against it than does the null value of 1. Thus, the graphs illustrate how the Brown et al. analysis provides absolutely no basis for claiming the study found "no association." Instead, their analysis exhibits an association similar to that seen in earlier studies and should have been reported as such, even though it leaves open the question of what *caused* the association (e.g., a drug effect, a bias, a positive random error, or some combination) and whether a clinically important effect is present.

Discussion

We now discuss several basic issues in the use of the methods we have described. The Supplement discusses several more technical topics mentioned earlier and below: Different units for the *S*-value besides base-2 logs (bits); the importance of uniformity (validity) of the *P*-value for interpretation of the *S*-value; and the relation of the *S*-value to other measures of statistical information about a test hypothesis or model.

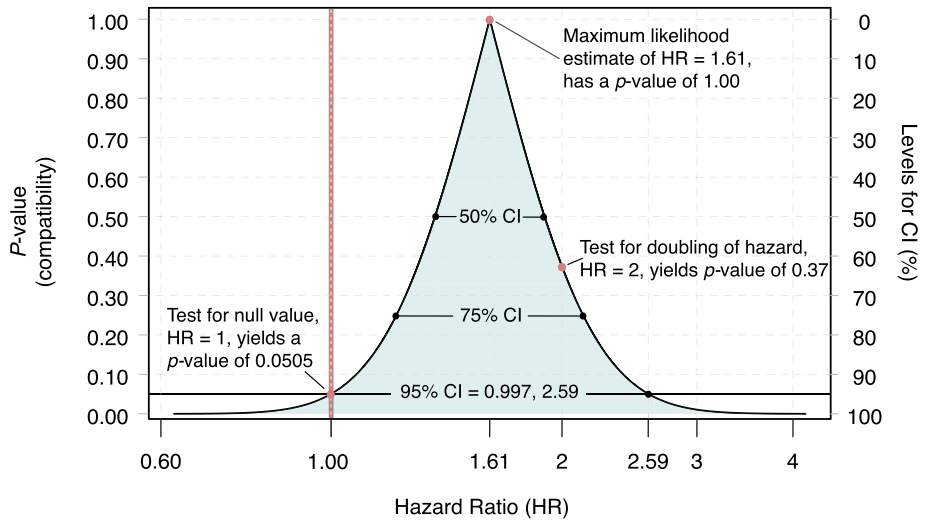

Fig. 2 *P*-Values for a range of hazard ratios (HR). A compatibility graph in which *P*-values are plotted across alternative hazard ratios. Computed from results in Brown et al. [34]. Compatibility intervals (CI) in percents can be read moving from the right-hand axis to the bottom (HR) axis. HR = 1 represents no association

Moving forward

Most efforts to reform statistical reporting have promoted interval estimates [19, 21] or Bayesian methods [26] over *P*-values. There is nonetheless scant empirical evidence that these or any proposals (including ours) have improved or will improve reporting without accompanying editorial and reviewer efforts to enforce proper interpretations. Instead, the above example and many others [31, 71, 72] illustrate how, without proper editorial monitoring, interval estimates are often of no help and can even be harmful when journals force dichotomous interpretations onto results, for example as does *JAMA* [73].

Cognitive psychology and its offshoot of behavioral economics (the "heuristics and biases" literature) have been studying misperceptions of probability for at least a half-century (e.g., see the anthologies of [74, 75]), with increasing attention to the harms of null-hypothesis significance testing (e.g., [2, 76]). Informal classroom observations on the devices we discuss have been encouraging

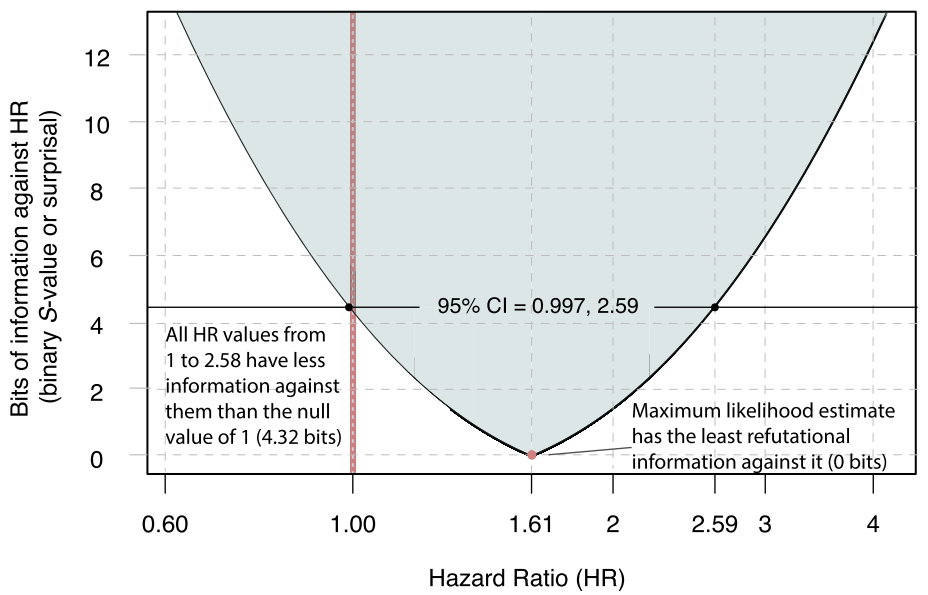

Fig. 3 *S*-Values (surprisals) for a range of hazard ratios (HR). An information graph in which *S*-values are plotted across alternative hazard ratios. Computed from results in Brown et al. [34]. HR = 1 represents no association

(both our own and those reported to us anecdotally by colleagues), leading to the present exposition.

We would thus encourage formal experiments to study cognitive devices like those we discuss. To justify such effort, the devices must be well-grounded in statistical theory (as reviewed in the prequel to this article [31]), and should be clearly operationalized, as the current article attempts to do. These preliminaries are especially important because prevailing practice is cemented into nearly a century of routine teaching and journal demands; thus, any comparison today will be dramatically confounded by tradition and exposure. Addressing this imbalance will require detailed instruction in the graphical information perspective, as illustrated here.

Tests of model fit

For simplicity we have focused on tests of specific hypotheses given a set of assumptions (the background model). The *S*-value can also be used to measure information against a data model, as supplied by the *P*-value from general tests of fit of a model to the data (such as the Pearson [44] chi-squared test of fit). In those tests, all deviations of the data from the model predictions contribute to lack of fit and are cumulated as evidence against the model. In yet another unfortunate misnaming, these tests have come to be called "goodness of fit" tests, when in fact the test statistics are measuring misfit (in Pearson's case, squared distances between the predictions and observations). The *P*-value accounts for the residual degrees of freedom for the misfit, but as discussed before, is scaled in a nonintuitive way: It shrinks to zero as misfit increases, even when misfit can increase indefinitely. The *S*-value restores the proper relation to the fit as seen in the original test statistic, where the cumulative information against the model growing larger without bound as misfit increases without bound.

Connections to Bayesian and information statistics

Our development has been based on conventional frequentist statistics, which focus on probabilities of various statistical observations (data features). There are several connections of *P*-values and compatibility intervals to Bayesian statistics, which are expressed in terms of hypothesis probabilities; for a basic review see [77]. These in turn lead to connections to *S*-values. Consider for example a one-sided *P*-value p for a directional hypothesis; under certain assumptions p is a lower bound on the posterior probability that the hypothesis is false, and the *S*-value $s = -\log_2(p)$ can be interpreted as the maximum surprisal in finding the hypothesis is false, given the data and assumptions. The Supplement describes a connection to Bayes factors, "safe testing", and testing by betting scores.

Some cautions

Demands for more statistical evidence against test hypotheses increase the need for numerical accuracy, especially because traditional normal Z-score (Wald) approximations (used by most software to derive *P*-values and compatibility intervals under nonlinear models) deteriorate as the *P*-value or α-level becomes smaller [78]. Adding that approximation error to the usual study uncertainties, we do not expect *P*-values below 0.001 from Z-scores to have more than 2-digit accuracy, and thus (outside of numeric illustrations) advise rounding *S*-values above $-\log_2(0.001) \approx 10$ to the nearest integer.

The *S*-values for testing the same hypothesis from K independent studies can be summed to provide a summary test statistic for the hypothesis (see Supplement). A caution is needed in that the resulting sum will have an expectation equal to K under the hypothesis and background assumptions. Thus its size must be evaluated against a distribution that increases with K (specifically, by doubling the sum and comparing it to a χ^2 distribution with $2K$ degrees of freedom) [31, 79].

As discussed in the Supplement, in Bayesian settings one may see certain *P*-values that are not valid frequentist *P*-values, the primary example being the posterior predictive *P*-value [80, 81]; unfortunately, the negative logs of such invalid *P*-values do not measure surprisal at the statistic given the model, and so are not valid *S*-values.

As mentioned earlier, one purpose of converting *P*-values to *S*-values is to thwart the fallacy of mistaking data probabilities like a *P*-value for hypotheses probabilities. It is often said that this fallacy is addressed by Bayesian methods because they give the reader hypothesis probabilities. A problem with such probabilities is that deriving them requires the analyst to supply a prior distribution ("prior") that supplies initial probabilities for competing hypotheses. In many serious applications, there is no simple, universal, and reliable guide to choosing a prior (other than as a shrinkage/penalization/regularization device to improve certain frequency properties), and thus posterior probability statements can vary considerably across analysts even when there is no disagreement about frequentist results [82]. That problem is precisely why frequentists reject Bayesian methods as a general foundation for data analysis.

In sharp contrast, frequency models for the data can be enforced by experimental devices, producing information that can be quantified even without agreement about a prior distribution for targeted quantities. This quantification does not preclude a further analysis which combines the experimental information with external information encoded in a penalty function or prior distribution (which may be partial [83]). Nor does it free data analysts from responsibility to weaken their interpretations when using methods derived from

devices or assumptions that are not known to be operative [33]. For example, explanations for results from randomization tests in nonrandomized studies must include not only treatment effects and random error among possible explanations, but also effects of randomization failure [84, 85].

Finally, we caution that Gelman and Carlin [86] refer to erroneously inferring the wrong sign of a parameter as "type-S error", an entirely different usage of "S".

Tests of different values for a parameter vs. tests of different parameters

Even if all background assumptions hold, no single number (whether a *P*-value, *S*-value, or point estimate) can by itself provide an adequate measure of sample information about a targeted parameter, such as a mean difference, a hazard ratio (HR), or some other contrast across treatment groups. We have thus formulated our description to allow the test hypothesis **H** to refer to different values for the same parameter. For example, **H** could be "HR = 1", the traditional null hypothesis of no change in hazard rate across compared groups; but **H** could just as well be "HR = 2", or "HR ≤ 2", or even "½ ≤ HR ≤ 2" [31]. In all these variations, the set of auxiliary assumptions (background model) used to compute the statistics stay unchanged; only **H** is changing. Unconditionally, the *S*-values for the different **H** are measuring information against different restrictions on HR beyond the background assumptions, which stay the same.

A similar comment applies when, in a model, we test different coefficients: The background assumptions are unchanged, only the targeted test hypothesis **H** is changing, although now the change is to *another parameter* (rather than another value for the same parameter). For example, in a model for effects of cancer treatments we might compute the *P*-value and *S*-value from a test of $\mathbf{H_r}$ = "the coefficient of radiotherapy is zero" and another *P*-value and *S*-value from a test of $\mathbf{H_c}$ = "the coefficient of chemotherapy is zero." Conditionally, these 2 *S*-values are giving information against different target hypotheses $\mathbf{H_r}$ and $\mathbf{H_c}$ using the same background model; for example, using a proportional-hazards model, that background includes the assumption that the effects of different treatments on the hazard multiply together to produce the total effect of all treatments combined. Unconditionally, these *S*-values are measuring information against different test models: a model with no effect of radiotherapy but allowing an effect of chemotherapy, and a model allowing an effect of radiotherapy but no effect of chemotherapy; all other assumptions are the same in both models (including possibly unseen and inappropriate assumptions about causal ordering [87]).

Testing different parameters with the same data raises issues of multiple comparisons (also known as simultaneous inference). These issues are very complex and controversial, with opinions about multiple-comparison adjustment ranging from complete dismissal of adjustments to demands for mindless, routine use, and extend far beyond the present scope; see [88, 89] for a recent commentary and review. We can only note here that the devices we recommend can also be applied to adjusted comparisons; for example, the *S*-value computed from an adjusted *P*-value becomes the information against a hypothesis penalized (reduced) to account for multiplicity.

We caution however against confusing the problem of testing multiple parameters with the testing of multiple values of the *same* parameter, as we recommend here: Tests of the same parameter are logically dependent in a manner eliminating the need for adjustment. This dependency can be seen in how a *P*-value for HR ≤ 1 must be less than the *P*-value for the less restrictive HR ≤ 2 (using a test derived from the same method and assumptions). Note also that a compatibility interval requires selection of values based on multiple tests of the parameter, namely the values for which $p > \alpha$; this selection does not harm any frequency property of the interval (e.g., coverage of the true parameter value at a rate $1 - \alpha$ if all background assumptions are correct).

Conclusion

Ongoing misinterpretations of important medical research demonstrate the need for simple reforms to traditional terms and interpretations. As lamented elsewhere, [29, 39, 57, 90], those traditions have led to overinterpretations and misinterpretations becoming standards of reporting in leading medical journals, with ardent defense of such malpractice by those invested in the traditions. Especially when there is doubt about conventional assumptions, overconfident terms like "significance," "confidence," and "severity" and decisive interpretations should be replaced with more cautiously graded unconditional descriptions such as "compatibility"; narrowly compressed probabilities like *P*-values can be supplemented with quantitative-information concepts like *S*-values; and requests can be made for tables or graphs of *P*-values and *S*-values for multiple alternative hypotheses, rather than forcing focus onto null hypotheses [31, 40, 61, 91]. These reforms need to be given a serious chance via editorial encouragement in both review and instructions to authors.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12874-020-01105-9.

Additional file 1: Appendix. Technical details for computations of figures and tables.

Additional file 2: Supplement. Technical issues in the interpretation of S-values and their relation to other information measures.

Additional file 3: Figure S1. Relative likelihoods for a range of hazard ratios. A relative likelihood function that corresponds to Fig. 2, the *P*-value function. Also plotted is the 1/6.83 likelihood interval (LI), which corresponds to the 95% compatibility interval. Computed from results in Brown et al. [34]. MLR = Maximum-Likelihood Ratio. HR = 1 represents no association.

Additional file 4: Figure S2. Deviance statistics for a range of hazard ratios. A deviance function, which corresponds to Fig. 3, the *S*-value function. Also plotted is the likelihood interval (LI), which corresponds to the 95% compatibility interval. Computed from results in Brown et al. [34]. MLR = Maximum-Likelihood Ratio. HR = 1 represents no association.

Abbreviations

ASD: Autism spectrum disorder; CI: Compatibility/confidence interval; HDPS: High-dimensional propensity score; HR: Hazard ratio; LI: Likelihood interval; LR: Likelihood ratio; MLR: Maximum-likelihood ratio; NHST: Null-hypothesis significance test; S-value: Surprisal (Shannon-information) value

Acknowledgements

We are most grateful for the generous comments and criticisms on our initial drafts offered by Andrew Althouse, Valentin Amrhein, Andrew Brown, Darren Dahly, Frank Harrell, John Ioannidis, Daniël Lakens, Nicole Lazar, Gregory Lopez, Oliver Maclaren, Blake McShane, Tim Morris, Keith O'Rourke, Kristin Sainani, Allen Schirm, Philip Stark, Andrew Vickers, Andrew Vigotsky, Jack Wilkinson, Corey Yanofsky, and the referees. We also thank Karen Pendergrass for her help in producing the figures in this paper. Our acknowledgment does not imply endorsement of our views by these colleagues, and we remain solely responsible for the views expressed herein.

Authors' contributions

Both authors (ZR and SG) wrote the first draft and revised the manuscript, read and approved the submitted manuscript, and have agreed to be personally accountable for their own contributions related to the accuracy and integrity of any part of the work.

Funding

This work was produced with no funding.

Availability of data and materials

The datasets generated and analyzed in the current paper are available in the Open Science Framework | DOI: https://doi.org/10.17605/OSF.IO/6W8G9 and on figshare | DOI: https://doi.org/10.6084/m9.figshare.9202211.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Population Health, NYU Langone Medical Center, 227 East 30th Street, New York, NY 10016, USA. [2]Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA, USA.

Received: 18 June 2020 Accepted: 25 August 2020
Published online: 30 September 2020

References

1. Greenland S. Invited commentary: the need for cognitive science in methodology. Am J Epidemiol. 2017;186:639–45.
2. Gigerenzer G. Mindless statistics. J Socio-Econ. 2004;33:587–606.
3. Stark PB, Saltelli A. Cargo-cult statistics and scientific crisis. Significance. 2018;15:40–3.
4. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci. 2011;22:1359–66.
5. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349:aac4716.
6. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biol. 2015;13:e1002165.
7. Camerer CF, Dreber A, Forsell E, Ho T-H, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. Science. 2016;351:1433–6.
8. Lash TL, Collin LJ, Van Dyke ME. The replication crisis in epidemiology: snowball, snow job, or winter solstice? Curr Epidemiol Rep. 2018;5:175–83.
9. Cassidy SA, Dimova R, Giguère B, Spence JR, Stanley DJ. Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. Adv Methods Pract Psychol Sci. 2019. https://doi.org/10.1177/2515245919858072.
10. Leek JT, Peng RD. Statistics: *P* values are just the tip of the iceberg. Nat News. 2015;520:612.
11. Lang JM, Rothman KJ, Cann CI. That confounded *P*-value. Epidemiology. 1998;9:7–8.
12. Pearson KV. Note on the significant or non-significant character of a sub-sample drawn from a sample. Biometrika. 1906;5:181–3.
13. Boring EG. Mathematical vs. scientific significance. Psychol Bull. 1919;16:335–8.
14. Tyler RW. What is statistical significance? Educ Res Bull. 1931;10:115–42.
15. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2017;2:6–10.
16. Lakens D, Adolfi FG, Albers CJ, Anvari F, Apps MAJ, Argamon SE, et al. Justify your alpha. Nat Hum Behav. 2018;2:168–71.
17. Lakens D, Scheel AM, Isager PM. Equivalence testing for psychological research: a tutorial. Adv Methods Pract Psychol Sci. 2018;1:259–69.
18. Mayo DG. Statistical inference as severe testing: how to get beyond the statistics wars: Cambridge University Press; 2018. https://doi.org/10.1017/9781107286184.
19. Rothman KJ. A show of confidence. N Engl J Med. 1978;299:1362–3.
20. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999;8:135–60.
21. Cumming G. Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis. Routledge. 2012. https://doi.org/10.4324/9780203807002.
22. Colquhoun D. The false positive risk: a proposal concerning what to do about p-values. Am Stat. 2019;73:192–201.
23. Goodman SN. Introduction to Bayesian methods I: measuring the strength of evidence. Clin Trials. 2005;2. https://doi.org/10.1191/1740774505cn098oa.
24. Held L. A new standard for the analysis and design of replication studies. J R Stat Soc Ser A Stat Soc. 2020;183:431–48.
25. Matthews RAJ. Moving towards the post p < 0.05 era via the analysis of credibility. Am Stat. 2019;73:202–12.
26. Sellke T, Bayarri MJ, Berger JO. Calibration of ρ values for testing precise null hypotheses. Am Stat. 2001;55:62–71.
27. Wang MQ, Yan AF, Katz RV. Researcher requests for inappropriate analysis and reporting: A U.S. survey of consulting biostatisticians. Ann Intern Med. 2018;169:554.
28. Gelman A. The problems with P-values are not just with P-values. Am Stat. 2016; 70 https://stat.columbia.edu/~gelman/research/published/asa_pvalues.pdf.
29. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567:305.
30. Greenland S. Are confidence intervals better termed "uncertainty intervals"? No: call them compatibility intervals. BMJ. 2019;366. https://doi.org/10.1136/bmj.l5381.
31. Greenland S. Valid P-values behave exactly as they should: some misleading criticisms of P-values and their resolution with S-values. Am Stat. 2019;73:106–14.
32. Cole SR, Edwards JK, Greenland S. Surprise! Am J Epidemiol. 2020; doi: https://doi.org/10/gg63md.
33. Greenland S, Rafi Z. To aid scientific inference, emphasize unconditional descriptions of statistics. ArXiv190908583 StatME. 2020; https://arxiv.org/abs/1909.08583.
34. Brown HK, Ray JG, Wilton AS, Lunsky Y, Gomes T, Vigod SN. Association between serotonergic antidepressant use during pregnancy and autism spectrum disorder in children. J Am Med Assoc. 2017;317:1544–52.
35. Brown HK, Hussain-Shamsy N, Lunsky Y, Dennis C-LE, Vigod SN. The association between antenatal exposure to selective serotonin reuptake inhibitors and autism: a systematic review and meta-analysis. J Clin Psychiatry. 2017;78:e48–58.

36. Yasgur B. Antidepressants in pregnancy: no link to autism. Medscape: ADHD; 2017. https://medscape.com/viewarticle/878948. Accessed 21 Aug 2019.
37. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995;311:485.
38. Amrhein V, Korner-Nievergelt F, Roth T. The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. PeerJ. 2017;5: e3544.
39. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon Statistical Significance. Am Stat. 2019;73:235–45.
40. Poole C. Beyond the confidence interval. Am J Public Health. 1987;77:195–9.
41. Rothman KJ. Significance Questing. Ann Intern Med. 1986;105:445–7.
42. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "$p < 0.05$". Am Stat. 2019;73:1–19.
43. Fisher RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1925. https://books.google.com/books?id=GmNAAAAAIAAJ&q.
44. Pearson KX. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Lond Edinb Dublin Philos Mag J Sci. 1900;50:157–75.
45. Stigler SM. Attempts to Revive the Binomial. In: The history of statistics: the measurement of uncertainty before 1900: Harvard University Press; 1986. https://books.google.com/books?id=M7yvkERHIIMC.
46. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31:337–50.
47. Perezgonzalez JD. *P*-values as percentiles. Commentary on: "Null hypothesis significance tests. A mix–up of two different theories: the basis for widespread confusion and numerous misinterpretations". Front Psychol. 2015;6. https://doi.org/10.3389/fpsyg.2015.00341.
48. Vos P, Holbert D. Frequentist inference without repeated sampling. ArXiv190608360 StatOT. 2019; https://arxiv.org/abs/1906.08360.
49. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27:379–423.
50. Good IJ. The surprise index for the multivariate normal distribution. Ann Math Stat. 1956;27:1130–5. Corrigendum Ann Math Stat. 1957;28:1055.
51. Cousins RD. The Jeffreys–Lindley paradox and discovery criteria in high energy physics. Synthese. 2017;194:395–432.
52. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. Genet Epidemiol. 2008;32:227–34.
53. Hand DJ. The improbability principle: why coincidences, miracles, and rare events happen every day: Macmillan; 2014. https://books.google.com/books?id=raZRAQAAQBAJ.
54. Bowley AL. Discussion on Dr. Neyman's Paper. P. 607–610 in: Neyman J. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection (with discussion). J R Stat Soc. 1934;4:558–625.
55. Cox DR, Hinkley DV. Chapter 7, interval estimation. In: Theoretical Statistics: Chapman and Hall/CRC; 1974. p. 207–49. https://doi.org/10.1201/b14832.
56. Cox DR. Principles of statistical inference: Cambridge University Press; 2006. https://doi.org/10.1017/cbo9780511813559.
57. Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. Am Stat. 2019;73:262–70.
58. Poole C. Confidence intervals exclude nothing. Am J Public Health. 1987;77:492–3.
59. Gelman A, Stern H. The difference between "significant" and "not significant" is not itself statistically significant. Am Stat. 2006;60:328–31.
60. Birnbaum A. A unified theory of estimation. I Ann Math Stat. 1961;32:112–35.
61. Sullivan KM, Foster DA. Use of the confidence interval function. Epidemiology. 1990;1:39–42.
62. Rothman KJ, Greenland S, Lash TL. Precision and statistics in epidemiologic studies. In: Rothman KJ, Greenland S, Lash TL, editors. Modern Epidemiology. 3rd edition: Lippincott Williams & Wilkins; 2008. p. 148–67. https://books.google.com/books/about/Modern_Epidemiology.html?id=Z3vjT9ALxHUC.
63. Rafi Z, Vigotsky AD. concurve: Computes and Plots Consonance (Confidence) Intervals, P-Values, and S-Values to Form Consonance and Surprisal Functions. R. CRAN; 2020. https://cran.r-project.org/package=concurve.
64. Rücker G, Schwarzer G. Beyond the forest plot: the drapery plot. Res Synth Methods. 2020. https://doi.org/10.1002/jrsm.1410.
65. Fraser DAS. The *P*-value function and statistical inference. Am Stat. 2019;73: 135–47.
66. Whitehead J. The case for frequentism in clinical trials. Stat Med. 1993;12: 1405–13.
67. Xie M, Singh K. Confidence distribution, the Frequentist distribution estimator of a parameter: a review. Int Stat Rev. 2013;81:3–39.
68. Singh K, Xie M, Strawderman WE. Confidence distribution (CD) – distribution estimator of a parameter; 2007.
69. Schweder T, Hjort NL. Confidence, likelihood, probability: statistical inference with confidence distributions: Cambridge University Press; 2016. https://books.google.com/books/about/Confidence_Likelihood_Probability.html?id=t7KzCwAAQBAJ.
70. Rubenstein S. A new low in drug research: 21 fabricated studies. WSJ. 2009; https://blogs.wsj.com/health/2009/03/11/a-new-low-in-drug-research-21-fabricated-studies/.
71. Schmidt M, Rothman KJ. Mistaken inference caused by reliance on and misinterpretation of a significance test. Int J Cardiol. 2014;177:1089–90.
72. Greenland S. A serious misinterpretation of a consistent inverse association of statin use with glioma across 3 case-control studies. Eur J Epidemiol. 2017;32:87–8.
73. Bauchner H, Golub RM, Fontanarosa PB. Reporting and interpretation of randomized clinical trials. J Am Med Assoc. 2019;322:732–5.
74. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science. 1974;185:1124–31.
75. Gilovich T, Griffin D, Kahneman D. Heuristics and biases: the psychology of intuitive judgment: Cambridge University Press; 2002. https://books.google.com/books/about/Heuristics_and_Biases.html?id=FfTVDY-zrCoC.
76. Gigerenzer G, Marewski JN. Surrogate science: the idol of a universal method for scientific inference. J Manag. 2015;41:421–40.
77. Greenland S, Poole C. Living with p values: resurrecting a Bayesian perspective on frequentist statistics. Epidemiology. 2013;24:62–8.
78. Greenland S, Rothman KJ. Fundamentals of epidemiologic data analysis. In: Rothman KJ, Greenland S, Lash TL, editors. Modern Epidemiology. 3rd edition: Lippincott Williams & Wilkins; 2008. p. 213–37. https://books.google.com/books/about/Modern_Epidemiology.html?id=Z3vjT9ALxHUC.
79. Cox DR, Hinkley DV. Chapter 3, pure significance tests. In: Theoretical Statistics: Chapman and Hall/CRC; 1974. p. 64–87. https://doi.org/10.1201/b14832.
80. Bayarri MJ, Berger JO. P values for composite null models. J Am Stat Assoc. 2000;95:1127–42.
81. Robins JM, van der Vaart A, Ventura V. Asymptotic distribution of P values in composite null models. J Am Stat Assoc. 2000;95:1143–56.
82. Stark PB. Constraints versus priors. SIAMASA J Uncertain Quantif. 2015;3:586–98.
83. Cox DR. A note on partially Bayes inference and the linear model. Biometrika. 1975;62:651–4.
84. Greenland S. Randomization, statistics, and causal inference. Epidemiology. 1990;1:421–9.
85. Greenland S, Robins J. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol. 1986;15:413–9.
86. Gelman A, Carlin J. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspect Psychol Sci. 2014;9:641–51.
87. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am J Epidemiol. 2013;177:292–8.
88. Greenland S, Hofman A. Multiple comparisons controversies are about context and costs, not frequentism versus Bayesianism. Eur J Epidemiol. 2019. https://doi.org/10.1007/s10654-019-00552-z.
89. Greenland S. Analysis goals, error-cost sensitivity, and analysis hacking: essential considerations in hypothesis testing and multiple comparisons. Ped Perinatal Epidemiol. 2020; in press.
90. McShane BB, Gal D. Statistical significance and the dichotomization of evidence. J Am Stat Assoc. 2017;112:885–95.
91. Folks L. Ideas of statistics: Wiley; 1981. https://books.google.com/books/about/Ideas_of_statistics.html?id=Bn8pAQAAMAAJ.

Publisher's Note