British Journal for the Philosophy of Science

Severe Testing: Error Statistics versus Bayes Factor Tests

Deborah G. Mayo

In the face of today's statistical crisis of science, it is often recommended that statistical significance tests be replaced with Bayes factor tests. In this article, I examine this recommendation. Bayes factor tests, unlike statistical significance tests, only depend on the probability of the data under H_0 and a competitor H_1. They are insensitive to a method's error probabilities such as significance levels, type 1 and type 2 errors, and confidence levels. It might be thought that if a method is insensitive to error probabilities that it escapes the inferential consequences of inflated error rates at the heart of obstacles to replication. I will argue that this is not the case, and that Bayes factor tests can accord strong evidence to a claim H, even though little has been done to rule out H's flaws. There are two reasons: their insensitivity to biasing selection effects, and the fact that H and its competitor need not exhaust the space of relevant possibilities. I will show how this results in a disconnect between Bayes factor tests and error control protocols that are being called for by replication reforms. To solve the problem, I propose that commonly used Bayes factor tests be supplemented with a post-data severity concept in the frequentist error statistical sense. The question is not whether 'severity' can be redefined Bayesianly—of course it can—the question is whether the resulting concept can address today's concerns behind obstacles to replication. I will also respond to criticisms of the severity reformulation of statistical significance tests, and show how it enables avoiding fallacies of statistical tests.

1. Introduction

A main source of handwringing in today's 'statistical crisis of science' (Gelman and Loken [2014]) is that high powered methods and researcher flexibility make it easy to find an impressive-looking effect in a particular study even though it is spurious. The data aligns with a hypothesized effect H, but the test H has passed fails to be stringent or severe. According to severe testers (Mayo [1996], [2018]; Mayo and Spanos [2006], [2011]; Mayo and Cox [2006]; Mayo and Hand [2022]):

Severity Requirement: Data $\boldsymbol{x}$ are evidence for a hypothesis H only to the extent that

H passes a test that probably would have found evidence that H is false (or specifically flawed) just if it is.

This probability is the severity with which H has passed. Statistical significance tests are intended 'as a first line of defense against being fooled by randomness' (Benjamini [2016], p. 1). However, data dredging, multiple testing and cherry picking can result in frequently inferring there is a genuine effect erroneously. The error probability associated with such an inference is high. Unsurprisingly, such data dredged effects often disappear in attempted replications with stricter protocols, and the random variation goes a different way. Methods where inferential assessments require knowing the relevant error probabilities of the method producing $\boldsymbol{x}$ may be called 'error statistical methods' (for example, significance levels, type 1 and type 2 errors, confidence levels). While the replication crisis is fostering preregistration and other protocols to control error probabilities, some also take it as grounds to replace statistical significance tests with alternative methods, some of which are insensitive to error probabilities.[1]

It might be thought that if a method is insensitive to error probabilities then it escapes inflated error rates due to biasing selection effects.[2] I will argue that this is not the case. Insensitivity to error probabilities has serious consequences when it comes to inferring genuine as opposed to spurious effects. My focus in examining this issue is an article advocating subjective Bayes factor tests by van Dongen, Sprenger, and Wagenmakers (VSW), two philosophers of science and a mathematical psychologist (van Dongen et al. [2023]). My arguments are relevant whenever Bayes factors are used as tests, as they typically are. The Bayes factor, B_{10}, is the probability (or density) of observed data $\boldsymbol{x}$ under statistical hypothesis H_1 as opposed to another, H_0. It is the likelihood ratio: $\Pr(\boldsymbol{x}; H_1)/\Pr(\boldsymbol{x}; H_0)$.[3] Following thresholds advocated by Jeffreys ([1961]), Bayes factor testers define test rules moving from a likelihood ratio of H_1 and a competitor H_0 to inferring weak, strong, very strong evidence for H_1 (or H_0).

VSW are clear that the Bayes factor test is insensitive to error probabilities: 'the Bayes factor only depends on the probability of the data in light of the two competing hypotheses. As Mayo emphasizes [...] the Bayes factor is insensitive to variations [of] the sampling protocol that affect the error rates i.e., optional stopping of the experiment' (van Dongen et al. [2023], p. 522). As a result, VSW acknowledge: 'many Bayesians deny

[1] The literature is huge. Some multi-authored sources are (Wasserstein and Lazar [2016] and supplementary comments; Benjamin et al. [2018]; Lakens et al. [2018]; Wasserstein et. al. [2019]; Benjamini et al. [2021]).

[2] 'Biasing Selection Effects: when data or hypotheses are selected or generated (or a test criterion is specified), in such a way that the minimal severity requirement is violated, seriously altered, or incapable of being assessed' (Mayo [2018], p. 92).

[3] B_{10} is the factor by which a ratio of prior probabilities, $\Pr(H_1)/\Pr(H_0)$, could be revised to obtain a ratio of posterior probabilities $\Pr(H_1|\boldsymbol{x})/\Pr(H_0|\boldsymbol{x})$ (see note 28). The ';' is generally used by frequentists, whereas Bayesians use '|'.

that severity should matter at all in inference. They refer to the Likelihood Principle [. . .] According to this line of response, Popper, Mayo and other defenders of severe testing are just mistaken when they believe that severity should enter the (post-experimental) assessment of a theory' ([2023], p. 517). The authors rightly observe that 'much of the "statistics wars" [(Mayo [2018], p. xi)] between Bayesians and frequentists revolve around this controversy' as to whether a method's error probabilities should enter the (post-experimental) assessment of evidence (Cox [1958], [1977], [1978]; Edwards et al. [1963]; Berger and Wolpert [1988]; Royall [1997], [2000a], [2000b]; Lindley [2000]; Mayo [2018]). The controversy turns on a basic principle of evidence: the likelihood principle.[4] On the likelihood principle, once the data are in hand, all of the evidence (for a statistical hypothesis in a model) is contained in the ratio of likelihoods of hypotheses. With likelihoods, the data are fixed, the statistical hypotheses vary.

VSW's standpoint creates a puzzle. While it adheres to the likelihood principle, it also 'acknowledges Popper's and Mayo's argument that severity needs to be accounted for' (van Dongen et al. [2023], p. 517). How can they account for severity while admitting 'the Bayesian ex-post evaluation of the evidence stays the same regardless of whether the test has been conducted in a severe or less severe fashion' (van Dongen et al. [2023], p. 522)? If they are to avoid inconsistency, I will argue, they must deny that severity enters in the post-experimental assessment of evidence. The question is not whether the term severity can be redefined Bayesianly, but whether the resulting concept will address today's concerns behind obstacles to replication. VSW argument is notable because it purports to give a Bayesian redefinition of severity that can capture 'the ideas of severe testing and error control' while bypassing error probabilities (van Dongen et al. [2023], p. 517). I disagree.

I will argue that the Bayes factor test rule can accord strong evidence to a claim H, even though little has been done to rule out H's flaws. There are two reasons: insensitivity to error control introduced by biasing selection effects, and the fact that H and its competitor need not encompass all relevant possibilities. This is especially problematic when Bayes factor tests are used as replacements for statistical significance tests—the main concern of this article. To mitigate this, I recommend Bayes factor tests be supplemented with a report of how severely H has passed, in the error statistical sense. Wagenmakers, a developer of Bayes factor software in psychology, might consider adding such an assessment. As practitioners accustomed to the error statistical guarantees of frequentist methods try out Bayesian tests, Bayesians are called on to reconcile their

[4] The likelihood principle follows from inference by Bayes theorem. Savage ([1962], p. 17) states it: 'According to Bayes's Theorem, $\Pr(\boldsymbol{x}|\mu)$ [. . .] constitutes the entire evidence of the experiment, that is it tells all that the experiment has to tell. More fully and more precisely, if $\boldsymbol{y}$ is the datum of some other experiment, and if it happens that $\Pr(\boldsymbol{x}|\mu)$ and $\Pr(\boldsymbol{y}|\mu)$ are proportional functions of μ (that is, constant multiples of each other), then each of the two data $\boldsymbol{x}$ and $\boldsymbol{y}$ have exactly the same thing to say about the value of μ' (notation altered).

support for popular reforms focused on promoting error probability controls with their advocacy of methods that are insensitive to them.

The article is structured as follows: Section 2 sets out preliminary notions, namely, the (error statistical) severity requirement, and the crux of the rival notions of evidence at issue. Section 3 explains the key elements of statistical significance tests and my proposed severity reformulation of tests. Section 4 responds to VSW's criticisms of the severity reformulation discussed in section 3. Section 5 introduces Bayes factor tests, and shows the severity consequences of their insensitivity to error probabilities. Section 6 critically examines the viability of VSW's Bayesian redefinition of severity.

2. Severe Testing: Preliminaries

2.1. Severity and corroboration

The rival approaches to severity invoke Popper. Granted, 'severity' is Popper's term, but he never worked it out adequately, and his proposals shifted over the years (Lakatos [1978]; Worrall [1978]; Mayo [1991], [2018]). He did not develop it for statistical inference, which is my focus. Still, severity has several 'Popperian' features. First, as a part of speech, 'severely tested' aligns with Popper's 'corroborated'—a claim is corroborated if it passes a severe test. Severity is post-data, focusing on how well tested a claim is with a given method and data. To say H is severely tested always means H passed severely, not just that it is undergoing stringent testing.[5]

Second, severity, is intended to measure, not how probable (or comparatively probable) claims are, but how well or poorly probed they are with data $\boldsymbol{x}$. It applies to the overall method. My account (Mayo [1996], pp. 178–87) begins with the framework from the Popper–Lakatos school:

> Severity Requirement: For data to warrant hypothesis H requires not just that
>
> (S1) H agrees with the data (H passes the test), but also
>
> (S2) with high probability, H would not have passed the test so well, were H false. (Mayo [2018], p. 92)

Popperians recognize the inadequacy of formal confirmation measures to capture non-adhocness and novel evidence.[6] Popper regards the formal measures he gives as merely 'potential' measures of corroboration. He warns we 'shall simply deceive ourselves if we

[5] Nor can it suffice that one 'sincerely' tried to find flaws, although some construe Popper's remarks that way (Popper [1959], p. 418).

[6] For Popper, the accordance in S1 is between statements. The data statement is 'constructed in such a way that it follows (or almost follows) from' H (Popper [1959], p. 412). In error statistical terms, the data, or really the test statistic, falls in the relevant non-rejection region. Mayo and Spanos ([2011]) require $\Pr(\boldsymbol{x}; H) > \Pr(\boldsymbol{x}; \sim H)$ for accordance. Leaving the accordance measure in S1 open allows applying severity to measures of fit used in other accounts. The issue then becomes S2.

think we can interpret [a formal measure $C(H, \boldsymbol{x})$] as degree of corroboration' unless it reports how stringently the falsity of H has been probed (Popper [1959], p. 418)—an assessment he denies can be fully formalized.

Third, like Popper, severe testers learn by conjecture and refutation. However, the falsification is nearly always statistical: if H_0 very probably would have 'survived' the test, if true, and yet the test yields a result discordant with H_0, then we infer evidence of the denial of H_0. By swapping out the claim to be inferred, the single notion of severity encompasses rejecting and failing to reject a claim. Severe testers use modern-day statistics to cash out Popper's requirements by means of a test's error probabilities. To address an anticipated criticism: I am not saying statistical significance warrants severely passing substantive scientific theories. Rather, the overall error statistical methodology, including experimental design, provides the types of tools (formal and informal) as well as the reasoning for assessing a method's capacity to uncover relevant errors in the context of a specific problem.[7]

To avoid confusion, in this article 'severity' will refer to the error statistical notion rejected by VSW. I call their Bayesian notion 'Bayes factor severity'. Their idea is to use the Bayes factor as the measure of agreement in S1. The question is in what sense they can satisfy S2.

2.2. Error statistical versus Bayes factor notions of statistical evidence

The first concern with Bayes factor tests is their insensitivity to error probability control. In a Bayes factor test, 'the Bayesian ex-post evaluation of the evidence stays the same regardless of whether the test has been conducted in a severe or less severe fashion' (van Dongen et al. [2023], p. 522). Data $\boldsymbol{x}$ gives strong evidence for hypothesis H_1 if the likelihood of H_1 sufficiently exceeds that of H_0, that is, Bayes factor$_{10}$ is high. The problem is, as statistician Barnard ([1972], p. 129) remarks, for any hypothesis H_0, 'there always is such a rival hypothesis viz., that things just had to turn out the way they actually did', unless $\Pr(\boldsymbol{x}; H_0)$ is one. 'The likelihood principle implies [...] the irrelevance of predesignation, of whether an hypothesis was thought of beforehand or was introduced to explain known effects' (Rosenkrantz [1977], p. 122). Since the import of the evidence is through the likelihood ratio, the evidence for H_1 is the same whether H_1 is constructed or predesignated. But it is not the same evidence for an error statistician. There can be a high probability of selecting or constructing a better fitting alternative H_1 even when H_0 correctly describes the data generation.[8] For the error statistician, this

[7] Other relevant discussions relating Bayesian accounts to Popper and severity are (Rosenkrantz [1977]; Gillies [2001]; Gelman [2011]; Gelman and Shalizi [2013]; Sprenger and Hartmann [2019]; Vanpaemel [2019]; Peden [2020]; Dienes [2023]).

[8] In ordinary English, 'likelihood' is used interchangeably with 'probability', but this leads to trouble when talking about the formal concept of likelihood. Any hypothesis that perfectly fits data $\boldsymbol{x}$ gets a

alters the evidence afforded by an observed $\boldsymbol{x}$.[9] The error statistical notion of evidence considers not just the observed $\boldsymbol{x}$ but how the test would behave under outcomes that could have been observed, but were not. This is called the sampling protocol.

2.3. Severe testing versus contrasting non-exhaustive hypotheses

The second concern with Bayes factor tests is that H and its competitor need not encompass all relevant possibilities. For the Bayes factor tester, strong evidence for H accrues to the extent that H is comparatively much more likely than some alternative, given the data. However, data can be much more probable under hypothesis H than under a non-exhaustive competitor H', even though H is poorly warranted. This leads to a problem:

Problem (or Fallacy) of Non-exhaustive Contrastivism: Taking $\boldsymbol{x}$ as good evidence for H on grounds that a non-exhaustive alternative H' is much less likely than is H given $\boldsymbol{x}$.

Consider an informal example given by VSW: Data $\boldsymbol{x}$, a white swan, is taken as strong evidence for H_1: 90% of swans are white, because $\boldsymbol{x}$ falsifies H_0: all swans are black. The BF_{01}, $\Pr(\boldsymbol{x}|H_0)/\Pr(\boldsymbol{x}|H_1)$, is zero; BF_{01} would be infinite. However, finding a single white swan has hardly ruled out the ways that H_1 can be false. For H_1 to pass severely, there needs to be grounds to rule out H_1's denial—here, all of the other possible percentages of white swans. Suppose we had observed 30% white swans. H_1 would still be much more likely than H_0, but there are more likely alternatives that have not been considered, such as 30% of swans in the population are white. The better supported (more likely) of two hypotheses can be very poorly tested.

Let me make two qualifications, to be fleshed out as we proceed. First, error probabilities, likelihood ratios and Bayes factors are computed within statistical models. I take 'exhaustiveness' of H_0 and H_1, relative to a statistical model and background, to mean that H_1 is the complement of H_0: they partition the parameter space Ω of the model (see section 3.1). For example, Ω might be the different proportions of white (or black) swans in a given population. Exhaustiveness is necessary though not sufficient for severity.[10] Assumptions of the models would require separate checks. Second, the problem of non-exhaustive contrastivism concerns Bayes factor tests, not Bayesian inference in general.[11] However, 'Bayes factors are the primary tool used in Bayesian inference for

likelihood of one, and many rival hypotheses can satisfy this. But rival hypotheses could not all have a probability of one.

[9] Strictly, this should be written $\boldsymbol{x}_0$, but for simplicity, I use $\boldsymbol{x}$, emphasizing, if pertinent, that it is observed.

[10] A discussion of two-sided, exhaustive, Bayes factor tests is in section 5.4.

[11] Some Bayesians, such as Gelman ([2011]), object to Bayes factor tests for their discreteness: 'To me,

hypothesis testing' (Berger and Pericchi [2015]) and they are recommended replacements for significance tests. The problem of non-exhaustive contrastivism is especially relevant to VSW's approach because they advertise specific and contrastive (non-exhaustive) hypotheses as a preferred way to deliver strong (Bayes factor severe) tests. VSW deny that 'Mayo's Severity Principle and Severity Requirement [...] are good operationalizations of the function of severity in scientific inference' (van Dongen et al. [2023], p. 521). With this in mind, let us begin with statistical significance tests.

3. Statistical Tests and Severity

3.1. Elements of statistical significance tests

Severity is 'not part of any statistical methodology as of yet' (Mayo [2018], p. 9). Existing formal error statistical methods, whether statistical significance tests, confidence intervals or others, are generally couched in terms of good error control performance in a series of applications. Good performance is a necessary, but not a sufficient, condition for severity. Performance has acquired new importance in AI and machine learning, but there is still a distinct role for making inferences about the specific claim. The severity principle explains the relevance of a method's performance for warranting a specific inference whenever such performance captures the method's error-probing capacity—that is, its 'probativeness'.

A statistical hypothesis H is a claim about some aspect of the process that might have generated the data $\boldsymbol{x}$, viewed as observed values of a random variable, $\boldsymbol{X}$. Data $\boldsymbol{x}$ are used to learn about the probability distribution of $\boldsymbol{X}$ by testing statistical hypotheses, generally about a parameter θ, in a model M—an idealized representation of the data generation. The parameter θ is viewed as fixed. Thus, except for special cases where θ may be seen as a random variable with its own (frequentist) probability distribution, no prior probability assignments are given to θ. Fisher ([1935]) called the reference hypothesis the null hypothesis, while Neyman and Pearson used a more satisfactory term, the test hypothesis, denoted by H_0. Here, I use 'null' for convenience. Worse, I will usually abbreviate 'statistical significance tests' as 'significance tests'.

Statistical significance tests consist of, first, a test (or null) hypothesis H_0, couched in terms of θ, and, second, a test statistic $d(\boldsymbol{X})$, which reflects how well or poorly data $\boldsymbol{x}$ accord with null hypothesis H_0. The larger the value of $d(\boldsymbol{x})$ the more improbable the outcome is from what is expected under H_0, with respect to the particular mistake being probed. An important aspect of an error statistical test is its ability to ensure the distribution of the test statistic $d(\boldsymbol{X})$—called the sampling distribution—can be computed

Bayes factors correspond to a discrete view of the world, in which we must choose between models A, B, or C (or a weighted average of [them])' (Gelman [2011], p. 74). Kass ([2009]), co-author of a leading paper on Bayes factors (Kass and Raftery [1995]), subsequently rejects their use because of their strong dependence on priors even for large samples.

under H_0 and under hypotheses discrepant from H_0. This enables computing the statistical significance level or p-value associated with $d(\boldsymbol{x}_0)$, where $\boldsymbol{x}_0$ is the observed data point, p-value $= \Pr\Big(d(\boldsymbol{X}) > d(\boldsymbol{x}_0);\ H_0\Big) = p$. Note that this is not the probability of the particular observation, $d(\boldsymbol{x}_0)$, under H_0. Any continuous observation is improbable under H_0. Were a result declared statistically significant simply because it is improbable under H_0, we would very probably declare results statistically significant erroneously, violating the error probability guarantees.[12]

The goal of the simple statistical significance test is to test if data $\boldsymbol{x}$ are reasonably consistent with a single hypothesis or model, or if discrepancies exist. Neyman and Pearson ([1933]) placed Fisherian tests on firmer ground by introducing the alternative statistical hypothesis H_1 (NP test). The two hypotheses partition the parameter space Ω: together they exhaust Ω. This is apt for severe testing which calls for a piecemeal approach to probing mistakes.[13] Formally, exhausting Ω is required for significance tests to have optimal error probabilities.[14] On the informal side, non-exhaustive tests license inferences that violate severity.

The NP test prespecifies a value of $d(\boldsymbol{x})$ beyond which the data are sufficiently discordant with H_0 to lead to its rejection; all other data points accord with H_0. Neyman and Pearson also allowed for an 'undecidable' region, but for simplicity, it is generally omitted. Inferences are to inequalities or composite hypotheses such as $\mu > 150$, not to points, such as $\mu = 150$.[15] (The same is true for confidence interval estimates, which are inversions of tests.) Rejection may be interpreted in many ways. I recommend construing it as 'evidence against H_0'. Requiring a low p-value before finding evidence against H_0 controls the type 1 error probability, the probability of erroneously rejecting H_0. Good NP tests also ensure low type 2 errors: failing to reject H_0 when there is a genuine discrepancy δ from H_0: equivalently, high power to detect discrepancies. To avoid confusion, 'discrepancy' will always refer to a population or parametric effect size, not an observed one. The severity interpretation goes beyond pre-data type 1 and type 2 error probabilities to assess, post-data, those discrepancies that are well or poorly warranted

[12] The P-value can itself be seen as a statistic: In continuous cases, $\Pr(P < p) = p$.

[13] Elsewhere, I identify four main mistakes: (1) taking background variability (noise) as genuine, (2) inferring unwarranted effect sizes, (3) violating statistical model assumptions, and (4) erroneously linking statistical measurements to substantive claims (Mayo [1996], [2018]). These, especially the fourth mistake, would split into a great many flaws.

[14] While the Neyman–Pearson lemma (Neyman and Pearson [1933]) is framed for a special case of two point hypotheses, it is uniformly most powerful because it limits the parameter space to $\{\mu_0, \mu_1\}$. Optimality is extended beyond this highly artificial starting point for tests that satisfy additional conditions: monotonicity and convexity. For a full discussion and references, see (Spanos [2019], sec. 13.4). It was found that Fisherian tests nearly always 'lead to the same destination' (Cox [2006], p. 25) as NP tests, starting from Fisher's requirements for an appropriate test statistic. Severity blends the two rationales (Mayo and Cox [2006]).

[15] A hypothesis is composite not only by dint of covering a range of parameter values. It is also composite when other parameters are required to determine the likelihoods, as is typical.

with data $\boldsymbol{x}$. This is the key to avoiding classic fallacies in both Fisherian and NP tests (Mayo [1996], [2018]; Mayo and Cox [2006]; Mayo and Spanos [2006], [2011]; Haig [2018], [2020]; Mayo and Hand [2022]).

3.2. Severity construal and extension, 1: Test T+

Mayo and Spanos ([2006], p. 329) abbreviate 'the severity with which inference H passes test T with outcome $\boldsymbol{x}$' as SEV(test T, outcome $\boldsymbol{x}$, claim H). Given the test specifics, we simplify with SEV(claim H). In the severe testing interpretation, a very small p-value, such as 0.01, indicates evidence for H_1 because H_1 has passed a severe test, provided the p-value is properly computed. I call it merely a severity 'indication' until the assumptions of the statistical model are themselves checked.[16] The severity (or severity indication) requirement is instantiated because: (S1) $\boldsymbol{x}_0$ accords with H_1, and (S2) there is a high probability $(1-p)$ that a less statistically significant difference would have resulted, were H_0 true.

Inferring the existence of some discrepancy is important, but rarely adequate. First, an indication of a discrepancy is not yet evidence of a genuine experimental effect. From the start, Fisher emphasized that finding an isolated small p-value was inadequate. Second, the p-value is not an effect size measure, neither is a Bayes factor. But we can determine how large an inconsistency is indicated by using several p-values. The leads to the first severity extension of tests.

Consider test T+: Suppose we are testing the mean μ of a normal distribution, $H_0: \mu \leq \mu_0$ versus $H_1: \mu > \mu_0$, with a random sample of size n. In this one-sided test, only positive discrepancies from μ_0 are being probed. (It is the same NP test if $H_0: \mu = \mu_0$: when it is rejected, so are all the points in $\mu < \mu_0$.) We observe, $\bar{x}$, the value of the sample mean, $\bar{X}$, a sufficient statistic, and compute how much it differs from μ_0 in standard error (SE) units. The test statistic $d(\bar{X})$ is $(\bar{X} - \mu_0)/\text{SE}$; abbreviate its value as d. Extension 1 allows inferring effect sizes, and avoiding magnitude errors.

Extension 1 (Severity for inferring warranted and unwarranted discrepancies in T+): If there is a fairly high (low) probability that $d(\bar{X})$ would have been larger than d, if μ is no greater than μ_1, then d is a poor (good) indication that $\mu > \mu_1$, where $\mu_1 = \mu_0 + \gamma$ (with $\gamma > 0$). The severity associated with $\mu > \mu_1$ is low (high).

This extension is equivalent to computing the p-value for several discrepancies from H_0, taken as a reference, and inferring those well or poorly warranted. Severity calls for reporting at least one unwarranted inference, in relation to the error being probed. A good benchmark for an unwarranted inference in test T+ is $\mu > \bar{x}$. The p-value (associated

[16] 'Auditing' a test also requires checking for biasing selection effects.

with $\bar{x}$), if we were testing, $H_0 : \mu = \bar{x}$ versus $H_1 : \mu > \bar{x}$, would be 0.5. In applying severity, the original hypotheses are not changed. Rather, the severity concept is being used to fill a gap in existing p-values.[17]

A Numerical Example: $H_0 : \mu \leq 150$ versus $H_1 : \mu > 150$.

Take a simple example of test T+ from (Mayo [2018], pp. 142–46) with a random sample of size $n = 100$, and known standard deviation, $\sigma = 10$. This has a standard error 1, and a convenient test statistic: $d(\bar{X}) = (\bar{X} - 150)$.[18] The 2-standard error cut-off for rejection is: Reject H_0 if and only if $\bar{X} \geq 152$. This ensures that the probability of rejecting H_0 when $\mu = 150$ (type 1 error) is only ∼0.02 (the p-value is 0.02). In this toy example, one is testing if μ, the mean temperature in a body of water in degrees Fahrenheit, exceeds 150. In effect, H_0 predicts $\bar{X} < 152$, which occurs with probability 0.98 in a world adequately described by H_0. Say the observed result, $\bar{x}$, is 152, disagreeing with H_0. So (S1) $d(\bar{x})$ accords with the alternative (it is in the test's rejection region), and (S2) were μ no higher than 150, then at least 98% of the outcomes would have produced a lower mean temperature than observed. Thus, $\bar{x} = 152$ is evidence of a positive discrepancy from H_0. What subsequent actions to take is distinct.

Suppose one wants to assess the severity of $C : \mu > 151$, that is, to compute SEV(T, $\bar{x} = 152$, $C : \mu > 151$). Severity for C requires at least a reasonable probability of a worse fit with C, if C is false. Here, 'worse fit with C' means $\bar{x} \leq 152$ ($d(\bar{x}) \leq 2$).[19] Severity is evaluated at point $\mu = 151$ since the probability is even greater for $\mu < 151$. Severity for C is $\Pr(\bar{X} \leq 152; \mu = 151) = \Pr(d \leq 1) = 0.84$. The severity decreases as the inferred discrepancies grow larger. In particular, SEV($\mu > 153$) is 0.16.[20, 21] We cannot infer $\mu > 153$ with severity because, if the sample came from a world where $\mu = 153$, it is fairly probable that the test would yield a higher mean than observed. Figure 1 shows severity for different claims as computed in (Mayo [2018], p. 144).

3.3. Severity extension for statistical non-significance, 2: Test T+ continued

While failing to find evidence against a claim is not evidence for it, it is misleading to aver, as do VSW, that if the p-value is not small, 'no conclusion is drawn' (van Dongen et

[17] While often overlooked, NP theory also recommends reporting the attained p-value. According to Lehmann, the key expositor of NP tests, 'One should routinely report the p value and [...] combine this with a statement on significance at any stated level' (Lehmann [1993], p. 1247).
[18] The SE of $\bar{X}$ is σ divided by the square root of n.
[19] Given $\boldsymbol{X}$ is continuous, $<$ or $\leq$ give the same result.
[20] $\Pr(\bar{X} \leq 152; \mu = 153) = \Pr(d \leq -1) = 0.16$.
[21] Severity apps by Richard Morey (<richarddmorey.shinyapps.io/severity>) and Marcos Jiménez (<marcosjnez.shinyapps.io/Severity/>) provide some of the computations needed; see <errorstatistics.com/sev-app/>.

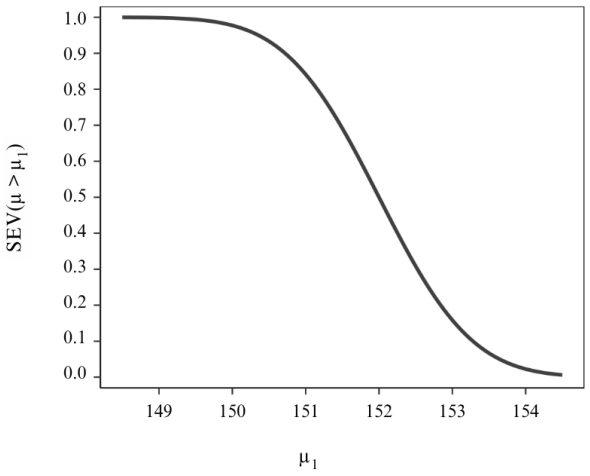

Figure 1. Severity curve for T+: $H : \mu > 150$; $\bar{x} = 152$.

al. [2023], p. 518). Even with simple Fisherian tests, failing to reject H_0 is informative—especially if the test is well specified. Minimally, it tells us that the results cannot be taken as evidence of a non-chance effect. In their discussion of significance tests, VSW overlook an essential test component: power (or, for Fisher, sensitivity). If a test has a high power to detect meaningful effects (where 'meaningful' is determined by context), then failure to reject H_0 is evidence of the absence of a meaningful effect. The severity interpretation uses a post-data version of power, leading to a companion severity extension. Still keeping to test T+:

Extension 2 (Severity for avoiding fallacies of non-significance): If there is a low (high) probability that $d(\bar{X})$ would have been larger than the observed d, if μ is as great as μ_1, then d is a poor (good) indication that $\mu < \mu_1$, where $\mu_1 = \mu_0 + \gamma$ (with $\gamma > 0$).

'This avoids unwarranted interpretations of consistency with H_0 with insensitive tests [and] is more relevant to specific data than is the notion of power' (Mayo and Cox [2006], p. 89).[22] (The power against μ_1 is $\Pr\big(d(\bar{X}) > c_\alpha; \mu_1\big)$, c_α the α-level cut-off for rejection.) 'Negative' results are essential to identify replication failures, rule out theoretical pathways, and validate probability models. As Fisher ([1935], p. 14) remarks, 'a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result'. Reporting failures, not hiding them in file drawers, is crucial for progress.

[22] Another way to handle this is with equivalence tests (Wellek [2010]; Lakens [2017]).

4. Van Dongen, Sprenger, and Wagenmakers's Criticisms of Error Statistical Severity

4.1. Severity is claimed to ignore a default condition

VSW claim: 'One of the key objectives of this paper is to provide a critical analysis of Mayo's account' (van Dongen et al. [2023], p. 517). They criticize my use of severity to avoid a magnitude error (extension 1). Why? They agree that if μ were 153, that 'it is fairly probable (84% of the time) that the test would yield an even larger mean temperature than we got' (van Dongen et al. [2023], p. 520). Thus to infer μ is as great as (let alone greater than) 153 is to follow a procedure that is wrong 84% of the time. (The p-value in testing $\mu \leq 153$ is 0.84.) How then could they object? They aver: 'this result is independent of what is considered the normal or default state of affairs. [Here] this is 150 [...] but this value is irrelevant for her SEV function [...] Suppose that we had observed the same mean temperature of 152 degrees Fahrenheit, though the normal mean temperature is 100 degrees. Then one would [...] draw the same conclusions as before' (van Dongen et al. [2023], p. 521). But this is not so. The report is that the result is statistically significantly larger than 150. The test hypothesis taken as the reference value does not disappear, even once we extend the analysis to different discrepancies from 150. It is part of the test statistic.

Nevertheless, VSW take this to show: 'The error-statistical analysis of the case goes against any intuition one might have about the concept of severity, treatments of this concept by other philosophers and methodologists [...] and even against a reasonable interpretation of Mayo's Severity Principle' (van Dongen et al. [2023], p. 521). The severe tester begs to disagree. It would be contrary to a severe tester's position to endorse a method that would be wrong over 50% of the time, violating Cox's weak repeated sampling principle.[23]

Suppose the test hypothesis had been $\mu = 100$, as VSW propose. Now the observed mean is 50 SE above what is deemed typical! The p-value would be 0, not 0.02. Such a result would lead to suspecting the measuring apparatus was systematically flawed, and the results thrown out. Invoking $\mu = 100$ to argue evidence that $\mu > 153$ commits the fallacy of non-exhaustive contrastivism (section 2.3). Are VSW saying that for the Bayes factor tester, there is evidence that μ is as great as 153 so long as one can select a much lower value, say $\mu = 100$, under which the probability of the observed mean, 152 is much smaller (essentially zero) than under $\mu = 153$? Such a gambit would make it too easy to find evidence for exaggerated effect sizes.

One does not need severity to see this; confidence interval estimation illustrates the

[23] '[...] we should not follow procedures which for some possible parameter values would give, in hypothetical repetitions, misleading conclusions most of the time' (Cox and Hinkley [1974], pp. 45–46).

same point due to its duality with tests (Neyman [1937]). Test T+, at statistical significance level α, corresponds to the $(1-\alpha)$ lower confidence interval (CI) estimate: $\mu > \bar{x} - c_\alpha$ SE. With $\bar{x} = 152$, the one-sided lower 0.975 CI estimate is $\mu > 150$, the same as the inference from the one-sided test (still assuming known SE). The CI contains all of the μ values which $\bar{x} = 152$ is not statistically significantly greater than at level 0.025. VSW's reasoning would also allow inferring μ is as great as 154 (as $\mu = 154$ is more likely than $\mu = 100$, with $\bar{x} = 152$). The two-sided 95% interval with $\bar{x} = 152$ is $[150 < \mu < 154]$. Values as large as 154 are inconsistent with the data.[24]

This response also highlights a key advantage of significance tests: their ability to test statistical assumptions (Mayo and Spanos [2004]; Spanos [2019]). Bayesians such as Box ([1983], p. 57) acknowledge needing 'diagnostic checks and tests of fit which [he argues] require frequentist theory significance tests for their formal justification'.[25] Observing $\bar{x} = 152$ falsifies a presumption that $\mu = 100$ is typical, avoiding unwarranted claims of ecological disaster![26]

4.2. Other criticisms

According to VSW, 'without explicit reference to alternatives to the claim, it is not clear how a claim can survive "stringent scrutiny"' (van Dongen et al. [2023], p. 519). For any inferred claim C, there is an alternative, namely, C's denial, within the model. How can a composite hypothesis like $H_1 : \mu > 150$ pass with severity they ask? We do so by rejecting $\mu \leq 150$ as in test T+. Bayes factor testers must assign prior probabilities over the points in the composite alternative; significance testers do not.

VSW allege that the severity principle does not require 'the riskiness or specificity of a hypothesis' (van Dongen et al. [2023], p. 519). To reply, first note that a hypothesis is risky, not in itself, but in relation to a test. Riskiness enters significance tests via a test's error probabilities. Any claim that is severely tested is one such that, were it false, then with high probability it would not have passed, or passed so well, with the method used. That a p-value must be computable under H_0 without free parameters requires that it be assessed at a specific point. This is the key to falsifying (statistically) a single statistical hypothesis or model, and inferring (with severity) its denial. What we do not want is comparing specific, non-exhaustive hypotheses.

VSW criticize error statistical tests for not according strong evidence to a universal theory (represented by a point hypothesis), whereas they do, by assigning it a Jeffreys-

[24] Severity improves on standard CIs by supplying, for each μ' in the interval, a distinct assessment of the severity for claims of form $\mu < \mu'$ or $\mu > \mu'$. Moreover, it gives an inferential, not merely a performance, rationale; see (Mayo [2018], pp. 13–14, 193–95).

[25] See also (Gelman [2011]; Gelman and Shalizi [2013]).

[26] One might instead test: $H_0 : \mu \geq 150$ versus $H_1 : \mu < 150$, if the 'more serious error' is seen as failing to find an increase. Thanks to extensions 1 and 2, and unlike an accept–reject construal, the severely tested inference is the same. Note that in test T+, SEV($\mu \leq \mu_1$) is $1 - \text{SEV}(\mu > \mu_1)$.

type prior (see section 5.4). For the error statistician, this does not align well with the context of statistical testing. The severe tester sees the role of statistical inference as building theories by local data analysis and stringent probes of specific errors. Fuller discussions of how statistical tests are linked to theories are in (Mayo [2010], [2018]). The toy illustration of the one-sided test immediately follows a realistic example: the famous eclipse tests of Einstein's light deflection effect. That test, thanks to extensive knowledge of star positions, may be framed as the one-sided test of H_0: the deflection effect is no greater than Newton's half deflection: $H_0 : \lambda \leq 0.87$ versus $H_0 : \lambda > 0.87$. Statistical reasoning enabled asking, first, if there is a genuine effect (how large) and then, second, if it is attributable to the sun's gravitational field (as stipulated in the general theory of relativity). This required testing numerous, plausible, Newton-saving theories—many that provided better fits to the data than the general theory of relativity's prediction. Each was separately falsified. Even so, 'Nothing like a stringent estimate of the deflection effect emerged until the field was rescued by radioastronomical data from quasars' in the 1960s (Mayo [2018], pp. 124–25). A zoo of rival relativistic theories emerged, prompting a new series of local parametric tests. At any stage, only limited aspects of theories are severely warranted—adequate accounts should reflect this.

5. The Error Statistical Consequences of Adhering to the Likelihood Principle: Why They Matter

Ensuring the validity of statistical significance tests requires supplements to adjust (or avoid) types of multiple testing and biasing selection effects—or at least to report them. Well-known methods include Bonferroni corrections and false discovery rate approaches (Benjamini and Hochberg [1995]; Benjamini [2020]). With current uses of big data and 'omics' research, addressing 'post-data selection' is now a critical and active area of research. By contrast, 'the Bayes factor is insensitive to variations of the sampling protocol that affect the error rates' (van Dongen et al. [2023], p. 522). As Kruschke and Liddell ([2018], p. 190) note: 'Bayesian analysis does not base decisions on error control. Indeed, Bayesian analysis does not use sampling distributions'.

5.1. Bayes factors in Bayesian testing

VSW present the Bayes factor as follows:

> The posterior probability of a theory H_0 given data $\boldsymbol{x}$ is calculated as $p(H_0|\boldsymbol{x}) = p(H_0) \cdot p(\boldsymbol{x}|H_0)/p(x)$. This allows us to write the posterior odds in favor of H_0 over H_1 as
>
> $$\frac{p(H_0|\boldsymbol{x})}{p(H_1|\boldsymbol{x})} = \frac{p(H_0)}{p(H_1} \cdot \frac{p(\boldsymbol{x}|H_0)}{p(\boldsymbol{x}|H_1)}.$$
>
> (van Dongen et al. [2023], p. 522)

Bayes Factor BF_{10}	**Interpretation**
> 100	Extreme evidence for H_1
30–100	Very strong evidence for H_1
10–30	Strong evidence for H_1
3–10	Moderate evidence for H_1
1–3	Anecdotal evidence for H_1
1	No evidence for either hypothesis H_1

Table 1. Classification of Bayes factors adjusted from (Jeffreys [1961]).

The Bayes factor can be seen as the ratio of the posterior probabilities of H_0 and H_1, divided by the ratio of their prior probabilities. An advantage of computing the Bayes factor, rather than a posterior probability, is that the most challenging term in the denominator, the probability of $\boldsymbol{x}$, cancels out: 'One of the attractive features of using a Bayes factor [...] is precisely that it does not depend on [the priors to the two hypotheses] which can vary considerably among consumers of a study' (Berger and Pericchi [2015], p. 2). But as they emphasize, the Bayes factor still depends on the prior probabilities that must be assigned to any free parameters in the two hypotheses: within model priors.

VSW employ a standard Bayes factor test rule, taken from Jeffreys ([1961]), for moving from Bayes factors to different strengths of evidence for H_1 (see table 1; Van Dongen et al. [2023], p. 522). Evidence for H_0 proceeds in the same way, inverting the numbers.

Replacing significance tests with Bayes factor tests is becoming so widespread that they are often called 'null hypothesis Bayesian tests' (NHBT). In some formulations, in a Bayesian hypothesis test 'the alternative hypothesis is accepted if $BF_{10} > k$' (Johnson [2013], p. 1726; notation altered), for sufficiently high k. The call to 'redefine significance' (Benjamin et al. [2018]), endorsed by over eighty practitioners (including Wagenmakers), shows how to better align significance tests and Bayes factor tests by using a p-value threshold of 0.005.[27] The test output, while based on a comparative appraisal of H_0 and H_1, infers strong evidence for the favoured hypothesis, although the alternative used in the comparison should be reported (Wong et al. [2022]; Tendeiro et al. [2024]). Whether Bayes factor testers infer the strength of evidence for H_0 (H_1), or move to accepting or rejecting based on a posterior probability, inferences are made.[28] This raises concerns about error probabilities (Lakens [2019], [2022]). If Bayes factor testers wish to promote trust

[27] For test T+, Johnson ([2013], p. 1716) recommends computing the Bayes factor using the alternative $H_1 : \mu = \bar{x}_\alpha$, where $\bar{x}_\alpha$ is the cut-off for rejecting H_0 at level α. The test will 'maximize the probability that the Bayes factor against a fixed null hypothesis exceeds a specified threshold'. Then, H_0 and H_1 are each given 0.5 priors, and hypotheses are accepted if their posterior probability is high. With $\alpha = 0.005$, the Bayes factor is $\sim$28 and the posterior for H_1 is $\sim$0.97 (see note 28). But the error statistician would not infer so large an effect size from merely $\bar{x}_\alpha$; $\text{SEV}(\mu > \bar{x}_\alpha)$ is only 0.5.

[28] The posterior probability is:

in declaring that 'Science needs to abandon p-values and adopt Bayes factors' (Berger and Pericchi [2015], p. 11), I argue, they should care about these consequences. VSW do not discuss how they ameliorate these problems, undermining their claim to offer a satisfactory substitute for significance tests.

5.2. Example 1: Finding your hypothesis in your data

Consider data dredging. A classic problem is a clinical trialist, failing to find a benefit in a (double-blind) randomized control trial on a medical treatment, ransacks the unblinded data until a subgroup is found where treateds do better than controls, in some way. They might search for patterns among patients with different characteristics (age, sex, employment, education, medical condition, and so on), and try different proxy variables to use in measuring benefit ('outcome switching'). All clinical trials in the US legally require prespecification of endpoints. Let H_1^{PD} be a post-data dredged hypothesis under which $\boldsymbol{x}$ is much more probable than under H_0^{PD} that there is zero benefit. The severe tester, like the significance tester, must go beyond the ratio of likelihoods (the S1 portion of severity) to the overall method that finds evidence for H_1^{PD} (the S2 portion). The severe tester asks: what is the probability the method would locate some subgroup or other that gives a high Bayes factor, even if spurious—an error probability. If it is high, she denies the large Bayes factor counts as evidence for H_1^{PD}. Why consider some subgroup or other? Because the statistical significance test is probing this general type of error: mistaking spurious effects as genuine. But the Bayes factor tester purports to bypass error probabilities. All the evidence is in the Bayes factor for the selected hypothesis: that is where the accounts of evidence differ.

Cannot data-dredged claims be blocked in the final inference by assigning them low prior belief? Maybe, but this 'misidentifies what the problem really is. The influence of the biased selection is not on the believability of H, but rather on the capability of the test to have unearthed errors' (Mayo [2018], p. 40). Drawing a line around treateds who show some beneficial effect is akin to the proverbial Texas Marksman circling a cluster of closely placed bullet holes, and regarding it as evidence of his marksmanship. The accordance between data and hypothesis is due to the biasing selection effects, not the truth of the hypotheses (about the drug's benefit or his marksmanship).[29] Moreover, altering priors because of the experiment to be performed, strictly speaking, violates Bayesian coherence: 'Why should one's knowledge, or ignorance, of a quantity depend

$$\Pr(H_0 \,|\, \boldsymbol{x}) = \frac{1}{1 + \text{BF}_{10}\frac{\Pr(H_1)}{\Pr(H_0)}}.$$

[29] Not all cases of data dredging are pejorative. For example, persevering to explain a known effect such as DNA matching, is warranted with severity (Mayo and Cox [2006]; Mayo [2008]).

on the experiment being used to determine it?' (Lindley [1972], p. 71).

5.3. Example 2: Optional stopping

For a second type of multiplicity, consider optional stopping. In Savage's ([1962]) famous example, two-sided testing of a normal mean with known σ : $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$, if the researcher keeps sampling until reaching a 'nominally' significant result 'then with probability 1 he will eventually reject any sharp null hypothesis, even though it be true' (Edwards et al. [1963], p. 239).[30] For the error statistician, the actual or overall significance level is the probability of finding a 0.05 nominally significant result at some stopping point or other, up to the point it stops. The p-value accumulates. To be clear: sequential stopping rules are regularly used by error statisticians, but the multiplicity must be taken account of to ensure error probability control (Wald [1947]; Armitage [1975]). By contrast, holders of the likelihood principle aver 'the import of the sequence of n data actually observed will be exactly the same as it would be had you planned to take exactly n observations in the first place' (Edwards et al. [1963], pp. 238–39). They call it the stopping rule principle, and it follows from the likelihood principle. Whether this insensitivity is a good thing or cheating depends on whether evidence is being viewed from the perspective of the likelihood principle or the perspective of error probabilities. According to Wagenmakers ([2007], p. 785), 'if the sampling plan is ignored, the researcher is able to always reject the null hypothesis, even if it is true. This example is sometimes used to argue that any statistical framework should somehow take the sampling plan into account. Some people feel that 'optional stopping' amounts to cheating [...] This feeling is, however, contradicted by a mathematical analysis'.[31] However, his mathematical analysis presupposes a measure of evidence (the Bayes factor) that obeys the likelihood principle.[32] While the Bayesian assessment of the evidence remains the same, the probability of erroneously attaining such an assessment grows. While 'many believe [optional stopping] exerts no more than a trivial influence on false-positive rates [...] Contradicting this intuition' the probability of erroneous rejections balloons (Simmons et al. [2011], pp. 1361–62).

To parallel Wagenmakers: if the sampling plan is ignored (in optional stopping), the researcher is able to always find strong Bayes factor evidence for alternative H_1, even if H_0 is true. The same thing happens with posterior probabilities. As Armitage ([1962], p. 72) points out: 'The departure of the mean by two standard errors [...] corresponds to

[30] This is a proper stopping rule: the probability it will stop in a finite number of trials is one. It is also uninformative, not affecting the priors.

[31] A famous argument by Birnbaum ([1962]) purports to show the likelihood principle follows from principles that frequentists hold. I argue it is unsound (Mayo [2014]).

[32] BF_{10} may be computed comparing the likelihood of μ_0 to that of maximally likely alternative $\mu_{\max}$ where $\mu_{\max}$ sets μ equal to $\bar{x}$. BF_{10} is $= \exp[z^2/2]$ for both fixed sample size and optional stopping. (Here, z is a standard normal variable.)

the null hypothesis being at the five per cent point of the posterior distribution'—using a uniform prior probability for μ. Since in this case the posterior for H_0 matches the significance level, if a Bayesian allows optional stopping (with this prior), she is assured of assigning a low posterior probability to H_0, even though it is true.[33] This error statistical implication is one even many Bayesians are not (usually) too happy about. Thus, some Bayesians try to find ways to avoid it.[34]

5.4. Spike priors: A central advantage of Bayes factors?

One way Bayesians avoid a low posterior probability on H_0 is to assign it a high enough prior: 'if one is concerned about the high probability of rejecting H_0, it must be because some possibility of its truth is being entertained [. . .] The appropriate way to entertain the possibility of the truth of H_0 is to assign a non-zero prior probability to H_0, as originally suggested by Jeffreys (for other reasons)' (Cornfield [1966], p. 580).[35] Jeffreys's test places a non-zero 'spike' of prior probability on the point null $H_0 : \mu = \mu_0$, often 0.5, and spreads the remaining 0.5 over a wide range of parameter values in the alternative $H_1 : \mu \neq \mu_1$—often called a 'spike and smear prior'. Jeffreys, an objective Bayesian, developed it to enable an already well-corroborated universal scientific theory to obtain a non-zero posterior probability. Notice, here the alternative H_1 formally covers all non-zero values, so the claims are 'exhaustive'. The trouble is, the likelihood for H_1 requires averaging over the priors of all the points in H_1. The result, as a leading (objective) Bayesian remarks, is that with spike and smear priors, 'the Bayes factor for the null may be arbitrarily large for sufficiently large n, however relatively unlikely the data may be under H_0' (Bernardo [2011], p. 59). This introduces a lack of control over the type 2 error. Ironically, far from subjecting $H_0 : \mu = \mu_0$ to risk of falsification, it allows the probability of failing to pick up on population discrepancies to be high, or even guaranteed if the discrepancies are small. (Popper assigned zero probability to all such universals in infinite domains: no spikes on the point universal could be kosher.)

Even statistically significant results at low p-values can correspond to a high Bayes factor in favour of the null hypothesis. This is the famous Jeffreys–Lindley paradox, or the Fisher–Jeffreys disagreement (Lindley [1957]; Senn [2001]]; Bernardo [2011]; Spanos [2013]; Cousins [2017]; Mayo [2018]).[36] It underscores the impact of how one chooses to

[33] The same thing happens with confidence intervals and corresponding Bayesian credible intervals, as Berger and Wolpert ([1988]) admit. The interval is guaranteed to exclude the true value.

[34] Objective Bayesians concede they 'have to live with some violations of the likelihood and stopping rule principles' (Ghosh et al. [2010], p. 148), since their priors are influenced by the sampling distribution.

[35] Unlike Cornfield, the severe tester's objection to optional stopping is the potential for high type 1 error—not that she holds H_0 as probable.

[36] Here is how it happens: The probability of $\boldsymbol{x}$ under a composite H_1 is an average over all the values in H_1. Having spread out the priors over the values in H_1, the most likely values of μ in H_1 are given a very low prior, so that the likelihood of H_1 is much less than the likelihood of H_0 (Senn [2001]).

compute the Bayes factor. If BF_{01} is computed with the spike and smear prior, there is strong evidence for H_0; if BF_{01} is computed comparing $H_0 : \mu = \mu_0$ and $H_1 : \mu = \bar{x}$, then H_0 is comparatively unlikely! Why are prior probabilities entering when we are only computing Bayes factors? The Bayes factor, recall, requires within model priors when the test involves composite hypotheses: 'Quite often, [the] alternative hypothesis will depend on the prior distribution of the parameter of interest, e.g., when we test a point null hypothesis $H_0 : \mu = \mu_0$ against an unspecific alternative $H_1 : \mu \neq \mu_0$' (van Dongen et al. [2023], p. 522).

Use of the spike and smear prior enables finding evidence for a point null hypothesis—often advertised as an advantage Bayes factor tests enjoy over statistical significance tests. This benefit, however, is due to the high prior for H_0, invoked to avoid too easily finding evidence against H_0. Edwards et al. ([1963], p. 235) aver that for Bayesian statisticians, 'no procedure for testing a sharp null hypothesis is likely to be appropriate unless the null hypothesis deserves special initial credence'. Yet it is fairly standard in Bayes factor tests. Casella and Berger ([1987], p. 111) remark: 'Concentrating mass on the point null hypothesis is biasing the prior in favor of H_0 as much as possible'. This is concerning if Bayes factor tests are used in place of significance testing. When error statisticians use a point null, it is just as a reference point. Following Cox and Hinkley ([1974]), they view the two-sided test as combining two one-sided tests, doubling the p-value for a selection effect.[37]

6. Bayes Factor Severity: Can the Puzzle Be Resolved?

Our examples in section 5 show what Bayes factor theorists know and accept: the Bayes factor inferential assessment is insensitive to gambits that violate error probability control: all of the evidence, post-data, is in the reported Bayes factor. Recall, however, that VSW claim their article 'acknowledges Popper's and Mayo's argument that severity needs to be accounted for' (van Dongen et al. [2023], p. 517). In this section, I will examine this claim. Near the end of their article, VSW propose reinterpreting the strong severity principle from (Mayo [2018], p. 14). They place one below the other:

> **Severity Principle (Strong)**: We have evidence for a claim C just to the extent it survives a stringent scrutiny. If C passes a test that was highly capable of finding flaws or discrepancies from C, and yet none or few are found, the passing result, $\boldsymbol{x}$, is evidence for C.
>
> Reinterpreting Mayo from a Bayesian viewpoint, we have evidence for the claim C if and only if (a) we observe a Bayes factor in favor of C beyond a context-sensitive threshold, and (b) the probability of finding misleading evidence for C (under the

[37] An exception would be for cases where a significant result is taken to infer there is an effect in either direction, without saying which. There is almost always interest in the direction of the effect.

> assumption that C's competitor is true) is low. (van Dongen et al. [2023], pp. 528–29)

They aver that their reinterpretation 'is exactly what Mayo alludes to in her Severity Principle' ([2023], p. 528). But it is quite different. The first big difference is that for severity, the parenthetical clause in condition b would be '(under the assumption that C is false)':

(b′) our method very probably would have found evidence for the denial of C, were C false.

Still, even condition b as stated seems to be an error probability—at least pre-data. What gives?

6.1. First way to resolve van Dongen, Sprenger, and Wagenmakers's puzzle: Pure likelihoodism

Their reinterpretation echoes the title of simple likelihoodist Royall's ([1997], [2000a]) 'On the probability of observing misleading statistical evidence', and whom they cite. The simple likelihoodist account is restricted to comparing the likelihoods of two simple hypotheses given data $\boldsymbol{x}$. Hacking ([1965]), Barnard ([1972]), and Birnbaum ([1969]) are early likelihoodists, but they all came to reject it by around 1972 (Hacking [1972], [1980]). The main reason they reject it is that 'the likelihood concept cannot be construed so as to allow useful appraisal and thereby possible control, of probabilities of erroneous interpretations [except] in the severely restricted case of a parameter space of just two points' (Birnbaum [1969], p. 128).[38] An example is $H_0 : \mu = \mu_0$, and $H_1 : \mu = \mu_1$, within a model, requiring no additional parameters (both are simple). It is further required that one of them be true, and that they be predesignated. If it is assumed, say, that H_0 is true, then the probability of obtaining a result that makes H_1 c times more likely than H_0 is less than $1/c$: $\Pr(\text{the likelihood ratio} > c; H_0) \leq 1/c$.[39] As Royall ([2000b], p. 776) remarks: 'The probabilities of weak and of misleading evidence are certainly relevant to the planning of studies [...] But after the experiment is completed, these probabilities are not relevant for interpreting the results'. This gives a first potential way to resolve the VSW puzzle: Pre-data, one may retain the likelihood principle and still uphold frequentist error probabilities—or something like them. Post-data, the evidence

[38] Another concern is whether likelihood ratios mean the same thing in different problems (Hacking [1972], p. 136).

[39] See (Neyman [1952]; Barnard [1962]; Savage [1962]; Kerridge [1963]; Birnbaum [1969]; Royall [2000a]; Mayo and Kruse [2001]; Sanborn and Hills [2013]).

is entirely in the likelihood ratio which Royall regards as an objective measure of what the data say, but he rejects using it as a test.[40]

However, this cannot be a plausible construal of VSW's Bayesian reinterpretation of severity. First, b′ holds for a very limited case of two predesignated simple point hypotheses, one of which must be true. This is sometimes called the 'universal bound' (Royall [2000a]). It would not hold for composite hypotheses and, as shown in our examples, it is readily violated for cases with biased selection effects. Generally, the parameter space is infinite, and to impose a restriction to two point hypotheses would scarcely capture hypothesis testing in practice. An account fails to be adequate for severe testing if it licenses insevere inferences; it is not enough that it could avoid them in certain cases. Nor is it enough that there is a 'limit' to error in the sense that foregone conclusions (to always being wrong) are avoided. The error probabilities must be controlled at small values.

Second, VSW cannot require condition b to hold in order to have strong evidence. There is no such post-data critique for the Bayes factor tester. The pre-data operational properties, even if known, do not enter into the appraisal of evidence: Once the Bayes factor is reported, all of the import of the evidence would have been given.[41] Conjoining the likelihood principle with VSW's conditions a and b is contradictory. The likelihood principle tells us condition b (or b′) is not required for a post-data assessment of the evidence from $\boldsymbol{x}$.[42]

6.2. Second way to resolve van Dongen, Sprenger, and Wagenmakers's puzzle: Expected high Bayes factor

VSW may intend a second line of argument. In their view: 'Bayes factor severity' means specifying contrasting point statistical hypotheses that are expected to give a high Bayes factor (for example, expected log likelihood ratio), for or against a hypothesis, on average. Good ([1983]) called this weight of evidence. Van Dongen et al. ([2023], p. 527) refer to Bayes factor design analysis (Schoenbrodt and Wagenmakers [2018]). Bayes factor severity promotes evidential value in the sense of promoting high Bayes factors. On this construal, VSW's Bayesian reinterpretation becomes: $\boldsymbol{x}$ is strong evidence for claim if

[40] On Royall's notion, evidence is 'misleading' because only a sample is observed. It is not intended to capture how the researcher can mislead by biasing selection effects. The same is true for VSW's notion of 'misleading'.

[41] With the typical mix of subjective and default priors for nuisance parameters, computing condition b might be difficult.

[42] In special cases, it is possible to match, or get close to matching, error probabilities and posteriors, as with types of objective or conventional Bayesian methods (Berger [2003], [2006]) and frequentist matching priors (Fraser and Reid [2002]; Fraser [2011]). There are also a variety of types of unifications of Bayesian and frequentist methods. A few of these are: calibrated Bayesians (Rosenkrantz [1977]; Rubin [1984]; Little [2006], [2025]; Williamson [2013]); pragmatic Bayesians (Kass [2011]; Gelman and Shalizi [2013]), and empirical Bayesians (Efron [2010]).

and only if (a) there is a high Bayes factor in favour of C and, (b) (something like) the test is designed so that it is expected to yield high Bayes factors either for H or for a chosen competitor to H on average. The 'if and only if' claim still cannot hold because the evidential assessment remains insensitive to the design properties in condition b. But even putting that aside, the account of testing is untenable.

Let us grant, for argument, that requiring highly contrastive, point hypotheses advances high Bayes factors, and thereby 'contributes to the [Bayes factor] severity of a test' (van Dongen et al. [2023], p. 522; see note 27).[43] The trouble is this opens the door to the problem of non-exhaustive contrastivism. Hypothesis H_1 can be accorded strong evidence on a Bayes factor test even though very little has been done to probe and rule out ways H_1 can be false. H_1 has passed the Bayes factor test, but it is not severe. This is not only because the Bayes factor is relative to a particular (non-exhaustive) alternative, but because of its insensitivity to the sampling protocol, and thus to error probabilities. I do not see how the most concerning problems in using significance tests are avoided rather than licensed. The same insensitivity to gambits that alter error probabilities would appear in the pre-data planning. The drive for high Bayes factors promotes the researchers' goal of obtaining a high Bayes factor for some hypothesis. It is not intended as assurance for the statistical consumer that the specific inference has been well probed.[44]

7. Conclusion

In this article, I have argued that the severity concept directs a reformulation of statistical significance tests while retaining their key features. These are: ensuring that regardless of which, if any, hypothesis in the parameter space is true, H_0 is rejected with small probability α when H_0 is true, and with much larger probability $(1 - \beta)$ when H_0 is false, increasing as discrepancies from H_0 increase. 'H is true' means it adequately captures some aspect of the process that could have generated the data, or that it solves a given statistical problem, including prediction. Error probability control is not merely to ensure good overall performance in inquiries, although that is important. Error control, I argued, plays a crucial role post-data in evaluating which claims have been well or poorly tested in a given case. The reformulation extends the conclusions beyond those of standard formal tests, but those extensions rely on a test's error probability control.

[43] Recall that a point hypothesis differs from a simple hypothesis. With the latter, there are no other parameters. The error statistician estimates the additional parameters rather than assigning them prior probabilities.

[44] The consumer can compute a posterior probability, but this only adds (allegedly) pre-data priors to the two hypotheses, with the consequence of yielding 0 posterior to any excluded parameter values:

$$\Pr(H_1 \mid \boldsymbol{x}) = \frac{1}{1 + \mathrm{BF}_{01}\frac{\Pr(H_0)}{\Pr(H_1)}}.$$

Statistical significance tests are only one of many error statistical methods, (which also include confidence interval estimation, resampling, randomization, and more), but they are important.

VSW assert that they deliver severe testing with Bayes factor tests, even maintaining that 'Bayesian inference is suited best to implement [the severe tester's] philosophical stance into experimental practice' (van Dongen et al. [2023], p. 529). To substantiate this claim, they need to address the error statistical implications of biasing selection effects and assure proponents of statistical significance testing that their error statistical concerns are preserved. I recommend VSW add a post-data report of how severely claims of interest have passed with the data, in the error statistical sense—even if only qualitative.[45] VSW's interest in severity is welcome. Their pursuing a Bayesian redefinition of severity suggests they are dissatisfied with their own rule that finds all the evidence post-experiment is in the Bayes factor.[46] The goal of statistical reforms is to mitigate questionable practices fueled by strong incentives in a reward-oriented research culture. At present, there is a serious disconnect between Bayes factor tests and error control protocols that are called for by replication reforms and by sceptical consumers of statistics.

Acknowledgements

I would like to thank the following people for illuminating exchanges on issues in this article: Jim Berger, George Chatfield, David Cox, Robert Kass, Richard Morey, Mark Rubin, Aris Spanos, Jan Sprenger, Noah van Dongen, and Eric-Jan Wagenmakers. I am especially grateful to Jean Miller and Wendy Parker for very useful comments and corrections on earlier drafts. I thank Marcos Jiménez for producing the figure, the BJPS editors and two anonymous reviewers for their astute insights and constructive recommendations.

Department of Philosophy
Virginia Tech
Blacksburg, VA, USA
mayod@vt.edu

References

Armitage, P. [1962]: 'Prepared Contribution', in G. A. Barnard and D. R. Cox (*eds*), *The Foundations of Statistical Inference*, Methuen, pp. 62–103.

[45] In this spirit, regulatory agencies perform a hybrid Bayesian–frequentist computation (using simulations) of a type 1 error probability (Ryan et al [2020]) to examine Bayesian sequential trials in exploratory inquiry.

[46] The striking omission of Bayes factor tests in Berger et al. ([2024])—despite Berger's long-standing advocacy—is telling.

Armitage, P. [1975]: *Sequential Medical Trials*, Wiley.

Barnard, G. [1962]: 'Prepared Contribution', in G. A. Barnard and D. R. Cox (*eds*), *The Foundations of Statistical Inference*, Methuen, pp. 62–103.

Barnard, G. [1972]: 'The Logic of Statistical Inference', *British Journal for the Philosophy of Science*, **23**, pp. 123–32.

Benjamin, D., Berger, J., Johannesson, M. et al. [2018]: 'Redefine Statistical Significance', *Nature Human Behaviour*, **2**, pp. 6–10.

Benjamini, Y. [2016]: 'It's Not the *P*-values' Fault: Supplement to Wasserstein and Lazar, "The ASA Statement on *p*-Values: Context, Process, and Purpose"', *The American Statistician*, **70**, available at <doi.org/10.1080/00031305.2016.1154108>.

Benjamini, Y. [2020]: 'Selective Inference: The Silent Killer of Replicability', *Harvard Data Science Review*, **2**, available at <doi.org/10.1162/99608f92.fc62b261>.

Benjamini, Y., De Veaux, R., Efron, B. et al. [2021]: 'The ASA President's Task Force Statement on Statistical Significance and Replicability', *Annals of Applied Statistics*, **15**, pp. 1084–85.

Benjamini, Y. and Hochburg, Y. [1995]: 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society B*, **57**, pp. 289–300.

Berger, J. [2003]: 'Could Fisher, Jeffreys, and Neyman Have Agreed on Testing?', *Statistical Science*, **18**, pp. 1–12.

Berger, J. [2006]: 'The Case for Objective Bayesian Analysis', *Bayesian Analysis*, **1**, pp. 385–402.

Berger, J. O., Bernardo, J. and Sun, D. [2024]: *Objective Bayesian Inference*, World Scientific.

Berger, J. and Pericchi, L. [2015]: 'Bayes Factors', in N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J. L. Teugels (*eds*), *Wiley StatsRef: Statistics*, available at <doi.org/10.1002/9781118445112.stat00224.pub2>.

Berger, J. and Wolpert, R. [1988]: *The Likelihood Principle*, Institute of Mathematical Statistics.

Bernardo, J. M. [2011]: 'Integrated Objective Bayesian Estimation and Hypothesis Testing', in J. M. Bernardo, M. J. Bayarri and J. O. Berger (*eds*), *Bayesian Statistics 9*, Oxford University Press, pp. 1–68.

Birnbaum, A. [1969]: 'Concepts of Statistical Evidence', in S. Morgenbesser, P. Suppes and M. White (*eds*), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, St Martin's Press, pp. 112–43.

Box, G. [1983]: 'An Apology for Ecumenism in Statistics', in G. Box, T. Leonard, and D. Wu (*eds*), *Scientific Inference, Data Analysis, and Robustness*, Academic Press, pp. 51–84.

Casella, G. and Berger, R. [1987]: 'Reconciling Bayesian and Frequentist Evidence in the

One-Sided Testing Problem', *Journal of the American Statistical Association*, **82**, pp. 106–11.

Cornfield, J. [1966]: 'A Bayesian Test of Some Classical Hypotheses, with Applications to Sequential Clinical Trials', *Journal of the American Statistical Association*, **61**, pp. 577–94.

Cousins, R. [2017]: 'The Jeffreys–Lindley Paradox and Discovery Criteria in High Energy Physics', *Synthese*, **194**, pp. 395–432.

Cox, D. R. [1958]: 'Some Problems Connected with Statistical Inference', *Annuals of Mathematical: Statistics*, **29**, pp. 357–72.

Cox, D. R. [1977]: 'The Role of Significance Tests', *Scandinavian Journal of Statistics*, **4**, pp. 49–70.

Cox, D. R. [1978]: 'Foundations of Statistical Inference: The Case for Eclecticism', *Australian Journal of Statistics*, **20**, pp. 43–59.

Cox, D. R. [2006]: *Principles of Statistical Inference*, Cambridge University Press.

Cox, D. R. and Hinkley, D. V. [1974]: *Theoretical Statistics*, Chapman and Hall.

Dienes, Z. [2023]: 'Testing Theories with Bayes Factors', in A. L. Nichols and J. Edlund (*eds*), *The Cambridge Handbook of Research Methods and Statistics for the Social and Behavioral Sciences, 1: Building a Program of Research*, Cambridge University Press, pp. 494–512.

Edwards, W., Lindman, H. and Savage, L. [1963]: 'Bayesian Statistical Inference for Psychological Research', *Psychological Review*, **70**, pp. 193–242.

Efron, B. [2010]: *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge University Press.

Fisher, R. A. [1935]: *Design of Experiments*, Oliver and Boyd.

Fraser, D. A. S. [2011]: 'Is Bayes Posterior Just Quick and Dirty Confidence?', *Statistical Science*, **26**, pp. 299–316.

Fraser, D. A. S. and Reid, N. [2002]: 'Strong Matching of Frequentist and Bayesian Parametric Inference', *Journal of Statistical Planning and Inference*, **103**, pp. 263–85.

Gelman, A. [2011]: 'Induction and Deduction in Bayesian Data Analysis', *Rationality, Markets, and Morals*, **2**, pp. 67–78.

Gelman, A. and Loken, E. [2014]: 'The Statistical Crisis in Science', *American Scientist*, **2**, pp. 460–65.

Gelman, A. and Shalizi, C. [2013]: 'Philosophy and the Practice of Bayesian Statistics', *British Journal of Mathematical and Statistical Psychology*, **66**, pp. 8—38.

Ghosh, J. Delampady, M. and Samanta, T. [2010]: *An Introduction to Bayesian Analysis: Theory and Methods*, Springer.

Gillies, D. [2001]: 'Bayesianism and the Fixity of the Theoretical Framework', in D. Corfield and J. Williamson (*eds*), *Foundations of Bayesianism 24*, Springer, pp. 363–79.

Good, I. J. [1983]: *Good Thinking: The Foundations of Probability and Its Applications*, University of Minnesota Press.

Hacking, I. [1965]: *Logic of Statistical Inference*, Cambridge University Press.

Hacking, I. [1972]: 'Likelihood', *British Journal for the Philosophy of Science*, **23**, pp. 132–37.

Hacking, I. [1980]: 'The Theory of Probable Inference: Neyman, Peirce, and Braithwaite', in D. Mellor (*ed.*), *Science, Belief, and Behavior: Essays in Honour of R. B. Braithwaite*, Cambridge University Press, pp. 141–60.

Haig, B. D. [2018]: *The Philosophy of Quantitative Methods*, Oxford University Press.

Haig, B. D. [2020]: 'What Can Psychology's Statistics Reformers Learn from the Error-Statistical Perspective', *Methods in Psychology*, **2**, available at <doi.org/10.1016/j.metip.2020.100020>.

Jeffreys, H. [1961]: *Theory of Probability*, Oxford University Press.

Johnson, V. [2013]: 'Uniformly Most Powerful Bayesian Tests', *The Annals of Statistics*, **41**, pp. 1716–41.

Kass, R. E. [2009]: 'The Importance of Jeffreys's Legacy (Comment on Robert, Chopin, and Rousseau)', *Statistical Science*, **24**, pp. 179–82.

Kass, R. E. [2011]: 'Statistical Inference: The Big Picture', *Statistical Science*, **26**, available at <doi.org/10.1214/10-STS337>.

Kass, R. E. and Raftery, A. [1995]: 'Bayes Factors', *Journal of the American Statistical Association*, **90**, pp. 773–95.

Kerridge, D. [1963]: 'Bounds for the Frequency of Misleading Bayes Inferences', *The Annals of Mathematical Statistics*, **34**, pp. 1109–10.

Kruschke, J. and Liddell, T. [2018]: 'The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-analysis, and Power Analysis from a Bayesian Perspective', *Psychonomic Bulletin and Review*, **25**, pp. 178–206.

Lakatos, I. [1978]: *The Methodology of Scientific Research Programmes*, Cambridge University Press.

Lakens, D. [2017]: 'Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-analyses', *Social Psychological and Personality Science*, **8**, pp. 355–62.

Lakens, D. [2019]: 'The Value of Preregistration for Psychological Science: A Conceptual Analysis', *Japanese Psychological Review*, **62**, pp. 221–30.

Lakens, D. [2022]: 'Improving Your Statistical Inferences', available at <doi.org/10.5281/zenodo.6409077>.

Lakens, D., Adolfi, F. G., Albers, C. J. et al. [2018]: 'Justify Your Alpha', *Nature Human Behavior*, **2**, pp. 168–71.

Lehmann, E. [1993]: 'The Fisher, Neyman–Pearson Theories of Testing Hypotheses: One Theory or Two?', *Journal of the American Statistical Association*, **88**, pp. 1242–49.

Lindley, D. V. [1957]: 'A Statistical Paradox', *Biometrika*, **44**, pp. 187–92.

Lindley, D. V. [1972]: *Bayesian Statistics: A Review*, SIAM.

Lindley, D. V. [2000]: 'The Philosophy of Statistics', *Journal of the Royal Statistical Society D*, **49**, pp. 293–337.

Little, R. [2006]: 'Calibrated Bayes: A Bayes/Frequentist Roadmap', *The American Statistician*, **60**, pp. 213–23.

Little, R. [2025]: *Seminal Ideas and Controversies in Statistics*, Chapman and Hall/CRC.

Mayo, D. G. [1991]: 'Novel Evidence and Severe Tests', *Philosophy of Science*, **58**, pp. 523–52.

Mayo, D. G. [1996]: *Error and the Growth of Experimental Knowledge*, University of Chicago Press.

Mayo, D. G. [2008]: 'How to Discount Double-Counting When It Counts: Some Clarifications', *British Journal for the Philosophy of Science*, **59**, pp. 857–79.

Mayo, D. G. [2010]: 'Learning from Error, Severe Testing, and the Growth of Theoretical Knowledge', in D. Mayo and A. Spanos (*eds*), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, Cambridge University Press, pp. 28–57.

Mayo, D. G. [2014]: 'On the Birnbaum Argument for the Strong Likelihood Principle', *Statistical Science*, **29**, pp. 227–39.

Mayo, D. G. [2018]: *Statistical Inference as Severe Testing: How to Get beyond the Statistics Wars*, Cambridge University Press.

Mayo, D. G. and Cox, D. R. [2006]: 'Frequentist Statistics as a Theory of Inductive Inference', in J. Rojo (*ed.*), *Optimality: The Second Erich L. Lehmann Symposium*, Institute of Mathematical Statistics, pp. 77–97.

Mayo, D. G. and Hand, D. [2022]: 'Statistical Significance and Its Critics: Practicing Damaging Science or Damaging Scientific Practice?', *Synthese*, **200**, available at <doi.org/10.1007/s11229-022-03692-0>.

Mayo, D. G. and Kruse, M. [2001]: 'Principles of Inference and Their Consequences', in D. Cornfield and J. Williamson (*eds*), *Foundations of Bayesianism*, Kluwer, pp. 381–403.

Mayo, D. G. and Spanos, A. [2004]: 'Methodology in Practice: Statistical Misspecification Testing', *Philosophy of Science*, **71**, pp. 1007–25.

Mayo, D. G. and Spanos, A. [2006]: 'Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction', *British Journal for the Philosophy of Science*, **57**, pp. 323–57.

Mayo, D. G. and Spanos, A. [2011]: 'Error Statistics', in P. S. Bandyopadhyay and M. R. Forster (*eds*), *Philosophy of Statistics*, North-Holland, pp. 153–98.

Neyman, J. [1967]: 'Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability', in J. Neyman (*ed.*), *Early Statistical Papers of J. Neyman*, University of California Press, pp. 250–90.

Neyman, J. [1952]: *Lectures and Conferences on Mathematical Statistics and Probability*, Graduate School of US Department of Agriculture.

Neyman, J. and Pearson, E. [1933]: 'On the Problem of the Most Efficient Tests of Statistical Hypotheses', *Philosophical Transactions of the Royal Society of London A*, **231**, pp. 289–337.

Peden, W. [2020]: 'The Bayesian Era in the Philosophy of Science', *Studies in History and Philosophy of Science*, **80**, pp. 123–27.

Popper, K. [1959]: *The Logic of Scientific Discovery*, Routledge.

Royall, R. [1997]: *Statistical Evidence: A Likelihood Paradigm*, Chapman and Hall/CRC.

Royall, R. [2000a]: 'On the Probability of Observing Misleading Statistical Evidence', *Journal of the American Statistical Association*, **95**, pp. 760–68.

Royall, R. [2000b]: 'Rejoinder', *Journal of the American Statistical Association*, **95**, pp. 773–80.

Rosenkrantz, R. [1977]: *Inference, Method, and Decision: Towards a Bayesian Philosophy of Science*, D. Reidel.

Rubin, D. [1984]: 'Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician', *The Annals of Statistics*, **12**, pp. 1151–72.

Ryan, E. G., Brock, K., Gates, S. and Slade, D. [2020]: 'Do We Need to Adjust for Interim Analyses in a Bayesian Adaptive Trial Design?', *BMC Medical Research Methodology*, **20**, available at <doi.org/10.1186/s12874-020-01042-7>.

Sanborn, A. and Hills, T. [2013]: 'The Frequentist Implications of Optional Stopping on Bayesian Hypothesis Tests', *Psychonomic Bulletin and Review*, **21**, pp. 283–300.

Savage, L. J. [1962]: 'Subjective Probability and Statistical Practice, in G. A. Barnard and D. R. Cox (*eds*), *The Foundations of Statistical Inference*, Methuen, pp. 9–35.

Schoenbrodt, F. and Wagenmakers, E.-J. [2018]: 'Bayes Factor Design Analysis: Planning for Compelling Evidence', *Psychonomic Bulletin and Review*, **25**, pp. 128–42.

Senn, S. [2001]: 'Two Cheers for *P*-values?', *Journal of Epidemiology and Biostatistics*, **6**, pp. 193–204.

Simmons, J., Nelson, L. and Simonsohn, U. [2011]: 'False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allow Presenting Anything as Significant', *Psychological Science*, **22**, pp. 1359–66.

Spanos, A. [2013]: 'Who Should Be Afraid of the Jeffreys–Lindley Paradox?', *Philosophy of Science*, **80**, pp. 73–93.

Spanos, A. [2019]: *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge University Press.

Sprenger, J. and Hartmann, S. [2019]: *Bayesian Philosophy of Science*, Oxford University Press.

Tendeiro, J. N., Kiers, H. A. L., Hoekstra, R., Wong, T. K. and Morey, R. D. [2024]: 'Diagnosing the Misuse of the Bayes Factor in Applied Research',

Advances in Methods and Practices in Psychological Science, **7**, available at <doi.org/10.1177/25152459231213371>.

van Dongen, N., Sprenger, J. and Wagenmakers, E.-J. [2023]: 'A Bayesian Perspective on Severity: Risky Predictions and Specific Hypotheses', *Psychonomic Bulletin and Review*, **30**, pp. 516–33.

Vanpaemel, W. [2019]: 'The Really Risky Registered Modeling Report: Incentivizing Strong Tests and Honest Modeling in Cognitive Science', *Computational Brain and Behavior*, **2**, pp. 218–22.

Wagenmakers, E.-J. [2007]: 'A Practical Solution to the Pervasive Problems of p Values', *Psychonomic Bulletin and Review*, **14**, pp. 779–804.

Wald, A. [1947]: *Sequential Analysis*, John Wiley.

Wasserstein, R. L. and Lazar, N. A. [2016]: 'The ASA's Statement on p-values: Context, Process, and Purpose', *The American Statistician*, **70**, pp. 129–33.

Wasserstein, R. L., Schirm, A. L. and Lazar, N. A. [2019]: 'Moving to a World beyond "$p < 0.05$"', *The American Statistician*, **73**, pp. 1–19.

Wellek, S. [2010]: *Testing Statistical Hypotheses of Equivalence and Noninferiority*, Chapman and Hall/CRC.

Williamson, J. [2013]: 'Why Frequentists and Bayesians Need Each Other', *Erkenntnis*, **78**, pp. 293–318.

Wong T. K., Kiers H., Tendeiro J. [2022]: 'On the Potential Mismatch between the Function of the Bayes Factor and Researchers' Expectations', *Collabra: Psychology*, **8**, available at <doi.org/10.1525/collabra.36357>.

Worrall, J. [1978]: 'Research Programmes, Empirical Support, and the Duhem Problem: Replies to Criticism', in G. Radnitzky and G. Andersson (*eds*), *Progress and Rationality in Science*, D. Reidel, pp. 321–38.