The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://tandfonline.com/loi/utas20

Valid *P*-Values Behave Exactly as They Should: Some Misleading Criticisms of *P*-Values and Their Resolution With *S*-Values

Sander Greenland

To cite this article: Sander Greenland (2019) Valid *P*-Values Behave Exactly as They Should: Some Misleading Criticisms of *P*-Values and Their Resolution With *S*-Values, The American Statistician, 73:sup1, 106-114, DOI: 10.1080/00031305.2018.1529625

To link to this article: https://doi.org/10.1080/00031305.2018.1529625

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 106–114: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1529625

Valid *P*-Values Behave Exactly as They Should: Some Misleading Criticisms of *P*-Values and Their Resolution With *S*-Values

Sander Greenland

Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA

Abstract

The present note explores sources of misplaced criticisms of *P*-values, such as conflicting definitions of "significance levels" and "*P*-values" in authoritative sources, and the consequent misinterpretation of *P*-values as error probabilities. It then discusses several properties of *P*-values that have been presented as fatal flaws: That *P*-values exhibit extreme variation across samples (and thus are "unreliable"), confound effect size with sample size, are sensitive to sample size, and depend on investigator sampling intentions. These properties are often criticized from a likelihood or Bayesian framework, yet they are exactly the properties *P*-values *should* exhibit when they are constructed and interpreted correctly within their originating framework. Other common criticisms are that *P*-values force users to focus on irrelevant hypotheses and overstate evidence against those hypotheses. These problems are not however properties of *P*-values but are faults of *researchers* who focus on null hypotheses and overstate evidence based on misperceptions that $p = 0.05$ represents enough evidence to reject hypotheses. Those problems are easily seen without use of Bayesian concepts by translating the observed *P*-value p into the Shannon information (*S*-value or surprisal) $-\log_2(p)$.

ARTICLE HISTORY
Received March 2018
Revised September 2018

KEYWORDS
Compatibility; Dichotomania; Evidence; Information; Logworth; Nullism, *P*-values; S-values; Significance testing; Surprisal

1. Introduction

There are many reasons to be critical of traditional testing (which tethered *P*-values to decision rules); nonetheless, they provide no basis for blaming *P*-values for behaving as they were designed to do, or for their misuse and misinterpretation (Benjamini 2016). As has been argued before (e.g., Senn 2001, 2002), the problems are instead a failure of textbooks and tutorials to describe correctly the inferential meaning of *P*-values, and a failure to describe test hypotheses appropriate for practical needs. There is also an egregious tendency to blame users for these problems, even though the statistics literature up to the highest levels displays descriptions and definitions founded on jargon that is inconsistent across sources and which violate ordinary language meanings. These problems require far more expertise to sort out than ordinary users could be reasonably expected to have, and set the stage for confusions and misinterpretations such as those discussed by Gigerenzer (2004), Hoekstra et al. (2006), Hurlbert and Lombardi (2009), Greenland et al. (2016), Amrhein et al. (2017), McShane et al. (2017, 2018), Wasserstein and Lazar (2016), and many other sources.

The present note describes how authoritative definitions of "significance levels" and "*P*-values" vary, leading to misinterpretations of *P*-values as error probabilities. It then discusses several criticisms of *P*-values that are instead problems of poor teaching and terminology, and several properties of *P*-values that are often presented as fatal flaws but reflect instead how valid *P*-values should behave. The underlying view is that, while *P*-values are certainly limited in scope and difficult to understand properly, many strident criticisms judge these statistics according to inappropriate criteria or in comparison to methods that are subject to parallel criticisms. Especially pernicious are criticisms that overlook how demands for conclusive inferences can undermine any method, including confidence intervals and Bayesian statistics.

Concepts will be illustrated with a cohort study which compared adverse event rates among infants receiving ibuprofen (Advil™, Motrin™), alone or in addition to other drugs, to rates among those receiving only acetaminophen (paracetamol, Tylenol™) (Walsh et al. 2018). A highlighted result for renal (kidney) adverse events was an adjusted estimated rate ratio (*RR*) of 1.84 with 95% confidence limits of 0.66, 5.19, which correspond to an event-rate increase of 84% for ibuprofen compared to acetaminophen alone, an interval ranging from a 34% decrease to a 419% increase, and a *P*-value of 0.25 for testing the null hypothesis of no association (rate ratio of 1, a 0% difference). Theoretical descriptions will focus on testing of a hypothesis **H** about a coefficient β embedded in a regression model **A** (which encodes the set of background assumptions used for the test), such as **H**: $\beta = b$ where b is a fixed, hypothesized value for β; in multiplicative rate models $\beta = \ln(RR)$. The tested value b is usually *but need not be* zero. The discussion will also consider the more general case of tests of fit of the embedding model **A**.[1] For simplicity, however, I will assume the dataset is large enough so that all the usual continuous

[1] This case is more general in that a test of $\beta = b$ given **A** can be treated as a test of the fit of the model in which $\beta = b$ holds relative to the model in which β is unconstrained but the embedding model **A** holds.

CONTACT Sander Greenland *lesdomes@ucla.edu* Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA

large-sample approximations hold. The most technical asides will be in footnotes. As will be described, some of the problems in understanding *P*-values as evidence measures can be reduced by converting the observed *P*-value p to an information measure such as the *S*-value $s = \log_2(1/p) = -\log_2(p)$.

2. The Confusing and Unacknowledged Variation in Basic Definitions and Descriptions

Some problems can be traced to variation in definitions and terminology across authoritative sources with no mention of the variation by those sources. At least two definitions of "the observed *P*-value" are in wide use. In the usual inferential (Fisherian) definition, a *P*-value is the tail probability p under **H** that a test statistic (such as an absolute *Z*-score or χ^2) would be as large or larger than what was observed, given the embedding model **A** (Cox and Donnelly 2011, sec. 8.4). Fisher (1925) called p the "significance level" or "value of *P*," and "significance level" is common in British sources thereafter (e.g., Cox and Hinkley 1974). In the decision-theoretical (Neyman–Pearsonian) definition, the observed p is often defined as the smallest α level (testing cutoff) that would allow rejection in an α-level decision rule (Neyman–Pearson hypothesis test) which rejects **H** when $p \leq \alpha$ (e.g., Lehmann 1986). While the two definitions appear superficially distinct, they are mathematically equivalent and thus represent logically a single definition stated in two different ways.

Unfortunately, some authors use "significance level" to refer to α rather than p (Lehmann 1986, p. 70). This second usage of "significance level" contradicts the original usage and leads to confusion of p and α, often in very subtle ways (as discussed in the next section). In the face of such inconsistent usage among authoritative texts, it should hardly be surprising when basic textbooks as well as researchers confuse p and α. The observed p is a sample feature relating the observed data to the hypothesis **H** and model **A** used to compute p. Specifically, $100p$ is the percentile location of the observed test statistic in a distribution computed from **H** and **A** (Perezgonzalez 2015), and in this special sense p can be considered as describing something about the data. In contrast, α is a fixed known number (like 0.05) that tells us nothing about the data.

Despite the common definition of "*P*-value" as a probability, many decision-theoretical authors instead focus only on repeated-sampling properties, and thus define a "*P*-value" not as the observed value p, but instead as the random variable P whose value (realization) in a given sample is the observed p (e.g., Kuffner and Walker 2017; Murdoch et al. 2008). Thus, we have two logically distinct definitions of "*P*-value." But, as with the conflicting terminology, this conflict in definitions is rarely noted—so it should be hardly be surprising when basic texts and researchers confuse the random P with the observed p, which is similar to confusing the name of a variable with an unspecified value for it (e.g., confusing the variable X called "weight in kg" with the unspecified value "x kg" that one might observe upon weighing someone).

The distinction between the two definitions is important, not the least because frequentists further define validity of the *P*-value in terms of the random P: The random variable P is said to be *valid for testing* **H** given **A** (or properly calibrated, or a U-value) if it is uniformly distributed when **H** and **A** are correct; in that case, for every α the rule "reject **H** when $p \leq \alpha$" will falsely reject **H** with frequency α (Bayarri and Berger 2004).[2] Of course, issues of power must enter into choice of the test statistic from which p is computed, but those issues are outside the present scope.

Confusion of the observed *P*-value p with the random variable P may be a major contributor to some of the fallacies described below and earlier (e.g., Greenland et al. 2016). Worse, however, is that the tail-probability definition of the observed p is often equated to or replaced by wholly incorrect descriptions such as "p is the probability that chance alone produced the association," which reflects confusion beyond mere terminology (since the probability of chance alone is none other than the probability that **H** and **A** are correct; see item no. 2 in Greenland et al. 2016). Setting such outright errors aside, more subtle problems arise when the observed p is described as measuring evidence against **H**, because it is inversely related to that evidence: *smaller* values of p correspond to *more* evidence against **H** in the data, given the model **A**.

In an attempt to forestall misinterpretations, p can be described as a measure of the degree of statistical compatibility between **H** and the data (given the model **A**) bounded by 0 = complete incompatibility (data impossible under **H** and **A**) and 1 = no incompatibility apparent from the test (Greenland et al. 2016). Similarly, in a test of fit of **A**, the resulting p can be interpreted as a measure of the compatibility between **A** and the data.[3] The scaling of p as a measure is poor, however, in that the difference between (say) 0.01 and 0.10 is quite a bit larger geometrically than the difference between 0.90 and 0.99. For example, using a test statistic that is normal with mean zero and standard deviation (SD) of 1 under **H** and **A,** a p of 0.01 vs. 0.10 corresponds to about a 1 SD difference in the statistic, whereas a p of 0.90 vs. 0.99 corresponds to about a 0.1 SD difference.

One solution to both the directional and scaling problems is to reverse the direction and rescale *P*-values by taking their negative base-2 logs, which results in the *S*-value $s = -\log_2(p)$. Larger values of s do correspond to more evidence against **H**. As discussed below this leads to using the *S*-value as a measure of evidence against **H** given **A** (or against **A** when p is from a test of fit of **A**).

3. A *P*-value is Not an Error Probability (Except in a Useless Hypothetical Sense)

Describing or defining an observed *P*-value p as a minimum α-level for rejection along with references to both p and α as "significance levels" seem to have led to p being misinterpreted as an error probability for actual decision problems. The latter interpretation is just a mistake (Goodman 1999; Sellke et al. 2001; Hubbard and Bayarri 2003; see also Greenland et al. 2016, item 9): The theory of α-level hypothesis tests requires an α

[2] In parallel, *P* is *conservatively valid* if the rule "reject **H** when $p \leq \alpha$" will reject **H** with frequency no greater than α when **H** and **A** are correct; conservative validity is usually the best one can do with discrete data.

[3] One could also say "p measures the compatibility of **A** with the data" when the subject is **A**, or "p measures the compatibility of the data with **A**" when the subject is the data.

that is a constant (such as 0.05) specified by the analyst *before* seeing the data as the desired upper bound for the error rate of a rejection rule over an entire sequence of datasets generated from the model (Neyman 1977; Lehmann 1986). Specifically, α is determined independently of the data, based on the cost of false rejections: Higher α would be used by those for whom false positives are of minor consequence, and lower α would be used by those for whom false positives are of major consequence (Lakens et al. 2018). Since decision consequences could vary for different stakeholders in the same setting, different readers of the same study report may well have different losses from false positives and false negatives, and thus different α-levels, in turn leading to conflicting decisions based on seeing the same data and thus the same observed *P*-value p.

Nonetheless, no such α is needed for the definition or presentation of *P*-values, nor does comparison of p to α make the number p an error rate. We can imagine a decision rule that has an error rate α equal to the observed p, but α and p are different quantities conceptually since p varies across samples and thus cannot be prespecified. In the example, the rate ratio of 1.84 for renal adverse events corresponds to an estimated ibuprofen coefficient of $\hat{\beta} = \ln(1.84) = 0.610$ in an exponential-rate model, with standard error $\sigma = \ln(5.19/0.66)/2(1.96) = 0.526$ and test statistic for **H**: $\beta =$ of $\hat{\beta}/\hat{\sigma} = 1.159$, yielding a *P*-value p of 0.25. We can imagine a rule defined by "reject $\beta = 0$ if its *P*-value is less than $\alpha = 0.25$," but no one stated that rule *before* seeing the data let alone derived it from error costs. To describe the observed p as a 25% Type-I error rate is thus at best a statement referring to someone who had a prespecified α of 0.25 which the *P*-value just happened by chance to equal. There is surely no such person in the example, and thus the statement is a completely unnecessary distraction: An accurate description of the test result is simply that p for $\beta = 0$ was 0.25. Readers who wanted to base a decision on that *P*-value alone could immediately see whether 0.25 was above or below their own cutoff α.

A small related point is the common confusion of the α-level α with the actual probability that the test rejects **H** when **H** is true (the Type-I error rate of the test, or test size). The actual Type-I error rate of a test of the hypothesis **H** given the assumptions **A** is often unknown to the investigator, because it may deviate from α due to defects in **A** or discreteness of the data.[4] In contrast, α is defined as the maximum tolerable Type-I error rate, which is set by the investigator and thus is known; p is then compared to this α to make decisions, on the assumption that the corresponding random *P* is valid (which makes α equal to the Type-I error rate).

4. Misleading Criticisms of *P*-values

4.1. *P*-Values Do Not Force Users to Focus on Null Hypotheses—But Nullistic Jargon Does

A common criticism of statistical testing is that it forces users to focus on irrelevant null hypotheses. There is no question that many null hypotheses are indeed scientifically irrelevant (Cohen 1994). This irrelevancy problem is not, however, a fault of *P*-values, but is instead a product of traditional training and an academic environment that makes users focus on such hypotheses. While there are ongoing calls to abolish misleading jargon involving "significance" (Amrhein et al. 2018; McShane et al. 2018), almost no attempt has been made to correct Fisher's mistake of using "null hypothesis" for any tested hypothesis **H**, ignoring that in ordinary English "null" is a synonym for zero or nothing.[5] This tradition has led users to think and certain experts to claim that statistical science is only about testing "null hypotheses," where "null" means "no association" or "no effect" rather than any and all hypotheses of importance or concern (Greenland 2004, 2017).

Breaking from Fisher's misleading terminology, Neyman (1977, pp. 104–106) instead called **H** the targeted or *tested* hypothesis. But Fisher's jargon has prevailed with the back-rationalization that **H** is the hypothesis to be "nullified" by the test, and was sustained by attempts to distinguish hypotheses like $\beta = 0$ as "nil hypotheses" (Cohen 1994). Unsurprisingly, then, the ill effects of Fisher's usage continue to be seen in the justly maligned cult of null-hypothesis significance testing (NHST) (Ziliak and McCloskey 2008) in which *P*-values are computed only for hypotheses of no effect, when they should also be given for alternatives of relevance (e.g., the hypotheses used to compute power in funding applications).

A more technically involved problem is excessive focus on point hypotheses, which will be discussed below under sensitivity to sample size.

4.2. *P*-Values Are "Unreliable"—Exactly as They Should Be

It is often noted that *P*-values vary dramatically from sample to sample even under ideal experimental replications (Goodman 1992; Senn 2001,2002; Gelman and Stern 2006; Murdoch et al. 2008; Boos and Stefanski 2011). As a consequence, some researchers criticize *P*-values for being "unreplicable" or "unreliable" or "noisy," or for failing to converge to some constant upon replication or increase in sample size, as if a *P*-value is estimating a parameter or "statistical significance" is a state of nature.

Parameter measurement is not, however, done by the *P*-value, but instead by the estimator $\hat{\beta}$: This estimator and its standard error extract information on the systematic component of data variation represented by β in the model **A**. The observed p is computed from the remaining variation, as captured by the absolute *Z*-score $|\hat{\beta} - b|/\hat{\sigma}$ for testing $\beta = b$; this test statistic can be viewed as a summary standardized residual for comparing the submodel of **A** in which **H** is correct ($\beta = b$) to the embedding model **A** which has no restriction on β. Upon recognizing that a *P*-value is a function of what would be purely random error under **H** and **A**, it should be seen that *by definition* it does not measure any parameter or population quantity—quite the opposite: the *P*-value is a measure of random error in the estimate, *given* **H** and **A**, rescaled (standardized) to a uniform distribution. Similarly, a *P*-value for a test of fit of the embedding model **A** may be based on a residual sum of squares which is a rescaled measure of the noise or random

[4] When the test is a conservatively valid test of the entire model (assumption set) **A**, discreteness becomes the only source of discrepancy.

[5] See, e.g., "null" in Oxford 2017: adj. 2. Having or associated with the value zero; noun 1. Zero. Merriam-Webster 2017: adj. 6. of, being, or relating to zero; 7. zero.

error left after extracting what the model **A** says is the signal or systematic variation (the fitted regression equation or table of fitted values, which is the signal remaining after the noise is filtered out).

Thus, if **H** and **A** are correct, the random P will (if valid) bounce around uniformly across the unit interval, producing $p \leq \alpha$ about $100\alpha\%$ of the time. This is exactly what P should do, because in that case it should be pure uniform noise (random error) which is supposed to vary unpredictably from study to study (Senn 2001, 2002). If across replications of data collection and analysis P does *not* look like uniform noise, we are warned that at least one of **H** or **A** is incorrect, i.e., we are using a wrong hypothesis or wrong model for inference (signal extraction) or decision. That is exactly what statistical tests are for. If instead P is uniform, we can only say that this particular test is insensitive to whatever violations of **H** or **A** are present (in decision-theoretical terms: for these violations, the test based on P has power equal to its Type-I error rate, meaning it conveys no information about those violations). Thus, P-values provide diagnostic or warning mechanisms for hypothesis or model problems, and like all such mechanisms are fallible.

4.3. P-Values Confound Effect Size with Sample Size—Exactly as They Should

The concept of "statistical compatibility" has been a major obstacle to proper understanding and interpretation of P-values, as the concept involves comparing the magnitude of discrepancies between observations and expectations against the estimated random variability in those discrepancies. It is in fact often lamented that P-values confound effect size with sample size, and they are even banned on account of this (Lang et al. 1998). That is ironic because this "confounding" reflects how P-values are doing their job correctly: The distance of (say) an estimate from a model prediction should constitute evidence against the model; nonetheless, *how much* evidence that distance corresponds to should depend on the precision of the estimate.

In particular, the information against **H**: $\beta = b$ given the model **A** is a function of both the absolute deviation (distance from estimate to hypothesized value) $|\hat{\beta} - b|$ and the standard error $\hat{\sigma}$ where $\hat{\beta}$ and $\hat{\sigma}$ correspond to the imputed signal and noise level based on the auxiliary (decoding) assumptions in **A**. These quantities are combined in the test statistic $|\hat{\beta} - b|/\hat{\sigma}$ for **H** given **A**. Because the precision $\hat{\sigma}^{-2}$ of $\hat{\beta}$ is proportional to the sample size n, the test statistic depends directly on n and on the distance $|\hat{\beta} - b|$ from the estimate $\hat{\beta}$ to the hypothesized value b, and so p is inversely related to both n and $|\hat{\beta} - b|$, exactly as a consistent measure of evidence against $\beta = b$ should be. In the example, for **H**: $\beta = 0$ the distance is $|\hat{\beta} - 0| = 0.610$, again with a p of 0.25; if, however, the precision of the study were increased fourfold (e.g., by quadrupling the size n of the cohort), $\hat{\sigma}$ would be halved and the same distance of 0.610 would give a p of 0.02, reflecting the fact that evidence against **H** represented by a given deviation *should* increase with n.

4.4. P-Values Do not Overstate Evidence Against Hypotheses—People Do

Among fair criticisms of P-values are that they are too easily confused with posterior probabilities, and that they are distortive evidence measures that need logarithmic transformation to gauge properly (e.g., Bayarri and Berger 1999; Boos and Stefanski 2011; Greenland 2017, 2018). Thus, consider again the Shannon-information or S-value transform of the observed P-value, $s = -\log_2(p)$, which is a measure of the information against **H** encoded in the test statistic (the refutational information supplied by the test given the model **A**).[6] The negative log transform of a probability is also known as the *self-information* or *surprisal* for observing an event that has probability p (Shannon 1948; MacKay 2003; Fraundorf 2017). With base-2 logs, the units for measuring this information are *bits* (binary digits); the first integer larger than s is the number of binary digits (indicator variables) needed to encode p.[7]

There are other formal definitions of statistical information, but the S-value s is a simple cognitive device for appreciating the information conveyed by p without reference to contextual details outside of those used to specify **H** and **A**. Notably, the information measure s refers to the observed tail probability p (rather than the random P) and thus represents a relation between the observed data and the model formed by combining **H** with **A**. A useful consequence of its log scaling is that it makes the information additive across independent tests, a fact used to create meta-analytic P-values (Cox and Hinkley 1974, p. 80). To provide an intuitive interpretation of the information conveyed by s, let k be the nearest integer to s. We may then say that p conveys roughly the same information or evidence against the tested hypothesis **H** given **A** as seeing all heads in k independent tosses of a coin conveys against the hypothesis that the tosses are "fair" (each independent with chance of heads $= 1/2$) versus loaded for heads; k indicator variables all equal to 1 would be needed to represent this event.

As an illustration, the chance of seeing all heads in 4 fair tosses is $1/2^4 = 0.0625$. Thus, under the model **A**, observing a P-value of 0.05 conveys only $s = -\log_2(0.05) = 4.3$ bits of information against **H**: $\beta = b$, which is hardly more surprising than seeing all heads in 4 fair tosses. In the ibuprofen example, the P-value of 0.25 is no more surprising if $\beta = 0$ than seeing 2 heads in 2 fair tosses, since $s = -\log_2(0.25) = 2$. For contrast, $\beta = \ln(2)$ corresponds to a doubling of the adverse-event rate with ibuprofen; the P-value from the statistic $|\hat{\beta} - \ln(2)|/\hat{\sigma} = 0.158$ is 0.87, for which $s = -\log_2(0.87) = 0.19$, showing that there is even less information against a doubling of the adverse-event rate with ibuprofen than against no difference ($\beta = 0$).

By this measure, and contrary to certain commentaries (e.g., Goodman 1999; Sellke et al. 2001; Hubbard and Lindsay 2008), P-values do *not* overstate evidence against hypotheses or models: The observed p is just a hypothetical probability or percentile computed from **H** and **A**. Any overstatement of the evidence conveyed by p is from those who (based on the entrenchment of $\alpha = 0.05$ in automated decision rules) mistakenly think a p of 0.05 represents just enough evidence to reject the tested hypothesis or model. Furthermore, under this interpretation any changes in the P-value p and thus the S-value s as a result

[6] Good 1956, pp. 1132; 1983, 146 suggested using this measure with centering around its mean $E\{-\log_2(P)\}$, the *Shannon entropy* of P (which for valid P is itself is maximized when **H** and **A** are correct).

[7] In base-10 logs the units are called *Hartleys*; when **H**: $\beta = 0$ this measure is sometimes called the *logworth* of β. The first integer larger than $-\log_{10}(p)$ is the number of decimal digits needed to encode p.

of changing the embedding model or tested hypothesis (e.g., from changing from a single to a multiple comparison) correctly reflects the differences in the information supplied by the data against the different models or hypotheses. Yet, the *S*-value also corroborates the commentaries by revealing that a p of 0.05 maps into only 4.3 bits of refutational information, showing that interpreting a p of 0.05 as "borderline evidence" is nothing more than a bad cultural habit. This is an underlying cognitive issue which can be seen without use of machinery such as prior spikes or Bayes factors (which even some Bayesian sympathizers find objectionable, e.g., Casella and Berger 1987ab; Gelman 2013; Greenland and Poole 2013).

4.5. P-values are Sensitive to Sample Size—Exactly as They Should Be

Using the *S*-value $s = -\log_2(p)$ to measure information in a test statistic, *P*-values obey the refutational version of the *P*-postulate (less accurately termed "the α postulate"): Equal *P*-values correspond to equal refutational information against the tested hypothesis **H** or the tested model **A** (Royall 1986). This property is often criticized, e.g., because two studies may provide the same *P*-values for the same test hypothesis and yet may show very different observed associations (Greenland et al. 2016, point 17). This criticism points to the necessity of considering additional background (contextual) information about what violations of **H** are of practical importance. A related criticism is that the embedding model **A** consists of all auxiliary assumptions used for inference, including explicit assumptions such as homogeneity of effects and random sampling or treatment randomization, and less-often stated assumptions such as no database errors or selective analysis reporting. As such, **A** is never perfectly correct; hence, for large enough samples p will become very small and thus s will become very large even if **H** is correct.

It follows that the test may now indicate that at least one or both of **H** and **A** are strictly false even if both are good enough for practical purposes; thus the criticism could be restated as "In large samples, *P*-values become too sensitive to small deviations from **H** or **A**." Bayesians extend this criticism by placing a point mass of prior probability on **H** and spreading the remainder over a restricted family of alternatives (usually the family of all models with β a known constant in the same embedding model **A**). A consequence of this artifice is that data supplying considerable information against **H** according to some criterion α for p or $-\log(\alpha)$ for s may still increase the posterior probability of **H** given **A** (the Jeffreys–Lindley paradox; see, e.g., Royall 1997; Senn 2001; Spanos 2013).

A crude fix for this large-sample sensitivity of *P*-values to unimportant discrepancies is to lower the α-level for rejecting **H** as the sample-size increases (Royall 1986). This fix is, however, only a demand for more information to reject **H** or **A** as the sample size increases, which ignores actual error costs on which α should be based. Furthermore, these costs and demands are irrelevant to measuring the statistical information given by the test. In fact, any valid and efficient test *should* detect model imperfections when given enough data information, even if those are of no practical consequence. Instead the defect lies in the traditional focus on point test hypotheses like $\beta = b$ (Casella and Berger 1987ab), and a failure to pay due attention to the size of the observed discrepancy $|\hat{\beta} - b|$ in practical terms.

Accounting for practical importance instead requires specification of tolerances for imperfections, such as a maximum tolerance for the actual discrepancy $|\beta - b|$. One way to do so is to replace point targets such as **H**: $\beta = b$ with interval targets such as **H**: $|\beta - b| \leq c$, where $[-c, c]$ represents an interval deemed practically equivalent to no discrepancy ($\beta = b$). It may even make more sense contextually to reverse the role of the test and alternative hypothesis, so that the tested **H** becomes $|\beta - b| \geq c$, as in equivalence testing (Berger and Hsu 1996; Senn 2008; Wellek 2010).[8] Other examples of role reversal include risk-limiting audits (Lindeman and Stark 2012). In all these cases, the observed *P*-value now measures data compatibility with the composite hypothesis **H** given **A**, and the *S*-value measures information against that **H** given **A**. Upon recognizing these hypotheses as more contextually relevant than $\beta = b$, sensitivity to sample size ceases to be a valid objection to *P*-values; it is rather an objection to point test hypotheses, reminding us that all our models have imperfections that become noticeable with enough data. It should be reassuring that *P*-values and *S*-values conform to this sound intuition.

An opposite objection is that *P*-values or *S*-values are *insensitive* insofar as they ignore features of the data that may be taken as evidence against the model (such as implausible estimates). This is true, but reflects nothing more than the limited information capacity of any single number such as a point estimate or confidence limit. A one-dimensional summary of multidimensional information is simply not a sufficient statistic for inference; hence, further information (such as residual plots and *P*-values for alternative hypotheses or models) will also be needed to make sensible inferences. A small p only warns that something may be wrong with the hypothesis **H** or embedding model **A**, not what is wrong or that they are unsafe to use. The kind of model updating that would follow from an initial check depends on focused prior information, diagnostics, and suspicions about model violations; for example, poor fit of the embedding model **A** is often addressed by relaxing (expanding) that model to address detected imperfections (e.g., by adding higher-order terms) (Box 1980).

Conversely, a large p only means the test did not provide much information against the tested hypothesis or model, and is *not* a "safety signal." The *S*-value reveals that no parameter value inside a 95% confidence interval has more than 4.3 bits of information against it, supporting recommendations to view the interval interior as a region of hypotheses highly compatible with the data, rather than overconfidently viewing its exterior as a region ruled out by the data (Poole 1987ab; Greenland et al. 2016). In the ibuprofen example, this region of high compatibility for the rate ratios $\exp(\beta)$ extended from 0.66 to 5.19, revealing that the study (even if otherwise perfect) was

[8] There are intricate logical and technical issues in extending *P*-values to composite **H**, e.g., see Berger and Boos 1994; Berger and Hsu 1996; Bayarri and Berger 1999, 2000, 2004. Some extensions may be rejected on the grounds that they lead to logical incoherencies, but Schervish (1996) also rejects the coherent extension based on maximizing p over **H**; the negative log of that supremum does, however, provide a lower bound on the information against **H**.

in fact practically uninformative about possible adverse effects of ibuprofen on renal outcomes, since anything from a 1/3 rate decrease ($RR = 1/3$) to a fivefold rate increase ($RR = 5$) has less than 4.3 bits of information against it. Yet, the study abstract concluded that rates of renal adverse events among infants "were not different between the ibuprofen (+/–acetaminophen) and acetaminophen-only groups," displaying a common type of nullistic cognitive blindness (McShane and Gal 2017; Greenland 2017) to the practical uninformativeness of the quoted results. In contrast, an accurate report would have said "Our study lacked sufficient information to reach any useful inference about adverse renal events comparing ibuprofen to acetaminophen alone; much more data would be needed to address ibuprofen safety concerns"—albeit under current journal publication criteria such an honest conclusion would make publication difficult. One could further remark that there is almost 10 bits of information against a 2/3 rate decrease ($RR = 1/3$) and a 10-fold rate increase ($RR = 10$) for each have $p \approx 2^{-10}$, but that would only distract from the inadequacy of the study for safety assurances.

4.6. P-Values Depend on Investigator Sampling Intentions—Exactly as They Should

Nowhere do statistical principles seem to clash more persistently than in the role of intentions or analysis plans (protocols) in inference (see, e.g., Bayarri and Berger 2004, sec. 5). Consider the claim that "if we stick to the short-run perspective when measuring evidence, identical data produce identical evidence regardless of the experimenters' intentions" (Goodman 1999, p. 1000). Without possibly questionable qualifications, this claim is false. Consider one experiment (which is a very short run) that reports $p = 0.004$ for $\mathbf{H}$: $RR = 1$ with a point estimate for RR of $\widehat{RR} = \exp(\hat{\beta}) = 1.8$ and a likelihood ratio of over 50 for comparing $RR = 2$ to $RR = 1$. If the experimenter intended to follow best practices and reported doing so, we might well regard those results as providing some credible evidence ($s = -\log_2(0.004) = 8$ bits) against $\mathbf{H}$. But suppose we discover that the experimenter's intention was to produce data that gave $p < 0.005$, if necessary by postrandomization reallocation and exclusions; we should then take the resulting data as providing little or no evidence regarding $\mathbf{H}$ or RR in general.

One could say the experimenter's intentions do matter in this example because in the first case they leave the sample space restricted only by the initial valid study design, whereas in the second case they further restrict the sample space to samples with $p < 0.005$. The only way we might reasonably discount these intentions in measuring the data evidence is if we knew *with certainty* that the intentions produced no action that changed the sample, e.g., if we knew the initial treatment allocation happened to give $p = 0.004$ and so led to no manipulation by the ill-motivated experimenter. If, however, we had no such information, we would have to rely on the probability that manipulation would prove unnecessary, which would be 0.005 if $\mathbf{H}$ were correct. Seeing $p = 0.004$ (or a correspondingly small Bayes factor for $\mathbf{H}$) would then only provide us with evidence against the combined hypothesis that "$\mathbf{H}$ is correct *and* there was no experimental misconduct." Thus, intentions matter even though their effects must be mediated through experimenter actions and reporting.

Discovering deceptive intentions of the sort just described should lead us to alter our embedding model in a way that changes the likelihood function for RR from its form under honest intentions (e.g., the probability of $\widehat{RR} = 1.8$ when $RR = 1$ would be much higher with deceptive than with honest intent); thus accounting for such intentions need not be controversial. The ongoing controversy instead concerns intentions which alter p but leave the likelihood function unchanged, as in optional-stopping problems. This is a vast topic, but briefly: The dependence of p on samples not in fact observed (counterfactual data) leads some to reject it as an evidence measure in favor of pure likelihood or Bayesian statistics (Berger and Wolpert 1988; Edwards 1992; Royall 1997; Goodman 1999). For many statisticians, however, the full sampling distribution (including counterfactual datasets) encodes important information about the data-generating mechanism (e.g., Cox and Hinkley 1974; LeCam 1988; Lane 1988; Ritov et al. 2014, p. 635)[9] leaving P-values as basic tools for checking assumptions even if the final results might be reported as posterior probabilities (Box 1980).

The dependence of P-values on counterfactual data is thus seen as a means of allowing for assumption uncertainties without specifying all alternatives (as in tests of fit that allow for unforeseen types of model violations). From this perspective, the insensitivity of likelihood ratios to counterfactual data reflects an information loss which can be extreme in high-dimensional problems (where likelihood-based procedures may provide no consistent inference yet valid P-values can be constructed from the full sampling distribution; see Robins and Wasserman 2000, sec. 5; Ritov et al. 2014).

5. Conclusions

P-values are often criticized for having properties they *should* possess under correct interpretations, and for encouraging misuse and misinterpretation of data. Unquestionably, P-values have proven problematic for correct teaching, description, and usage, as evidenced by the many common misinterpretations (Greenland et al. 2016). These problems are aggravated by the distortive scaling and unwarranted dichotomization of P-values, the misleading and inconsistent jargon surrounding their definitions and description, and the misplaced criticisms which blame P-values for ill-conceived traditions surrounding their use and interpretation (Hurlbert and Lombardi 2009). These problems can be reduced by

1. focusing on definitions of P-values that make no reference to α levels or decisions, in which the observed p is a tail probability and its random counterpart P is a uniform variate under the hypothesis and model used for its construction;

[9] In terms used in this literature, P-values can violate the likelihood principle (which says proportional likelihood functions should yield the same inferences), but dissenters from the principle do not in practical terms regard the likelihood function as a sufficient statistic outside of some narrowly specified circumstances.

2. computing P-values for contextually important alternatives to null (nil or "no effect") hypotheses, such as minimal important differences;
3. rescaling p to the Shannon information (S-value) $s = -\log_2(p)$ to provide a better scale for measuring the amount of information the test supplies against the hypothesis;

and most of all

4. avoiding dichotomization of p or comparison of p to α-levels (see, e.g., Poole 1987a; Hoekstra et al. 2006; Hurlbert and Lombardi 2009; Greenland et al. 2016, 2017; Amrhein et al. 2017, 2018; McShane et al. 2018, 2017; Wasserstein and Lazar 2016).

It may be argued that exceptions to (4) arise, for example when the central purpose of an analysis is to reach a decision based solely on comparing p to an α, especially when that decision rule has been given explicit justification including both false-acceptance (Type-II) error costs under relevant alternatives and false-rejection (Type-I) error costs (Lakens et al. 2018), as in quality-control applications. Such thoroughly justified applications are, however, uncommon in observational research settings.

It may then seem ironic that many writers who have campaigned vigorously against rigid α-level hypothesis testing promote in its place a version of it in the form of the 95% confidence interval. As with the 5% α-level from which it is derived, the conventional 95% confidence level is divorced from any consideration of error costs, and the resulting interval is typically nothing more than the set of all values b for β for which the test of $\beta = b$ yields $p > 0.05$. It should thus be no surprise that numerous examples (like the ibuprofen study) show that confidence intervals have not provided the hoped-for cure for testing abuse, but have instead perpetuated the dichotomania and excessive certainty that plague research reports. And there is no basis for expecting Bayesian tests or posterior intervals to be treated more wisely.

At best, a 95% confidence interval roughly indicates an entire region of high compatibility between the data and possible parameter values within a given model, as judged for example by having less than 4.3 bits of information against the values inside the interval (Using S-values, a simpler and arguably better range for claiming "high compatibility" would be a 5-bit interval ($\approx$ 97% "confidence".). Thus it would be less misleading to refer to these intervals as compatibility intervals rather than confidence intervals, where "compatibility" only means that the data supply limited information (not even 5 coin-tosses worth) against the parameter values in the interval under the model used to construct the interval. Parallel cautions would apply to posterior probability intervals, with the rephrasing that "compatibility" only means the data supply limited information against the parameter values in the interval under the model *and prior distribution* used to construct the interval.

It should be emphasized that this limitation represents a paucity of information about the parameter in the data and model, rather than decisive evidence in favor of values in the interval or a refutation of values outside the interval. In particular, values not far outside a 95% interval also have limited information against them, and may easily fall inside an interval produced from another, equally plausible model. Such caution is especially important when the model underlying the interval-estimation method (including so-called robust methods) may be incorrect in ways not accounted for by the model (such as unmodeled or mismodeled measurement errors). In such cases (which are the norm in health and social sciences), a 95% coverage or posterior probability description does not adequately incorporate model uncertainty, and thus the interval so described becomes an *overconfidence* interval (A more complete discussion of these issues is given in Greenland 2018.).

Acknowledgements

I thank Valentin Amrhein, Sameera Daniels, Michael Fay, Lawrence McCandless, Mohammad Mansournia, Keith O'Rourke, Philip Stark, David Trafimow, and an anonymous reviewer for helpful comments on earlier drafts of this paper. I also owe a special debt of gratitude to Allen Schirm for exceptionally detailed comments, corrections and suggestions on the original and revised versions.

References

Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017), "The Earth is Flat ($p > 0.05$): Significance Thresholds and the Crisis of Unreplicable Research," *Peer J*, 5, e3544. [106,112]

Amrhein, V., Trafimow, D., and Greenland, S. (2018), "Inferential Statistics are Descriptive Statistics," *The American Statistician*, this issue. [108,112]

Bayarri, M. J., and Berger, J. O. (1999), "Quantifying Surprise in the Data and Model Verification," in *Bayesian Statistics 6*, eds. J. M. Bernardo, J.O. Berger, A.P. Dawid, and A. F. M. Smith, Oxford, UK: Oxford University Press, pp. 53–82. [109,110]

Bayarri, M. J., and Berger, J. O. (2000), "Values for Composite Null Models," *Journal of the American Statistical Association*, 95, 1127–1142. [110]

Bayarri, M. J., and Berger, J. O. (2004), "The Interplay of Bayesian and Frequentist Analysis," *Statistical Science*, 19, 58–80. [107,110,111]

Benjamini, Y. (2016), "It's Not the P-values' Fault," *The American Statistician*, Online Supplement to ASA Statement on P-values. 70, online supplement 1, available at *http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/xxxx*. [106]

Berger, J. O., and Sellke, T. M. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of P-values and Evidence" (with discussion), *Journal of the American Statistical Association*, 82, 112–139.

Berger, J. O., and Wolpert, R. L. (1988), "The Likelihood Principle" (with discussion) (2nd ed.), IMS Lecture Notes-Monograph Series, 6, 1–199. [111]

Berger, R. L., and Boos, D. D. (1994), "P Values Maximized Over a Confidence Set for the Nuisance Parameter," *Journal of the American Statistical Association*, 89, 1012–1016. [110]

Berger, R. L., and Hsu, J. C. (1996), "Bioequivalence Trials, Intersection-Union Tests, and Equivalence Confidence Sets," *Statistical Science*, 11, 283–319. [110]

Boos, D. D., and Stefanski, L. A. (2011), "P-Value Precision and Reproducibility," *The American Statistician*, 65, 213–221. [108,109]

Box, G. E. P. (1980), "Sampling and Bayes Inference in Scientific Modeling and Robustness," *Journal of the Royal Statistical Society,* Series A, 143, 383–430. [110,111]

Casella, G., and Berger, R. L. (1987), "Reconciling Bayesian and Frequentist Evidence in the 1-sided Testing Problem" (with discussion), *Journal of the American Statistical Association*, 82, 106–135.

Casella, G., and Berger, R. L. (1987), "Comment," *Statistical Science*, 2, 344–417. [110]

Cohen, J. (1994), "The Earth is Round ($p < 0.05$)," *American Psychology*, 47, 997–1003. [108]

Cox, D. R., and Donnelly, C. A. (2011), *Principle of Applied Statistics*, Cambridge, UK: Cambridge University Press. [107]
Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, New York: Chapman and Hall. [107,109,111]
Edwards, A. W. F. (1992), *Likelihood* (2nd ed.), Baltimore, MD: Johns Hopkins University Press. [111]
Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh, UK: Oliver and Boyd. [107]
Fraundorf, P. (2017), "Examples of Surprisal," available at *http://www.umsl.edu/~fraundorfp/egsurpri.html*. [109]
Gelman, A. (2013), "*P* Values and Statistical Practice," *Epidemiology*, 24, 69–72. [110]
Gelman, A., and Stern, H. (2006), "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant," *The American Statistician*, 60, 328–331. [108]
Gigerenzer, G. (2004), "Mindless Statistics," *Journal of Socio-Economics*, 33, 587–606. [106]
Good, I. J. (1956), "The Surprise Index for the Multivariate Normal Distribution," *The Annals of Mathematical Statistics*, 27, 1130–1135. [109]
Good, I. J. (1983), "Some Logic and History of Hypothesis Testing," in *Philosophical Foundations of Economics*, ed. J. C. Pitt, Dordrecht: D. Reidel, pp. 149–174. Reprinted as Ch. 14 in Good, I.J. (1983), *Good Thinking*, Minneapolis, MN: University of Minnesota Press, pp. 129–148. [109]
Goodman, S. N. (1992), "A Comment on Replication, p-values and Evidence," *Statistics in Medicine*, 11, 875–879. [108]
Goodman, S. N. (1999), "Towards Evidence-Based Medical Statistics, I: The P-value Fallacy," *Annals of Internal Medicine*, 130, 995–1004. [107,109,111]
Greenland, S. (2004), "The Need for Critical Appraisal of Expert Witnesses in Epidemiology and Statistics," *Wake Forest Law Review*, 39, 291–310. [108]
Greenland, S. (2017), "The Need for Cognitive Science in Methodology," *American Journal of Epidemiology*, 186, 639–645. [108,109,111,112]
——— (2018), "The Unconditional Information in P-values, and Its Refutational Interpretation via S-values," manuscript. [109,112]
Greenland, S., and Poole, C. (2013), "Living with Statistics in Observational Research," *Epidemiology (Cambridge, Mass.)*, 24, 73–78. [110]
Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.C., Poole, C., Goodman, S.N., and Altman, D.G. (2016), "Statistical Tests, Confidence Intervals, and Power: A Guide to Misinterpretations," *The American Statistician*, 70, online supplement 1, available at *http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/utas_a_1154108_sm5368.pdf*; reprinted in the *European Journal of Epidemiology*, 31, 337–350. [106,107,110,111,112]
Hoekstra, R., Finch, S., Kiers, H. A. L., and Johnson, A. (2006), "Probability as Certainty: Dichotomous Thinking and the Misuse of p-values," *Psychonomic Bulletin & Review*, 13, 1033–1037. [106,112]
Hubbard, R., and Bayarri, M. J. (2003), "Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing," *The American Statistician*, 57, 171–177. [107]
Hubbard, R., and Lindsay, R. M. (2008), "Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing," *Theory & Psychology*, 18, 69–88. [109]
Hurlbert, S. H., and Lombardi, C. M. (2009), "Final Collapse of the Neyman–Pearson Decision Theoretic Framework and Rise of the neoFisherian," *Annales Zoologici Fennici*, 46, 311–349. [106,111,112]
Kuffner, T. A., & Walker, S. G. (2017), "Why Are p-values Controversial?" *The American Statistician, in Press*, 1. [107]
Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., Cross, E. S., Daniels, S., Danielsson, H., DeBruine, L., Dunleavy, D. J., Earp, B. D., Feist, M. I., Ferrell, J. D., Field, J. G., Fox, N. W., Friesen, A., Gomes, C., Gonzalez-Marquez, M., Grange, J. A., Grieve, A. P., Guggenberger, R., Grist, J., van Harmelen, A.-L., Hasselman, F., Hochard, K. D., Hoffarth, M. R., Holmes, N. P., Ingre, M., Isager, P. M., Isotalus, H. K., Johansson, C., Juszczyk, K., Kenny, D. A., Khalil, A. A., Konat, B., Lao, J., Larsen, E. G., Lodder, G. M. A., Lukavský, J., Madan, C. R., Manheim, D., Martin, S. R., Martin, A. E., Mayo, D. G., McCarthy, R. J., McConway, K., McFarland, C., Nio, A. Q. X., Nilsonne, G., de Oliveira, C. L., de Xivry, J.-J. O., Parsons, S., Pfuhl, G., Quinn, K. A., Sakon, J. J., Saribay, S. A., Schneider, I. K., Selvaraju, M., Sjoerds, Z., Smith, S. G., Smits, T., Spies, J. R., Sreekumar, V., Steltenpohl, C. N., Stenhouse, N., Świątkowski, W., Vadillo, M. A., Van Assen, M. A. L. M., Williams, M. N., Williams, S. E., Williams, D. R., Yarkoni, T., Ziano, I., & Zwaan, R. A.) (2018), "Justify Your Alpha: A Response to 'Redefine Statistical Significance,'" *Nature Human Behaviour*, 2, 168–171. [108,112]
Lane, D. (1988), "Discussion of Berger and Wolpert," *IMS Lecture Notes-Monograph*, 6, 175–181. [111]
Lang, J. M., Rothman, K. J., and Cann, C. I. (1998), "That Confounded P-value," *Epidemiology (Cambridge, Mass.)*, 9, 7—8. [109]
LeCam, L. (1988), "Discussion of Berger and Wolpert," *IMS Lecture Notes-Monograph*, 6, 182–185. [111]
Lehmann, E. L. (1986), *Testing Statistical Hypotheses*, New York: Wiley. [107,108]
Lindeman, M., & Stark, P. B. (2012), "A Gentle Introduction to Risk-limiting Audits," *IEEE Security & Privacy*, 10, 42–49. [110]
MacKay, D. J. C. (2003), *Information Theory, Inference, and Learning Algorithms*, Cambridge, Cambridge University Press, sec. 2.4, available at *http://www.inference.org.uk/mackay/itila/book.html* [109]
McShane, B. B., and Gal, D. (2017), "Statistical Significance and the Dichotomization of Evidence" (with discussion), *Journal of the American Statistical Association*, 112, 885–908. [106,111]
McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2018), "Abandon Statistical Significance," *The American Statistician*, this issue. [106,108,112]
Merriam-Webster Dictionary (2017), "Null," available at *https://www.merriam-webster.com/dictionary/null*. [112]
Murdoch, D. J., Tsai, Y.-L., and Adcock, J. (2008), "P-Values are Random Variables," *The American Statistician*, 62, 242–245. [107,108]
Neyman, J. (1977), "Frequentist Probability and Frequentist Statistics," *Synthese*, 36, 97–131. [108]
Oxford Living Dictionary (2017), "Null," available at *https://en.oxforddictionaries.com/definition/null*.
Perezgonzalez, J. D. (2015), "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'," *Frontiers in Psychology*, 6, 341. [107]
Poole, C. (1987a), "Beyond the Confidence Interval," *American Journal of Public Health*, 77, 195–199. [110]
——— (1987b), "Confidence Intervals Exclude Nothing," *American Journal of Public Health*, 77, 492–493. [112]
Ritov, Y., Bickel, P. J., Gamst, A. C., and Kleijn, B. J. K. (2014), "The Bayesian Analysis of Complex, High-Dimensional Models: Can It Be CODA?" *Statistical Science*, 29, 619–639. [111]
Robins, J. M., and Wasserman, L. (2000), "Conditioning, Likelihood, and Coherence: A Review of Some Foundational Concepts," *Journal of the American Statistical Association*, 95, 1340–1346. [111]
Royall, R. R. (1986), "The Effect of Sample Size on the Meaning of Significance Tests," *The American Statistician*, 40, 313–315. [110]
——— (1997), *Statistical Inference: A Likelihood Paradigm*, New York: Chapman and Hall. [110,111]
Schervish, M. J. (1996), "P-values: What They Are and What They Are Not," *The American Statistician*, 50, 203–206. [110]
Sellke, T. M., Bayarri, M. J., and Berger, J. O. (2001), "Calibration of p Values for Testing Precise Null Hypotheses," *The American Statistician*, 55, 62–71. [107,109]
Senn, S. J. (2001), "Two Cheers for P-Values," *Journal of Epidemiology and Biostatistics*, 6, 193–204. [106,108,109,110]
——— (2002), "Letter to the Editor re: Goodman 1992," *Statistics in Medicine*, 21, 2437–2444. [106,108,109]
——— (2008), *Statistical Issues in Drug Development* (2nd ed.), New York: Wiley. [110]
Shannon, C.E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27, 379–423, 623–656. [109]
Spanos, A. (2013), "Who Should Be Afraid of the Jeffreys–Lindley Paradox?" *Philosophy of Science*, 80, 73–93. [110]

Walsh, P., Rothenberg, S. J., and Bang, H. (2018), "Safety of Ibuprofen in Infants Younger than Six Months: A Retrospective Cohort Study," *PLoS One*, 13, e0199493, available at *https://doi.org/10.1371/journal.pone.0199493* [106]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on p-values: Context, Process and Purpose," *The American Statistician*, 70, 129–133. [106,112]

Wellek, S. (2010), *Testing Statistical Hypotheses of Equivalence and Noninferiority* (2nd ed.), New York: Chapman & Hall. [110]

Ziliak, S. T., and McCloskey, D. N. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives,* Ann Arbor, MI: University of Michigan Press. [108]