ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Three-Decision Methods: A Sensible Formulation of Significance Tests—and Much Else

Kenneth M. Rice[1] and Chloe A. Krakauer[2]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, USA; email: kenrice@uw.edu

[2]Kaiser Permanente Washington Health Research Institute, Seattle, Washington, USA; email: chloe.a.krakauer@kp.org

Annu. Rev. Stat. Appl. 2023. 10:525–46

First published as a Review in Advance on October 6, 2022

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-033021-111159

Keywords

statistical tests, significance tests, decision theory, power, multiple testing, sensitivity analysis

Abstract

For real-valued parameters, significance tests can be motivated as three-decision methods, in which we either assert the sign of the parameter above or below a specified null value, or say nothing either way. Tukey viewed this as a "sensible formulation" of tests, unlike the widely taught null hypothesis significance testing (NHST) system that is today's default. We review the three-decision framework, collecting the substantial literature on how other statistical tools can be usefully motivated in this way. These tools include close Bayesian analogs of frequentist power calculations, *p*-values, confidence intervals, and multiple testing corrections. We also show how three-decision arguments can straightforwardly resolve some well-known difficulties in the interpretation and criticism of testing results. Explicit results are shown for simple conjugate analyses, but the methods discussed apply generally to real-valued parameters.

1. INTRODUCTION

Statistical testing has long been a controversial topic (Fisher 1935b, Neyman & Pearson 1933, Jeffreys 1935), and debate on it continues (Wasserstein & Lazar 2016, Benjamini et al. 2021). In particular, the formal reject/accept framework of hypothesis tests, though widely taught, has been criticized for lack of relevance to underlying scientific aims (McShane et al. 2019). In this article we review and explore an alternative, much less well-known approach to statistical tests—viewed by John Tukey as a "sensible formulation" of them (Jones & Tukey 2000)—that views tests as decisions about the sign of an underlying univariate parameter. The signs can be asserted to be positive or negative, or, in an important third option, we can make no decision either way.

Elements of this "three-decision" approach have long been known (Lehmann 1950, Bahadur 1952, Cox 1958), but to date no single reference draws together the depth of developments to which it leads, unpicking many of the difficulties that afflict standard tests. This article groups these developments into three core areas. In Section 2 we review the basic three-decision approach and its properties, as well as Bayesian and decision-theoretic versions of it, that provide measures defining its optimality. Section 3 extends the approach to motivate other well-known tools used with tests (e.g., power calculations, *p*-values, multiple testing corrections), giving them "sensible" formulations, too. Finally, Section 4 introduces some novel tools for assessing the reliability of testing-based inferences. We conclude with a contemporary worked example and a short discussion, including scope for further work in this area.

2. REVIEW OF VIEWING TESTS AS THREE-DECISION PROBLEMS

2.1. Defining the Three Decisions

Using the three-decision approach, we denote the results of a test as one of the options in **Table 1**, where (as throughout this article) θ is the univariate parameter of interest and θ_0 is a prespecified null value, where the sign of θ around θ_0 is of scientific interest. For simplicity—and following the arguments of, e.g., Jones & Tukey (2000) about its scientific relevance—we assume θ cannot plausibly be exactly equal to θ_0 and so do not entertain conclusions of that form.

Initial formulations of the three-decision problem (Lehmann 1950) considered decisions where one might accept a central null region, but compatibility of inference with intuition about evidence then becomes challenging (Schervish 1996, Hansen & Rice 2022). Moreover, as argued by Bohrer (1979), by considering small enough effect sizes we can never completely rule out situations where, using standard hypothesis tests, the probability of correctly determining the sign would approach only 0.5—i.e., no better than a coin toss. To avoid this, the third "no decision" option is required. It was proposed initially by Bahadur (1952), who suggested a view of the *t*-test where, with nonsignificant results, one should just reserve judgement (see also Kaiser 1960). Separately, Esteves et al. (2016) show how the option of a nondecision is required if tests are to have certain logical coherence properties. Key subsequent references on the three-decision problem are the work of Harris (1997), who stresses how viewing tests as sign decisions can avoid many problems of

Table 1 Notation and informal names for potential decisions under the three-decision framework

Notation	Informal name	Assertion
$d = A$	Above	$\theta > \theta_0$
$d = N$	No decision	None
$d = B$	Below	$\theta < \theta_0$

Assertions made concern the sign of unknown parameter θ relative to null value θ_0.

misinterpretation, and Hurlbert & Lombardi (2009), for whom three-decision approaches are an important element of their "neo-Fisherian" system. As noted by Harris (1997), viewing testing results in this way is not far from current practice: Reporting the estimated sign of θ (albeit also with an estimate of its magnitude) is standard, as is reserving judgement about the sign of θ around θ_0 when tests do not reject the point null hypothesis that $\theta = \theta_0$.

However, default justification of hypothesis tests of point nulls (Barnett 1999, p. 196) provides neither of these two elements. First, using classical tests a rejection of the point null hypothesis says nothing about the sign of θ, unless we augment it by considering the possibility of a type III error—rejecting the null hypothesis with an erroneous sign estimate, i.e., making the right decision "for the wrong reason" (Mosteller 1948, p. 61). Second, simply reserving judgement is not an option whenever we consider type II error rates or, equivalently, power. As power is defined in terms of accepting the null—not remaining agnostic about it—using power to justify making no assertion at all is at best a mismatch between motivation and practice. Jonsson (2013) considers frequentist alternatives to power that are compatible with the three-decision approach; Berg (2004) develops optimality criteria for three-decision tests.

For balance, we note that even this close connection with practice is not compelling to some authors; the three-decision approach was derided as impractical by Hunter (1997), whose call for statistical testing to be abandoned is still being made (e.g., Longford 2020).

For completeness, we also note that if we omit one of the possible decisions—i.e., we remove a row in **Table 1**—then one-sided significance and hypothesis tests result (Rice et al. 2020). Much of what follows in this article can also be developed for one-sided tests, but we focus on the two-sided version as, following, e.g., Bland & Altman (1994), it is almost always important to entertain making sign decisions in either direction, and not just one's preferred direction.

2.2. Decision-Theoretic and Bayesian Approaches

With the set of three decisions given in Section 2.1, if we constrain interest in θ to just its sign around θ_0 (positive or negative), then evaluation of testing methods relies on, at most, $3 \times 2 = 6$ quantities. This makes standard decision-theoretic approaches particularly compelling, as an exhaustive set of losses can be considered; all that one needs is the relative loss for decisions A, N, or B under θ of either sign. [Jonsson (2013) avoids specifying relative costs by—unconventionally—maximizing overall rates of correct sign decisions subject to uniformly most powerful unbiased testing of two null hypotheses, that θ is above or below θ_0, and uniform minimization of nondecision rates. The rationale for this precise form of constraint is not specified.]

Further restrictions on the six values in the loss function limit the form of tests that need to be considered; a straightforward commitment to veracity means we should use losses that penalize wrong decisions more than correct ones (see also Duncan 1965, section 5.1). Rice et al. (2020) further show how, with losses that penalize $d = N$ equally regardless of the underlying true sign of $\theta - \theta_0$ and are symmetric with regard to sign errors in either direction, the loss must be equivalent to

$$L(d, \theta) = 1_{d=\text{above} \cap \theta<\theta_0} + \frac{\alpha}{2} 1_{d=\text{no decision}} + 1_{d=\text{below} \cap \theta>\theta_0} \qquad 1.$$

for some $\alpha \in (0, 1)$ chosen by the user. The loss may be more informally written as

$$1_{\text{make sign error}} + \frac{\alpha}{2} 1_{\text{make no decision}},$$

i.e., a trade-off between a potentially-wrong sign decision where errors are costly and a cheaper but fixed-cost nondecision. Trade-off rate $\alpha/2$ states how many times cheaper it is to make no

decision than to make a sign error, where both quantities are measured relative to the zero loss we incur for correct sign decisions. While we expect that $\alpha = 0.05$ will be a natural default for many users, considering the trade-off being made may instead help users choose a rate α based on the scientific setting relevant to their analysis. We return to this topic in Section 4.3.

In Bayesian decision theory, under an assumed prior and model describing knowledge of θ and how data update that knowledge, the Bayes rule is chosen to minimize the posterior expected loss (Bernardo & Smith 2009, p. 448). This further automates the connection of scientific goals with statistical methods. For three-decision problems, Bayes rules must depend on only the posterior tail area, $\mathbb{P}[\theta < \theta_0]$, or equivalently $\mathbb{P}[\theta > \theta_0]$. [For formal proof, consider special cases of the results of Bansal & Sheng (2010); see also Thulin (2014).] For any three-decision loss satisfying the coherence conditions, the Bayes rule asserts the sign of θ about θ_0 only if the corresponding tail area is below some critical threshold. Further assuming symmetry around θ_0, the critical quantity is the smaller of the two tail areas, and we only make a sign decision when this critical quantity is smaller that $\alpha/2$—very much like a standard non-Bayesian two-sided test. By the Bernstein–von Mises theorem (Van der Vaart 2000, chapter 10), under mild regularity conditions, this means the Bayesian test also provides large-sample control of the type I error rate, and with Equation 1's symmetric loss, we control it at level α. Less formally, the Bayesian test achieves the usual frequentist definition of a valid test. Sign decision tests therefore avoid the Jeffreys-Lindley paradox (Lindley 1957), in which results from default Bayesian tests—based on priors with atoms of probability at exactly $\theta = \theta_0$—can strongly disagree with default frequentist tests.

The minimized expectation of Equation 1's symmetric loss can be written as

$$\frac{1}{2}\min(\tilde{P}, \alpha), \text{ where } \tilde{P} = 2\min(\mathbb{P}[\theta < \theta_0], \mathbb{P}[\theta > \theta_0]).$$

We see that, when a sign decision occurs, the minimized loss is just the smaller of the two tail areas, strongly analogous to frequentist methods that double the smaller tail. Similarly, $\tilde{P}$ gives a close Bayesian analog of the standard frequentist p-value from a two-sided test that rejects the point null when $p < \alpha$. Importantly, this shows that while such p-values are "irreconcilable" with Bayesian measures of support for point null hypotheses (Berger & Sellke 1987), they are straightforwardly compatible with other forms of Bayesian tests. Moreover, the Bayesian three-decision motivation requires no statement of hypothetical replications, which appears to be a notable challenge to making correct use of p-values intuitive (McShane & Gal 2017).

For frequentist evaluation, the "risk" of a given rule is the average, over repeated experiments, of its realized loss. For the Bayes rule, the risk is

$$\text{risk}(\theta) = \mathbb{P}_\theta\left[\tilde{P} < \alpha \cap \hat{\theta} > \theta_0\right] 1_{\theta<\theta_0} + \frac{\alpha}{2}\mathbb{P}_\theta\left[\tilde{P} > \alpha\right] + \mathbb{P}_\theta\left[\tilde{P} < \alpha \cap \hat{\theta} < \theta_0\right] 1_{\theta>\theta_0}, \qquad 2.$$

where $\hat{\theta}$ denotes the posterior median, used here just to indicate which side of θ_0 has greater support. In all cases the outer expectation is frequentist, over repeated experiments for which θ is the true parameter. Informally the risk may be stated as

$$\text{risk} = \text{rate}\left[\text{make sign error}\right] + \frac{\alpha}{2}\text{rate}\left[\text{make no decision}\right].$$

The Bayes rule's risk for a setting where we have independent $N(\theta, 1)$ observations and a zero-centered Normal prior on θ (a Normal-Normal setup) is given in **Figure 1**; we note that risk is monotone in the magnitude of Normal location parameter θ but not monotone decreasing in sample size, nor monotone decreasing in prior precision, when keeping all other factors fixed. The nonmonotonicity reflects a combination of the two components that contribute to the risk. Increasing the sample size, the contribution from nondecisions shrinks straightforwardly. However, the contribution from sign errors is more complex; it is small when the sample size is tiny,

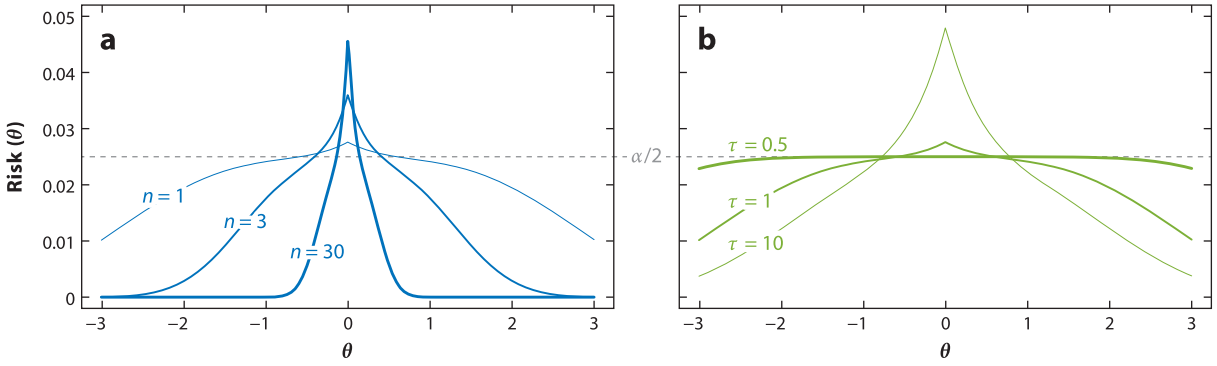

Figure 1

Frequentist risk of Bayes rules for independent $N(\theta, 1)$ observations using Equation 1's loss with $\alpha = 0.05$ and prior $\theta \sim N(0, \tau^2)$. We fix $\tau^2 = 1$ and vary n in panel *a*; in panel *b* we fix $n = 1$ and vary τ^2. The risk is monotonic decreasing in $|\theta|$ for fixed n and τ, but not monotonic decreasing in n for all fixed θ and τ, or monotonic decreasing in τ for all fixed θ, n.

as the prior dominates and sign errors are very rare. It is also small with large sample sizes, as the likelihood dominates and sign errors are rare. Yet in between, there is a "twilight zone" in which the rate of sign errors peaks sufficiently strongly to overwhelm the shrinking contribution from nondecisions. Similar arguments explain the nonmonotonicity with respect to prior precision.

3. EXTENSIONS

3.1. Power and Related Issues

When planning studies or analyses, or comparing classical testing methods, the default measure considered is power or, equivalently, type II error rate. This presents a challenge to the three-decision approach; while some measure of performance is clearly needed, as noted in Section 2.1 [and known going back to at least Neyman (1952, chapter 1, part 3)], power is not an obvious metric when we never accept null hypotheses.

Various alternatives to power are available (O'Hagan & Stevens 2001, Gelman & Carlin 2014, Bayarri et al. 2016), but we focus on those most directly connected to Equation 1's loss. Specifically, these are its corresponding risk, defined in Equation 2, and the Bayes risk, the expectation of risk with respect to the prior.

Informally, the risk tells us how bad a testing procedure is on average, assuming nondecisions are cheaper than sign errors by a factor of $\alpha/2$ and that correct sign decisions incur no loss. This is similar in spirit to the motivation of Neyman & Pearson (1933, p. 291) for using power—trying to ensure that we are "not too often wrong"—though here α appears as part of the statement of losses, not as a distinct rate that we choose to control. Close connections between risk and power can be expected whenever the contribution from nondecisions, the central term in Equation 2, is much greater than that from wrong sign decisions; the risk will be essentially just a rescaled probability of nondecisions, equivalent in the standard approach to a type II error rate. (Still stronger connections between risk and power occur when, as in **Figure 1**, both are monotonic in effect size.) Familiar formulae for power calculations can be expected to also largely determine the risk of three-decision methods at a given effect size or sample size.

Notable differences do exist, however. In the classical setting, when we consider decreasing absolute effects or sample sizes, for reasonable tests we can expect power to tend to α. But no such lower bound applies to the risk under Equation 1's loss; if a large enough proportion of active

sign decisions are incorrect—described by Gelman & Tuerlinckx (2000) as having a high "type S error rate"—then risk will exceed $\alpha/2$, i.e., the test will perform worse than simply ignoring the data and making no decision regardless. Such testing procedures have been labeled "futile" (Rice et al. 2020). Futile tests need not be pathological or unrealistic situations: For the Normal location problem of **Figure 1**, using the standard Z test (which is uniformly most powerful unbiased; see Casella & Berger 2021, pp. 390–91), at level $\alpha = 0.05$, futility occurs whenever the test has less than 12% power. If prior knowledge strongly supported such small effects, then the corresponding Bayes rule would still lead to futility for some parameter values, but with risk at those small effect sizes being closer to α. Averaging over the prior on effect sizes, the corresponding Bayes rule will (by its definition) always achieve Bayes risk below $\alpha/2$.

Moving away from futile tests, the goal of reducing risk can motivate calculations much like those familiar for sample size/minimum effect sizes. The key choice here is how much lower than $\alpha/2$ the risk has to be for a study to have merit. No default level of reduction exists, but in closely related work, Shafer (2021) suggests that learning enough to improve performance by a factor of 5 merits attention; one might view this as the testing setup being "worthwhile." Of note, for **Figure 1**'s Normal location problem using Z tests with $\alpha = 0.05$, reducing risk to the worthwhile level of $0.05/10 = 0.005$ or less corresponds to requiring sample size/effect size yielding at least 80% power (Krakauer & Rice 2021)—a familiar criterion in current practice.

Rather than specifying a particular value of θ, one can also average risk(θ) over the prior, giving the Bayes risk. Ensuring that the Bayes risk is below a prespecified threshold (e.g., $\alpha/10$) also determines a sample size calculation, for analysts who share the relevant prior beliefs. This is essentially the approach used in "Bayesian assurance," although it averages power, not risk, over the prior (O'Hagan & Stevens 2001). Following reasoning used to set minimum thresholds for Bayesian assurance, to set a threshold for sufficient Bayes risk reduction, one should consider the cost of the study: A modest reduction in risk (i.e., having a moderate chance that any sign decisions are wrong, at a priori plausible θ values) may be worthwhile if the study is cheap to conduct. With still more explicit assumptions about trade-offs of cost per data point versus reduction in risk, one could (following Lindley 1997) determine an optimal sample size, not just a sufficiently large one. Section 4.3 discusses posterior estimation of the risk and how this may help with interpretation of results.

3.2. Two-Sided *p*-Values

As noted in Section 1, controversy over statistical testing is widespread, with corresponding criticism of *p*-values (Nuzzo 2014). Nevertheless, various authors have defended *p*-values, for example, as being in practice "too familiar and useful to ditch" (Matthews et al. 2017, p. 41, quoting David Spiegelhalter). *P*-values also have a formal interpretation as a summary of potential tests, as the highest α at which the null would not be rejected. Rafi & Greenland (2020) also review how *p*-values are a legitimate measure of data-model conflict even without doing tests.

Decision-theoretic views of *p*-values can clarify their connections with tests. Following the arguments of Rice et al. (2020, appendix B), we consider the loss function

$$L(s, a, \theta) = \frac{1}{\sqrt{a}}\left(2s1_{\theta<\theta_0} + a + 2(1-s)1_{\theta>\theta_0}\right)$$

for decisions $a \in (0, 1)$ and $s \in \{0, 1\}$, which is related to but distinct from three-decision Equation 1's loss. Ratio a describes how much more willing we are to accept a unit of loss for the indicator of the sign of θ around θ_0 (which is known imperfectly) versus a fixed cost, and the choice of sign is given by binary decision s. The Bayes rule sets $s_B = 0, 1$ according to whether there is greater posterior support for $\theta < \theta_0$ or $\theta > \theta_0$, respectively, and a_B is just $\tilde{P}$, i.e., twice the

minimum tail area and a direct Bayesian analog of the two-sided p-value. This Bayesian analog of two-sided p-values is obtained without test being done—however, it does require that we consider quantities that (as seen in Sections 2.2 and 3.1) are strongly relevant to testing decisions, so some connection remains. Formally, we are viewing the p-value as the Bayes rule for a problem that is the dual of the testing problem, and in which we decide the optimal rate for making trade-offs between functions of the sign of θ. A distinct advantage of the three-decision approach is how it avoids needing to consider long-run frequency properties under the point null, the relevance of which is strongly criticized by, e.g., Jeffreys (1980).

3.3. Credible Intervals

A widely used formulation of confidence sets comes from inverting tests, describing the, e.g., 95% confidence set as the collection of null values θ_0 that would not be rejected by tests with level $\alpha = 0.05$. Rice et al. (2020) show how decision theory permits the same construction for univariate parameters. One simply integrates the sign-testing Equation 1's loss over all possible null values θ_0, making a distinct decision for each null value. Formally, this gives

$$L(A, B, N, \theta; \pi) = \pi\left(A \cap \{\theta_0 : \theta < \theta_0\}\right) + \frac{\alpha}{2}\pi(N) + \pi\left(B \cap \{\theta_0 : \theta > \theta_0\}\right), \qquad 3.$$

where decisions A, B, and N are the sets of all null values about which we report that θ is above or below or make no decision, and user-chosen measure π states the relative importance given to each contributing value of θ_0. Informally, Equation 3's loss may be written as

$$\text{area(sign errors)} + \frac{\alpha}{2}\text{area(no decision)},$$

where the areas—subsets of the space of possible null values—are calculated with regard to measure π. We trade the area of null values about which we make sign errors for the area of null values about which we make no decision, with no decision being cheaper per unit area by a factor of $\alpha/2$.

Assuming, reasonably, that π provides nonzero support to all $\theta_0 \in \Theta$, then the Bayes rules for A and B, respectively, report simply the sets of values above and below the $1 - \alpha/2$ and $\alpha/2$ posterior quantiles, and the central quantile interval is the Bayes rule for N. Of note, the Bayes rules are the same for all measures π with nonzero support everywhere, so beyond this regularity condition, the choice of measure π can be ignored for making set decisions. Frequentist agreement between credible and confidence intervals follows by the usual Bernstein–von Mises results (see Section 2.2), but notably, the agreement happens at a faster rate for quantile-based intervals than for other credible sets (Hartigan 1966). (Specifically, the quantile-based interval is a second-order confidence interval, with asymptotic accuracy of its coverage for sample size n shrinking with $1/n$, compared with other first-order intervals where accuracy only shrinks with $1/\sqrt{n}$.)

Having the sets be connected comes for free in this motivation; Equation 3's loss does not specify it directly. This is not the case for other motivations, specifically the loss given by Schervish (1995, section 5.2.5) for the ends of a connected interval, which balances penalties for interval width and for being further from the true θ, when it is not covered by the interval.

Use of the overall expected loss (which can be expected to depend on measure π) for evaluation and criticism of credible sets—similarly to Section 3's discussion of test trustworthiness—remains unexplored. Extensions in which we divide the real line into more than three sets, bringing in concerns about the consonance (Gabriel 1969) of different decisions, also have yet to be implemented. A further unexplored but intriguing connection is with point estimation. Unlike implementing credible intervals as accompaniments to point estimates (e.g., Rice & Ye 2022), the inverting-tests approach here yields credible intervals without any intermediate estimation step. Nevertheless, in the limit as α approaches 1, the Bayes rule for set N reduces to a single point, the posterior median, a widely used point estimate.

3.4. Multiple Testing

When multiple hypotheses are to be tested, the controversy noted in Sections 1 and 3.2 intensifies; prominently expressed views range from the stance that no corrections are needed (Rothman 1990) to claims that unaccounted-for multiple testing is a major reason for false positive findings (Forstmeier et al. 2017) and hence the replication crisis. There is also a large and pragmatic middle ground, however. For example, most statisticians would agree that when a large number of hypotheses are tested, reporting only the smallest p-values is quite likely to be misleading (Cox 2006, p. 86). When inference for multiple parameters is considered, using just their signs has been recognized as a helpfully robust approach, in terms of being widely applicable and also leading to estimation of error rates with reduced sensitivity to modeling assumptions (Stephens 2017). Formal decision theory approaches date to Lehmann (1957a,b) and Duncan (1965).

In contrast to Section 2.2's single sign decision, for which all loss functions can be considered in a small table, joint losses for tests of multiple parameters quickly become challenging. In particular, different rules will be preferred depending on whether we measure performance based on a simple summation of univariate testing losses, versus some measure of the rate of errors among sign decisions that are made [see Shaffer (2002) and Lewis & Thayer (2004, 2013), but also the earlier work of Robbins (1951) and its extensions by Sun & Cai (2007)]. Without forethought, this can lead to reversals between one-at-a-time tests of multiple parameters and multiplicity-corrected versions of them (Perlman & Wu 1999) that might be expected to be more stringent.

Illustrating this in the sign decision setting, we first simply add together a copy of Equation 1's loss for each parameter, giving loss

$$\#\{d_j = A \cap \theta_j < \theta_{0j}\} + \sum_{j:d_j=N} \frac{\alpha_j}{2} + \#\{d_j = B \cap \theta_j > \theta_{0j}\}. \tag{4.}$$

By the results of Lehmann (1957a) (and see also Lewis & Thayer 2013), this leads to testing at level α_j for each θ_j, with no correction for multiplicity. However, the simple count of incorrect sign decisions is likely inappropriate in many settings, where making each additional sign error need not incur the same additional penalty. A simple—but extreme—form of interaction between the sign errors is given for m parameters by loss

$$\sum_{j:d_j=N} \frac{\alpha_j}{2} + 1_{\bigcup_{j=1}^{m}\{(d_j=A\cap\theta_j<\theta_{0j})\cup(d_j=B\cap\theta_j>\theta_{0j})\}}, \tag{5.}$$

in which the sum of all costs for nondecisions is traded against an indicator that any sign error has occurred. This means that making one sign error is as bad as making any larger number of them, up to and including sign errors for all m parameters.

It is notable that with Equation 5's loss, not all sets of α_j will make sense. In particular if $\sum_{j=1}^{m} \alpha_j/2 > 1/2$, then deciding $d_j = N$ for all j will always incur more expected loss than making sign decisions $d_j = A$ or B throughout, with the choice made according to which sign has greater posterior support. The expected loss is then bounded above by 1/2, and hence we would never set all $d_j = N$, regardless of the observed data. To avoid this, we can correct the α_j levels, relative to their use in single parameter tests, and constrain $\sum_{j=1}^{m} \alpha_j < 1$. For the simple case where all α_j are equal, we can set $\alpha_j = \alpha/m$ for some $\alpha < 1$, recognizable as a Bayesian version of Bonferroni correction (Bland & Altman 1995). As with classical use of Bonferroni controlling family-wise error rate, the criterion employed is conservative. It can similarly be shown that using a conservative approximation of the Bayes rule for Equation 5's loss with $\alpha_j = \alpha/m$ is essentially a Bayesian version of applying Bonferroni correction to the p-values used for testing (Rice et al. 2020).

Extending the multiple testing losses above leads to other criteria and methods. Lewis & Thayer (2013), following Sarkar & Zhou (2008), note how the Bayes risk of Equation 4's loss—the

expectation over repeated experiments of the minimized posterior loss—is bounded:

$$\frac{1}{2}\mathbb{E}\left[\sum_{j:\tilde{P}_j<\alpha_j}\tilde{P}_j+\sum_{j:\tilde{P}_j>\alpha_j}\alpha_j\right]\leq\frac{1}{2}\sum_{j=1}^{m}\alpha_j,$$

and rearranging this, we must have

$$\mathbb{E}\left[\frac{\sum_{j:\tilde{P}_j<\alpha_j}\tilde{P}_j}{\#\{j:\tilde{P}_j<\alpha_j\}}\right]\leq\frac{\sum_{j:\tilde{P}_j<\alpha_j}^{m}\alpha_j}{\#\{j:\tilde{P}_j<\alpha_j\}}.$$

The left-hand quantity is the expectation (with respect to both prior and sampling uncertainty) of the #{sign errors}/ max(1, #{sign decisions}) and so is termed the "Bayesian directional false discovery rate" (Sarkar & Zhou 2008), which in turn is similar to the directional false discovery rate introduced by Williams et al. (1999). The right-hand quantity, minimized over selections of indices j, provides an upper bound on the Bayesian directional false discovery rate; for the simple situation with all $\alpha_j=\alpha$, it is just α.

The connection to frequentist false discovery rate (FDR) control is more direct if we instead adapt Equation 4's loss, now trading off the proportion of active decisions that are wrong for the proportion of parameters about which no decision is made, giving loss

$$\frac{\#\{\text{sign errors}\}}{1\vee\#\{\text{sign decisions}\}}+\frac{\alpha}{2}\frac{\#\{\text{sign nondecisions}\}}{m},$$

where we note that no parameter is upweighted compared with any other. As shown by Lewis & Thayer (2009), the Bayes rule is given by a Bayesian version of the Benjamini–Hochberg algorithm, making sign decisions for the parameters ordered by their increasing $\tilde{P}_j$, up until the largest value for which $\tilde{P}_{(j)}<\frac{\alpha j}{m}$. Further work remains on evaluating or adapting these losses for inference on selected parameters, or explicitly reflecting dependence between knowledge of different parameters, via either simple correlation or more complex graphical structures.

4. TOOLS TO ASSESS THREE-DECISION INFERENCE

Even with well-planned studies providing inference on clearly specified and relevant parameters, results will not always be clear-cut. A large part of the current crises in both statistical understanding and reproducibility (Gelman & Loken 2014, Goodman 2016, Wasserstein & Lazar 2016) may be due to overinterpretation of findings that are ambiguous or weaker. Using simple accept/reject (or accept/do not accept) dichotomies may fuel these problems—and tests have long been criticized as excessively crude summaries of data analyses, for example, the depiction of them as "mechanical rituals" (Cohen 1994, p. 1001). In fairness, providing a sign decision is only slightly more informative, so essentially the same criticisms apply if sign decisions alone are reported.

However, rather than retiring tests completely (Amrhein et al. 2019), a more complete picture of testing results may be produced by augmenting them with a corresponding measure of trustworthiness (Hand 2022), meaning their capacity to dispute anything other than true statements. While this is an active area of research for classical tests of point nulls (Mayo & Spanos 2006, Gelman & Carlin 2014, Bayarri et al. 2016, Hannig et al. 2016, Matthews 2018, Shafer 2021), beyond p-values and confidence intervals, measures of trustworthiness are rarely actually reported—and misinterpretations and misuse of these measures are widespread (Goodman 2016). While not a panacea, three-decision approaches are attractively simple in three distinct ways. First, they consider only a limited, easily understood set of decisions. Second, they omit point masses of support at the null value—unlike, e.g., postexperimental rejection odds (Bayarri et al. 2016). Third, they optimize a single criterion—the expectation of Equation 1's loss—without constraint,

instead of optimizing one criterion (power) subject to a strict bound on another (type I error rate). All three factors may help readers see more directly how tests are being assessed. Below, we describe how tests may be assessed using established reverse-Bayes approaches (i.e., evaluating which priors would lead to different decisions) and loss estimation in which the realized value of Equation 1's loss is estimated, to the extent that this is possible. We also consider risk estimation, a novel approach in which standard Bayesian tools update beliefs about the risk of the testing setup over replicate studies. This leads to surprising distinctions between the motivation to make a specific testing decision versus trust in the testing procedure that generated it.

4.1. Reverse-Bayes Assessment of Three-Decision Results

Dating at least to Good (1950), it has been recognized that Bayesian updating can be reversed, deducing the prior given a posterior and likelihood. This approach lends itself to assessment of analyses by reverse-Bayes methods, where, for a given likelihood and a set of posteriors—say, those consistent with a certain decision—we deduce the corresponding priors. Using subjunctive reasoning (Senn 2001), we then assess whether the priors needed to motivate a given decision are compatible with existing scientific insights. In other words, we determine if making a different decision would require one to have started from untenable beliefs. Where this occurs—and the elements providing the likelihood are trusted—the initial results appear warranted. Where it does not, the testing result, while formally valid, may nevertheless be uncompelling.

A thorough review of this general approach is given by Held et al. (2021). It is particularly well suited to the three-decision problem: We need only categorize priors according to which sign decision or nondecision they lead to, when combined with the likelihood. Moreover, with the implicit use of univariate parameters, the sets of priors we might consider—when characterized by a small set of key features such as mean and variance—is tractable.

Implementing reverse-Bayes methods for three-decision analysis, denoting the initial posterior density as post(θ) obtained using initial prior density prior(θ), we can obtain the alternate posterior under another prior by calculating

$$\frac{\text{post}(\theta)}{\text{prior}(\theta)}\text{prior}^*(\theta),$$

for alternate prior density prior$^*(\theta)$, and then normalizing so that this quantity integrates to 1. From there, we only need the minimum tail area to either side of θ_0 to calculate the corresponding alternate $\tilde{P}^*$ and alternate Bayes rule d^*. When there is conjugacy (e.g., the simple but useful setting where both prior and posterior are Normal), closed form evaluations are available; more generally, we can reweight Monte Carlo samples from the initial posterior to approximate values of alternate $\tilde{P}^*$ values.

We give four examples in **Figure 2**. We suppose all initial sign decisions use $\theta_0 = 0$, with $\alpha = 0.05$ and initial prior where $\theta \sim N(0, 25)$. The data (and hence posterior) for each case are chosen to give the specified initial $\tilde{P}$ value. In each row we do this in two different ways: one in which the posterior is precise and one where it is diffuse. Contours within the plot connect the different alternate $N(\mu^*, \sigma^{*2})$ priors that give particular $\tilde{P}^*$ values, with shading colors indicating the set of alternate priors for which the alternate Bayes rule is $d = B$, N, or A.

In the top row, with initial $\tilde{P} = 0.046$ and $d = A$, from the more precise posterior we see that the $d = A$ decision accords with the Bayes rules under priors that are more precise only if they indicate a slight bias for positive effects; for less precise priors, some degree of negative bias would also lead to $d = A$. Under the more diffuse initial posterior, a much smaller set of $\{\mu^*, \sigma^*\}$ combinations leads to alternate decisions that agree with the initial one. Clearly, $\tilde{P}$ alone does not tell us everything about how compelling the initial analysis may be.

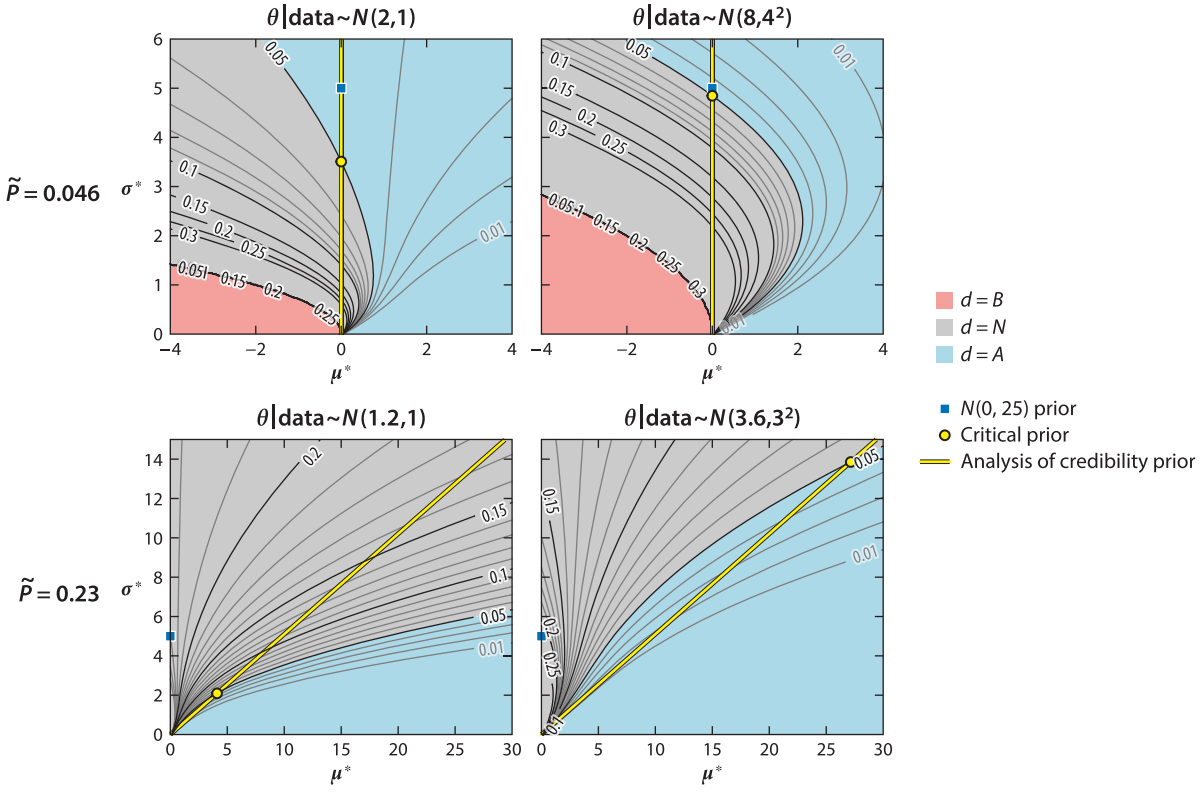

Figure 2

Contour plots of $\tilde{P}^*$ at alternate prior means μ^* and standard deviations σ^*, for reverse-Bayes analysis of the initial posteriors in each figure title, all with an initial $N(0, 25)$ prior and the stated initial $\tilde{P}$ values. The colored regions indicate the corresponding Bayes rule, for sign decisions with $\alpha = 0.05$. Also highlighted are the critical priors given by analysis of credibility, which considers only priors indicated by the yellow line. Contours of $\tilde{P}^*$ from 0.01 through 0.3 are shown, although we omit contours in regions where the initial sign decision is completely reversed. *A*, *B*, and *N* stand for the decision that would be made (above, below, or no decision; see **Table 1**) at the given alternate prior.

The vertical lines in the top row of **Figure 2** indicate alternate priors considered under Matthews's (2018) analysis of credibility (AnCred) (initially developed by Matthews 2001; see also Spiegelhalter et al. 2004, section 3.11). For initial sign decisions, AnCred only considers priors centered at the null value. It seeks the critical prior, which is the alternate prior giving results exactly on the boundary between a sign decision and nondecision, i.e., where $\tilde{P}^* = \alpha$. The critical priors are indicated on the plots; the formulae that give them are provided by Held et al. (2021). Matthews (2018) further suggests summarizing them by the extreme 2.5% quantiles of the prior, labeling these the skepticism limits. These limits represent the most extreme null values about which one would make no decision (under Equation 1's loss) using the alternate prior alone, where that alternate prior is chosen such that its posterior has equal expected loss under $d = A$ and $d = N$. In the examples of the top row of **Figure 2**, the distance between the original and critical prior similarly indicates how much/little skepticism about larger θ values would be needed to overturn the Bayes rule for the precise/diffuse posteriors, respectively.

In the lower row we give two results that have the same $\tilde{P}$ ($\tilde{P} = 0.23$, giving Bayes rule $d = N$) but have different interpretations when challenged with the use of alternate priors. For the precise posterior, we see that to change to $d = A$, one needs a prior with a mean several times bigger than

its standard deviation. For the diffuse posterior, the Bayes rule can be changed by priors with approximately equal mean and standard deviation, at least for θ values that are plausible under the initial prior.

For these initial $d = N$ decisions, AnCred obtains its critical prior by considering alternates for which, making decisions based on the prior alone, a sign decision just favors asserting the direction under consideration instead of making no decision. These priors correspond to the line through the origin with slope $1/z_{\alpha/2}$, i.e., $1/1.96$ for $\alpha = 0.05$. The critical prior on this line occurs where $\tilde{P}^* = \alpha$, i.e., the alternate Bayes rule, is also exactly borderline. In place of the skepticism limit, AnCred summarizes this prior by the advocacy limit, which is the 0.025 quantile in the direction of consideration. It represents the least extreme null value about which one would make a sign decision using the alternate prior alone, where the alternate prior is chosen such that its posterior yields equal expected loss under $d = A$ and $d = N$. In the examples in the bottom row of **Figure 2**, the horizontal difference in the critical prior under the precise/diffuse priors similarly illustrates how strongly shifted prior beliefs would have to be to overturn the initial posterior Bayes rule yet retain $d = N$ under the alternate prior.

Reverse Bayes offers a direct way to assess the sensitivity of decisions to prior assumptions, indicating the prior beliefs that (when updated rationally with the information in the data and assumed model) give results that agree or disagree with the initial Bayes rule. In application it does, however, require that the alternate priors can be concisely parameterized. This is not a major problem for univariate analyses with simple closed-form priors, but where more flexible classes of alternate priors are considered, it will be a challenge to summarize the set of $\tilde{P}^*$ indexed by larger numbers of parameters. Using testing-based properties to define a reduced set of alternate priors for consideration, as AnCred does, also risks confusion with the testing that is the primary purpose of the analysis. A final limitation of reverse Bayes is its propensity for considering priors (for example, unrealistically precise ones) with which the observed data are strongly in conflict (Evans & Moshonov 2006). Rather than Bayesian updating, faced with such priors, the data would provoke reappraisal of relevant modeling assumptions. One can rule out considering such priors, but defining them in the first place is itself a challenge. The notion of intrinsic credibility (Matthews 2018) assesses this concern in part, evaluating whether observed data are tenable under critical priors, but more general assessment of prior/data conflict and its impact on decisions remains unresolved.

4.2. Estimating Loss of Three-Decision Results

Even when prior assumptions on θ and the model are uncontroversial, with limited data there may still be concern that a specific decision—even an optimal Bayes rule—may still not be a good decision in absolute terms. A natural quantification of this is given by estimating the realized loss for the specific decision, i.e., the loss that was actually incurred. Given the decision, the realized loss is a function of the unknown θ alone, and hence amenable to Bayesian inference. For example, under squared-error loss for estimation decisions, the Bayes rule is the posterior mean and the realized loss is $(\theta - \mathbb{E}[\theta])^2$, typically estimated by its posterior expectation, the posterior variance. To estimate the realized loss more formally using decision theory, we can either expand the original loss function so it also estimates realized loss as an auxiliary decision (Rukhin 1988), or instead—fixing inference from the original decision—use an entirely separate loss function (Robert 2007, pp. 178–80).

For the three-decision problem this formalism seems excessive, as we can only make correct sign decisions, nondecisions, or incorrect sign decisions, and the realized loss can only take one of three values: 0, $\alpha/2$, and 1. Moreover, for nondecisions, we know with certainty that the loss is exactly $\alpha/2$, so the only uncertainty comes from whether sign decisions, if made, are correct or

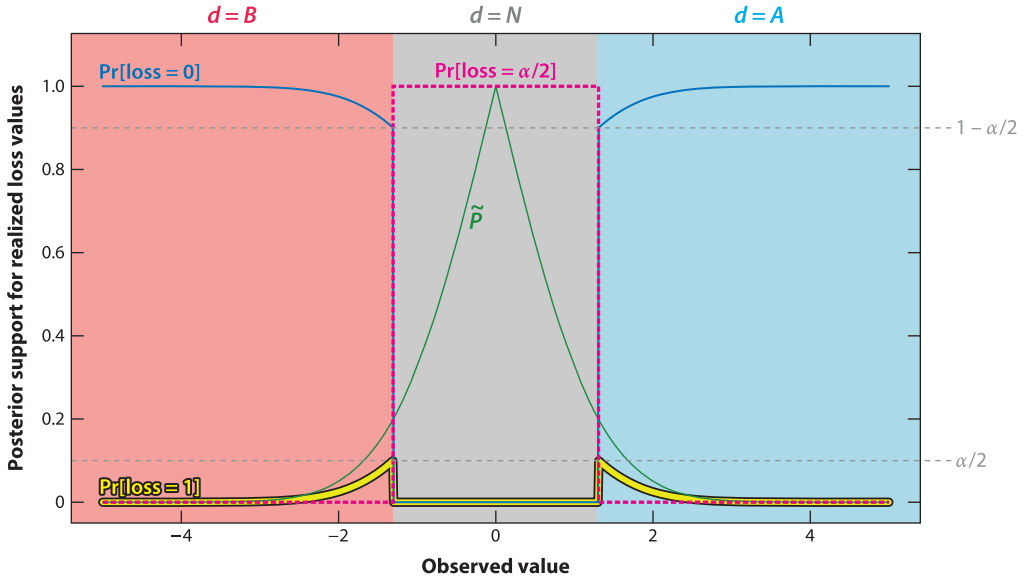

Figure 3

Probabilities of the three possible values of realized loss for a single $Y \sim N(\theta, 1)$ observation with prior $\theta \sim N(0, 25)$. $\tilde{P}$ is also shown. The shaded regions indicate the corresponding Bayes rule under Equation 1's loss with $\alpha = 0.2$. Where sign decisions are made, the posterior expectation of the realized loss, a binary variable, is just the probability that loss = 1, i.e., half the $\tilde{P}$ value. *A*, *B*, and *N* stand for the decision that would be made (above, below, or no decision; see **Table 1**).

not. **Figure 3** shows (using $\alpha = 0.2$ for clarity) the three posterior probabilities as functions of the observed sample value, when we have an $n = 1$ realization of $N(\theta, 1)$ with prior $\theta \sim N(0, 25)$. The shading color indicates the corresponding Bayes rule under Equation 1's loss. Where $d = A$ or $d = B$ and sign decisions are made, the posterior expectation of the realized loss is just the probability that loss = 1, i.e., half the $\tilde{P}$ value. In other words, when a sign decision is made, $\tilde{P}$ completely determines what is known about the realized loss.

Motivated in this way, $\tilde{P}$ values are an inevitable measure of vulnerability of sign decisions—i.e., their capacity to produce "a consequence which is substantially discrepant from what we would expect were an assertion true" (Hand 2022, p. 333). Interpreting $\tilde{P}$ as a loss estimate—and reporting it to assess the vulnerability of an initial testing decision—provides a Bayesian analog of the standard non-Bayesian practice (see e.g., Altman et al. 1983) of reporting full p-values and not just whether $p < \alpha$.

The redundancy, when we know $\tilde{P}$, of other posterior-based measures of vulnerability (i.e., realized loss) is rhetorically useful. Under the setup we have described, it forces discussions of how to assess a given test away from which measure(s) to report, focusing instead on interpretability of the assessment given by $\tilde{P}$. Given the challenges different audiences have in interpreting standard p-values (Royall 1986, Hubbard & Bayarri 2003, Goodman 2008), there seems value in having a variety of presentations.

The redundancy also provides a Bayesian analog of Hoenig & Heisey's (2001) celebrated critique of observed or post hoc power calculations. In these calculations, in an effort to understand whether a specific testing decision is well-motivated or not, users reuse the data used for the test to also estimate the parameter being tested, and plug that estimate (and perhaps also estimates of any nuisance parameters) into standard power formulae. Post hoc power calculations were common in earlier literature, and debates over them continue (Bababekov & Chang 2019, Heinsberg &

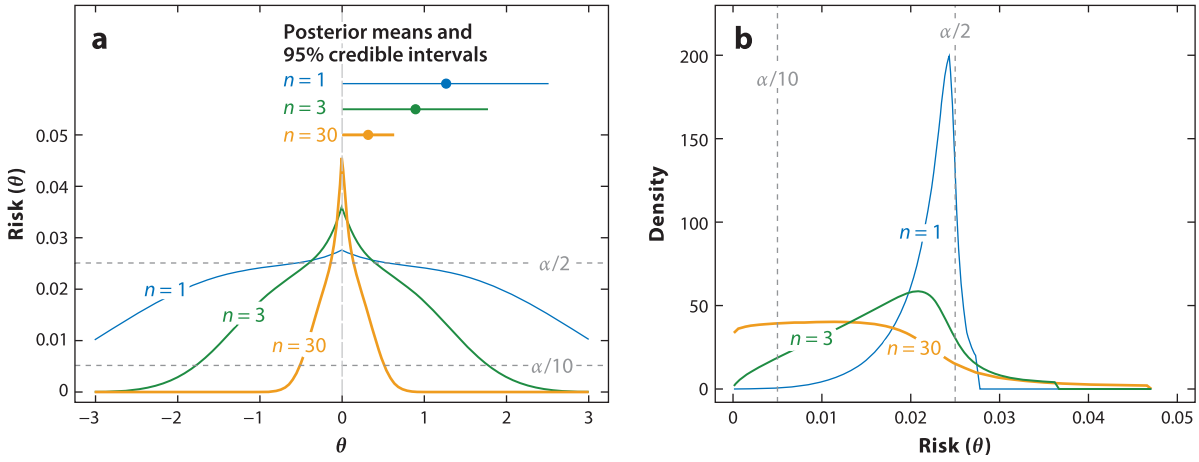

Figure 4

Risk at fixed θ and density of risk of Bayes rules for independent $N(\theta, 1)$ observations using Equation 1's loss with $\alpha = 0.05$ and prior $\theta \sim N(0, 1)$, for results with $\tilde{P} = 0.046$. (*a*) The posterior for θ is summarized and risk(θ) at each θ shown, as also shown in **Figure 1**. (*b*) The posterior density for risk(θ) is shown. Both plots show the suggested threshold risk levels of $\alpha/2$ and $\alpha/10$.

Weeks 2022), but as noted by Gelman (2019, p. e64, quoting a phrase from Bababekov & Chang 2019), "considering this noisy estimate as knowing the power after the fact is an invitation to overconfidence." Hoenig & Heisey (2001) showed that for several common settings, the post hoc power estimate is just a fixed function of the *p*-value, and so it provides zero further information about the chances of a type II error. In the Bayesian analog, with knowledge of the $\tilde{P}$ value, zero further information is available about whether a sign decision is a sign error. We also note that the Bayesian analog result is completely general for univariate parameters, holding for any posterior; Hoenig & Heisey's (2001) version is claimed to be general, but no rigorous proof of this is given.

4.3. Estimating Risk of Three-Decision Methods

As seen in Section 3.1, the risk of three-decision methods plays a central role in study planning. While it does not inform one of the plausible losses actually incurred in an analysis (as Section 4.2's loss estimates do), updating knowledge of the risk is nevertheless natural, using the available data to enable more precise statements of the long-run average costs incurred in repeated use of those methods.

Indeed, some knowledge of this form of frequentist calibration seems essential in practice, as those reading the results are unlikely to trust them if obtained from methods that "if used repeatedly, give systematically misleading conclusions" (Reid & Cox 2015, p. 295; Hand 2022, p. 333). More positively, establishing acceptable levels of risk helps build the case for the trustworthiness of the methods being used (Sekhon et al. 2014).

Compared with loss estimation, implementing risk estimation for three-decision approaches using Bayesian inference is more straightforward: Risk, for the Bayes rule of interest and under a specified study design and loss, is simply a function of θ, and so the posterior obtained for θ also conveys all posterior uncertainty about the risk at the true value of θ. As we describe below, this simplicity nevertheless provides straightforward motivations for why sign-testing decisions merit skepticism, particularly when results are borderline.

Figure 4*b* shows the distribution of risk(θ)—the long-run average loss incurred using the Bayes rule for Equation 1's loss with $\alpha = 0.05$ and $\theta_0 = 0$—under the $N(0, 1)$ prior and Normal location

model used in **Figure 1**, where the observed data in each case give $\tilde{P} = 0.046$, as in **Figure 2**. Less technically, **Figure 4*b*** shows beliefs about risk based on data that only just result in a sign decision; other cases are considered by Krakauer & Rice (2021). Notably, the posterior of the risk is not simply stretched or shifted as sample size varies; the shape differs. At the smallest sample size, situations with borderline $d = A$ or $d = B$ decisions give (nontrivial) 12.1% support to futility of the testing setup—i.e., risk(θ) $> \alpha/2$—with most of the rest of the support for only slightly lower risk. With increasing sample size, the risk curve gets steeper around $\theta = 0$, but the posterior support for θ also becomes more concentrated. These factors combine to make the posterior of the risk more uniform. More specifically, support for futility decreases, but (importantly) only to 10.9%, not vanishing to zero.

Considering low values of risk, we can follow Shafer (2021) and deem settings with risk(θ) $< (\alpha/2)/5$ as a worthwhile reduction in risk. With $n = 1$ we have support 0.0009, i.e., 0.09%, for the setting being worthwhile, given borderline $\tilde{P} = 0.046$. This is shown by **Figure 4*b***'s very light left tail of risk(θ)'s posterior, below $\alpha/10$. For larger n (e.g., $n = 30$) we see greater support in this region given the same borderline $\tilde{P}$, but support rises to only 21% as n increases without bound. In plain terms, we see that while borderline results may result in active sign decisions, they do not provide support that the testing process by which they arose is a reliable one that is "not too often wrong."

To meet this stricter standard of defensibility, a lower $\tilde{P}$ threshold can be introduced. For example, obtaining $\tilde{P} = 0.005$ or below, as suggested by Benjamin et al. (2018), with large n, we reduce support for futility (of the sign decision with $\alpha = 0.05$) to just 2.1%, while support for risk values low enough to be worthwhile (as defined above) rises to 50.1%, i.e., a coin toss.

Clearly, these considerations of risk link the sign decision framework to other approaches, notably the motivation of the $p < 0.005$ criterion via approximate Bayes Factor arguments (Benjamin et al. 2018, Johnson 2013), or via 80% replicability (Greenwald et al. 1996). While not considered here, links to still further approaches can be obtained by elaborating upon the calculation of risk. Specifically, when integrating loss over replicate datasets to obtain risk(θ), we can consider the proportion of them for which the data are more extreme than that actually observed, where extremity might be measured by the corresponding $\tilde{P}$ values. This proportion defines a form of severity (Mayo & Spanos 2006), a measure of the stringency of the test; severe tests are capable of uncovering falsehood in a false hypothesis. Severity is more specifically defined as the probability the test statistic would be more extreme than that observed under a range of true θ. Inference on the severity of a test can therefore be provided as we have done for risk, but its implicitly subjective choice of how to measure extremity would first need to be defended.

Finally, as noted by Rice (2010), there may be good scientific reasons to consider lower α with larger sample sizes; trade-off rates for different types of error may differ in small versus large studies. Wakefield (2009), following Cox & Hinkley (1974), shows how this adaptation can reconcile frequentist tests with measures of evidence for and against point null hypotheses. In considering risk, reducing α proportionally with $(n \log n)^{-1/2}$ (Cox & Hinkley 1974, p. 397) means that borderline results yield decreasing support for futility (at the relevant α) at larger sample sizes. Less technically, this means that even with borderline results, we eventually conclude the testing setup is not futile.

5. EXAMPLE: ANALYSIS OF THE ANDROMEDA-SHOCK TRIAL

To illustrate practical use of three-decision analyses, we consider data from the ANDROMEDA-SHOCK trial (Hernández et al. 2018). It studied treatments for septic shock, an exceptionally serious condition with anticipated four-week mortality up to 45%. ANDROMEDA-SHOCK

compared peripheral perfusion-targeted resuscitation (a novel treatment) to lactate-targeted resuscitation. The study enrolled $n = 424$ participants, sufficient to provide 90% power to detect a 15% reduction—or hazard ratio (HR) of 0.60—in the four-week mortality rate. On analyzing the data, the two-sided p-value of 0.06 just exceeded the prespecified $\alpha = 0.05$. However, the study did provide some evidence of benefit of the novel treatment, with estimated HR = 0.75 [95% CI 0.55–1.02] (Hernández et al. 2019).

Despite these borderline results, the investigators primarily concluded that the novel treatment did not reduce mortality (Hernández et al. 2019). This was controversial: Others claimed that "the new treatment clearly reduced mortality" (Hardwicke & Ioannidis 2019, p. 4), consistent with common interpretations of just-nonsignificant results as "trending towards significance" (Wood et al. 2014, p. 1). The ensuing public debate over interpreting the results (Ahuja 2019, Amrhein et al. 2019, Hardwicke & Ioannidis 2019, Spiegelhalter 2020) motivated several Bayesian analyses of the data, including one by the original study team (Spiegelhalter 2020, Zampieri et al. 2020).

To show how three-decision Bayesian approaches could help, we illustrate the approaches of Sections 2.1–4.3 for three Normal priors for θ, the log HR of interest. The priors are centered around the null, $\theta_0 = 0$, but have different prior variances ν^2. We consider (*a*) a weakly informative $N(0, \nu^2 = 10)$ prior; (*b*) a somewhat skeptical $N(0, \nu^2 = 0.134)$ prior, chosen to have a 95% prior credible interval identical to a prior used by some members of the original ANDROMEDA-SHOCK team (Zampieri et al. 2020); and (*c*) a very skeptical $N(0, \nu^2 = 0.032)$ prior used by Spiegelhalter (2020), with a 1 in 500 prior chance of the HR being below 0.60, the HR for which the original study was well-powered. Updating the prior generally requires defining the full data likelihood, which would, in this case, require defining the baseline hazard, a high-dimensional nuisance parameter. For simplicity and in keeping with the Cox proportional-hazards model's lack of dependence on baseline hazard, we instead update the prior using an approximation of the sampling distribution of the log HR (Tsiatis 1981), where $\hat{\Theta} \sim N(\hat{\theta}, \hat{s}^2)$, using estimate $\hat{s}$ of the standard error from the fully adjusted Cox regression output (Spiegelhalter et al. 2004, pp. 27–30).

For all three priors, the Bayes rule for Equation 1's loss with $\alpha = 0.05$ is to make no decision about mortality differences between the treatments, with $\tilde{P} = 0.15$, 0.08, and 0.06 for $\nu^2 = 0.032$, 0.135, and 10, respectively. As expected, $\tilde{P}$ most closely approximates non-Bayesian two-sided $p = 0.06$ for the most diffuse prior. For the same priors, the HR's quantile-based 95% credible intervals (developed in Section 3.3) are, respectively, (0.68–1.07), (0.59–1.04), and (0.55-1.02). The latter interval, from the most diffuse prior, is essentially identical to the original 95% frequentist confidence interval (Hernández et al. 2019).

The dataset is large enough that, even comparing values of ν^2 that differ by orders of magnitude, in addition to the Bayes rule being $d = N$, the $\tilde{P}$-values and credible intervals are broadly similar. In particular, the HR's intervals all contain 1, the null value, but also HR = 0.75, which is equivalent to an 8.5% reduction in absolute risk during the trial and deemed clinically significant by the investigators (Zampieri et al. 2020).

Using reverse Bayes via Matthews's (2018) AnCred recommendations (as discussed in Section 4.1), and considering only Normal priors, we note that the critical prior for θ that just leads to an active sign decision yet retains zero in the central 95% of the prior is $N(-111.72, 57^2)$. The great majority of this prior belief lies at implausibly small HRs ($e^{-111.72} \approx 3 \times 10^{-49}$), and so the trial results put no useful constraint on advocates of $\theta < 0$ who seek to challenge the original nondecision. In other words, there is insufficient evidence present for credibility of the initial nondecision and only more data can resolve this ambiguity. This issue persists across any study yielding borderline p-values (Matthews 2018).

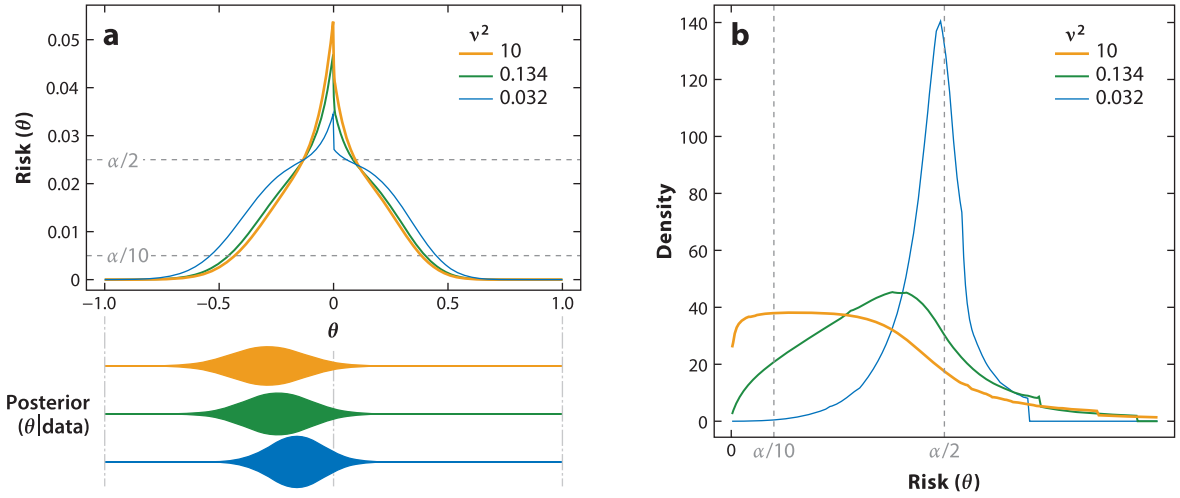

Figure 5

(*a*) Risk of Bayes rules at fixed θ, under the ANDROMEDA-SHOCK design, for various $N(0, \nu^2)$ priors on log hazard ratio θ, and using $\alpha = 0.05$. The violin plots below show the posteriors for θ given the ANDROMEDA-SHOCK data. (*b*) Posterior densities for risk(θ) for each of the posteriors and risks denoted in panel *a*. Both plots indicate where risk(θ) $= \alpha/2$, the threshold for futility, and risk(θ) $= \alpha/10$, the suggested threshold for the analysis being worthwhile, in the sense described in Section 4.3.

Finally, we estimate the risk of this testing setup, as described in Section 4.3. The risks of the Bayes rule test under the three priors are given in **Figure 5*a***, with the three posteriors for θ shown in the violin plots immediately below. Unlike Section 4.3's example, here we see that the risk is noncontinuous, fundamentally because of the mean–variance relationship our approximation of the sampling distribution of $\hat{\theta}$ induces. However, these discontinuities have little impact on subsequent calculations.

Combining the risk functions with the posteriors on θ, we get posterior distributions for risk(θ), shown in **Figure 5*b***. Both panels show the threshold risk(θ) $= \alpha/2$ (futility; see Section 3.1) and suggested threshold risk(θ) $= \alpha/10$ (for the testing setup being worthwhile; see Section 4.3).

With prior variance $\nu^2 = 0.032$, 0.134, and 10, support for futility is 36.4%, 20.7%, and 15.5%, respectively. These values do support the test being better than simply ignoring the data, but not overwhelmingly so, even for the most diffuse prior that mostly supports values of θ that are large in absolute terms. At the other extreme of the risk distribution, the same three priors give small-to-negligible support (0.07%, 6.8%, and 17.9%, respectively) for risk(θ) $< \alpha/10$, the suggested threshold for the testing setup being worthwhile enough for others to trust its conclusions. Combining these, we find that under all of the priors, it is reasonable to believe that the testing setup (the results of which, for the observed data, give $\tilde{P} = 0.06$ and $d = N$) would, over long-run replications, be only moderately less risky than ignoring the data altogether.

Consequently, it seems that no strong conclusions are justified about which treatment, if any, is better. This includes the original report of no benefit, but also strong claims of mortality reduction. Even if the treatment's effect is to reduce mortality, it is rational to also believe that the trial's design would not "rarely fail" (Fisher 1935a, p. 16) to give us correct results.

6. CONCLUSIONS AND FUTURE ISSUES

This review has shown that there is a substantial and diverse body of work on setting up tests as three-decision problems, and has discussed the benefits of this approach. Not only is Section 2's

motivation of tests themselves straightforward, but three-decision reasoning also leads directly to other widely used and familiar statistical tools and results, as in Section 3. Three-decision approaches also provide insights relevant to ongoing debates about statistical inference, for example, Section 4's distinction between inferring the performance of a specific test versus that of the testing process that generated it. Here, we describe some limitations of the approach and open areas for further research.

One difficulty may be the focus on signs alone. Sign(θ) likely captures overall qualities of the underlying reality—for example, that a drug is beneficial or harmful, on average, over some specified population under the circumstances that occurred in the study—but not more than this, such as the magnitude or transportability of the drug's effect. A form of inference on effect magnitude can be given by Section 3.3's intervals, and differences in the drug's effect between subgroups might be addressed in part by Section 3.4's multiple testing methods. However, where analysts are willing to expend the cognitive energy to construct an appropriate loss function—likely more complex than Equation 1—parameter estimates (in one or more dimensions) will more directly motivate correspondingly optimal summaries of the posterior. Credible intervals also follow directly from this specified loss function (Rice & Ye 2022), without testing as an interim step. However, in basing results around estimates, the analysis inevitably makes more assumptions than tests and can be expected to be more sensitive to them.

A further advantage of restricting attention to signs is that—together with the language in Section 2.1 that sign decisions are asserted—it may reduce the propensity to overinterpret test results as proof of a real effect, that in the traditional frameworks follows all too easily when the decision is presented as unequivocally rejecting the [point] null hypothesis. In three-decision approaches, even when the sign of the parameter does merit assertion, this result may come with substantial caveats about the reliability of the test itself (see Sections 4.3 and 5).

We distinguish the problem of overinterpretation of the test result from overinterpretation of the underlying parameter. In the traditional hypothesis testing framework, there is a strong tendency for analysts to accept their preferred alternative hypothesis when testing a null hypothesis that is not plausibly exactly true; Gelman (2016, p. 1) calls this a "parody of falsificationism." By not accepting the null—using θ_0 only as a reference point and not as a value we report as true—three-decision approaches may help somewhat; however, when multiple explanations are available for a sign decision result (e.g., $\theta > 0$ due to a true causal effect, association due to confounding, sampling bias, measurement errors, etc.) the temptation to focus on the one with the most appeal must still be avoided. This difficulty also depends on the purpose served by the test; for a useful catalog of these, readers are directed to Cox et al. (1977).

Another outstanding problem is the role of preliminary analytic choices—for example, choosing among models, or choosing which hypothesis to test. Where these choices can be identified ahead of time, they may (with some effort) be included in the loss function as auxiliary decisions, and there is flexibility to have the losses for testing decisions to vary based on them. However, when the preliminary choices are ignored, their impact on the resulting analysis (often leading to overly confident results) can be very difficult to assess (Gelman & Loken 2014), not least as the analysts may have made their choices essentially unconsciously.

Finally, we provide some areas where new tools might be developed by further expanding three-decision approaches. A first concern is what to do with nuisance parameters. For the sign decisions themselves, this is not a problem: Following standard Bayesian theory, one can integrate over the nuisance parameters in the posterior and use the tail areas of the marginal posterior for θ to give decisions, $\tilde{P}$ values, intervals, and reverse-Bayes assessments. Yet to evaluate risk we need to incorporate nuisance parameters. Rice et al. (2020, section 2.4) explore how this matters little when risk is extremely high or low, but a formal system for allowing for this extra uncertainty is needed.

There is also a clear practical need for tests of multivariate parameters; declaring signs of at least one of a set of possible univariate parameters or linear combinations of them could be considered as a basis for relevant loss functions. Fortunately loss functions for multivariate estimation are well-studied (see, e.g., Ye & Rice 2021) and will likely provide useful insights. Sequential testing (Ghosh & Sen 1991) also adds another outcome, in which interim decisions may be to make no decision but keep sampling, versus the terminal decisions to assert the sign of θ or make no decision and stop the study. Practical insights abound about what makes sequential test methods better or worse, but motivating methods via formal loss functions (and the appraisal of decisions made using them) remains relatively unexplored. Lastly, simply permitting more categories of decisions about a real-valued θ opens up further loss functions, though maintaining coherence of the resulting decisions can be expected to provide challenges (Hansen & Rice 2022).

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors are grateful to Professor Leonhard Held for his assistance, and to the ANDROMEDA-SHOCK investigators for supplying the data used in Section 5. An R script implementing all analyses described here is available at **https://faculty.washington.edu/kenrice/knowingthesigns/reviewpaperpics.R**.

LITERATURE CITED

Ahuja A. 2019. Scientists strike back on statistical tyranny. *Financial Times*, March 27. **https://www.ft.com/content/36f9374c-5075-11e9-8f44-fe4a86c48b33**

Altman DG, Gore SM, Gardner MJ, Pocock SJ. 1983. Statistical guidelines for contributors to medical journals. *Br. Med. J.* 286(6376):1489

Amrhein V, Greenland S, McShane B. 2019. Scientists rise up against statistical significance. *Nature* 2019:305–7

Bababekov YJ, Chang DC. 2019. Post hoc power: a surgeon's first assistant in interpreting "negative" studies. *Ann. Surg.* 269(1):e11–12

Bahadur RR. 1952. A property of the *t*-statistic. *Sankhyā* 12(1/2):79–88

Bansal NK, Sheng R. 2010. Bayesian decision theoretic approach to hypothesis problems with skewed alternatives. *J. Stat. Plan. Inference* 140(10):2894–903

Barnett V. 1999. *Comparative Statistical Inference*. New York: Wiley. 3rd ed.

Bayarri M, Benjamin DJ, Berger JO, Sellke TM. 2016. Rejection odds and rejection ratios: a proposal for statistical practice in testing hypotheses. *J. Math. Psychol.* 72:90–103

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, et al. 2018. Redefine statistical significance. *Nat. Hum. Behav.* 2(1):6–10

Benjamini Y, De Veaux RD, Efron B, Evans S, Glickman M, et al. 2021. The ASA president's task force statement on statistical significance and replicability. *Ann. Appl. Stat.* 15(3):1084–85

Berg N. 2004. No-decision classification: an alternative to testing for statistical significance. *J. Socio-Econ.* 33(5):631–50

Berger JO, Sellke T. 1987. Testing a point null hypothesis: the irreconcilability of *P* values and evidence. *J. Am. Stat. Assoc.* 82(397):112–22

Bernardo JM, Smith AF. 2009. *Bayesian Theory*. New York: Wiley

Bland JM, Altman DG. 1994. Statistics notes: one and two sided tests of significance. *BMJ* 309(6949):248

Bland JM, Altman DG. 1995. Multiple significance tests: the Bonferroni method. *BMJ* 310(6973):170

Bohrer R. 1979. Multiple three-decision rules for parametric signs. *J. Am. Stat. Assoc.* 74(366a):432–37
Casella G, Berger RL. 2021. *Statistical Inference*. Independence, KY: Cengage
Cohen J. 1994. The earth is round ($p < .05$). *Am. Psychol.* 49(12):997–1003
Cox DR. 1958. Some problems connected with statistical inference. *Ann. Math. Stat.* 29(2):357–72
Cox DR. 2006. *Principles of Statistical Inference*. Cambridge, UK: Cambridge University Press
Cox DR, Hinkley D. 1974. *Theoretical Statistics*. Boca Raton, FL: Chapman and Hall/CRC
Cox DR, Spjøtvoll E, Johansen S, van Zwet WR, Bithell J, et al. 1977. The role of significance tests [with discussion and reply]. *Scand. J. Stat.* 4(2):49–70
Duncan DB. 1965. A Bayesian approach to multiple comparisons. *Technometrics* 7(2):171–222
Esteves LG, Izbicki R, Stern JM, Stern RB. 2016. The logical consistency of simultaneous agnostic hypothesis tests. *Entropy* 18(7):256
Evans M, Moshonov H. 2006. Checking for prior-data conflict. *Bayesian Anal.* 1(4):893–914
Fisher R. 1935a. *The Design of Experiments*. Edinburgh, UK: Oliver & Boyd
Fisher R. 1935b. The logic of inductive inference (with discussion). *J. R. Stat. Soc.* 98:39–82
Forstmeier W, Wagenmakers EJ, Parker TH. 2017. Detecting and avoiding likely false-positive findings—a practical guide. *Biol. Rev.* 92(4):1941–68
Gabriel KR. 1969. Simultaneous test procedures–some theory of multiple comparisons. *Ann. Math. Stat.* 40(1):224–50
Gelman A. 2016. The problems with *p*-values are not just with *p*-values. Supplemental material to the ASA statement on statistical significance and *p*-values. *Am. Stat.* 70(suppl.):1–2
Gelman A. 2019. Comment on "Post-hoc power using observed estimate of effect size is too noisy to be useful". *Ann. Surg.* 270(2):e64
Gelman A, Carlin J. 2014. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspect. Psychol. Sci.* 9(6):641–51
Gelman A, Loken E. 2014. The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *Am. Sci.* 102(6):460
Gelman A, Tuerlinckx F. 2000. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Comput. Stat.* 15(3):373–90
Ghosh BK, Sen PK. 1991. *Handbook of Sequential Analysis*. Boca Raton, FL: Chapman and Hall/CRC
Good I. 1950. *Probability and the Weighing of Evidence*. London: Charles Griffin
Goodman S. 2008. A dirty dozen: twelve *p*-value misconceptions. *Semin. Hematol.* 45(3):135–40
Goodman SN. 2016. Aligning statistical and scientific reasoning. *Science* 352(6290):1180–81
Greenwald A, Gonzalez R, Harris RJ, Guthrie D. 1996. Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology* 33(2):175–83
Hand DJ. 2022. Trustworthiness of statistical inference. *J. R. Stat. Soc. Ser. A* 185:329–47
Hannig J, Iyer H, Lai RC, Lee TC. 2016. Generalized fiducial inference: a review and new results. *J. Am. Stat. Assoc.* 111(515):1346–61
Hansen S, Rice K. 2022. Coherent tests for interval null hypotheses. *Am. Stat.* In press. **https://doi.org/10.1080/00031305.2022.2050299**
Hardwicke TE, Ioannidis JP. 2019. Petitions in scientific argumentation: dissecting the request to retire statistical significance. *Eur. J. Clin. Investig.* 49(10):e13162
Harris RJ. 1997. Reforming significance testing via three-valued logic. In *What If There Were No Significance Tests?*, ed. LL Harlow, SA Mulaik, JH Steiger, pp. 145–74. London: Routledge
Hartigan J. 1966. Note on the confidence-prior of Welch and Peers. *J. R. Stat. Soc. Ser. B* 28(1):55–56
Heinsberg LW, Weeks DE. 2022. Post hoc power is not informative. *Genet. Epidemiol.* 46(7):390–94
Held L, Matthews R, Ott M, Pawel S. 2021. Reverse-Bayes methods for evidence assessment and research synthesis. *Res. Synthesis Methods* 13:295–314
Hernández G, Cavalcanti AB, Ospina-Tascón G, Dubin A, Hurtado FJ, et al. 2018. Statistical analysis plan for early goal-directed therapy using a physiological holistic view—the ANDROMEDA-SHOCK: a randomized controlled trial. *Rev. Bras. Ter. Intensiva* 30(3):253
Hernández G, Ospina-Tascón GA, Damiani LP, Estenssoro E, Dubin A, et al. 2019. Effect of a resuscitation strategy targeting peripheral perfusion status vs serum lactate levels on 28-day mortality among patients with septic shock: the ANDROMEDA-SHOCK randomized clinical trial. *JAMA* 321(7):654–64

Hoenig JM, Heisey DM. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am. Stat.* 55(1):19–24
Hubbard R, Bayarri MJ. 2003. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *Am. Stat.* 57(3):171–78
Hunter JE. 1997. Needed: a ban on the significance test. *Psychol. Sci.* 8(1):3–7
Hurlbert SH, Lombardi CM. 2009. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Ann. Zool. Fennici* 46(5):311–49
Jeffreys H. 1935. Some tests of significance, treated by the theory of probability. *Math. Proc. Camb. Philos. Soc.* 31:203–22
Jeffreys H. 1980. Some general points in probability theory. In *Bayesian Analysis in Econometrics and Statistics. Essays in Honor of Harold Jeffreys*, ed. A Zellner, J Kadane, pp. 451–53. Amsterdam: North-Holland
Johnson VE. 2013. Revised standards for statistical evidence. *PNAS* 110(48):19313–17
Jones LV, Tukey JW. 2000. A sensible formulation of the significance test. *Psychol. Methods* 5(4):411–14
Jonsson F. 2013. Characterizing optimality among three-decision procedures for directional conclusions. *J. Stat. Plan. Inference* 143(2):392–99
Kaiser HF. 1960. Directional statistical decisions. *Psychol. Rev.* 67(3):160–67
Krakauer C, Rice K. 2021. Discussion of "Testing by betting: a strategy for statistical and scientific communication" by Glenn Shafer. *J. R. Stat. Soc. Ser. A* 184(2):452–53
Lehmann EL. 1950. Some principles of the theory of testing hypotheses. *Ann. Math. Stat.* 21(1):1–26
Lehmann EL. 1957a. A theory of some multiple decision problems, I. *Ann. Math. Stat.* 28:1–25
Lehmann EL. 1957b. A theory of some multiple decision problems, II. *Ann. Math. Stat.* 28:547–72
Lewis C, Thayer DT. 2004. A loss function related to the FDR for random effects multiple comparisons. *J. Stat. Plan. Inference* 125(1–2):49–58
Lewis C, Thayer DT. 2009. Bayesian decision theory for multiple comparisons. In *Optimality: The Third Erich L. Lehmann Symposium*, ed. J Rojo, pp. 326–32. N.p.: Inst. Math. Stat.
Lewis C, Thayer DT. 2013. Undesirable optimality results in multiple testing? *Stat. Model.* 13(5–6):541–51
Lindley DV. 1957. A statistical paradox. *Biometrika* 44(1/2):187–92
Lindley DV. 1997. The choice of sample size. *J. R. Stat. Soc. Ser. D* 46(2):129–38
Longford N. 2020. Discussion on the meeting on 'Signs and sizes: understanding and replicating statistical findings.' *J. R. Stat. Soc. Ser. A* 183(2):451
Matthews R, Wasserstein R, Spiegelhalter D. 2017. The ASA's p-value statement, one year on. *Significance* 14(2):38–41
Matthews RA. 2001. Methods for assessing the credibility of clinical trial outcomes. *Drug Inform. J.* 35(4):1469–78
Matthews RA. 2018. Beyond 'significance': principles and practice of the analysis of credibility. *R. Soc. Open Sci.* 5(1):171047
Mayo DG, Spanos A. 2006. Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *Br. J. Philos. Sci.* 57(2):323–57
McShane BB, Gal D. 2017. Statistical significance and the dichotomization of evidence. *J. Am. Stat. Assoc.* 112(519):885–95
McShane BB, Gal D, Gelman A, Robert C, Tackett JL. 2019. Abandon statistical significance. *Am. Stat.* 73(suppl.):235–45
Mosteller F. 1948. A k-sample slippage test for an extreme population. *Ann. Math. Stat.* 19:58–65
Neyman J. 1952. *Lectures and Conferences on Mathematical Statistics and Probability*. Washington, DC: USDA
Neyman J, Pearson ES. 1933. IX. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. Ser. A* 231(694–706):289–337
Nuzzo R. 2014. Scientific method: statistical errors. *Nat. News* 506(7487):150
O'Hagan A, Stevens JW. 2001. Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Med. Decis. Making* 21(3):219–30
Perlman MD, Wu L. 1999. The emperor's new tests. *Stat. Sci.* 14(4):355–69
Rafi Z, Greenland S. 2020. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Med. Res. Methodol.* 20(1):244

Reid N, Cox DR. 2015. On some principles of statistical inference. *Int. Stat. Rev.* 83(2):293–308
Rice K. 2010. A decision-theoretic formulation of Fisher's approach to testing. *Am. Stat.* 64(4):345–49
Rice K, Bonnett T, Krakauer C. 2020. Knowing the signs: a direct and generalizable motivation of two-sided tests. *J. R. Stat. Soc. Ser. A* 183(2):411–30
Rice K, Ye L. 2022. Expressing regret: a unified view of credible intervals. *Am. Stat.* 76:248–56
Robbins H. 1951. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 2, ed. J Neyman, pp. 131–49. Berkeley: Univ. Calif. Press
Robert C. 2007. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York: Springer
Rothman KJ. 1990. No adjustments are needed for multiple comparisons. *Epidemiology* 1:43–46
Royall RM. 1986. The effect of sample size on the meaning of significance tests. *Am. Stat.* 40(4):313–15
Rukhin AL. 1988. Loss functions for loss estimation. *Ann. Stat.* 16:1262–69
Sarkar SK, Zhou T. 2008. Controlling Bayes directional false discovery rate in random effects model. *J. Stat. Plan. Inference* 138(3):682–93
Schervish MJ. 1995. *Theory of Statistics*. New York: Springer
Schervish MJ. 1996. *P* values: what they are and what they are not. *Am. Stat.* 50(3):203–6
Sekhon H, Ennew C, Kharouf H, Devlin J. 2014. Trustworthiness and trust: influences and implications. *J. Mark. Manag.* 30(3–4):409–30
Senn S. 2001. Two cheers for *p*-values? *J. Epidemiol. Biostat.* 6(2):193–204
Shafer G. 2021. Testing by betting: a strategy for statistical and scientific communication. *J. R. Stat. Soc. Ser. A* 184(2):407–31
Shaffer JP. 2002. Multiplicity, directional (type III) errors, and the null hypothesis. *Psychol. Methods* 7(3):356–69
Spiegelhalter DJ. 2020. Andromeda and 'appalling science': a response to Hardwicke and Ioannidis. *WintonCentre Blog*, Jan. 25. **https://medium.com/wintoncentre/andromeda-and-appalling-science-a-response-to-hardwicke-and-ioannidis-a79458efdba1**
Spiegelhalter DJ, Abrams KR, Myles JP. 2004. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York: Wiley
Stephens M. 2017. False discovery rates: a new deal. *Biostatistics* 18(2):275–94
Sun W, Cai TT. 2007. Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Stat. Assoc.* 102(479):901–12
Thulin M. 2014. Decision-theoretic justifications for Bayesian hypothesis testing using credible sets. *J. Stat. Plan. Inference* 146:133–38
Tsiatis AA. 1981. The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika* 68(1):311–15
Van der Vaart AW. 2000. *Asymptotic Statistics*, Vol. 3. Cambridge, UK: Cambridge Univ. Press
Wakefield J. 2009. Bayes factors for genome-wide association studies: comparison with *p*-values. *Genet. Epidemiol.* 33(1):79–86
Wasserstein RL, Lazar NA. 2016. The ASA statement on *p*-values: context, process, and purpose. *Am. Stat.* 70(2):129–33
Williams VS, Jones LV, Tukey JW. 1999. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *J. Educ. Behav. Stat.* 24(1):42–69
Wood J, Freemantle N, King M, Nazareth I. 2014. Trap of trends to statistical significance: likelihood of near significant P value becoming more significant with extra data. *BMJ* 348:g2215
Ye L, Rice K. 2021. Bayesian optimality and intervals for Stein-type estimates. *Stat* 11:e445
Zampieri FG, Damiani LP, Bakker J, Ospina-Tascón GA, Castro R, et al. 2020. Effects of a resuscitation strategy targeting peripheral perfusion status versus serum lactate levels among patients with septic shock. A Bayesian reanalysis of the ANDROMEDA-SHOCK trial. *Am. J. Respir. Crit. Care Med.* 201(4):423–29