Knowing the signs: a direct and generalizable motivation of two-sided tests

Kenneth Rice,
University of Washington, Seattle, USA

Tyler Bonnett
Frederick National Laboratory for Cancer Research, Frederick, USA

and Chloe Krakauer
University of Washington, Seattle, USA

[Read before The Royal Statistical Society at a meeting on ‘Signs and sizes: understanding and replicating statistical findings’ at the Society’s 2019 annual conference in Belfast on Wednesday, September 4th, 2019, the President, Professor D. Ashby, in the Chair]

Summary. Many well-known problems with two-sided p-values are due to their use in hypothesis tests, with ‘reject–accept’ conclusions about point null hypotheses. We present an alternative motivation for p-value-based tests, viewing them as assessments of only the sign of an underlying parameter, where we can conclude that the parameter is positive or negative, or simply say nothing either way. Our approach is decision theoretic, but—unusually—we consider the whole set of possible utility functions available. Doing this we show how, in a specific sense, close analogues of familiar one- and two-sided tests are always the optimal decision. We argue that this simplicity could aid non-experts’ understanding and use of tests—and help them to think critically about whether or not tests are appropriate tools for answering their questions of interest. Several extensions are also considered, showing that the simple idea of determining the signs of parameters yields a rich framework for inference.

Keywords: Bayesian; Decision theory, Frequentist; Hypothesis tests; Significance tests

1. Introduction

As the recent American Statistical Association statement on p-values (Wasserstein et al., 2016) made clear, statistical testing remains a controversial area among statisticians that is widely misunderstood by analysts applying it and by the wider audience reading scientific reports and papers. These long-standing difficulties have motivated multiple efforts to alter or discard the use of statistical tests; recent examples are given in Benjamin et al. (2018), Woolston (2015) and the body of work that was summarized by Wasserstein et al. (2019). However, other statisticians argue that p-values are ‘too familiar and useful to ditch’ (Spiegelhalter, 2017) and so should be retained as tools, with statisticians helping non-experts to use them more appropriately. Providing this help is a considerable challenge; Robinson (2018) has given an excellent review of the many difficulties in testing and pointed out the lack of clarity on what question is actually being addressed by widely used tests. In this paper we aim to provide a framework for better

Address for correspondence: Kenneth Rice, Department of Biostatistics, University of Washington, Seattle, WA 98195-7232, USA.
E-mail: kenrice@uw.edu

© 2019 Royal Statistical Society
understanding $p$-values and tests. Our approach, though quite general, is deliberately simple, to help its communication to non-experts. We are also careful to specify explicitly questions being addressed when we do use statistical tests. Our approach uses decision theory, where statements of loss for particular decisions describe what we value learning about, and how relatively valuable or costly different conclusions might be.

Developing tests by using decision theory is controversial; see for example Cox’s (1982) statement that tests exist only to assess the available information, not to choose actions. We reject this on two grounds, the first of which is pedagogic. Framing the process as a decision moves it from being entirely abstract (i.e. to quantify the available information) to concrete (i.e. to decide what to do based on that information). Abstract reasoning is facilitated by realistic context (Venet and Markovits, 2001), so viewing tests as decisions provides this context and should help analysts to think more clearly about what tests provide. Even though a final action may not be taken, clear thinking may be aided by subjunctive reasoning—users can gain insights by considering what concrete actions would be motivated given different losses. Second, by starting with decision theory we provide the user with tools where they can make decisions about what actions to take based on information—a topic largely absent from standard presentations of inference. Following on from considering how standard testing decisions can be made, we also hope to encourage users to realize the limitations of those tests—and if standard tests are inadequate for the situation at hand, to help them to identify alternatives.

Several researchers have considered testing via decision theory. Although we attempt to cite prior results appropriately throughout the paper, some key comprehensive references are worth noting directly. Lehmann (1957a) and the work of DeGroot (1970), chapter 11, discussed decision theoretic testing losses, whereas Berger (1985) provided extensive discussion of parallels between decision theory and Bayesian and frequentist testing. Ferguson (1967), chapter 5, reviews how complete class theorems prove that Bayes rules for testing are, essentially, the only technically admissible decisions. Several other texts consider decision-based testing, e.g. Parmigiani and Inoue (2009), chapter 7, and Liese and Miescke (2008), chapter 8. However, building a framework of results all from the same elemental decisions, concerned only with the sign of $\theta$, does appear to be novel. We believe that it is also novel to argue that, under some conditions described below, the testing losses that we have given are essentially the only losses that could be considered for such decisions.

1.1. Loss functions for hypothesis and significance tests
To illustrate how this framework motivates the use of $p$-values, we begin with a simple testing scenario in which we are concerned with testing a hypothesis about the sign of a real univariate parameter $\theta$. For simplicity here we consider the sign of $\theta$ relative to 0, but without loss of generality. This reference point is generalized to $\theta_0$ in Section 2.2. Also, formal proofs are removed to Appendix A. Following the arguments against testing point nulls provided by for example Berkson (1938) or William (2000)—primarily that point null hypotheses are usually scientifically implausible and hence only a straw man—we do not consider decisions that accept $\theta$ being exactly 0. We instead consider situations where $\theta$ being exactly 0 is not supported and we choose between reporting $\theta > 0$ or $\theta < 0$. For recent work arguing in favour of calibrating tests via incorrect sign decisions, see Stephens (2016).

Our initial example, in the style of a hypothesis test (Barnett (2009), page 136) permits only two possible decisions: deciding that $\theta > 0$ (‘$d=$ above’ (A)) or $\theta < 0$ (‘$d=$ below’ (B)). With the truth represented only through the sign of $\theta$ and only binary choices for decision $d$, there can be only four possible losses to consider in any loss function. The only condition that we impose is that any incorrect decisions about $\theta$ incur more loss than any correct decisions. (Losses with this
property of being minimized by correct reports of the truth are known as proper losses (Hwang et al., 1992). As shown in Appendix A, then without loss of generality any loss function is equivalent to one of the form

\[
\begin{array}{ccc}
\text{Loss when} & \text{Decision} & \alpha \\
\theta > 0 & d = A & 0 \\
\theta < 0 & d = B & \alpha \\
\end{array}
\]  

(1)

where \(\alpha \in (0, 1)\) is a positive constant, and \(\alpha/(1 - \alpha)\) states the relative cost of an incorrect \(d = B\) decision versus an incorrect \(d = A\) decision. Such losses have appeared before, e.g. Lehmann (1957a), page 549, but with different motivations.

The decisions that are motivated by our framework are those suggested by a Bayes rule—the rule which minimizes the posterior expected loss. The Bayes rule for this loss sets \(d = A\) if the posterior’s left-hand tail area \(P(\theta < 0)\) is below \(\alpha\), and \(d = B\) otherwise. This is partly intuitive: we decide on a sign based on having enough posterior support for that sign—but in the decision theoretic approach \(\alpha\)’s role is to state precisely the relative cost that determines what ‘enough’ actually means. Importantly, this statement is made without reference to repeated sampling or error rates under the (often implausible) point null hypothesis—issues that routinely confuse non-experts (Little, 2016). However, the usual frequentist interpretation of \(\alpha\) also holds in large sample settings, by Bernstein-von Mises theorem—where in large samples and under correct-model assumptions and mild regularity conditions the posterior and likelihood tend to being directly proportional. For our purposes this means that the tail area is approximately the one-sided \(p\)-value (Hinkley and Cox (1979), page 392) and the type I error rate under the null that \(\theta = 0\) tends to \(\alpha\). The close connection between one-sided tail areas and frequentist measures was further explored by Casella and Berger (1987). So using this motivation for \(\alpha\), frequentist concerns can be addressed but need not be the focus: the threshold \(\alpha\) has an interpretation from loss (1) that connects directly to statements about \(\theta\), without requiring any intermediary steps where we consider error rates.

Viewed through our approach, the implications of standard choices of \(\alpha\) can be seen directly. For example, the threshold of \(\alpha = 0.05\), which is used widely in biomedical work and much else, corresponds to the cost of an incorrect \(d = B\) decision being 1/20th that of an incorrect \(d = A\) decision, or, in other words, that an incorrect \(d = A\) decision is roughly 20 times worse than an incorrect \(d = B\) decision. Compare this with, say, physics’s more stringent threshold of \(5\sigma\), corresponding to \(\alpha = 2.9 \times 10^{-7}\), and hence a cost ratio of 3.5 million. This discrepancy may seem alarming, but the stage of scientific research at which the testing is done should also be considered. If determining just the sign of \(\theta\) in a biomedical context represents an early stage of investigation—e.g. establishing whether a treatment is on average harmful or beneficial before studying the magnitude of its effect and how variable it is in different circumstances—then this relatively low ratio may be appropriate. In contrast, physics’s \(5\sigma\)-threshold is explicitly for ‘discovery’ (Lamb, 2012) and so may instead reflect the cost of erroneously rewriting physical laws. For further consideration of the role of multiple testing on choice of \(\alpha\), see Section 2.3.

Our approach also makes it easier to assess whether the form of loss (1) is relevant to scientific questions of interest. To illustrate this, consider instead a significance test decision, where we either decide that the hypothesis \(\theta < 0\) should be rejected, or (following Fisher (1935), page 16) make no decision about the sign of \(\theta\). We note that this approach differs from the hypothesis testing framework, which does not allow for making no decision. For the general importance of ‘no decision’ as an potential conclusion see Tukey (1960): in practice tests are routinely used
in this way when ‘screening’ large numbers of potential signals (Rice, 2010). We denote these decisions as \(d = A\) and \(d = N\) respectively. As in the hypothesis test, the active \(d = A\) decision can be correct (incurring zero loss) or not (incurring positive loss). But the alternative \(d = N\) decision can never be ‘correct’ so receiving zero loss when \(\theta < 0\) does not reflect our goals. Instead, we may argue that making no decision is equally bad regardless of the true sign of \(\theta\), and although it should receive more loss than correctly stating the truth, its loss should be less than an incorrect active \(d = A\) decision. As shown in Appendix A, with these conditions any loss function is equivalent to one of the form

\[
\text{Loss when} \quad \begin{array}{c|c|c}
\theta > 0 & d = A & d = N \\
\theta < 0 & 0 & \alpha \\
\end{array}
\]

(2)

where \(\alpha \in (0, 1)\) is again a positive constant. Although the loss differs from loss (1), the Bayes rule is functionally identical, returning \(d = A\) if and only if \(P(\theta < 0) < \alpha\). Although the connection with large sample frequentist properties obtains as before, the direct interpretation of \(\alpha\) differs: here it represents the relative cost of making no decision, compared with making an incorrect active \(d = A\) decision.

We view these dual views of \(\alpha\) as a strength: they make clear that the same decision (essentially a \(p\)-value-based test) can have distinct interpretations—here as significance and hypothesis tests—which may or may not be appropriate depending on the context. Conversely, this also illustrates directly that attempts to classify \(p\)-value-based tests (or specific \(\alpha\)-levels) as universally ‘right’ or ‘wrong’ are misguided: criticism of any test must account for the use to which that test is put.

To extend this decision theoretic approach to more complex decisions, we consider two-sided significance tests. Extending loss (2), we still consider only the sign of \(\theta\) but now with potential decisions \(d = A, d = B\) or \(d = N\). As before we incur the minimal loss for any correct decisions, and constant loss for making no decision. We also restrict the losses such that any incorrect decision incurs more loss than making no decision. Without loss of generality, this provides loss function

\[
\text{Loss when} \quad \begin{array}{c|c|c|c}
\theta > 0 & d = A & d = N & d = B \\
\theta < 0 & \alpha_A & \alpha_B & 0 \\
\end{array}
\]

(3)

for \(\alpha_A, \alpha_B \in (0, 1)\). To ensure that decision \(d = N\) can never be ruled out without knowledge of the posterior—i.e. that at least some rules taking decisions \(d = N\) are admissible—we shall also require that \(\alpha_A + \alpha_B < 1\). A proof is given in Appendix A; for related discussions of restricting loss functions for admissibility and coherence see Fossaluza et al. (2017) and Schervish (1996). For prior work featuring losses for this ‘three-decision problem’ see for example Duncan (1965).

Under these conditions, the Bayes rule sets \(d = A\) if \(P(\theta < 0) < \alpha_A\), \(d = B\) if \(P(\theta > 0) < \alpha_B\) and otherwise makes no decision. This is again mostly intuitive—choosing a sign based on having sufficient posterior support for that sign—but the thresholds are determined by directly considering relative losses. The lower–upper thresholds, \(\alpha_A\) and \(\alpha_B\), simply state the relative costs of incorrectly deciding \(d = A\) or \(d = B\) versus making no decision. As before, the Bayesian justification is simple but the familiar frequentist interpretation follows as a large sample result: \(\alpha_A\) and \(\alpha_B\) are the threshold significance values for \(p\)-values by using a standard two-sided test. When the losses are symmetric with regard to above–below decisions, we have \(\alpha_A = \alpha_B\). With
the constraint that $\alpha_A + \alpha_B < 1$ it is natural to define $\alpha_A = \alpha_B = \alpha/2$ for some $\alpha \in (0, 1)$, giving a symmetric version of loss (3). An important feature of the Bayes rule in this case is that it only makes an active decision when
\[ \hat{P} = 2 \min \{P(\theta < 0), P(\theta > 0)\} \] (4)
is below $\alpha$. This summary of the Bayes rule has the ‘doubling the smallest tail’ behaviour of $p$-values for two-sided tests (or equivalently halving the significance threshold) that is not present in one-sided tests. Informally, this factor of 2 may be seen in the Bayesian approach as reflecting the fact that the expected loss for any rule now receives contributions from both $d = A$ and $d = B$ decisions; if the maximum of these two terms is to be compared against the $d = N$ contribution at a cost ratio equivalent to that seen in the one-sided test (which considers only $d = A$ versus $d = N$, say) then the costs for incorrect decisions $d = A$ and $d = B$ must be increased. Also of note, although it is known that two-sided but not one-sided $p$-values are incompatible with standard Bayesian measures of evidence against the null (Berger and Sellke, 1987; Casella and Berger, 1987), we see that no such difficulty arises when both one- and two-sided tests are instead viewed as decisions about the sign of $\theta$.

### 1.2. Riskiness of significance tests

A common issue in teaching statistical tests is the difference in interpretation between obtaining a significant result in well and poorly powered situations: see for example the discussion by Royall (1986). As we now show, the decision theoretic approach that we used to motivate testing-based decisions also straightforwardly provides tools for assessing the reliability (or unreliability) of these decisions.

Having defined a loss function, we consider the corresponding risk of different rules in different situations. The risk of a specific rule is the expected loss that is incurred by applying that rule for some specified underlying true value of $\theta$, where this expectation is over repeated samples.

To illustrate the risks of testing decisions, in Fig. 1 we consider the frequentist expected risk in the classic normal location problem where $n$ observations are sampled independently from $N(\theta, \sigma^2)$, with $\sigma > 0$ known. Within Fig. 1 we consider this problem from the perspective of one-sided significance testing (Fig. 1(a), where the risk shown is for loss (2)) and two-sided significance testing (Fig. 1(b), where for simplicity the risk shown is for the symmetric version of loss (3) scaled so that the loss for no decision is $\alpha$). We illustrate the problem in a weak prior–strong data setting (displayed in green, with the diffuse shape of the prior shown below the risk curve) and in a strong prior–weak data setting (displayed in red, with a more peaked prior also shown below the risk curve). Finally, the values of the Bayes risks—i.e. the risk function averaged over the prior—are displayed as broken lines. For a fair comparison, all priors are normal and centred at zero. In both examples, $\alpha = 0.05$.

In Fig. 1’s weak prior–strong data setting, for both one-sided and two-sided tests we see a peak in the risk near $\theta = 0$, reaching approximately $\alpha(2 - \alpha)$. This region near $\theta = 0$ is where deciding a sign is unreliable (because of the small effect sizes for which we have prior support), meaning that active decisions—when made—are often wrong. However, for large $n$ this peak is concentrated in a very small region near $\theta = 0$, unless the true $\theta$ lies there the risk will be small.

In the strong prior–weak data setting, the Bayes rule has a much more diffuse peak near $\theta = 0$. This is because the strong prior means that for most samples the posterior is heavily centred at $\theta = 0$, so no decision is made and cost $\alpha$ is typically incurred. This reluctance to make an active decision makes the risk notably greater, compared with a weaker prior, when $\theta$ truly lies between 1 and 5 standard errors away from 0. With even stronger priors (or weaker data) the risk function has a larger plateau around $\theta = 0$ where the risk—even of this optimal test—is
Fig. 1. Expected risk for the Bayes rules for loss (2) and the symmetric version of loss (3) scaled so that the loss for no decision is $\alpha$, with $\alpha=0.05$, for the classic normal location problem with independent $N(\theta, \sigma^2)$ samples where $\sigma$ is known (for the weak prior–strong data setting the prior for location $\theta$ is $N(0, 3^2 \sigma^2/n)$, and for the strong prior–weak data setting it is $N(0, 0.8^2 \sigma^2/n)$; also shown are the values of the Bayes risks (---, ---)---i.e. the risk function averaged over the prior—and $\alpha$, the cost associated with ignoring the data and not making any active decision): (a) one-sided significance test; (b) two-sided significance test.

extremely close to $\alpha$. This illustrates that data in such settings provide only a minor benefit over simply making no decision regardless of almost any plausible data. In short, the data can be expected to be close to useless for learning about the sign of $\theta$—the study is nearly futile.

Interestingly, the specification of the loss also means that this notion of futility extends to the classic $Z$-test—for which the risk is given by letting the weak prior in Fig. 1 tend to its flat limiting case. Using the weak prior shown, the risk for the $Z$-test is graphically almost indistinguishable from this limiting case. For the $Z$-test with $\alpha=0.05$, the region where the risk is greater than $\alpha$ occurs when $\theta \in (-0.78 \sigma/\sqrt{n}, 0.78 \sigma/\sqrt{n})$—corresponding to tests with no more than 12% power. In these settings, the zero losses that come from correct statements about the sign of $\theta$ (up to 12% of the time) are more than offset by the rarer but much larger losses that are incurred when incorrect sign decisions are made, and the total risk exceeds the loss for making no decision. In other words with power below 12%, rather than implementing the $Z$-test, it is always less risky to ignore the data and just to report no decision. This further emphasizes that, even if a significant result is achieved, and even with other issues such as ‘$p$-hacking’ and selective reporting (Wasserstein et al., 2016), weak data are unlikely to provide a strong rationale for correct decisions about the sign of $\theta$.

The behaviour in Fig. 1 also illustrates issues with extremely diffuse priors, which despite careful discussion in the literature (e.g. Royall (1997), pages 174–175) are often used in an attempt to make the prior minimally informative. By making the posterior essentially proportional to the
likelihood, these replicate likelihood-based inference—in this case the $Z$-test. But considering the Bayes risk—the regular risk averaged over the prior for $\theta$—shows how these priors actually inform analysis. Using a diffuse prior, almost all of the support is for extremely large values of $\theta$ (where the risk is low) so the Bayes risk is also low. As the prior becomes increasingly diffuse the Bayes risk goes to zero, meaning that the prior indicates increasing certainty that the sign can be determined with little risk—and exactly how little risk depends strongly on how diffuse the prior is. We see that, although the choice of prior may not much affect the testing result $d$, for assessing that test’s risk—which should matter when using a test in practice—the choice of prior is strongly informative.

To evaluate the reliability of the test we have assessed the average loss incurred from a loss such as loss (2) or (3), in which a categorical decision $d$ is made. For assessing the ability of the tail area (such as $P$ in loss (4)) as a basis for decisions we can similarly assess the risk of loss functions for the dual problem, of setting a trade-off rate between making no decision and an incorrect decision about the sign of $\theta$. This is explored in Appendix B: in particular its close connection with measures of severity (Mayo and Spanos, 2006).

2. Extensions

2.1. Tests based on Bayes factors

Perhaps the most contentious issue in comparisons of Bayesian and frequentist testing is the use of Bayes factors (BFs), which compare prior and posterior odds in support of a hypothesis (and thereby serve to measure the information gained through data), versus hypothesis tests. For priors with point masses at null values, see for example Aitkin (1991) for a discussion of the possibly extreme sensitivity to the prior at non-null values, and the possibility of the Jeffreys–Lindley paradox (Lindley, 1957), where the two approaches can totally conflict. For tests focused on the sign of a parameter, we show that using BFs appears as a modest extension of the results in Section 1.1.

The BF for the sign of $\theta$ (in support of $\theta > 0$) is an odds ratio, comparing the posterior odds for $\theta > 0$ with the prior odds for $\theta > 0$. It can be variously written as

$$BF = \frac{\text{posterior odds}(\theta > 0)}{\text{prior odds}(\theta > 0)} = \frac{P(\theta > 0)/P(\theta < 0)}{P^*(\theta > 0)/P^*(\theta < 0)} = \frac{P(\theta > 0)/P^*(\theta < 0)}{P(\theta < 0)/P^*(\theta > 0)},$$

where $P(E)$ and $P^*(E)$ respectively denote the probability of event $E$ under the posterior and prior distributions. The BF can also be represented as the ratio of the integrated likelihoods of the data, where the integration is with respect to the prior constrained to $\theta > 0$ and $\theta < 0$ in the numerator and denominator respectively. Intuitively, if the BF is large, the data provide reason to believe more strongly in $\theta > 0$ than we did a priori—but some threshold value is needed if testing decisions are to be based on the BF.

Losses by using $\theta$ only through its sign—such as loss (3)—do not provide this form of test, as only the posterior determines the decision, not the prior. We note that the relative costs of decisions are assumed to be free of dependence on study design or other features not reflected by $\theta$—the argument being that we should make the same decisions about $\theta$ regardless of how this information was obtained. But with two modifications to the framework of Section 1.1 we can obtain these tests as decisions. First, as well as $\theta$ we consider another parameter $\theta^*$, that we shall call a clone parameter, that has the same prior as $\theta$ and comes from the same parameter space, but which is independent of $\theta$ and which is not updated by the data. This new parameter is critical to introducing tests based on BFs—it provides the reference point without which there can be no notion of measuring the relative strength of prior and posterior beliefs. Second, we
switch from decisions on the sign of $\theta$ to those on the relative magnitude of $\theta$ and its clone $\theta^*$, and the decisions that we now make are that $d = \text{larger}(L)$ (that $\theta > \theta^*$), $d = N$ (as before) or $d = \text{smaller}(S)$ (that $\theta < \theta^*$). But, importantly, we consider the state of $\theta$ and $\theta^*$ through their signs only—the difference of the signs, not the sign of the differences.

With three possible decisions and four possible values of the pair $\{\text{sign}(\theta), \text{sign}(\theta^*)\}$—each sign can be positive or negative—we must consider 12 possible loss values. With the loss depending only on the signs of $\theta$ and $\theta^*$, the system described provides no way of identifying a correct or incorrect ordering of $\theta$ and $\theta^*$ when they both have the same sign, and so we accord all three decisions the same loss: $l_{PP}$ and $l_{NN}$ for respectively positive and negative $\theta$ and $\theta^*$. This step is key if decisions based on BF's alone are to result. For the other six losses, two are 'correct' decisions which incur zero loss, as in Section 1.1, and we similarly accord $d = N$ decisions identical loss regardless of the truth. Finally, we insist that incorrect decisions are worse than any $d = N$ decision. This provides the following loss:

<table>
<thead>
<tr>
<th>Loss when</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d = L$</td>
<td>$d = N$</td>
</tr>
<tr>
<td>$\theta^* &gt; 0, \theta &gt; 0$</td>
<td>$l_{PP}$</td>
</tr>
<tr>
<td>$\theta &lt; 0$</td>
<td>$1 + B_L$</td>
</tr>
<tr>
<td>$\theta^* &lt; 0, \theta &gt; 0$</td>
<td>$0$</td>
</tr>
<tr>
<td>$\theta &lt; 0$</td>
<td>$l_{NN}$</td>
</tr>
</tbody>
</table>

where $B_L$ and $B_S$ are positive constants. Under this loss, the BF emerges automatically as the statistic on which a decision is based: the Bayes rule returns decision $d = L$ if $BF > B_L$, $d = S$ if $BF < 1/B_S$, and $d = N$. To constrain the loss so rules allowing $d = N$ are admissible, we must insist that $B_SB_L > 1$; in practice it would be common to set both $B_S, B_L > 1$. This decision is partly intuitive—we decide on an ordering only when the posterior odds of $\theta$ having one sign sufficiently exceed the odds of $\theta^*$ having the other sign. The thresholds $B_L$ and $B_S$ state, relative to the cost of making no decision, exactly how much worse it is to make a wrong $d = L$ or $d = S$ decision respectively. This direct interpretation avoids the arbitrariness of current recommendations—described by Kass and Raftery (1995) as ‘rough descriptive statements’—for interpreting BF values.

2.2. Testing many null values or interval decisions

We now consider changing the null value from a fixed $\theta = 0$ to the general case of $\theta = \theta_0$. Considering testing decisions for all values of $\theta_0$ leads to straightforward motivations for credible sets. For simplicity we consider only the symmetric two-sided setting of loss (3). For a general null $\theta = \theta_0$ and factoring out the common factor $\alpha/2$, this loss can be written as

$$L(d, \theta; \theta_0) = 1_{d=A, \cap \theta < \theta_0} + \frac{\alpha}{2} 1_{d=N} + 1_{d=B, \cap \theta > \theta_0}.$$  

To generalize it to a loss that makes a testing decision for each $\theta_0$—denoted $d_{\theta_0}$—we integrate pointwise contributions $L(d_{\theta_0}, \theta; \theta_0)$ over all $\theta_0$, where this integration is done with respect to some measure $\pi$. The decision can also be viewed as making set-valued decisions $A, B$ and $N$ respectively denoting the sets of $\theta_0$-values where we decide that $\theta$ is above those $\theta_0$, below those $\theta_0$ or for which we make no decision. This can be written as

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

where $\pi(S)$ denotes the measure of set $S$. 
For loss functions that are additive in distinct decisions, the overall Bayes rule is just the Bayes rule for each component considered individually (Lehmann, 1957b). Hence the Bayes rule for this loss—for any measure $\pi$ that provides positive support at all points of the parameter space for $\theta$—sets $A$ as the left-hand $(100\alpha/2)^\%$ tail of the posterior, $B$ as the corresponding upper tail and $N$ as the central $100(1-\alpha)^\%$ credible interval. We note that the precise choice of $\pi$ does not affect the decisions for $A$, $B$ and $N$, as at each $\theta_0$ the support of $\pi$ acts as a multiplier of loss (5), and scaling losses by positive constants does not affect the Bayes rule. Also of note, the decision sets $A$, $B$ and $N$ are all connected, although the loss function makes no specific reference to such a property.

We have thus provided a direct motivation for the central credible interval: it is the set of values for which we make no decision about the ordering of $\theta$ and those values. As with the component decisions about the sign of $\theta$, our motivation of the interval does not use the notion of hypothetical replicate studies; nor does it specify that we accept any single $\theta$-values as ‘true’. This simplicity is in contrast with the subtlety of the standard confidence interval, which is widely misinterpreted (Hoekstra et al., 2014). However, by the Bernstein–von Mises theorem, the credible interval does provide an approximate $100(1-\alpha)^\%$ confidence interval under correct-model assumptions and regularity conditions, so the simpler interpretation could often be applied as an approximate interpretation of confidence intervals.

2.3. Testing many parameters

Lehmann’s additive loss functions, which was used in Section 2.2, can also be used when we consider decisions about the signs of multiple parameters, which we here denote as $d_j$ and $\theta_j$ for $1 \leq j \leq m$. For notational simplicity here we consider only one-sided decisions, though the arguments generalize to two-sided methods directly. To construct a loss function for these decisions, one approach would be simply to add up $m$ contributions $L(d_j, \theta_j; \theta_0)$, which results in $m$ tests totally uncorrected for multiplicity. But we now show that similar approaches can take multiplicity into account, by instead stating that the loss for many incorrect sign decisions is not simply a count of those wrong decisions.

One extreme form of dependence is where the loss for any set of mistakes about $\theta_j$ is identical—so any one wrong sign decision is as bad as any two, or three, up to $m$. Keeping the rest of the loss structure identical to loss (2)—i.e. no loss for a correct decision, loss $\alpha_j$ for no decision on parameter $\theta_j$—results in overall loss function

$$L(\theta, d) = \left( \sum_{j:d_j=N} \alpha_j \right) + 1_{\{j:d_j=A \text{ and } \theta_j<0\}}. \quad (6)$$

Because the second term is never more than 1, as written the loss function permits situations where making every $d_j = A$, even if these decisions are all wrong, to incur less loss than setting them all to $d_j = N$. To avoid this situation, we must force $\sum_{j:d_j=N} \alpha_j = \alpha$ for some $\alpha < 1$. The value of this $\alpha$ has a straightforward interpretation: it states the relative cost of making no decisions at all, versus making at least one sign error. Of course, if all $\alpha_j$ are equal, this means setting all $\alpha_j = \alpha/m$, and we recover a Bayesian analogue of the Bonferroni correction of the level of significance, where the classical type I error rate threshold is adjusted by dividing by the number of tests performed. More generally, selecting different $\alpha_j$ for different parameters $\theta_j$ such that their total $\sum_{j=1}^m \alpha_j = \alpha$ reflects a simple form of ‘$\alpha$-spending’ (Lan and DeMets, 1983). Notably, our derivation of these multiple-testing corrections requires no consideration of averaging losses over either the posterior or over random samples. It is thus simpler than the standard motivation of Bonferroni correction, as a way to control the familywise error rate.
The connections to Bonferroni correction go deeper than this. The posterior expected loss under equation (6) is

\[
\left( \sum_{j:d_j=N} \alpha_j \right) + \mathbb{P}[\bigcup\{ j: d_j = A \text{ and } \theta_j < 0 \}] \leq \sum_{j:d_j=N} \alpha_j + \sum_{j:d_j=A} \mathbb{P}(\theta_j < 0),
\]

where the bound follows by Bonferroni’s inequality. Hence we can conservatively approximate the true Bayes rule by setting \( d_j = A \) if and only if \( \mathbb{P}(\theta_j < 0) < \alpha_j \), and hence we have established a motivation for Bonferroni-corrected tests.

Bonferroni’s correction is often criticized for being ‘conservative’ (Bland and Altman, 1995), but this Bayesian analogue makes it clear that there are two distinct senses in which the approach is conservative. First, it is conservative to treat any number of sign errors as having equal loss, where making one sign error is as bad as making \( m \) of them, and the cost of this error is more than making non-decisions for all \( m \) parameters. This forces non-decisions to be cheap, and active decisions to be expensive.

Second, when the events \( \{ \theta_j < 0 \} \) for different \( \theta_j \) are strongly dependent in the posterior, the bound that is established by Bonferroni’s inequality will be loose, and the approximate Bayes rule (that sets \( d_j = A \) if and only if \( \mathbb{P}(\theta_j < 0) < \alpha_j \)) will be a poor approximation of the true Bayes rule. Moreover, the extra terms that are added to the posterior expected loss by use of Bonferroni’s inequality are all in \( d_j = A \), so the approximation inflates the losses for sign errors while leaving non-decision losses unchanged. We see that as well as being a poor approximation it is also seen to be a conservative approximation; it will set more decisions \( d_j = N \) than will the true Bayes rule.

The risk of conflating these two forms of conservatism—a real problem when non-experts consider multiple-testing problems—can be reduced when we consider a slightly different loss function, in which we trade off an average of weighted ‘no decision’ losses against the sum of losses for sign errors. The loss is

\[
\frac{1}{m} \sum_{j=1}^{m} \alpha_j \mathbf{1}_{d_j=N} + \sum_{j=1}^{m} \mathbf{1}_{d_j=A} \mathbf{1}_{\theta_j<0}
\]

and we constrain each \( 0 < \alpha_j < m \), so that making no decision for any \( \theta_j \) incurs less loss than making an incorrect sign statement about \( \theta_j \). The Bayes rules here sets \( d_j = A \) if and only if \( \mathbb{P}(\theta_j < 0) < \alpha_j/m \)—i.e. below the Bonferroni-corrected threshold. With this motivation the only conservatism comes from trading off average no-decision losses versus the sum of losses for sign errors—there is no conservative approximation of the Bayes rule. Analogous non-decision theoretic results were presented by Gordon et al. (2007), who used them to argue that Bonferroni’s ‘notorious’ conservatism can be viewed as a misconception. Lewis and Thayer (2009) provided a strongly related decision theory approach that motivates the Benjamini–Hochberg algorithm (Benjamini and Hochberg, 1995) as a conservative approximation to a Bayes rule.

2.4. Allowing for nuisance parameters

Inference in the presence of nuisance parameters is a key topic in practical systems of inference, although it seems conceptually simpler in the Bayesian approach than elsewhere (Berger et al., 1999). In this section we describe and illustrate how nuisance parameter uncertainty can be included when we assess the risk of a test.

A key observation is that the values of nuisance parameters, when present, determine the risk of the test along with the parameter \( \theta \) of interest and the study design. Propagating just the
nuisance parameter uncertainty is therefore sufficient to tell us how uncertain our risk assessment is, at any particular \( \theta \).

As an example, we again consider the normal location problem from Section 1.2, using the symmetric version of loss (3), but now with unknown \( \sigma \) in the analysis. For convenience we assume a gamma distribution on \( \sigma^2 \), but standard numeric integration approaches could be used for general settings. Our example is again normal location testing of \( \theta \), now specifying \( n = 25 \) and with true \( \sigma = 1 \). We highlight the effect of less or more precise information on \( \sigma \) by two gamma distributions for \( \sigma \). Both have the same median (\( \sigma = 1 \)) but we use shape parameters 25 and 100 to reflect less or more precise knowledge of \( \sigma \). The less precise distribution provides approximately the accuracy with which \( \sigma \) would be known from a sample of size 25 alone, whereas the more precise distribution could be the result of having strong prior knowledge of \( \sigma \).

To show the uncertainty in the riskiness of tests, in Fig. 2 we have plotted the frequentist expected risk curves for 200 values of \( \sigma \) randomly sampled from the two gamma distributions. These expected risk curves are shown as light blue curves, the amalgamation of which appears as a region where the true risk curve could reasonably be thought to lie. We also show the risk at median \( \sigma \) (i.e. at \( \sigma = 1 \), by construction) and at the 2.5% and 97.5% quantiles of \( \sigma \). At the bottom of Fig. 2 we display the shape of the prior on \( \theta \), which is the same as the prior on \( \theta \) displayed in the weak prior–strong data setting of Fig. 1.

We see that uncertainty around the value of \( \sigma \) has little effect on the assessment of risk if \( \theta \) is either close to 0 or many standard errors from 0. But for intermediate values of \( \theta \) the risk is uncertain. For example, at \( \theta = 0.25 \), the propagated uncertainty bounds on the expected risk are (0.009, 0.051) for shape 25 and (0.028, 0.048) for shape 100. With imprecise knowledge of \( \sigma \) we are left with imprecise knowledge of whether the test is a low risk advance versus being futile in the sense of Section 1.2 where simply making no decision regardless of the data incurs less expected risk. With more precise knowledge of \( \sigma \) the risk assessment indicates that the test will provide a modest benefit, under plausible values of the nuisance parameters.
3. Discussion

We have provided a simple decision theoretic framework that motivates one- and two-sided significance tests, hypothesis tests, p-values, BF-based decisions, intervals, multiple-testing correction and measures of a test’s reliability. Although not explored here, more methods can be similarly developed in this way.

These properties lend themselves to forms of efficiency in understanding statistical methods, compared with many current presentations. Once the basic idea of declaring the sign of \( \theta \) has been grasped, extensions such as those listed above—some of which would be considered advanced—can be made with little extra insight, but also with little mathematical complexity. The normative element of decision theory, moving automatically to an optimal rule given a loss function, also means that how to do the analysis follows directly from stating what we want the analysis to address—there is no need to choose test statistics, or to confirm test unbiasedness or to examine any uniform optimality criteria. Moreover, justifying why these choices are appropriate is not required, assuming that the loss function in use describes a decision that we actually want to make. Finally and perhaps most importantly, the simplicity of focusing on just the sign of \( \theta \) and limited decisions about it should emphasize that tests provide only a limited form of inference. A decision that some \( \theta > 0 \) (versus making no decision or deciding that \( \theta < 0 \)) clearly leaves aspects of \( \theta \) unexplored—which contrasts with the accept–reject finality that is often currently taught. If some property of \( \theta \) beyond its sign is of interest (say its magnitude) then the testing losses that are studied here are seen to be inappropriate, as \( \theta \) enters them by its sign only. In this way, criticism of the approach becomes efficient as, when the losses are found wanting, we can immediately then ask what is of interest—and to move on to inference based on a more relevant loss function.

Of course, the approach has limitations. As no natural ordering of parameters values exists in higher dimensions, building methods by deciding whether \( \theta > \theta_0 \) may not be straightforward for multivariate \( \theta \) and \( \theta_0 \). Extremely high dimensional parameters may also cause more general difficulties when Bernstein–von Mises results may not apply, and where complete-class results may not always permit Bayesian criticisms to inform us about the appropriateness of (admissible) frequentist methods. Well-documented concerns about Bayesian use of prior information should also be considered, particularly when priors are difficult to elicit, or where there are concerns that the use of priors may introduce opportunities to game a regulatory system.

In conclusion, we feel that the direct connections that we have described, between scientific questions about the signs of parameters and statistical methods provide a useful framework on which to build better understanding of existing approaches. Naturally, they also provide a framework where novel methods could be developed.

Acknowledgements

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services; nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

Appendix A: Proofs

A.1. General case of hypothesis testing loss (1)

The general form of the loss and its expectation (for either decision) under the posterior can be written as
Loss when $\theta > 0$ $d = A$ $d = B$

\[
\begin{array}{ccc}
\theta > 0 & l_{TA} & l_{FB} \\
\theta < 0 & l_{TB} + 1 - \alpha & l_{TA} + \alpha \\
\text{Expected posterior loss} & l_{TA} + (l_{TB} - l_{TA} + 1 - \alpha)\mathbb{P}(\theta < 0) & l_{TA} + \alpha + (l_{TB} - l_{TA} - \alpha)\mathbb{P}(\theta < 0)
\end{array}
\]

where $l_{TA}, l_{TB}, l_{FA}$ and $l_{FB}$ denote the losses incurred when we make a true $T$ or false $F$ decision that $\theta$ is above $A$ or below $B$. We restrict the loss as follows. To reflect that it is preferable to make a correct versus an incorrect decision, we insist that $l_{TA} < l_{FA}$ and $l_{TB} < l_{FB}$. More specifically, we penalize incorrect decisions more than correct decisions regardless of whether those decisions are that $\theta > 0$ or that $\theta < 0$, which equates to insisting that $\max(l_{TA}, l_{TB}) < \min(l_{FA}, l_{FB})$. Based on these assumptions alone, the expected posterior loss under the two actions is to set $d = A$ if and only if

\[\mathbb{P}(\theta < 0) < \frac{l_{FA} - l_{TA}}{l_{FA} - l_{TA} + l_{FB} - l_{TB}}.\]

To obtain the version presented as loss (1), we note that the decision problem is unchanged by three actions: adding a constant to all cells, scaling all cells by a positive constant or adding a function of $\theta$. Without loss of generality we can scale the loss so that $l_{FA} - l_{TA} + l_{FB} - l_{TB} = 1$, and reparameterize so that $l_{FB} - l_{TA} = \alpha \in (0, 1)$. This leaves loss–expected loss and decisions that set $d = A$ if and only if $\mathbb{P}(\theta < 0) < \alpha$.

By further adding row-specific constants—that set $l_{TA}$ and $l_{TB}$ to 0—then the decision is unchanged but the version that is shown as loss (1) is obtained without loss of generality.

A.2. General case of one-sided significance testing loss (2)

Making the evaluation that the loss for making no decision is constant regardless of $\theta$, the general forms are

Loss when $\theta > 0$ $d = A$ $d = N$

\[
\begin{array}{ccc}
\theta > 0 & l_{TA} & l_{N} \\
\theta < 0 & l_{FA} & l_{N} \\
\text{Expected posterior loss} & l_{TA} + (l_{FA} - l_{TA})\mathbb{P}(\theta < 0) & l_{N}\mathbb{P}(\theta < 0)
\end{array}
\]

To reflect that making no decision is both worse than correctly deciding $d = A$ and better than incorrectly deciding $d = A$, we assume that $l_{TA} < l_{N} < l_{FA}$. This results in the Bayes rule setting $d = A$ if and only if

\[\mathbb{P}(\theta < 0) < \frac{l_{N} - l_{TA}}{l_{FA} - l_{TA}}.\]
Rescaling the loss so that $l_{FA} - l_{TA} = 1$, and parameterizing $l_N - l_{TA} = \alpha$, we obtain

$$
\begin{array}{ccc}
\text{Loss when} & \text{Decision} \\
\hline
\theta > 0 & l_{TA} & l_{TA} + \alpha \\
\theta < 0 & l_{TA} + 1 & l_{TA} + \alpha \\
\hline
\text{Expected posterior loss} & l_{TA} + \mathbb{P}(\theta < 0) & l_{TA} + \alpha
\end{array}
$$

which provides the Bayes rule as setting $d = A$ if and only if $\mathbb{P}(\theta < 0) < \alpha$. Adding a constant to the loss we can obtain $l_{TA} = 0$, and loss (2) is obtained without loss of generality.

### A.3. General case of two-sided significance testing loss (3)

Making the evaluation that the loss for making no decision is constant regardless of $\theta$, and that the losses for any correct response are equal, the general forms are

$$
\begin{array}{ccc}
\text{Loss when} & \text{Decision} \\
\hline
\theta > 0 & l_T & l_N & l_FB \\
\theta < 0 & l_{FA} & l_N & l_T \\
\hline
\text{Expected posterior loss} & l_T + (l_{FA} - l_T)\mathbb{P}(\theta < 0) & l_N & l_FB + (l_T - l_FB)\mathbb{P}(\theta < 0)
\end{array}
$$

As in Appendix A.1, we insist that $l_T < l_{FA}$ and $l_T < l_FB$. We also expand the previous requirement that $l_T < \min(l_{FA}, l_FB)$ to reflect that making no decision is no worse than making an incorrect decision and no better than making a correct decision. This equates to insisting that $l_T < l_N < \min(l_{FA}, l_FB)$.

To ensure that rules that set $d = N$ are admissible—i.e. to ensure that we cannot guarantee that it is always better to follow some rule which only ever takes $d = A$ or $d = B$ decisions—we must insist that $l_N$ is sufficiently low that

$$
l_N < \min\{l_T + (l_{FA} - l_T)\mathbb{P}(\theta < 0), l_FB + (l_T - l_FB)\mathbb{P}(\theta < 0)\}
$$

for at least some value of $\mathbb{P}(\theta < 0)$. This condition states that the difference in loss between no decision and any true decision is less than the harmonic mean of the differences between false above and below decisions and no decision. The two terms being considered are (respectively) increasing and decreasing linear functions of $\mathbb{P}(\theta < 0)$, so the requirement is equivalent to insisting that

$$
l_N < l_T + \frac{(l_{FA} - l_T)(l_FB - l_T)}{l_FB - l_T + l_{FA} - l_T}.
$$

Under these conditions, the Bayes rule sets $d = A$ if and only if

$$
\mathbb{P}(\theta < 0) < \frac{l_N - l_T}{l_{FA} - l_T},
$$

sets $d = B$ if and only if

$$
\mathbb{P}(\theta < 0) > 1 - \frac{l_N - l_T}{l_FB - l_T},
$$

and otherwise—i.e. for some intermediate $\mathbb{P}(\theta < 0)$—sets $d = N$.

Again following Section A.1, we now reparameterize in terms of the thresholds just established, writing

$$
\alpha_A = \frac{l_N - l_T}{l_{FA} - l_T}.
$$
and
\[ \alpha_B = \frac{l_N - l_T}{l_{FB} - l_T}, \]
or equivalently
\[ l_{FA} = l_T + \frac{l_N - l_T}{\alpha_A} \]
and
\[ l_{FB} = l_T + \frac{l_N - l_T}{\alpha_B}. \]
This provides

\[ \text{Loss when } \theta > 0 \quad \text{Decision} \]
\[ d = A \quad l_T \quad l_N \quad l_T + \frac{l_N - l_T}{\alpha_B} \]
\[ d = N \quad l_T + \frac{l_N - l_T}{\alpha_A} \quad l_N \quad l_T \]
\[ d = B \quad l_T + \left( \frac{l_N - l_T}{\alpha_A} \right) \mathbb{P}(\theta < 0) \quad l_N \quad l_T + \frac{l_N - l_T}{\alpha_B} + \left( \frac{l_T - l_N}{\alpha_B} \right) \mathbb{P}(\theta < 0) \]

and the condition above on \( l_N \) now states that
\[ l_N < l_T + \frac{(l_N - l_T)^2}{\alpha_A \alpha_B \{ (l_N - l_T)/\alpha_A + (l_N - l_T)/\alpha_B \}}, \]
which after some cancellation equates to insisting that \( \alpha_A + \alpha_B < 1 \). Scaling this loss by \( \alpha_A \alpha_B / (l_N - l_T) \) and subtracting \( l_T \alpha_B \alpha_B / (l_N - l_T) \) from all cells, we obtain the form (3), without loss of generality.

The insistence that both forms of true response incur the same loss can be relaxed to obtain the same rules, but slightly different expected losses, as we now describe. With the general form

\[ \text{Loss when } \theta > 0 \quad \text{Decision} \]
\[ d = A \quad l_T \quad l_N \quad l_T + \frac{l_N - l_T}{\alpha_B} \]
\[ d = N \quad l_T + \frac{l_N - l_T}{\alpha_A} \quad l_N \quad l_T \]
\[ d = B \quad l_T + (l_{FA} - l_{TA}) \mathbb{P}(\theta < 0) \quad l_N \quad l_{FB} + (l_{TB} - l_{FB}) \mathbb{P}(\theta < 0) \]

we can insist that \( l_{TA} < l_{FA} \) and \( l_{TB} < l_{FB} \) as before, and that \( \max(l_{TA}, l_{TB}) < \min(l_{FA}, l_{FB}) \) to reflect that making no decision is no worse than making an incorrect decision and no better than making a correct decision, which equates to insisting that \( \max(l_{TA}, l_{TB}) < l_N < \min(l_{FA}, l_{FB}) \). To ensure that rules which set \( d = N \) are admissible we must insist that
\[ l_N < \frac{l_{FA}(l_{FB} - l_{TB}) + l_{TB}(l_{FA} - l_{TA})}{l_{FB} - l_{TB} + l_{FA} - l_{TA}} = \frac{l_{FA}(l_{FA} - l_{TA}) + l_{TA}(l_{TB} - l_{FB})}{l_{FB} - l_{TB} + l_{FA} - l_{TA}}. \]

Under these conditions, the Bayes rule sets \( d = A \) if and only if
\[ \mathbb{P}(\theta < 0) < \frac{l_N - l_{TA}}{l_{FA} - l_{TA}}, \]
sets \( d = B \) if and only if
\[ \mathbb{P}(\theta < 0) > 1 - \frac{l_N - l_{TB}}{l_{FB} - l_{TB}}. \]
and otherwise—i.e. for some intermediate $\mathbb{P}(\theta < 0)$—sets $d = N$.

Again following Appendix A.1, we now reparameterize in terms of the thresholds just established, writing

$$\alpha_A = \frac{l_N - l_{TA}}{l_{FA} - l_{TA}}$$

and

$$\alpha_B = \frac{l_N - l_{TB}}{l_{FB} - l_{TB}},$$

or equivalently

$$l_{FA} = l_{TA} + \frac{l_N - l_{TA}}{\alpha_A},$$

and

$$l_{FB} = l_{TB} + \frac{l_N - l_{TB}}{\alpha_B}.$$  

It is also convenient to rescale $l_{FA} - l_{TA} + l_{FB} - l_{TB} = 1$ and reparameterize so that $l_{FA} - l_{TA} = \gamma$ and $l_{FB} - l_{TB} = 1 - \gamma$, meaning that $\gamma$ describes how strongly we wish to penalize different decisions when $\theta > 0$ compared with when $\theta < 0$. Combining all these steps we obtain

<table>
<thead>
<tr>
<th>Loss when Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d = A$</td>
</tr>
<tr>
<td>$\theta &gt; 0$</td>
</tr>
<tr>
<td>$\theta &lt; 0$</td>
</tr>
<tr>
<td>Expected posterior loss</td>
</tr>
<tr>
<td>Bayes rule: do $d$ if and only if $\mathbb{P}(\theta &lt; 0) &lt; \alpha_A$</td>
</tr>
</tbody>
</table>

Under these losses the tail areas $\mathbb{P}(\theta < 0)$ and $\mathbb{P}(\theta > 0)$ enter the expected loss multiplied by factors of only $\gamma$, so we see that risk evaluations will depend on its value as well the values of thresholds $\alpha_A$ and $\alpha_B$.

**Appendix B: p-values as Bayes rules**

As is well known, p-values are not tests, and so their reliability cannot be assessed by using the risk of testing losses, as presented in Section 1.2. But $p$-values do summarize tests: they are the lowest significance level $\alpha$ such that a ‘reject’ decision would be made with the available data. We can therefore assess their reliability by considering them as the result of a decision-based approach where we aim to summarize potential tests. Specifically, we develop decisions that set the trade-off rate between a reject decision, versus the alternative of making no decision. To aid with later comparisons with severity—which is typically presented as a criticism of one-sided inferences—we do this for one-sided significance tests, although its extension to two-sided tests is straightforward.

We denote the data-dependent decision trade-off rate as $a$, in contrast with the fixed trade-off rate denoted $\alpha$ in Section 1.1. Decision $a$ sets the trade-off rate between the constant loss that we would incur for making no decision, versus the loss for making an active but incorrect decision on the sign of $\theta$ around some null value $\theta_0$. The size of the two potential losses is a key part of this trade-off but, to encourage conservative inference, we also weight the entire loss function by a decreasing function of $a$—making large trade-off values less costly. A loss function that provides this is

$$L(a, \theta) = \frac{1}{\sqrt{a}}(a + 1_{\theta < \theta_0})$$  

for decision $a$ between 0 and 1. The Bayes rule sets $a = \mathbb{P}(\theta < \theta_0)$, i.e. the left-hand tail area that is the basis for earlier testing decisions. The minimized expected posterior loss is $2\sqrt{\mathbb{P}(\theta < \theta_0)}$: a function of this
Fig. 3. Proportion of risk from more extreme data sets and $P(p^* < p)$ (standard and Bayesian analogue) under (a) the same tail area of 0.05 and (b) the frequentist $p$-value of 0.05 (in (b), $P(p^* < p)$ is identical across priors and is plotted only once): — , flat prior; — , weak prior; — , strong prior; — — , $Pr(p^* < p)$.
tail area. Although not shown here, a two-sided version of the same loss leads to the same Bayes rule and minimized expected posterior loss, except substituting $P$ from loss (4) for $P(\theta < \theta_0)$.

The risk of the decision for $a$ is
\[
\text{risk}(a, \theta) = \mathbb{E} \left[ \sqrt{P(\theta < \theta_0)} + \frac{1_{\theta < \theta_0}}{\sqrt{P(\theta < \theta_0)}} \right],
\]
where the outer expectation is over repeated samples, generated under some true $\theta$. The risk can be broken into two components: first, where samples would provide decisions $a$ (or equivalently minimized posterior losses) that are at least as small as that actually observed for the data at hand, and a second component where they would not.

For such a fixed $\theta$, if this first component is large, most of the risk that is associated with decisions about the sign of $\theta$ comes from more likely data sets that yield smaller losses, i.e. more extreme data sets. The second component comes from data sets with smaller likelihood that yields much higher loss, i.e. less extreme data sets. Where there is a higher proportion of risk from more extreme data sets it is suggested that the evidence for the conclusions stated should be stronger.

In reality of course, $\theta$ is unknown, so we instead calculate the first component (for a fixed observed $p$-value or Bayesian left-hand tail area) across a range of $\theta$, thereby quantifying the strength of evidence across a range of alternative hypotheses. In this way, the ability to find fault in the null hypothesis across a range of alternatives may be evaluated. Naturally, the aforementioned proportion increases as $\theta$ increases in this one-sided testing scenario. If the plot remains somewhat flat as $\theta$ increases, the true $\theta$ must be quite large to find fault in the null hypothesis; therefore, across a wide range of $\theta$, the observed data (used through its frequentist $p$-value or Bayesian left-hand tail area) hold little sway.

This reasoning is similar to calculating power at a range of fixed $\theta$ to aid data interpretation, though we note that this is not the same as naively plugging in some estimate $\hat{\theta}$ as the true value of $\theta$ in a ‘post hoc’ power calculation. Across a range of $\theta$, this reasoning is also consistent with the notion of ‘severity’, by which one concludes that there is lack of evidence for a claim if there is little or no capability of finding faults in the said claim should it be false (Mayo and Spanos, 2006). Severity is only defined under a frequentist framework and is calculated as $P(p^* > p)$ at fixed values of $\theta$, where $p^*$ denotes the $p$-value that is calculated under a replicate study. In particular, the first component of risk defined above is the same as 1 minus severity (i.e. $P(p^* < p)$) if the loss $L(a, \theta)$ is assumed to be constant across all possible data sets. Severity and the first component will also be similar when $L(a, \theta)$ varies little, when using the Bayes rule for decision $a$, over typical data sets.

Fig. 3 displays both the proportion of risk from more extreme data sets under loss (7) and multiple forms of 1 minus the typical severity measure under multiple prior distributions for parameter $\theta$ with different amounts of evidence from data. Normal priors for $\theta$ centralized at 0, i.e. $\theta \sim N(0, \tau^2)$, are updated by independent $N(\theta, \sigma^2)$ data, where $\sigma^2$ is assumed known, to test $H_0: \theta = \theta_0 = 0$ versus $H_A: \theta > \theta_0 = 0$. Prior distributions on $\theta$ are uninformative, weak or strong (i.e. $\tau^2 = \infty, 3^2, 0.8^2$ respectively).

In Fig. 3(a) the prior is assumed to be updated with observations such that the posterior left-hand tail area $P(\theta < \theta_0)$ is 0.05 (for example the data updating the weaker prior will be more extreme than the strong prior). In Fig. 3(b), we instead fix the standard frequentist $p$-value to be exactly 0.05. The full curves are proportions of risk from more extreme data sets under loss (7), whereas broken curves display $P(p^* < p)$, calculated by using standard frequentist $p$-values (the black curve) or as the proportion of Bayesian tail areas from replicate analyses that are more extreme than the tail area that is actually observed. Since left-hand tails areas are used here, the black broken curve (with the flat prior) displays 1 minus the typical severity measure (since the prior is flat), whereas broken curves of other colours display Bayesian analogues for severity (i.e. using left-hand tail areas with somewhat informative priors). In Fig. 3(a) $P(p^* < p)$ is plotted only once in black, since the curve is identical for all prior distributions.

The general similarity between the curves is apparent: the general behaviour for $\theta > 0$, with curves increasing monotonically up to 1 as the signal strength becomes large, is shared between severity and the proportion of risk that we have considered. For the loss (7) that is used there is no consistent pattern in how stronger or weaker priors lead to respectively better or worse agreement between the two approaches, though the differences between the curves do seem to be of the same order of magnitude as differences in severity when calculated by using standard $p$-values versus using left-hand tail areas from different priors. (See Fig. 3(a)). Just before $\theta = 0$, different assignments of loss for $\theta < 0$ and $\theta > 0$ lead to jumps in the proportion of risk seen under loss (7) but not seen with constant loss under severity.

Accepting severity and similar concepts as assessments of evidence, Fig. 3(a) suggests that, for data viewed as borderline in the Bayesian analysis, more faith should be placed in tests that use flat and weak
priors as opposed to strong priors. In Fig. 3(b), data viewed as borderline in the frequentist analysis lead to more faith in the claim on the sign of \( \theta \) if the prior is stronger. The choice of prior does not affect severity here.

References


