entropy

MDPI

Article

How the Post-Data Severity Converts Testing Results into Evidence for or against Pertinent Inferential Claims

Aris Spanos

Department of Economics, Virginia Tech, Blacksburg, VA 24061, USA; aris@vt.edu

Abstract: The paper makes a case that the current discussions on replicability and the abuse of significance testing have overlooked a more general contributor to the untrustworthiness of published empirical evidence, which is the uninformed and recipe-like implementation of statistical modeling and inference. It is argued that this contributes to the untrustworthiness problem in several different ways, including [a] statistical misspecification, [b] unwarranted evidential interpretations of frequentist inference results, and [c] questionable modeling strategies that rely on curve-fitting. What is more, the alternative proposals to replace or modify frequentist testing, including [i] replacing *p*-values with observed confidence intervals and effects sizes, and [ii] redefining statistical significance, will not address the untrustworthiness of evidence problem since they are equally vulnerable to [a]–[c]. The paper calls for distinguishing between unduly data-dependant 'statistical results', such as a point estimate, a *p*-value, and accept/reject H_0, from 'evidence for or against inferential claims'. The post-data severity (SEV) evaluation of the accept/reject H_0 results, converts them into evidence for or against germane inferential claims. These claims can be used to address/elucidate several foundational issues, including (i) statistical vs. substantive significance, (ii) the large n problem, and (iii) the replicability of evidence. Also, the SEV perspective sheds light on the impertinence of the proposed alternatives [i]–[iii], and oppugns [iii] the alleged arbitrariness of framing H_0 and H_1 which is often exploited to undermine the credibility of frequentist testing.

Keywords: replication; untrustworthy evidence; statistical misspecification; statistical vs. substantive significance; pre-data vs. post-data error probabilities; p-hacking; post-data severity evaluation; observed confidence intervals; effect sizes

Citation: Spanos, A. How the Post-Data Severity Converts Testing Results into Evidence for or against Pertinent Inferential Claims. *Entropy* **2024**, *26*, 95. https://doi.org/10.3390/e26010095

Academic Editors: Brian Dennis, Mark L. Taper, Jose Miguel Ponciano and Geert Verdoolaege

Received: 29 October 2023
Revised: 29 November 2023
Accepted: 29 December 2023
Published: 22 January 2024

1. Introduction

The replication crisis has dominated discussions on empirical evidence and their trustworthiness in scientific journals for the last two decades. The broad agreement is that the non-replicability of such empirical evidence provides prima facie evidence of their untrustworthiness; see National Academy of Sciences [1], Wasserstein and Lazar [2], Baker [3], Hoffler [4]. A statistical study is said to be replicable if its empirical results can be independently confirmed – with very similar or consistent results – by other researchers using akin data and modeling the same phenomenon of interest.

Using the Medical Diagnostic Screening (MDS) perspective on Neyman-Pearson (N-P) testing Ioannidis [5] declares "... most published research findings are false", attributing the untrustworthiness of evidence to several abuses of frequentist testing, such as p-hacking, multiple testing, cherry-picking and low power. This diagnosis is anchored on apparent analogies between the type I/II error probabilities and the false negative/positive probabilities of the MDS model, tracing the untrustworthiness to ignoring the Bonferroni-type adjustments needed to ensure that the *actual* error probabilities approximate the *nominal* ones. In light of that, leading statisticians in different applied fields called for reforms which include replacing *p*-values with observed Confidence Intervals (CIs), using effect sizes, and redefining statistical significance; see Benjamin et al. [6].

In the discussion that follows, a case is made that Ioannidis' assessment about the untrustworthiness of published empirical findings is largely right. Still, the veracity of viewing the MDS model as a surrogate for N-P testing, and its pertinence in diagnosing the untrustworthiness of evidence problem are highly questionable. This stems from the fact that the invoked analogies between the type I/II error probabilities and the false negative/positive probabilities of the MDS model are more apparent than real, since the former are hypothetical/unobservable/unconditional and the latter are observable conditional probabilities; see Spanos [7].

A more persuasive case can be made that the untrustworthiness of empirical evidence stems from the broader problem of *the uninformed and recipe-like implementation of statistical modeling and inference* that contributes to untrustworthy evidence in several interrelated ways, including:

[a] Statistical misspecification: invalid probabilistic assumptions imposed (implicitly or explicitly) on one's data, comprising the invoked statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

[b] 'Empirical evidence' is often conflated with raw 'inference results', such as point estimates, effect sizes, observed CIs, accept/reject H_0 results, and p-values, giving rise to (i) erroneous evidential interpretations of these results, and (ii) unwarranted claims relating to their replicability.

[c] Questionable modeling strategies that rely on curve-fitting of hybrid models—an amalgam of substantive subject matter and probabilistic assumptions—guided by error term assumptions and evaluated on goodness-of-fit/prediction grounds. The key weakness of this strategy is that excellent goodness-of-fit/prediction is neither necessary nor sufficient for the statistical adequacy of the selected model since it depends crucially on the invoked loss function whose choice is based on information other than the data. It can shown that statistical models chosen on goodness-of-fit/prediction grounds are often statistically misspecified; see Spanos [8].

Viewed in the broader context of [a]–[c], the abuses of frequentist testing represent the tip of the untrustworthy evidence iceberg. It also questions the presumption that replicability attests to the trustworthiness of empirical evidence. As argued by Leek and Peng [9]: "... an analysis can be fully reproducible and still be wron." (p. 1314). For instance, dozens of MBA students confirm the efficient market hypothesis (EMH) on a daily basis because they follow the same uninformed and recipe-like, implementation of statistics, unmindful of what it takes to ensure the trustworthiness of the ensuing evidence by addressing the issues [a]–[c]; see Spanos [10].

The *primary focus* of the discussion that follows is on [b] with brief comments on [a] and [c], but citing relevant published papers. The discussion revolves around the distinction between unduly data-specific 'inference results', such as point estimates, observed CIs, p-values, effect sizes, and the accept/reject H_0 results, and ensuing inductive generalizations from such results in the form of 'evidence for or against germane inferential claims' framed in terms of the unknown parameters $\boldsymbol{\theta}$. The crucial difference between 'results' and 'evidence' is twofold:

(a) the evidence is framed in terms of *post-data* error probabilities aiming to account for the uncertainty arising from the fact that 'inference results' rely unduly on the particular data $\mathbf{x}_0{:=}(x_1, x_2, ..., x_n)$, which constitutes a single realization $\mathbf{X} = \mathbf{x}_0$ of the sample $\mathbf{X}{:=}(X_1, ..., X_n)$, and

(b) the evidence, in the form of warranted inferential claims, enhances learning from data $\mathbf{x}_0$ about the stochastic mechanism that could have given rise to this data.

As a prelude to the discussion that follows, Section 2 provides a brief overview of Fisher's model-based frequentist statistics with special emphasis on key concepts and pertinent interpretations of inference procedures that are invariably misconstrued by the uninformed and recipe-like implementation of statistical modeling and inference. Section 3 discusses a way to bridge the gap between unduly data-specific inference results and an evidential interpretation of such results, in the form of the post-data severity (SEV) evaluation of the accept/reject H_0 results. The SEV evaluation is used to elucidate or/and

address several foundational issues that have bedeviled frequentist testing since the 1930s, including the large n problem, statistical vs. substantive significance and the replicability of evidence, as opposed to the replicability of statistical results. In Section 4 the SEV evaluation is used to appraise several proposed alternatives to (or modifications of) N-P testing by the replication literature, including replacing the p-value with effect sizes and observed CIs and redefining statistical significance. Section 5 compares and contrasts the evidential account based on the SEV evaluation with Royall's [11] Likelihood Ratio approach to statistical evidence.

2. Model-Based Frequentist Inference

2.1. Fisher's Statistical Induction

Fisher [12] pioneered modern frequentist statistics by viewing data $\mathbf{x}_0$ as a typical realization of a prespecified parametric statistical model whose generic form is:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}) = \{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\subset\mathbb{R}^m,\ \mathbf{x}\in\mathbb{R}^n_X\},\ m<n, \tag{1}$$

where Θ and $\mathbb{R}^n_X$ denote the parameter and sample space, respectively, $f(x;\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}^n_X$, refers to the (joint) *distribution of the sample* $\mathbf{X}$. The initial choice (specification) of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ should be a response to the question: "Of what population is this a random sample?" (Fisher, [12], p. 313), underscoring that: 'the adequacy of our choice may be tested a posteriori' (ibid., p. 314). This can be secured by establishing the statistical adequacy (approximate validity) of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ using thorough Mis-Specification (M-S) testing; see Spanos [13].

Selecting $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ for data $\mathbf{x}_0$ has a twofold objective (Spanos [14]):

(i) $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is selected with a view to account for the chance regularity patterns exhibit by data $\mathbf{x}_0$ by accounting for these regularities using appropriate probabilistic assumptions relating to $\{X_k,\ t\in\mathbb{N}:=(1,2,...,n,...\}$ from three broad categories: **Distribution** (D), **Dependence** (M) and **Heterogeneity** (H).

(ii) $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is parametrized [$\boldsymbol{\theta}\in\Theta$] in a way that can shed light on the substantive questions of interest using data $\mathbf{x}_0$. When such questions are framed in terms of a *substantive model*, say $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$, $\boldsymbol{\varphi}\in\Phi$, one needs to bring out the implicit statistical model without restricting its parameters $\boldsymbol{\theta}$, and ensure that $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are related via a set of restrictions $\mathbf{g}(\boldsymbol{\varphi},\boldsymbol{\theta})=\mathbf{0}$ connecting $\boldsymbol{\varphi}$ to the data $\mathbf{x}_0$ via $\boldsymbol{\theta}$.

Example 1. Consider the well-known *simple Normal model*:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}):\quad X_t\backsim\mathsf{NIID}(\mu,\sigma^2),\ \mu\in\mathbb{R},\ \sigma^2>0,\ x_t\in\mathbb{R},\ t\in\mathbb{N}:=\{1,2,...,n,...\}, \tag{2}$$

where '$X_t\backsim$NIID' stands for 'X_t is Normal (D), Independent (M) and Identically Distributed (H)', $\mathbb{R}:=(-\infty,\infty)$. It is important to emphasize that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ revolves around $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}^n_X$ in (1) since it encapsulates all its probabilistic assumptions:

$$f_{SN}(\mathbf{x};\boldsymbol{\theta})\overset{\mathrm{I}}{=}\prod_{k=1}^{n}f_k(x_k;\boldsymbol{\theta}_k)\overset{\mathrm{IID}}{=}\prod_{k=1}^{n}f(x_k;\boldsymbol{\theta})\overset{\mathrm{NIID}}{=}\left(\tfrac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\{-\tfrac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2\},\ \mathbf{x}\in\mathbb{R}^n, \tag{3}$$

and provides the *cornerstone* for all forms of statistical inference.

The *primary objective* of model-based frequentist inference is to 'learn from data $\mathbf{x}_0$' about $\boldsymbol{\theta}^*$, where $\boldsymbol{\theta}^*$ denotes the 'true' $\boldsymbol{\theta}$ in Θ. This is shorthand for saying that there exists a $\boldsymbol{\theta}^*\in\Theta$ such that $\mathcal{M}_{\boldsymbol{\theta}^*}(\mathbf{x})=f(\mathbf{x};\boldsymbol{\theta}^*)$, $\mathbf{x}\in\mathbb{R}^n_X$, could have generated $\mathbf{x}_0$.

The main variants of statistical inference in frequentist statistics are: (i) point estimation, (ii) interval estimation, (iii) hypothesis testing, and (iv) prediction. These forms of statistical inference share the following features:

(a) They assume that the prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is valid vis-à-vis data $\mathbf{x}_0$.

(b) They aim is to learn about $\mathcal{M}_{\boldsymbol{\theta}^*}(\mathbf{x})$ using statistical approximations relating to $\boldsymbol{\theta}^*$.

(c) Their inferences are based on a *statistic* (estimator, test statistic, predictor), say $Y_n=g(\mathbf{X})$, whose sampling distribution, $f(y_n;\boldsymbol{\theta})$, $\forall y\in\mathbb{R}_Y$, ($\forall$ stands 'for all') is derived directly from the distribution of the sample $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}^n_X$, of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, using two different forms of reasoning with prespecified values of $\boldsymbol{\theta}$:

(a) *factual* (estimation and prediction): presume that $\boldsymbol{\theta}=\boldsymbol{\theta}^*$, and

(b) *hypothetical* (testing): H_0: $\boldsymbol{\theta}\in\Theta_0$ (what if $\boldsymbol{\theta}\in\Theta_0$) vs. H_1: $\boldsymbol{\theta}\in\Theta_1$ (what if $\boldsymbol{\theta}\in\Theta_1$). The crucial difference between these two forms of reasoning is that factual reasoning does not extend to post-data (after $\mathbf{x}_0$ is known) evaluations relating to evidence, but hypothetical reasoning does. This plays a key role in the following discussion.

The primary role of the sampling distribution of a statistic $f(y_n;\boldsymbol{\theta})$, $\forall y\in\mathbb{R}_Y$, is to frame the uncertainty relating to the fact that $\mathbf{x}_0$ is just one, out of all $\mathbf{x}\in\mathbb{R}_X^n$ realizations, of $\mathbf{X}$ so as to provide (i) the basis for the optimality of the statistic $Y_n=g(\mathbf{X})$, as well as (ii) the relevant error probabilities to 'calibrate' the capacity (optimality) of inference based on Y_n; how often the inference procedure errs.

The *statistical adequacy* (approximate validity) of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ plays a pivotal role in securing the reliability of inference and the trustworthiness of ensuing evidence because it ensures that the *nominal* optimality—derived by assuming the validity of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$—is also *actual* for data $\mathbf{x}_0$, and secures the approximate equality between the actual (based on $\mathbf{x}_0$) and the nominal error probabilities. In contrast, when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is *statistically misspecified*:

(a) the joint distribution of the sample $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}_X^n$, and the likelihood function $L(\boldsymbol{\theta};\mathbf{x}_0)\propto f(\mathbf{x}_0;\boldsymbol{\theta})$, $\boldsymbol{\theta}\in\Theta$, are both erroneous,

(b) all sampling distributions $f(y_n;\boldsymbol{\theta})$, derived by invoking the validity of $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}_X^n$, will be incorrect, (i) giving rise to 'non-optimal' estimators, and (ii) sizeable *discrepancies* between the actual and nominal error probabilities.

Applying a $\alpha=0.05$ significance level test when the actual type I error probability is 0.97 due to invalid probabilistic assumptions will yield untrustworthy evidence. Increasing the sample size will often worsen the untrustworthiness by increasing the discrepancy between actual and nominal error probabilities; see Spanos [15], p. 691. Hence, the best way to keep track of the relevant error probabilities is to establish the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. It is important to emphasize that other forms of statistical inference, including Bayesian and Akaike-type model selection procedures, are equally vulnerable to statistical misspecification since they rely on the likelihood function $L(\boldsymbol{\theta};\mathbf{x}_0)\propto f(\mathbf{x}_0;\boldsymbol{\theta})$, $\boldsymbol{\theta}\in\Theta$; see Spanos [16].

In the discussion that follows, it is assumed that the invoked statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is statistically adequate to avoid repetitions and digressions, but see Spanos [17] and [18] on why [a] statistical misspecification calls into question important aspects of the current replication crisis literature.

2.2. Frequentist Inference: Estimation

Point estimation revolves around an estimator, say $\widehat{\boldsymbol{\theta}}_n(\mathbf{X})$, that pinpoints (as closely as possible) $\boldsymbol{\theta}^*$. The clause 'as closely as possible' is framed in terms of certain 'optimal' properties stemming from the sampling distribution $f(\widehat{\boldsymbol{\theta}}_n(\mathbf{x});\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}_X^n$, including: unbiasedness, efficiency, sufficiency, consistency, etc.; see Casella and Berger [19]. Regrettably, the *factual reasoning*, presuming $\boldsymbol{\theta}=\boldsymbol{\theta}^*$, underlying the derivation of the relevant sampling distributions is often implicit in traditional textbook discussions, resulting in erroneous interpretations and unwarranted claims.

Example 1 (continued). The relevant sampling distributions associated with (2), are appositely stated as:

$$\begin{aligned}&\text{[i] } \overline{X}_n \overset{\boldsymbol{\theta}=\boldsymbol{\theta}^*}{\backsim} \mathsf{N}\left(\mu_*,\tfrac{\sigma_*^2}{n}\right), \quad \text{[ii] } \tfrac{(n-1)s^2}{\sigma_*^2} \overset{\boldsymbol{\theta}=\boldsymbol{\theta}^*}{\backsim} \chi^2(n-1), \quad \text{[iii] } \tau(\mathbf{X};\mu) \overset{\boldsymbol{\theta}=\boldsymbol{\theta}^*}{\backsim} \mathsf{St}(n-1),\\ &\overline{X}_n=\tfrac{1}{n}\textstyle\sum_{t=1}^n X_t, \qquad s^2=\tfrac{1}{(n-1)}\textstyle\sum_{t=1}^n (X_t-\overline{X}_n)^2, \qquad \tau(\mathbf{X};\mu)=\tfrac{\sqrt{n}(\overline{X}_n-\mu)}{s},\end{aligned} \tag{4}$$

where $\boldsymbol{\theta}^*:=(\mu^* and \sigma_*^2)$ denote the 'true' values of the unknown parameters, $\chi^2(n-1)$ denotes the chi-square distribution, and $\mathsf{St}(n-1)$ the Student's t distribution, with $(n-1)$ degrees of freedom. The problem is that without '$\boldsymbol{\theta}=\boldsymbol{\theta}^*$' the distributional results in (4) will not hold. For instance, what ensures in [i] that $E(\overline{X}_n)=\mu$ and $Var(\overline{X}_n)=(\sigma^2/n)$? The answer is the unbiasedness and full efficiency of $\overline{X}_n$, respectively, both of which are defined at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$. There is no such a thing as a sampling distribution $\overline{X}_n \backsim \mathsf{N}(\mu,\frac{\sigma^2}{n})$, $\forall\boldsymbol{\theta}:=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+$ since the NIID assumptions imply that each element of the sample $\mathbf{X}$

comes from a single Normal distribution with a unique mean and variance $\boldsymbol{\theta}^{*}:=(\mu^{*},\sigma^{2}_{*})$ around which all forms of statistical inference revolve. Hence, the claim by Schweder and Hjort [20] that "$\frac{\sqrt{n}(\mu-\overline{X}_n)}{s}$ has a fixed distribution regardless of the values of the interest parameter μ and the (in this context) nuisance parameter σ... " (p. 15), i.e.,

$$d(\mathbf{X};\mu)=\frac{\sqrt{n}(\mu-\overline{X}_n)}{s}\backsim \mathsf{St}(n-1),\ \forall\boldsymbol{\theta}:=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_{+}, \tag{5}$$

makes no sense from a statistical inference perspective.

Point estimation is very important since it provides the basis for all other forms of optimal inference (CIs, testing, and prediction) via the optimality of a point estimator $\widehat{\boldsymbol{\theta}}_n(\mathbf{X})$. A point estimate $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_0)$, by itself, however, is considered *inadequate* for learning from data $\mathbf{x}_0$ since it is unduly data specific; it ignores the relevant uncertainty stemming from the fact that $\mathbf{x}_0$ constitutes a single realization (out of $\forall\mathbf{x}\in\mathbb{R}^n_X$) as framed by the sampling distribution, $f(\widehat{\boldsymbol{\theta}}_n(\mathbf{x});\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$, of $\widehat{\boldsymbol{\theta}}_n(\mathbf{X})$; hence $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_0)$ is often reported as $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_0)\pm 2\sqrt{Var(\widehat{\boldsymbol{\theta}}_n)}$.

Interval estimation accounts for the relevant uncertainty in terms of an error probability of 'overlaying' the true value $\boldsymbol{\theta}^{*}$ of $\boldsymbol{\theta}$, based on $f(\widehat{\boldsymbol{\theta}}_n(\mathbf{x});\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$, in the form of the Confidence Interval (CI):

$$\mathbb{P}(L(\mathbf{X})\leq\boldsymbol{\theta}\leq U(\mathbf{X});\ \boldsymbol{\theta}=\boldsymbol{\theta}^{*})\quad=1-\alpha, \tag{6}$$

where the statistics $L(\mathbf{X})$ and $U(\mathbf{X})$ denote the lower and upper (random) bounds that 'overlay' $\boldsymbol{\theta}^{*}$ with probability $(1-\alpha)$. An $(1-\alpha)$ CI is optimal when its expected length $E[U(\mathbf{X})-L(\mathbf{X})]$ is the shortest and referred to as Uniformly Most Accurate (UMA); see Lehmann and Romano [21].

Example 1 (continued). For (2), the two-sided optimal $(1-\alpha)$ CI for μ is based on the pivot $\tau(\mathbf{X};\mu)$ in (4)-[iii] and takes the form:

$$\text{(a) }\mathbb{P}[CI(\mathbf{X});\alpha]=\mathbb{P}(\overline{X}_n-c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}})\leq\mu<\overline{X}_n+c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}});\ \boldsymbol{\theta}=\boldsymbol{\theta}^{*})=1-\alpha. \tag{7}$$

The analogous one-sided optimal CIs take the form:

$$\begin{aligned}&\text{(b) Lower: }\mathbb{P}[CI_L(\mathbf{X});\alpha]=\mathbb{P}_L(\mu\geq\overline{X}_n-c_{\alpha}(\tfrac{s}{\sqrt{n}}));\ \boldsymbol{\theta}=\boldsymbol{\theta}^{*})=1-\alpha,\\&\text{(c) Upper: }\mathbb{P}[CI_U(\mathbf{X});\alpha]=\mathbb{P}_U(\mu\leq\overline{X}_n+c_{\alpha}(\tfrac{s}{\sqrt{n}});\ \boldsymbol{\theta}=\boldsymbol{\theta}^{*})=1-\alpha.\end{aligned} \tag{8}$$

2.3. Frequentist Inference: Neyman-Pearson (N-P) Testing

The reasoning underlying hypothesis testing is *hypothetical*, based on prespecified values of $\boldsymbol{\theta}$, as they relate to H_0: $\boldsymbol{\theta}\in\Theta_0$ and H_1: $\boldsymbol{\theta}\in\Theta_1$.

Example 1 (continued). Consider testing the hypotheses of interest:

$$H_0\text{: }\mu\leq\mu_0\text{ vs. }H_1\text{: }\mu>\mu_0, \tag{9}$$

in the context of (2). An optimal N-P test for the hypotheses in (9) is defined in terms of a test statistic and the rejection region:

$$T_{\alpha}:=\{\tau(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{s},\ C_1(\alpha)=\{\mathbf{x}\text{: }d(\mathbf{x})>c_{\alpha}\}\}, \tag{10}$$

whose error probabilities are evaluated using:

$$\text{[i] what if }\mu=\mu_0\text{: }\tau(\mathbf{X})\overset{\mu=\mu_0}{\backsim}\mathsf{St}(n-1), \tag{11}$$

where $\mathsf{St}(n-1)$ is the Student's t distribution with $n-1$ degrees of freedom, and:

$$\text{[ii]: what if }\mu=\mu_1,\ \tau(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{s}\overset{\mu=\mu_1}{\backsim}\mathsf{St}(\delta_1,n-1),\ \forall\mu_1=\mu_0+\gamma_1,\ \gamma_1\geq 0, \tag{12}$$

where $\mathsf{St}(\delta_1,n-1)$ is a noncentral Student's t distribution with $\delta_1=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$. It is important to emphasize that (12) differs from (11) in terms of their mean, variance, and higher moments, rendering (12) non-symmetric for $\delta_1\neq 0$; see Owen [22].

The sampling distribution in (11) is used to evaluate the pre-data type I error probability and the *post-data* [$\tau(\mathbf{x}_0)$ is known] p-value:

$$\text{(a) } \alpha = \mathbb{P}(\mathbf{x}\text{: } \tau(\mathbf{X}) > c_\alpha; \mu = \mu_0), \quad p(\mathbf{x}_0) = \mathbb{P}(\mathbf{x}\text{: } \tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu = \mu_0). \tag{13}$$

The sampling distribution in (12) is used to evaluate the power of T_α:

$$\text{(b) } \mathcal{P}(\mu_1) = \mathbb{P}(\mathbf{x}\text{: } \tau(\mathbf{X}) > c_\alpha; \mu = \mu_1), \ \forall \mu \geq \mu_1 = \mu_0 + \gamma_1, \ \gamma_1 \geq 0, \tag{14}$$

as well as the type II error probability: $\beta(\mu_1) = 1 - \mathcal{P}(\mu_1)$, $\forall \mu \geq \mu_1$.

The test T_α in (10) is optimal in the sense of being Uniformly Most Powerful (UMP), i.e., T_α is the most effective α-level test for detecting any discrepancy ($\gamma > 0$) of interest from $\mu = \mu_0$; see Lehmann and Romano [21].

Why prespecify α at a low threshold, such as $\alpha = 0.05$? Neyman and Pearson [23] put forward two crucial stipulations relating to the framing of H_0: $\boldsymbol{\theta} \in \Theta_0$ and H_1: $\boldsymbol{\theta} \in \Theta_1$, $\Theta_i \subset \Theta$, $i = 0, 1$, to ensure the effectiveness of N-P testing and the informativeness of its results:

[1] Θ_0 and Θ_1 should form a *partition* of Θ (p. 293) to avoid $\boldsymbol{\theta}^* \notin [\Theta_0 \cup \Theta_1]$.

[2] Θ_0 and Θ_1 should be framed in such a way so as to ensure that the type I error *is the more serious* of the two.

To provide some intuition for [2], they use the analogy with a criminal trial where to ensure [2] one should use the framing, H_0: not guilty vs. H_1: guilty, to render the type I error of sending an innocent person to prison, more serious than acquitting a guilty person (p. 296). Hence, prespecifying α at a small value and maximizing the power over $\forall \mu \geq \mu_1 = \mu_0 + \gamma_1, \ \gamma_1 \geq 0$, requires deliberation about the framing. A moment's reflection suggests that stipulation [2] implies that high power is needed around the potential neighborhood of $\boldsymbol{\theta}^*$. Regrettably, stipulations [1]–[2] are often ignored, undermining the proper implementation and effectiveness of N-P testing; see Section 4.1.

Returning to the power of T_α, the noncentrality parameter δ_1 indicates that the power increases monotonically with $\sqrt{n}$ and $(\mu_1 - \mu_0)$ and decreases with σ. This suggests that the inherent trade-off between the type I and II error probabilities in N-P testing, in conjunction with the sample size n and α, plays a crucial role in determining the capacity of a N-P test. This means that the selection of the significance level α should always take into account the particular n for data $\mathbf{x}_0$, since an uninformed choice of α can give rise to two problems.

The small n problem. This arises when the sample size n is not large enough generate any learning from data about $\boldsymbol{\theta}^*$ since it has insufficient power to detect particular discrepancies γ of interest. To avoid underpowered tests the formula in (14) can be used *pre-data* (before $\tau(\mathbf{x}_0)$ is known) to evaluate the sample size n necessary for T_α to detect such discrepancies with high enough probability (power). That is, for a given α, there is always a small enough n that would accept H_0 despite the presence of a sizeable discrepancy $\gamma \neq 0$ of interest. This also undermines the M-S testing to evaluate the statistical adequacy of the invoked $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ since a small n will ensure that M-S tests do not have sufficient power to detect existing departures from the model assumptions; see Spanos [18].

The large n problem. This arises when a practitioner uses conventional significance levels, say $\alpha = 0.1, 0.05, 0.025, 0.01$, for very large sample sizes, say $n > 10{,}000$. The source of this problem is that for a given α, as n increases the power of a test increases and the p-value decreases, giving rise to over-sensitive tests. Fisher [24] explained why: "By increasing the size of the experiment [n], we can render it more sensitive, meaning by this that it will allow of the detection of ... quantitatively smaller departures from the null hypothesis." (pp. 21–22). Hence, for a given α, there is always a large enough n that would reject H_0 for any discrepancy $\gamma \neq 0$ (however small, say $\gamma = 0.0000001$) from a null value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$; see Spanos [25].

It is very important to emphasize at the outset that the pre-data testing error probabilities (type I, II, and power) are Spanos [7]:

(i) *hypothetical* and *unobservable* in principle since they revolve around $\boldsymbol{\theta}^*$,

(ii) not *conditional* on values of $\boldsymbol{\theta} \in \Theta$ since 'presuming $\boldsymbol{\theta} = \boldsymbol{\theta}_i$, $i = 0, 1$' constitute neither events nor random variables, and

(iii) assigned to the test procedure T_α to 'calibrate' its *generic* (for any $\mathbf{x}\in\mathbb{R}^n$) *capacity* to detect different *discrepancies* γ from $\mu=\mu_0$ for a prespecified α.

As mentioned above, the cornerstone of N-P testing is the in-built trade-off between the type I and II error probabilities, which Neyman and Pearson [23] addressed by pre-specifying α at a low value and maximizing $\mathcal{P}(\mu_1)$, $\forall\mu_1=\mu_0+\gamma_1\in\Theta_1$, $\gamma_1\pm 0$, seeking an optimal test; see Lehmann and Romano [21]. The primary role of the testing error probabilities is to operationalize the notions of 'statistically significant/insignificant' in terms of statistical approximations relating to $\boldsymbol{\theta}^*$, and framed in terms of the sampling distribution of a test statistic $\tau(\mathbf{X})$.

This relates directly to the replication crisis since for a misspecified $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ one cannot keep track of the relevant error probabilities to be able to adjust them for p-hacking, data-dredging, multiple testing and cherry-picking, in light of the fact that the actual error probabilities will be different from the nominal ones; see Spanos and McGuirk [26].

2.4. Statistical Inference 'Results' vs. 'Evidence' for or against Inferential Claims

Statistical results, such as a point estimate, say $\widehat{\boldsymbol{\theta}}(\mathbf{x}_0)$, an observed $(1-\alpha)$ CI, say $[L(\mathbf{x}_0),U(\mathbf{x}_0)]$, an effect size, a p-value and the accept/reject H_0 results, are not replicable in principle, in the sense that akin data do not often yield very similar numbers since they are unduly data-specific when contrasted with broader inferential claims relating to inductive generalizations stemming from such results. In particular, the accept/reject H_0 results, (i) are unduly data-specific, (ii) are too coarse to provide informative enough evidence relating to $\boldsymbol{\theta}^*$, and (iii) depend crucially on the particular statistical context:

$$\text{(i) } \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ \text{(ii) } H_0\text{: }\boldsymbol{\theta}\in\Theta_0 \text{ vs. } H_1\text{: }\boldsymbol{\theta}\in\Theta_1,\ \text{(iii) } T_\alpha:=\{d(\mathbf{X}),C_1(\alpha)\},\ \text{(iv) data } \mathbf{x}_0, \tag{15}$$

which includes the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ as well as the sample size n.

Example 1 (continued). It is often erroneously presumed that the optimality of the point estimators, $\widehat{\mu}(\mathbf{X})=\overline{X}_n=\frac{1}{n}\sum_{t=1}^n X_t$, $s^2(\mathbf{X})=\frac{1}{(n-1)}\sum_{t=1}^n(X_t-\overline{X}_n)^2$, can justify the following inferential claims for the particular data $\mathbf{x}_0$ when n is large enough.

(a) The point estimates $\widehat{\mu}(\mathbf{x}_0)=\overline{x}_n$ and $s^2(\mathbf{x}_0)=s_n^2$, based on data $\mathbf{x}_0$, 'approximate closely' ($\simeq$) the true parameter values μ^* and σ_*^2, i.e.,

$$\widehat{\mu}(\mathbf{x}_0)\simeq\mu^*,\ \text{and } s^2(\mathbf{x}_0)\simeq\sigma_*^2,\ \text{for } n \text{ large enough.} \tag{16}$$

Invoking limit theorems, such as strong consistency, will not alleviate the problem since, as argued by Le Cam [27], p. xiv: "... limit theorems 'as n tends to infinity' are logically devoid of content about what happens at any particular n." The inferential claims in (16) are unwarranted since $\widehat{\mu}(\mathbf{x}_0)$ and $s^2(\mathbf{x}_0)$ ignore the relevant uncertainty associated with their representing a single point, $\mathbf{X}=\mathbf{x}_0$, from the relevant sampling distributions: $f(\widehat{\mu}(\mathbf{x});\boldsymbol{\theta})$, $f(s^2(\mathbf{x});\boldsymbol{\theta})$, $\forall\mathbf{x}\in\mathbb{R}^n$; see Spanos [7].

(b) The inferential claim associated with an $(1-\alpha)$ optimal CI for μ in (7) relates to CI$(\mathbf{X};\mu)$ overlaying μ^* with probability $(1-\alpha)$, but its optimality does not justify the claim that the observed CI:

$$CI(\mathbf{x}_0)=[\overline{x}_n-c_{\frac{\alpha}{2}}(s(\mathbf{x}_0)/\sqrt{n}),\ \overline{x}_n+c_{\frac{\alpha}{2}}(s(\mathbf{x}_0)], \tag{17}$$

overlays μ^* with probability $(1-\alpha)$. As argued by Neyman [28]: "... valid probability statements about random variables usually cease to be valid if the random variables are replaced by their particular values." (p. 288). In terms of the underlying factual reasoning, post-data $\mathbf{x}_0$ has been revealed but it is unknowable whether μ^* is within or outside (17); see Spanos [29]. Indeed, one can make a case that the widely held impression that an effect size ([30]) provides more reliable information about the 'scientific effect' than p-values and observed CIs stems from the unwarranted inferential claim in (16), i.e., an optimal estimator $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ of $\boldsymbol{\theta}$ justifies the inferential claim $\widehat{\boldsymbol{\theta}}(\mathbf{x}_0)\simeq\boldsymbol{\theta}^*$ for n large enough.

(c) The N-P testing 'accept H_0' with a large p-value, and rejecting H_0 with a small p-value, do not entail evidence for H_0 and H_1, respectively, since such evidential interpretations are fallacious; see Mayo and Spanos [14].

3. Post-Data Severity Evaluation of Testing Results

3.1. Accept/Reject H_0 Results vs. Evidence for or against Inferential Claims

Bridging the gap between the binary accept/reject H_0 results and learning from data $\mathbf{x}_0$ about $\boldsymbol{\theta}^*$ using statistical approximations framed in terms of the sampling distribution of a test statistic $\tau(\mathbf{X})$, has been confounding frequentist testing. Mayo and Spanos [31] proposed the post-data severity (SEV) evaluation of the accept/reject H_0 results as a way to convert them into evidence for germane inferential claims. The SEV differs from other attempts to address this issue in so far as:

(i) The SEV evaluation constitutes a principled argument framed in terms of a germane inferential claim relating to $\boldsymbol{\theta}^*$ (learning from data $\mathbf{x}_0$).

(ii) The SEV evaluation is guided by the sign and magnitude of the observed test statistic, $\tau(\mathbf{x}_0)$, and not by the prespecified significance level α; see Spanos [32].

(iii) The SEV evaluation accounts fully for the relevant statistical context in (15).

(iv) Its germane inferential claim, in the form of the discrepancy from the null value, is warranted with high probability with $\mathbf{x}_0$ and T_α when all the different ways it can be false have been adequately probed and forfended (Mayo [33]).

The most crucial way to forfend a false accept/reject H_0 result is to ensure that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is statistically adequate for data $\mathbf{x}_0$, before any inferences are drawn. This is because the discrepancies induced by invalid probabilistic assumptions will render impossible the task of controlling (keeping track of) the relevant error probabilities in terms of which N-P tests are framed. Hence, for the discussion that follows it is assumed that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is statistically adequate for the particular data $\mathbf{x}_0$.

Example 2. Consider the simple Bernoulli (Ber) model:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}): \quad X_k \backsim \mathsf{BerIID}(\boldsymbol{\theta}, \boldsymbol{\theta}(1-\boldsymbol{\theta})), \;\; x_k = 0,1, \; 0 < \boldsymbol{\theta} < 1, \; k \in \mathbb{N}, \tag{18}$$

where $\boldsymbol{\theta} = E(X_k) = \mathbb{P}(X_k = 1)$, $\mathbb{P}(X_k = 0) = (1-\boldsymbol{\theta})$. Let the hypotheses of interest be:

$$H_0\text{: } \boldsymbol{\theta} \leq \boldsymbol{\theta}_0 \text{ vs. } H_1\text{: } \boldsymbol{\theta} > \boldsymbol{\theta}_0, \; \boldsymbol{\theta}_0 = 0.5, \tag{19}$$

in the context of (18). It can be shown that the t-type test:

$$T_\alpha^{>} := \{d(\mathbf{X}) = \tfrac{\sqrt{n}(\overline{X}_n - \boldsymbol{\theta}_0)}{\sqrt{\boldsymbol{\theta}_0(1-\boldsymbol{\theta}_0)}}, \; \mathcal{C}_1(\alpha) = \{\mathbf{x}\text{: } d(\mathbf{x}) > c_\alpha\}, \tag{20}$$

where $\overline{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$, is Uniformly Most Powerful (UMP); see Lehmann and Romano [21]. The sampling distribution of $d(\mathbf{X})$, evaluated under H_0 (hypothetical), is:

$$d(\mathbf{X}) = \tfrac{\sqrt{n}(\overline{X}_n - \boldsymbol{\theta}_0)}{\sqrt{\boldsymbol{\theta}_0(1-\boldsymbol{\theta}_0)}} \overset{\boldsymbol{\theta}=\boldsymbol{\theta}_0}{\backsim} \mathsf{Bin}(0,1;n) \simeq \mathsf{N}(0,1). \tag{21}$$

For $n \geq 40$ the 'standardized' Binomial distribution, $\mathsf{Bin}(0,1;n)$, can be approximated ($\simeq$) by the $\mathsf{N}(0,1)$. The latter can be used to evaluate the type I error probability and the p-value:

$$\alpha = \mathbb{P}(d(\mathbf{X}) > c_\alpha; \; \boldsymbol{\theta} = \boldsymbol{\theta}_0), \quad p(\mathbf{x}_0) = \mathbb{P}(d(\mathbf{X}) > d(\mathbf{x}_0); \; \boldsymbol{\theta} = \boldsymbol{\theta}_0). \tag{22}$$

The sampling distribution of $d(\mathbf{X})$ evaluated under H_1 (hypothetical) is:

$$\begin{gathered} d(\mathbf{X}) = \tfrac{\sqrt{n}(\overline{X}_n - \boldsymbol{\theta}_0)}{\sqrt{\boldsymbol{\theta}_0(1-\boldsymbol{\theta}_0)}} \overset{\boldsymbol{\theta}=\boldsymbol{\theta}_1}{\backsim} \mathsf{Bin}\Big(\delta(\boldsymbol{\theta}_1), \sqrt{V(\boldsymbol{\theta}_1)}; n\Big), \text{ for } \boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \gamma_1, \; \gamma_1 \geq 0, \\ \delta(\boldsymbol{\theta}_1) = \tfrac{\sqrt{n}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)}{\sqrt{\boldsymbol{\theta}_0(1-\boldsymbol{\theta}_0)}}, \; V(\boldsymbol{\theta}_1) = \tfrac{\boldsymbol{\theta}_1(1-\boldsymbol{\theta}_1)}{\boldsymbol{\theta}_0(1-\boldsymbol{\theta}_0)}, \end{gathered} \tag{23}$$

whose tail area probabilities can be approximated using:

$$\big(\sqrt{V(\boldsymbol{\theta}_1)}\big)^{-1}[d(\mathbf{X}) - \delta(\boldsymbol{\theta}_1)] \overset{\boldsymbol{\theta}=\boldsymbol{\theta}_1}{\backsim} \mathsf{Bin}(0,1;n) \simeq \mathsf{N}(0,1). \tag{24}$$

(24) is used to derive the type II error probability and the power of the test $T_\alpha^{>}$ in (20) which increases monotonically with $\sqrt{n}$ and $(\mu_1 - \mu_0)$ and decreases with $V(\boldsymbol{\theta}_1)$.

The post-data severity (SEV) evaluation transforms the 'accept/reject H_0 results' into 'evidence' for or against germane inferential claims framed in terms of $\boldsymbol{\theta}$. The post-data severity evaluation is defined as follows:

A hypothesis H (H_0 or H_1) *passes a severe test* T_α with data $\mathbf{x}_0$ if:

(C-1) $\mathbf{x}_0$ accords with H, and

(C-2) with very high probability, test T_α would have produced a result that 'accords less well' with H than $\mathbf{x}_0$ does, if H were false; see Mayo and Spanos [31,34].

Example 2 (continued). Consider data $\mathbf{x}_0$ referring to newborns during 1995 in Cyprus, 5152 boys $(X=1)$ and 4717 girls $(X=0)$. In this case, there is no reason to question the validity of the IID probabilistic assumptions since nature ensures their validity when such data are collected over a sufficiently long period of time in a particular locality. Applying the optimal test $T^{>}_{\alpha}$ in (20) with $\alpha=0.01$ (large n) yields:

$$d(\mathbf{x}_0)=\frac{\sqrt{9869}(\frac{5152}{9869}-0.5)}{\sqrt{0.5(1-0.5)}}=4.379, \text{ indicating 'reject } H_0\text{' with } p_{>}(\mathbf{x}_0)=0.000006.$$

Broadly speaking, this result indicates that the 'true' value $\boldsymbol{\theta}^*$ of $\boldsymbol{\theta}$ lies within the interval $(0.5,1)$ which is too coarse to engender any learning about $\boldsymbol{\theta}^*$.

The post-data severity outputs a germane evidential claim that revolves around a discrepancy $\gamma^{\ddagger}$ warranted by test $T^{>}_{\alpha}$ and data $\mathbf{x}_0$ with high probability. In contrast to pre-data testing error probabilities (type I, II, and power), severity is a *post-data error probability* that uses additional information in the form of the sign and magnitude of $d(\mathbf{x}_0)$, but shares with the former the underlying *hypothetical reasoning*: presuming that $\boldsymbol{\theta}=\boldsymbol{\theta}_1,\ \forall\boldsymbol{\theta}_1\in\Theta_1$.

Given that $d(\mathbf{x}_0)=4.379>c_\alpha$,

[C-1] $\mathbf{x}_0$ accords with H_1 and since $d(\mathbf{x}_0)>0$, the relevant inferential claim takes the form $\boldsymbol{\theta}>\boldsymbol{\theta}_1=\boldsymbol{\theta}_0+\gamma_1,\ \gamma_1\geq 0$.

[C-2] revolves around the event: "outcomes $\mathbf{x}$ that accord less well with $\boldsymbol{\theta}>\boldsymbol{\theta}_1$ than $\mathbf{x}_0$ does", i.e., event $\{\mathbf{x}\colon d(\mathbf{x})\leq d(\mathbf{x}_0)\},\ \forall\mathbf{x}\in\{0,1\}^n$, and its probability:

$$SEV(T^{>}_{\alpha};\boldsymbol{\theta}>\boldsymbol{\theta}_1)=\mathbb{P}(d(\mathbf{X})\leq d(\mathbf{x}_0);\ \boldsymbol{\theta}=\boldsymbol{\theta}_1),\ \forall\boldsymbol{\theta}_1\in\Theta_1=(0.5,1), \tag{25}$$

stemming from (23). This severity curve $\forall\boldsymbol{\theta}_1\in\Theta_1=(0.5,0.53)$ is depicted in Figure 1.

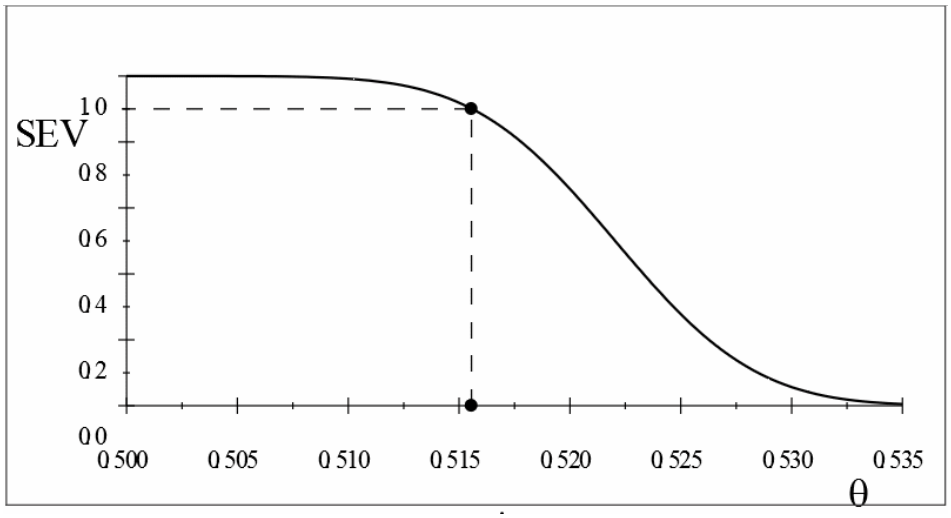

Figure 1. Severity curve for test $T^{>}_{\alpha}$ and data $\mathbf{x}_0$.

In the case of 'reject H_0' the objective is to evaluate the largest $\boldsymbol{\theta}_1=\boldsymbol{\theta}_0+\gamma_1$ such that any $\boldsymbol{\theta}$ less than $\boldsymbol{\theta}_1$ would very probably, at least 0.9, have resulted in a smaller observed difference, warranted by test $T^{>}_{\alpha}$ and data $\mathbf{x}_0$:

$$\max_{\boldsymbol{\theta}_1\in(0.5,1)} SEV(T^{>}_{\alpha};\mathbf{x}_0;\boldsymbol{\theta}>\boldsymbol{\theta}_1)=0.9\rightarrow\gamma^{\ddagger}_1\leq 0.01557$$

Like all error probabilities, the SEV evaluation is always attached to the procedure itself as it pertains to the inferential claim $\boldsymbol{\theta} > \boldsymbol{\theta}_1$. The inferential claim $\gamma_1^{\ddagger} \leq 0.01557$ warranted with probability .9, however, can be 'informally' interpreted as evidence for a germane neighborhood of $\boldsymbol{\theta}^*$, $\boldsymbol{\theta}_1 = 0.51557 \pm \varepsilon$, for some $\varepsilon \geq 0$, arising from the SEV evaluation narrowing down the coarse $\boldsymbol{\theta}^* \in (0.5, 1)$ associated with the 'reject H_0'. This narrowing down of the potential neighborhood of $\boldsymbol{\theta}^*$ enhances learning from data.

It is also important to emphasize that the SEV evaluation of the inferential claim $\boldsymbol{\theta} > \boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \gamma_1$, $\gamma_1 \geq 0$ with discrepancy $\gamma_1 = 0.0223$, based on $\overline{x}_n = 0.5223$, gives rise to $SEV(T_{\alpha}^{>}; \mathbf{x}_0; \boldsymbol{\theta} > 0.5223) = 0.5$, which implies that there is no evidence for $\boldsymbol{\theta}_1 \leq 0.5223$. More generally, the SEV evaluation will invariably provide evidence *against* the inferential claim $\boldsymbol{\theta}_1 \leq \overline{x}_n$. Hence, the importance of distinguishing between 'statistical results', such as $\overline{x}_n = 0.5223$, and 'evidence' for or against inferential claims relating to $\overline{x}_n$.

What is the *nature of evidence* the post-data severity (SEV) gives rise to? Since the objective of inference is to learn from data about phenomena of interest via learning about $\boldsymbol{\theta}^*$, the evidence from the SEV comes in the form of an inferential claim that revolves around the discrepancy γ_1 warranted by the particular data and test with high enough probability, pinpointing the neighborhood of $\boldsymbol{\theta}^*$ as closely as possible. In the above case, the warranted discrepancy is $\gamma_1^{\ddagger} \leq 0.01557$, or equivalently, $\boldsymbol{\theta}^* \leq 0.51557$, with probability 0.9. Although all probabilities are assigned to the inference procedure itself as it relates to the inferential claim $\boldsymbol{\theta} > \boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \gamma_1$, $\gamma_1 > 0$, the SEV evaluation can be viewed intuitively as narrowing the coarse reject H_0 result entailing $\boldsymbol{\theta}^* \in (0.5, 1)$ down to $\boldsymbol{\theta}^* \in (0.512, 0.5156)$.

The most important attributes of the SEV evaluation are:

[i] It is a *post-data* error probability stemming from *hypothetical* reasoning that takes into account the statistical context in (15) and is guided by $d(\mathbf{x}_0) \gtrless 0$.

[ii] Its evaluation is invariably based on a discrepancy $(\sqrt{V(\boldsymbol{\theta}_1)})^{-1}[d(\mathbf{x}_0) - \delta(\boldsymbol{\theta}_1)]$ relating to the noncentral distribution in (24), where $d(\mathbf{x}_0)$ and $\delta(\boldsymbol{\theta}_1)$ use the same n to output the warranted discrepancy $\gamma_1 = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) \geq 0$.

3.2. The Robustness of the Post-Data Severity Evaluation

To exemplify the robustness of the SEV evaluation with respect to changing $\boldsymbol{\theta}_0$, consider replacing $\boldsymbol{\theta}_0 = 0.5$ in (19) with the Nicolas Bernoulli value $\boldsymbol{\theta}_0 = (18/35) \simeq 0.5143$.

Example 2 (continued). Applying the same test $T_{\alpha}^{>}$ yields

$$d_B(\mathbf{x}_0) = \sqrt{9869}\left(\tfrac{5152}{9869} - 0.5143\right) / \sqrt{0.5143(1 - 0.5143)} = 1.541, \text{ indicating 'accept } H_0\text{'},$$

with $p(\mathbf{x}_0) = 0.062$. Is this 'accept H_0' result at odds with the previous 'reject H_0' result? In light of $d_B(\mathbf{x}_0) > 0$, the relevant inferential claim is identical to the case with $\boldsymbol{\theta}_0 = 0.5$, $\boldsymbol{\theta} > \boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \gamma_1$, $\gamma_1 > 0$, as it relates to the event $\{\mathbf{x}\colon\ d(\mathbf{x}) \leq d(\mathbf{x}_0)\}$, $\forall \mathbf{x} \in \{0,1\}^n$. Thus, the severity curve $SEV_B(T_{\alpha}^{>}; \boldsymbol{\theta} > \boldsymbol{\theta}_1)$ is identical $\forall \boldsymbol{\theta}_1 \in (0.5, 1)$ to one in Figure 1, but now defined with respect to $\boldsymbol{\theta}_0 = \frac{18}{35} = 0.5143$, i.e.,

$$\min_{\boldsymbol{\theta}_1 \in (0.5,1)} SEV(T_{\alpha}^{>}; \mathbf{x}_0; \boldsymbol{\theta} > \boldsymbol{\theta}_1) = 0.9 \rightarrow \gamma_1^{\ddagger} \leq 0.00127,$$

as shown in Table 1, a feature of a sound account of statistical evidence.

Table 1. SEV for 'reject $\boldsymbol{\theta}_0 = 0.5$' and 'accept $\boldsymbol{\theta}_0 = (18/35)$' with $(T_{\alpha}^{>}; \mathbf{x}_0)$.

$\boldsymbol{\theta}_0 = 0.5$: $\gamma_1 =$	0.01	0.014	0.015	0.01557	0.016	0.018	0.022	0.025	0.03
$\boldsymbol{\theta}_0 = \frac{18}{35}$: $\gamma_1 =$	−0.0043	−0.0003	0.0007	0.00128	0.0017	0.0037	0.0077	0.0107	0.016
$\boldsymbol{\theta}_1 =$	0.51	0.514	0.515	0.51557	0.516	0.518	0.522	0.525	0.53
$\mathsf{Sev}(\boldsymbol{\theta} > \boldsymbol{\theta}_1) =$	0.991	0.944	0.918	0.900	0.884	0.787	0.500	0.276	0.056

3.3. Post-Data Severity and the Replicability of Evidence

The post-data severity evaluation of the 'accept/reject H_0' results can also provide a more robust way to evaluate the replicability of empirical evidence based on comparing the discrepancies γ from the null value, $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ warranted with similar data with high enough severity. To illustrate this, consider the following example that uses similar data $\mathbf{x}_1$ from a different country more than three centuries apart.

Example 3. Data $\mathbf{x}_1$ refer to newborns during 1668 in London (England), 6073 boys, 5560 girls, $n = 11{,}633$; see Arbuthnot [35]. The optimal test in (20) yields $d(\mathbf{x}_1) = 4.756$, with $p(\mathbf{x}_1) = 0.0000001$, rejecting H_0: $\boldsymbol{\theta} \leq 0.5$. This result is almost identical to the result with data from Cyprus for 1995, but the question is whether the latter can be viewed as a successful replication with trustworthy evidence.

Evaluating the post-data severity with the same probability $SEV(T_\alpha^>; \mathbf{x}_1; \boldsymbol{\theta} > \boldsymbol{\theta}_1) = 0.9$, the warranted discrepancy from $\boldsymbol{\theta}_0 = 0.5$ by test $T_\alpha^>$ and data $\mathbf{x}_1$ is $\gamma_1^\ddagger \leq 0.01516$, which is very close to $\gamma_0^\ddagger \leq 0.01557$; a fourth decimal difference.

As shown in Figure 2, the severity curves for data $\mathbf{x}_0$ and $\mathbf{x}_1$ almost coincide. This suggests that, for a statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, the post-data severity could provide a more robust measure of replicability associated with trustworthy evidence, than point estimates, effect sizes, observed CIs, or p-values. Indeed, it can be argued that the warranted discrepancy γ with high probability provides a much more robust testing-based effect size for the scientific effect; see Spanos [7].

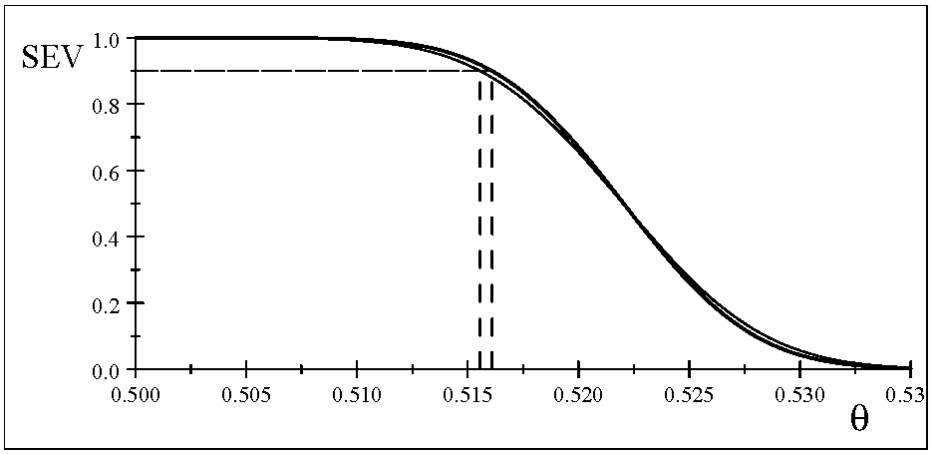

Figure 2. Severity curves for data $\mathbf{x}_i$, $i = 0, 1$.

3.4. Statistical vs. Substantive Significance

The post-data severity evaluation can also address this problem by relating the discrepancy $\gamma^\ddagger$ from $\boldsymbol{\theta}_0$ ($\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 \pm \gamma^\ddagger$) warranted by test T_α and data $\mathbf{x}_0$ with high probability, to the substantively (human biology) determined value $\varphi^\blacklozenge$. For that, one needs to supplement the statistical information in data $\mathbf{x}_0$ with reliable substantive subject matter information to evaluate the 'scientific effect'.

Example 2 (continued). Human biology informs us that the *substantive value* for the ratio of boys to all newborns is $\varphi^\blacklozenge \simeq 0.5122$; see Hardy [36]. Comparing $\varphi^\blacklozenge$ with the severity-based warranted discrepancy, $\gamma_1^\ddagger \leq 0.01557$ ($\boldsymbol{\theta}_1 \leq 0.51557$) indicates that the statistically determined $\gamma_1^\ddagger$ entails the substantive significance, since $\varphi^\blacklozenge \simeq 0.5122 < 0.51557$. Hence, the statistical value also implies substantive significance.

3.5. Post-Data Severity and the Large n Problem

To alleviate the large n problem, some textbooks in statistics advise practitioners to keep reducing α as n increases beyond $n > 200$; see Lehmann and Romano [21]. A less arbitrary method is to agree that, say, $\alpha = 0.05$ for $n = 100$, seems a reasonable standard, and then modify Good's [37] standardization of the p-values into thresholds $\alpha(n)$ using

the formula: $\alpha(n)$=$0.05/\sqrt{n/100}$, $n \geq 50$, as shown in Table 2. This standardization is also ad hoc since (i) it depends on an agreed standard, (ii) the simple scaling, but (iii) for $n \geq 1 \times 10^8$ the implied thresholds are tiny.

Table 2. Standardized $\alpha(n)$ relative to $(\alpha = 0.05, n = 100)$.

n	50	100	200	500	1000	5000	10,000	1×10^5	1×10^6	1×10^7	1×10^8
$\alpha(n)$	0.071	0.05	0.035	0.022	0.016	0.007	0.005	0.0016	0.0005	0.00016	0.00005

The post-data severity evaluation of the accept/reject H_0 results can be used to shed light on the link between α and n. Let us return to example 1 and assume that $n = 1000$ is large enough (a) to establish the statistical adequacy of the simple Bernoulli model in (18), (b) to avoid the small n problem, and (c) to provide a reliable enough estimate $\widehat{\boldsymbol{\theta}}(\mathbf{x}_0) = \overline{x}_n$ for $\boldsymbol{\theta}$. There are two possible scenarios one can consider.

Scenario 1 assumes that all different values of $n \geq 1000$ yield the same observed value of the test statistic $d(\mathbf{x}_0)$. This scenario has been explored in the context of the SEV evaluation by Mayo and Spanos [31].

Scenario 2 assumes that as n increases beyond $n = 1000$ the changes in the estimate $\widehat{\boldsymbol{\theta}}(\mathbf{x}_0) = \overline{x}_n$ will be 'relatively small' to render $[(\widehat{\boldsymbol{\theta}}(\mathbf{x}_0)-\boldsymbol{\theta}_0)/\sqrt{\boldsymbol{\theta}_0(1-\boldsymbol{\theta}_0)}]$ approximately constant when the IID assumptions are valid for $\mathbf{x}_n$.

To explore scenario 2, let us return to example 2, related to testing H_0: $\boldsymbol{\theta} \leq \boldsymbol{\theta}_0$ vs. H_1: $\boldsymbol{\theta} > \boldsymbol{\theta}_0$, $\boldsymbol{\theta}_0 = 0.5$, in the context of the simple Bernoulli model in (18), using data on newborns in Cyprus during 1995, 5152 (male) and 4717 (female). Particular values for n and $p_n(\mathbf{x}_0)$ are given in Table 3, indicating clearly that for $n > 20{,}000$ the p-value goes to zero very fast, and thus, the thresholds $\alpha(n)$ needed to counter the increase in n will be tiny.

Table 3. The p-value with increasing n ($\overline{x}_n$ constant).

n	2000	3000	3256	5000	10,000	20,000	50,000	100,000
$d_n(\mathbf{x}_0)$	1.971	2.414	2.515	3.117	3.652	6.234	9.856	13.939
$p_n(\mathbf{x}_0)$	0.0341	0.0128	0.0099	0.0019	0.000023	4.1×10^{-9}	3.9×10^{-20}	0.000000...

Figure 3 shows the p-value for different values of n, indicating that for $n \geq 3256$ the null hypothesis will be rejected, but for smaller n will be accepted. This indicates clearly that for a given α the accept/reject results are highly vulnerable to abuse stemming from manipulating the sample size to obtain the desired result. This abuse can be addressed by the SEV evaluation for such results.

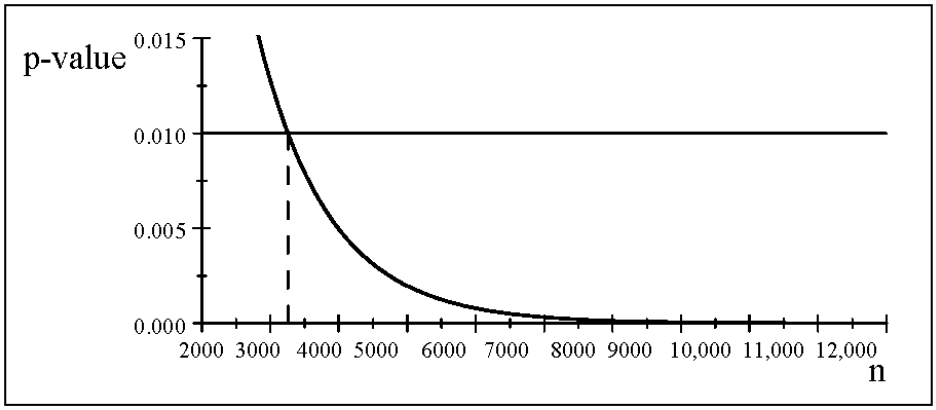

Figure 3. The p-value for $\alpha = 0.01$ and different n.

The other side of the large n coin relates to the increase in power for a given $\alpha = 0.01$ as n increases. Figure 4 shows that the power curve becomes steeper and steeper as n increases, reflecting the detection of smaller and smaller discrepancies from $\boldsymbol{\theta}_0 = 0.5$ with

probability 0.85. This stems from the fact that the power of the test in (20) is evaluated using the difference between a fixed c_α and $\delta(\boldsymbol{\theta}_1)$, which increases with $\sqrt{n}$, based on (24).

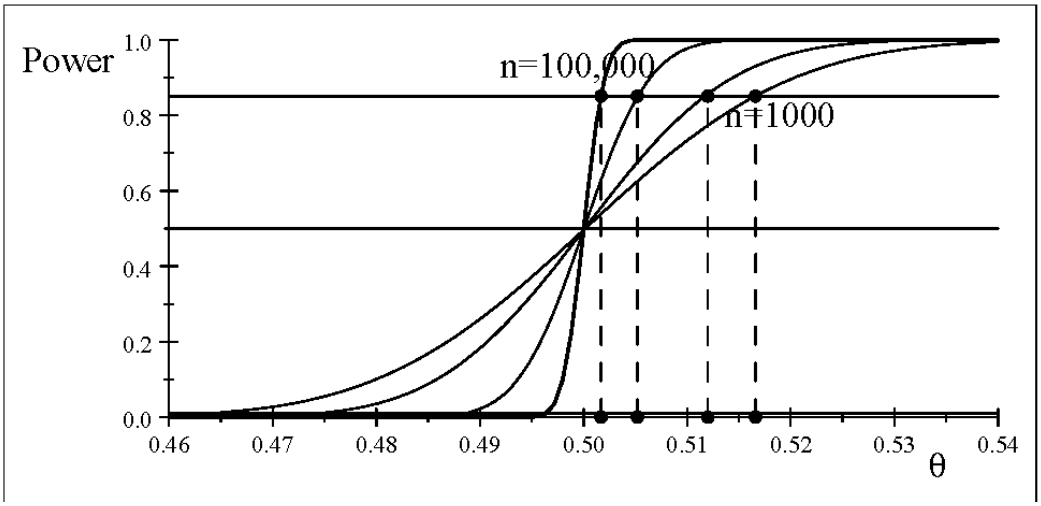

Figure 4. The power curves for $\alpha = 0.01$ and different n.

The above numerical examples relating to test (20) under scenario 2 suggest that rules of thumb relating to decreasing α as n increases, in an attempt to meliorate potentially spurious results, can be useful in tempering the trade-off between the type I and II error probabilities. They do not, however, address the large n problem, since they are ad hoc and their standardized thresholds decrease to zero beyond $n =$ 100,000.

The post-data severity evaluation (SEV) of the accept/reject H_0 results constitutes a principled argument that addresses the large n problem by ensuring that the same n is used in both terms $d_n(\mathbf{x}_n)$ and $\delta(\boldsymbol{\theta}_1)$ when the warranted discrepancy γ_1 is evaluated based on

$$(\sqrt{V(\boldsymbol{\theta}_1)})^{-1}[d_n(\mathbf{x}_n)-\delta(\boldsymbol{\theta}_1)],\ \forall \boldsymbol{\theta}_1 \in (.5,1), \tag{26}$$

using (24), in contrast to the power, which replaces $d_n(\mathbf{x}_n)$ with c_α in (26). To illustrate this argument, Table 4 reports the SEV evaluations using scenario 2, where $(0.52204-0.5)/\sqrt{0.5(1-0.5)}$ is retained and n is allowed to vary below and above the original $n = 9869$ for values that give rise to reject H_0. The numbers indicate that for a given $SEV(\boldsymbol{\theta} > \boldsymbol{\theta}_1; n) = 0.9$, as n increases, the warranted discrepancy $\gamma_n^{\ddagger}$ increases, or equivalently the $SEV(\boldsymbol{\theta} > 0.51557; n)$ for $\gamma_1^{\ddagger} = 0.01557$ increases with n.

Table 4. SEV with changing n ($\overline{x}_n = 0.52204$ is held constant).

n	3256	5000	10,000	12,000	20,000	50,000	100,000
$d_n(\mathbf{x}_n)$	2.515	3.117	4.441	4.828	6.234	9.856	13.938
$SEV(\boldsymbol{\theta} > \boldsymbol{\theta}_1; n) = 0.9$: $\boldsymbol{\theta} > \boldsymbol{\theta}_1$	0.5109	0.513	0.5156	0.5162	0.5175	0.5192	0.5200
$SEV(\boldsymbol{\theta} > 0.51557; n) =$	0.770	0.820	0.903	0.922	0.967	0.998	0.999

The severity curves $SEV(T_\alpha^{>}; \boldsymbol{\theta} > \boldsymbol{\theta}_1; n)$ for the different n in Table 4 are shown in Figure 5, with the original $n = 9869$ and $n = 1 \times 10^5$ in heavy lines. The curves confirm the results in Table 4, and provide additional details, indicating the need to increase the benchmark of how high the probability $SEV(\boldsymbol{\theta} > \boldsymbol{\theta}_1; n)$ should be to counter-balance the increase in n; hence the use of 0.9 for example 2 ($n = 9869$) and 3 ($n =$ 11,633), and 0.7 ($n = 20$) for example 4.

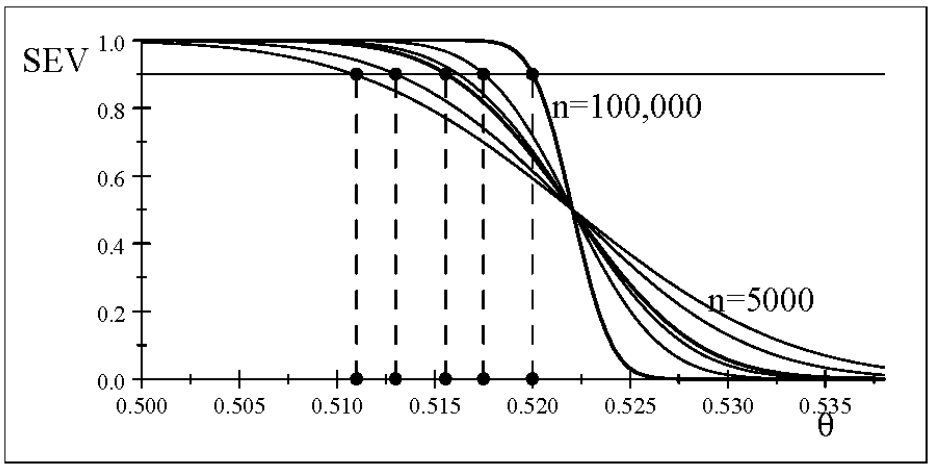

Figure 5. Severity curves for test $T_{\alpha}^{>}$ and data $\mathbf{x}_0$.

4. Post-Data Severity and the Remedies Proposed by the Replication Literature

4.1. The Alleged Arbitrariness in Framing H_0 and H_1

The issue of the framing of H_0 and H_1 in N-P testing has been widely misconstrued by the replication crisis literature, questioning its coherence and blaming the accept/reject H_0 results and the p-value, as providing misleading evidence and often non-replicable results. Contrary to that, in addition to Neyman and Pearson [23] warning against misinterpreting 'accept H_0' as evidence for H_0 and 'reject H_0' as evidence for H_1, they also put forward two stipulations (1–2) (Section 2.3) relating to the framing of H_0 and H_1, whose objective is to ensure the effectiveness of N-P testing and the informativeness of the ensuing results. The following example illustrates how a 'nominally' optimal test can be transformed into a wild goose chase when (1–2) are ignored.

Example 4. In an attempt to demonstrate the ineptitude of the p-value as it compares to Bayesian testing, Berger [38] p. 4, put forward an example of testing:

$$H_0\text{: } \boldsymbol{\theta} = 0.5 \text{ vs. } H_1\text{: } \boldsymbol{\theta} > 0.5, \tag{27}$$

in the context of the simple Bernoulli model in (18), with $\alpha = 0.05$, $c_{\alpha} = 1.645$, $n = 20$, and $\widehat{\boldsymbol{\theta}}(\mathbf{x}_0) = \overline{x}_n = 0.2$. Applying the UMP test $T_{\alpha}^{>}$ yields: $d(\mathbf{x}_0) = -2.683$, with a p-value $p_{>}(\mathbf{x}_0) = 0.996$, indicating 'accept H_0'. Berger avails this "ridiculous" result to make a case for Bayesian statistics: "A sensible Bayesian analysis suggests that the evidence indeed favors H_0, but only by a factor of roughly 5 to 1." (p. 4). Viewing this result in the context of the N-P stipulations (1–2) (Section 2.3) reveals that the real source of this absurd result is likely to be the framing in (27) since it flouts both stipulations. Assuming the statistical adequacy of the invoked $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in (18), $\widehat{\boldsymbol{\theta}}(\mathbf{x}_0) = 0.2$ gives a broad indication of the potential neighborhood of $\boldsymbol{\theta}^*$. In contrast, the framing in (27) ensures that the test $T_{\alpha}^{>}$ has no power to detect any discrepancies around $\boldsymbol{\theta}_1 = \widehat{\boldsymbol{\theta}}(\mathbf{x}_0) \pm \epsilon$, $\epsilon > 0$, since the implicit power is $\mathcal{P}(\boldsymbol{\theta}_1 = 0.2) = 0.00000003$, confirming that $p_{>}(\mathbf{x}_0) = 0.996$ is the result of ignoring stipulations (1–2).

How can one avoid such abuses of N-P testing and secure trustworthy evidence? When there is no reliable information about the potential neighborhood around $\boldsymbol{\theta}^*$, one should always use the two-sided framing

$$H_0\text{: } \boldsymbol{\theta} = 0.5 \text{ vs. } H_1\text{: } \boldsymbol{\theta} \neq 0.5,$$

that accords with the N-P stipulations (1–2). Applying the UMP unbiased test

$$T_{\alpha}^{\neq} := [d(\mathbf{X}) = \tfrac{\sqrt{n}(\overline{X}_n - \boldsymbol{\theta}_0)}{\sqrt{\boldsymbol{\theta}_0(1-\boldsymbol{\theta}_0)}},\ C_1(\alpha) = \{\mathbf{x}\text{: } |d(\mathbf{x})| > c_{\frac{\alpha}{2}}\}]$$

(Lehmann and Romano [21]), reject H_0 with $p_{\neq}(\mathbf{x}_0)=0.0073$, with $\mathcal{P}_{\neq}(\boldsymbol{\theta}_1=0.2\pm\epsilon)\geq 0.94$ for $\epsilon=0.1$. An even stronger rejection, with $p_{<}(\mathbf{x}_0)=0.004$, will result by testing the hypotheses

$$H_0\text{: } \boldsymbol{\theta}>0.5 \text{ vs. } H_1\text{: } \boldsymbol{\theta}\leq 0.5, \tag{28}$$

using the UMP test $T_{\alpha}^{<}:=\{d(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\boldsymbol{\theta}_0)}{\sqrt{\boldsymbol{\theta}_0(1-\boldsymbol{\theta}_0)}}\}$, $C_1(\alpha)=\{\mathbf{x}\text{: } d(\mathbf{x})<c_{\alpha}\}$.

Applying the SEV evaluation to the result 'reject H_0', with $d(\mathbf{x}_0)=-2.683$ based on (28), it is clear that the sign and magnitude of $d(\mathbf{x}_0)$ indicate that the relevant inferential claim is $\boldsymbol{\theta}>\boldsymbol{\theta}_1=0.5-\gamma_1$, $\gamma_1>0$. (C-2) implies that the relevant event is $\{\mathbf{x}\text{: } d(\mathbf{x})>d(\mathbf{x}_0)\}$, $\forall\mathbf{x}\in\{0,1\}^n$, to infer the warranted discrepancy with high probability, say, 0.7 ($n=20$):

$$\max_{\boldsymbol{\theta}_1\in(0,0.5)} SEV(T_{\alpha}^{>};\boldsymbol{\theta}>\boldsymbol{\theta}_1)=\mathbb{P}(d(\mathbf{X})<d(\mathbf{x}_0);\ \boldsymbol{\theta}=\boldsymbol{\theta}_1)=0.7,\ \forall\boldsymbol{\theta}_1\in\Theta_1=(0,0.5),$$

which yields $\gamma^{\ddagger}\leq 0.343$ ($\boldsymbol{\theta}\leq 0.157$), with the severity curve in Figure 6.

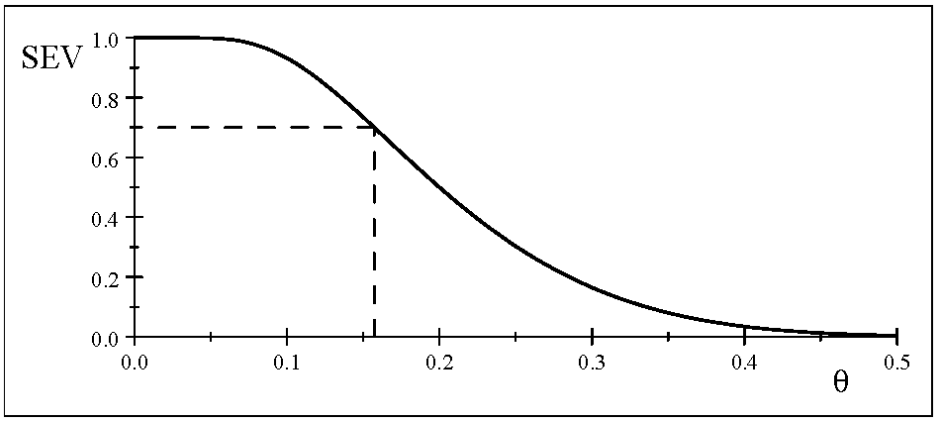

Figure 6. Severity curve for test $T_{\alpha}^{<}$ and data $\widehat{\boldsymbol{\theta}}(\mathbf{x}_0)=0.2$.

Note: the severity curve for the two-sided test would have been identical to that in Figure 6, since $d(\mathbf{x}_0)=-2.683$ determines the direction of departure, irrespective of $\pm c_{\frac{\alpha}{2}}$.

Applying the post-data severity evaluation for discrepancies $\gamma>-0.03$ ($\boldsymbol{\theta}_1>0.47$) gives $SEV(T_{\alpha}^{>};\boldsymbol{\theta}>\boldsymbol{\theta}_1=0.47)\leq 0.008$, which is the strongest possible evidence against $\boldsymbol{\theta}_0=0.5$, repudiating the Bayesian posterior odds of 5 to 1 *for* $\boldsymbol{\theta}_0=0.5$.

A potential counter-argument to the above discussion, claiming that the estimate $\overline{x}_n=0.2$ could be the result of $\mathbf{x}_0$ a *bad draw*, is not a well-grounded argument, since misspecification testing of the invoked $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ would reveal whether $\mathbf{x}_0$ is atypical, i.e., a bad draw from the sample $\mathbf{X}$. Recall that the adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ ensures that $\mathbf{x}_0$ is a 'typical' realization thereof. It is worth mentioning, however, that example 4 with $n=20$ could be vulnerable to the small n problem discussed in Section 2.3.

4.2. The Call for Redefining Statistical Significance

In light of the above discussion on the large n problem, the call by Benjamin et al. [6] "For fields where the threshold for defining statistical significance for new discoveries is P < 0.05, we propose a change to P < 0.005." (p. 6) seems visceral! It brushes aside the inherent trade-off between the type I and II error probabilities and the implied inverse relationship between the sample size n and the appropriate α to avoid the large/small n problems; see Section 2.3. The main argument used by Benjamin et al. [6] is that empirical evidence from large-scale replications indicates that studies with $p(\mathbf{x}_0)<0.005$ are more likely to be replicable than those based on $p(\mathbf{x}_0)<0.05$.

When this claim is viewed in the context of the broader problem of the uninformed and recipe-like implementation of statistical modeling and inference, in conjunction with its many different ways it can contribute to the untrustworthiness of empirical evidence, including (a–c), and the fact that replicability is neither necessary nor sufficient for the

trustworthiness of empirical evidence, the above argument is unpersuasive. The threshold $p(\mathbf{x}_0) < \alpha$ was never meant to be either arbitrary or fixed for all frequentist tests, and the above discussion of the large n problem shows that using $\alpha = 0.05$ for a large n, say $n > 10{,}000$, will often give rise to spurious significance results. Aware of the loss of power when $\alpha = 0.05$ decreases to $\alpha = 0.005$, Benjamin et al. [6] call for increasing n, to ensure a high power of 0.8 at some arbitrary $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. The problem with the proposed remedy is twofold. First, increasing n is often impracticable with observational data, and second, securing high power for arbitrary discrepancies $\gamma_1 = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)$ is not conducive to learning about $\boldsymbol{\theta}^*$.

Another argument for lowering the threshold put forward by Benjamin et al. [6] stems from a misleading comparison between the two-sided p-value for the thresholds 0.05 and 0.005, and the corresponding Bayes factor:

$$(\mathrm{a}) B_{01}(\mathbf{x}_0) = [\pi(\boldsymbol{\theta} \in \Theta_0 | \mathbf{x}_0) / \pi(\boldsymbol{\theta} \in \Theta_1 | \mathbf{x}_0)] / [\pi(\boldsymbol{\theta} \in \Theta_0) / \pi(\boldsymbol{\theta} \in \Theta_1)], \forall \boldsymbol{\theta} \in \Theta,$$

where $\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|\mathbf{x}_0)$ denote the prior and the posterior distributions. This is an impertinent comparison since the Bayesian perspective on evidence, based on $B_{01}(\mathbf{x}_0)$, has a meager connection to the p-value as an indicator of discordance between $\mathbf{x}_0$ and $\boldsymbol{\theta}_0 = 0.5$. This is because the presumed comparability (analogy) between the tail areas of $\tau(\mathbf{X}) \overset{\mu=\mu_0}{\backsim} \mathsf{St}(n-1)$ that varies over $\mathbf{x} \in \mathbb{R}^n_X$, and revolves around $\boldsymbol{\theta}^*$, with the ratio in $B_{01}(\mathbf{x}_0)$ that varies with $\forall \boldsymbol{\theta} \in \Theta$, is ill-thought-out! The uncertainty accounted for by the former has nothing to do with that of the latter, since the posterior distribution accounts for the uncertainty stemming from the prior distribution, weighted by the likelihood function, both of which vary over $\boldsymbol{\theta}$.

4.3. The Severity Perspective on the p-Value, Observed CIs, and Effect Sizes

The *p*-value and the accept/reject H_0 results. The real problem is their binary nature created by the threshold α, which gives rise to counter-intuitive results, such as for $\alpha = 0.05$ one rejects H_0 when $p(\mathbf{x}_0) = 0.49$, and accepts H_0 when $p(\mathbf{x}_0) = 0.51$. The SEV-based evidential account does away with this binary dimension since its inferential claim and the discrepancy γ from $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ warranted with high probability are guided by the sign and magnitude of the observed test statistic $d(\mathbf{x}_0)$. The SEV evaluation uses the statistical context in (15), in conjunction with the statistical approximations framed by the relevant sampling distribution of $d(\mathbf{X})$ to deal with the binary nature of the results, as examples 2–4 illustrate.

Example 2 (continued). Let us return to the data denoting newborns in Cyprus, 5152 boys and 4717 girls during 1995, where the optimal test $T^{>}_{\alpha}$ with $d(\mathbf{x}_0) = 4.379$, indicates 'reject H_0' with $p_{>}(\mathbf{x}_0) = 0.000006$. When $p_{>}(\mathbf{x}_0)$ is viewed from the severity vantage point, it is directly related to evaluating the post-data severity for a zero discrepancy, i.e., $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$ since (Mayo and Spanos, [31]):

$$\mathsf{Sev}(T^{>}_{\alpha}; \mathbf{x}_0; \boldsymbol{\theta} > \boldsymbol{\theta}_0) = \mathbb{P}(d(\mathbf{X}) < d(\mathbf{x}_0); \boldsymbol{\theta} \leq \boldsymbol{\theta}_0) = 1 - \mathbb{P}(d(\mathbf{X}) > d(\mathbf{x}_0); \boldsymbol{\theta} = \boldsymbol{\theta}_0) = 0.99999,$$

which suggests that a small p-value indicates the existence of *some* discrepancy $\gamma \geq 0$, but provides *no information* about its *magnitude* warranted by $\mathbf{x}_0$. The severity evaluation remedies that by outputting the missing magnitude in terms of the discrepancy γ warranted by data $\mathbf{x}_0$ and test $T^{>}_{\alpha}$ with high probability by taking into account the relevant statistical context in (15). The key problem is that the p-value is evaluated using $d(\mathbf{X}) \overset{\boldsymbol{\theta}=\boldsymbol{\theta}_0}{\backsim} \mathsf{N}(0,1)$, and thus, it contains no information relating to different discrepancies from $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, unlike the post-data severity evaluation, since it is based on $(\sqrt{V(\boldsymbol{\theta}_1)})^{-1}[d(\mathbf{X}) - \delta(\boldsymbol{\theta}_1)] \overset{\boldsymbol{\theta}=\boldsymbol{\theta}_0}{\backsim} \mathsf{N}(0,1)$.

Observed CIs and effect sizes. The question that arises is why the claim $\mu^* \in CI(\mathbf{x}_0) = [\overline{x}_n - c_{\frac{\alpha}{2}}(s(\mathbf{x}_0)/\sqrt{n}),\ \overline{x}_n + c_{\frac{\alpha}{2}}(s(\mathbf{x}_0))]$ with probability $(1-\alpha)$ is unwarranted. As argued in Section 2.4 this stems from the fact that *factual* reasoning is baseless post-data, and thus, one cannot assign a probability to $CI(\mathbf{x}_0)$. This calls into question calls by the reformers in the replication crises to replace the p-value with the analogous observed CI because the

latter is (i) less vulnerable to the large n problem, and (ii) more informative than the p-value since it provides a measure of the 'effect size'. Cohen's [39] recommendation is: "routinely report effect sizes in the form of confidence intervals" (p. 1002).

Claim (i) is questionable because a CI is equally as vulnerable to the large n problem as the p-value since the expected length of a consistent CI shrinks to zero as $n \to \infty$. In the case of (7) this takes the form

$E\Big([\overline{X}_n+c_{\frac{\alpha}{2}}(\frac{s}{\sqrt{n}})]-[\overline{X}_n-c_{\frac{\alpha}{2}}(\frac{s}{\sqrt{n}})]\Big) = 2c_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}}) \underset{n\to\infty}{\to} 0$, and thus, as n increases, the width of the observed $CI(\mathbf{x}_0) = [\overline{x}_n-c_{\frac{\alpha}{2}}(\frac{s}{\sqrt{n}}),\ \overline{x}_n+c_{\frac{\alpha}{2}}(\frac{s}{\sqrt{n}})]$ decreases.

This also questions claim (ii), that it provides a reliable measure of the 'effect size', since the larger the n the smaller the observed CI. Worse, the concept of the 'effect size' was introduced partly to address the large n problem using a measure that is free of n: "... the raw size effect as a measure is that its expected value is independent of the size of the sample used to perform the significance test." (Abelson [40], p. 46).

Example 2 (continued). The effect size for $\boldsymbol{\theta}$ is known as Cohen's $g = (\widehat{\boldsymbol{\theta}}(\mathbf{x}_0)-\boldsymbol{\theta}_0)$ (Ellis [30]). When evaluated using the Arbuthnot value $\boldsymbol{\theta}_0 = 0.5$, $g = 0.02204$, which is rather small, and when evaluated using the Bernoulli value $\boldsymbol{\theta}_0 = (18/35)$ it is even smaller, $g = 0.0078$. How do these values provide a better measure of the 'scientific effect'? They do not, since Cohen's g is nothing more than another variant of the unwarranted claim that $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_0) \simeq \boldsymbol{\theta}^*$ for a large enough n; see Spanos [7]. On the other hand, the post-data severity evaluation outputs the discrepancy $\gamma^\ddagger$ warranted by data $\mathbf{x}_0$ and test $T_\alpha^>$ with probability 0.9. This implies that the severity-based effect size is $\boldsymbol{\theta}_1^\ddagger \leq 0.5156$, which takes into account the relevant error probabilities that calibrate the uncertainty relating to the single realization $\mathbf{x}_0$. In addition, the SEV evaluation gives rise to identical severity curves (Figure 1) and the same evidence for both null values $\boldsymbol{\theta}_0 = 0.5$ and $\boldsymbol{\theta}_0 = (18/35)$.

Severity and observed CIs. The great advantage of hypothetical reasoning is that it applies both pre-data and post-data, enabling the SEV to shed light on several foundational problems relating to frequentist inference more broadly. In particular, the severity evaluation of 'reject H_0' relating to the inferential claim $\mu > \mu_1 = \mu_0+\gamma_1,\ \gamma_1 \geq 0$, has a superficial resemblance to $CI_L(\mathbf{x}_0) = (\mu \geq \overline{x}_n-c_\alpha(\frac{\sigma}{\sqrt{n}}))$ in (8)-(b), especially if one were to consider $\mu_1 = \overline{x}_n-c_\alpha(\frac{\sigma}{\sqrt{n}})$ as a relevant discrepancy of interest in the SEV evaluation; see Mayo and Spanos [31]. This resemblance, however, is more apparent than real since:

(i)The relevant sampling distribution for $CI_L(\mathbf{X};\alpha)$ is $\tau(\mathbf{X};\mu) \overset{\boldsymbol{\theta}=\boldsymbol{\theta}^*}{\backsim} \mathrm{St}(n-1)$, and that for the SEV evaluation is $\tau(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n-\mu_0)}{s} \overset{\mu=\mu_1}{\backsim} \mathrm{St}(\delta_1, n-1)$, $\forall \mu > \mu_1 = \mu_0 + \gamma_1,\ \gamma_1 \geq 0$;

(ii) They are derived under two different forms of reasoning, factual and hypothetical, respectively, which are not interchangeable. Indeed, the presence of δ_1 for the SEV evaluation renders the two assignments of probabilities very different. Hence, any attempt to relate the SEV evaluation to the illicit assignment $\mathbb{P}_L(\mu \geq \overline{x}_n-c_\alpha(\frac{\sigma}{\sqrt{n}})) = 1-\alpha$, is ill-thought-out since (a) it will (implicitly) impose the restriction $\delta_1 = 0$, since $\mu_1 = \overline{x}_n-c_\alpha(\frac{\sigma}{\sqrt{n}})$ leaves no room for the discrepancy γ_1, and (b) the assigned faux probability will be unrelated to the coverage probability; see Spanos [29]. Also, the SEV evaluation can be used to shed light on several confounds relating to different attempts to assign probabilities to different values of $\boldsymbol{\theta}$ within an observed CI, including the most recent attempt based on confidence distributions by Schweder and Hjort [20] above.

5. The SEV Evaluation vs. the Law of Likelihood

Royall [11] popularized an alternative approach to converting inference results into evidence using the likelihood ratio anchored on the Maximum Likelihood (ML) estimator $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_0)$. He rephrased Hacking's [41] *Law of Likelihood (LL)*: *Data* $\mathbf{x}_0$ *support hypothesis* H_0 *over hypothesis* H_1 *if and only if* $L(H_0;\mathbf{x}_0) > L(H_1;\mathbf{x}_0)$. *The degree to which* $\mathbf{x}_0$ *supports* H_0 *better than* H_1 *is given by the Likelihood Ratio (LR):*

$$LR(H_0, H_1; \mathbf{x}_0) = L(H_0;\mathbf{x}_0)/L(H_1;\mathbf{x}_0),$$

by replacing the second sentence with "The likelihood ratio $LR(H_0, H_1; \mathbf{x}_0)$ measures the strength of evidence for H_0". The term "strength of evidence" is borrowed from Fisher's [42] questionable claim about the p-value: "The actual value of p ... indicates the *strength of evidence* against the hypothesis" (p. 80).

To avoid the various contradictions arising from allowing H_0 or H_1 to be composite hypotheses (Spanos, [32,43]), Royall's account of evidence revolves around simple hypotheses $\boldsymbol{\theta}_0$ vs. $\boldsymbol{\theta}_1$. This, however, creates a problem with nuisance parameters that does not arise in the context of N-P testing where all the parameters $\boldsymbol{\theta}$ of the invoked $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ are viewed as an integral part of the inductive premises, irrespective of whether any one parameter is of particular interest.

To exemplify the Royall LR (RLR) account of evidence let us return to example 2.

Example 2 (continued). Consider the data 5152 boys $(X = 1)$ and 4717 girls $(X = 0)$ in the context of the simple Bernoulli model in (18), yielding the Maximum Likelihood (ML) estimate $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_0) = 0.52204$ of $\boldsymbol{\theta}$. For $y_n = \sum_{i=1}^{n} x_i$ the likelihood function takes the form:

$$L(\boldsymbol{\theta}; \mathbf{x}_0) \propto f(\mathbf{x}_0; \boldsymbol{\theta}) = \boldsymbol{\theta}^{y_n}(1-\boldsymbol{\theta})^{n-y_n} = \boldsymbol{\theta}^{5152}(1-\boldsymbol{\theta})^{4717}, \ \boldsymbol{\theta} \in (0,1)$$

The likelihood function scaled by the ML estimate is (Figure 7):

$$L_s(\boldsymbol{\theta}; \mathbf{x}_0) = \frac{(\boldsymbol{\theta})^{5152}(1-\boldsymbol{\theta})^{4717}}{(0.52204)^{5152}(1-0.52204)^{4717}}, \ \boldsymbol{\theta} \in (0,1)$$

Consider the hypotheses H_0: $\boldsymbol{\theta}_0 = 0.52204$ vs. H_1: $\boldsymbol{\theta}_1 = 0.51557$ whose likelihood ratio yields:

$$LR(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; \mathbf{x}_0) = \frac{(0.52204)^{5152}(1-0.52204)^{4717}}{(0.51557)^{5152}(1-0.51557)^{4717}} = 2.286.$$

The first issue that arises is how to interpret 2.286 in terms of the strength of evidence. Royall [11] proposes three thresholds:

$$\begin{aligned} \textbf{weak}: &\quad LR(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; \mathbf{x}_0) \leq 4, \\ \textbf{fairly strong}: &\quad 4 < LR(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; \mathbf{x}_0) \leq 8, \\ \textbf{very strong}: &\quad LR(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; \mathbf{x}_0) \geq 32. \end{aligned}$$

Irrespective of how justified these thresholds are in the proposer's mind, they can be easily disputed as ad hoc and arbitrary since the likelihood function $L(\boldsymbol{\theta}; \mathbf{x}_0) \propto f(\mathbf{x}_0; \boldsymbol{\theta})$, as well as the LR $LR(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; \mathbf{x}_0)$, depend crucially on the invoked statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. Hence, when $LR(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; \mathbf{x}_0)$ is used to determine the appropriate distance between the two likelihoods, the notion of universal thresholds is dubious. This is because every likelihood function has an in-built natural distance relating to each of its parameters known as *the score function*: the derivative of the log-likelihood $s(\boldsymbol{\theta}) = \frac{d \ln L(\boldsymbol{\theta}; \mathbf{x})}{d\boldsymbol{\theta}}$, $\mathbf{x} \in \mathbb{R}_X^n$, evaluated at a particular point $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. Its key role stems from the fact that its mean is zero and its variance is equal to Fisher's information; see Casella and Berger [19]. The score function indicates the sensitivity of $\ln L(\boldsymbol{\theta}; \mathbf{x})$ to infinitesimal changes of $\boldsymbol{\theta}_1$. The problem is that the score function differs for different statistical models. For instance, for the simple Normal in (2) the score function $s(\mu) = \frac{n(\overline{X}_n - \mu)}{\sigma^2}$ is linear in μ, but for the simple Bernoulli model in (18) the score function $s(\boldsymbol{\theta}) = \frac{n(\overline{X}_n - \boldsymbol{\theta})}{\boldsymbol{\theta}(1-\boldsymbol{\theta})}$ is highly non-linear in $\boldsymbol{\theta}$, even though both parameters denote the mean of their underlying distributions. This calls into question the notion of LR universal thresholds, and could explain why so many different such thresholds have been proposed in the literature; see Reid [44].

To shed light on Royall's strength of evidence notion, Table 5 reports $LR(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; \mathbf{x}_0)$ for $\boldsymbol{\theta}_0 = 0.52204$ and different values of $\boldsymbol{\theta}_1$, including the value related to the SEV warranted discrepancy, $\gamma_1^{\ddagger} = 0.01557$ [$\boldsymbol{\theta}_1 = 0.51557$] with probability 0.9, as shown in Figure 7.

Table 5. RLR strength of evidence for $\theta_0 = 0.52204$ vs. different values of θ_1.

$\theta_1 =$	**0.5085**	**0.51**	**0.513**	**0.514**	**0.51557**	**0.525**	**0.5285**	**0.532**	**0.534**	**0.536**
$LR(\theta_0, \theta_1) =$	37.36	17.5	5.02	3.59	2.286	1.189	3.51	6.57	17.02	47.59
$[1/LR(\theta_0, \theta_1)] =$	0.027	0.057	0.199	0.276	0.437	0.841	0.437	0.152	0.059	0.021

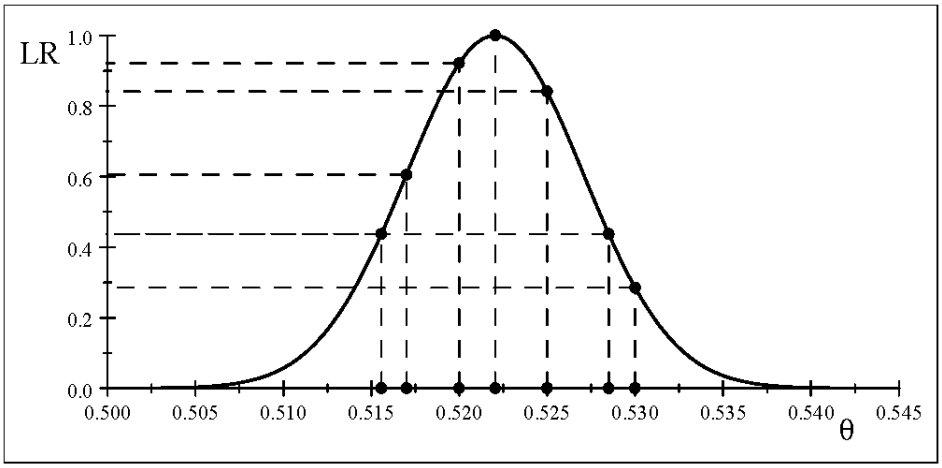

Figure 7. The RLR strength of evidence for $\theta_0 = 0.52204$ vs. $\theta_1 \in (0.5, 0.545)$.

Using the above thresholds, the result $LR(\theta_0, \theta_1; \mathbf{x}_0) = 2.269$ indicates that the strength of evidence for $\theta_0 = 0.52204$ vs. $\theta_1 = 0.51557$ is 'weak', and the evidence will be equal or weaker for any $\theta_1 \in (0.51557, 0.5285)$. If one were to take a firm stand on $\mathbf{x}_0$ providing 'fairly strong' evidence for θ_0, the relevant range of values for θ_1 will be $\theta_1 \notin (0.5117, 0.5323)$.

What is even more problematic for the RLR approach is that $\theta_0 = 0.52204$ has the same strength of evidence against the two values $\theta_1 = 0.51557$ and $\theta_2 = 0.5285$, shown in Figure 7. This calls into question the nature of evidence the RLR approach gives rise to since it undermines the primary objective of frequentist inference. Its strength of evidence for θ_0 is identical against two values on either side of the ML estimate value 0.52204, and as the threshold increases the distance between them increases. This derails any learning from data since it undermines the primary objective of narrowing down the relevant neighborhood for θ^* unless the choice is invariably the ML point estimate as it relates to the fallacious claim $\widehat{\theta}_n(\mathbf{x}_0) \simeq \theta^*$. This weakness was initially pointed out by an early pioneer of the likelihood ratio approach, Barnard [45], p. 129, in his belated review of Hacking [41]. He argued that for any prespecified value θ_1:

"... there always is such a rival hypothesis, viz. that things just had to turn out the way they actually did."

As evidenced in Figure 7, the RLR will pinpoint the ML estimate $\widehat{\theta}(\mathbf{x}_0)$ no matter what the other value in $(0, 1)$ happens to be since it is *the maximally likely value*; see Mayo [46]. No wonder, Hacking [47], p. 137, in his book review of Edward's [48] " Likelihood" changed his mind and abandoned the LR approach altogether:

"I do not know how Edwards's favoured concept [the difference of log-likelihoods] will fare. The only great thinker who tried it out was Fisher, and he was ambivalent. Allan Birnbaum and myself are very favourably reported in this book for things we have said about likelihood, but Birnbaum has given it up and I have become pretty dubious."

Indeed, Hacking [49], p. 141, not only rejected the Law of Likelihood but went a step further by reversing his original viewpoint and wholeheartedly endorsing the N-P testing:

"This paper will show that the Neyman-Pearson theories of testing hypotheses and of confidence interval estimation are sound theories of probable inference."

It is worth noting that even before the optimal N-P theory of testing was finalized in 1933, Pearson and Neyman [50] confronted the problem of how to construe $LR(\theta_0, \theta_1; \mathbf{x})$ by emphasizing the crucial role of the relevant error probabilities:

"But if we accept the criterion suggested by the method of the likelihood it is still necessary to determine its sampling distribution in order to control the errors involved in rejecting a true hypothesis, a knowledge of λ $[LR(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\mathbf{x}_0)]$ alone is not adequate to insure control of the error. ... In order to fix a limit between "small" and "large" value of λ we must know how often such values appear when we deal with a true hypothesis." (p. 106).

A critical weakness of RLR approach is that learning from data about $\boldsymbol{\theta}^*$ using two points $LR(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\mathbf{x}_0)$ at a time becomes untenable since the parameter space is usually uncountable, and thus the point $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_0)=\boldsymbol{\theta}^*$ will always belong to a set of measure zero; see Williams [51]. In addition, the 'maximally likely value' problem is compounded by the fact that the framing of $LR(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\mathbf{x}_0)$ runs afoul of the two crucial stipulations [1]–[2] introduced by Neyman and Pearson [23] in Section 2.3.

A likely counter-argument might be that the RLR approach ignores any potentially relevant error probabilities, but asymptotically the likelihood function will pinpoint $\boldsymbol{\theta}^*$, by invoking the Strong Law of Large Numbers (SLLN):

$$\mathbb{P}\Big(\lim_{n\to\infty}\Big[\ln\big(\tfrac{\frac{1}{n}L_n(\boldsymbol{\theta}^*;\mathbf{x})}{\frac{1}{n}L_n(\boldsymbol{\theta};\mathbf{x})}\big)\Big]>0\Big)=1,\ \forall\boldsymbol{\theta}\in\Theta,\ L_n(\boldsymbol{\theta};\mathbf{x})=\textstyle\sum_{i=1}^n\ln f(\mathbf{x};\boldsymbol{\theta}),\ \forall\mathbf{x}\in\mathbb{R}_X^n. \tag{29}$$

This argument is misplaced since it does not address the key problems relating to $LR(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\mathbf{x}_0)$ narrowing down the potential neighborhood of $\boldsymbol{\theta}^*$, to give rise to any learning from data. To begin with, as the above quote from Le Cam [27], such limits theorems are uninformative as to what happens with the particular data. Also, the probabilistic assignment in (29) relies on $\forall\mathbf{x}\in\mathbb{R}_X^n$, revolves around $\boldsymbol{\theta}^*$, and has nothing to do with $\mathbf{x}_0$ and $L_n(\boldsymbol{\theta}_i;\mathbf{x}_0)$, $i=0,1$. Indeed, for the same reason the Kullback-Leibler (K-L) divergence (Lele, [52] in [53]) defines the difference between $L_n(\boldsymbol{\theta}_0;\mathbf{x})$ and $L_n(\boldsymbol{\theta}_1;\mathbf{x})$, $\forall\mathbf{x}\in\mathbb{R}_X^n$, and provides a comparative account relative to $f(\mathbf{x}_0;\boldsymbol{\theta})$, and not $f(\mathbf{x};\boldsymbol{\theta}^*)$, $\mathbf{x}\in\mathbb{R}_X^n$, as specified by (29). This implies that the RLR and the K-L divergence invoke (implicitly) a variant of the fallacious claim $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_0)\simeq\boldsymbol{\theta}^*$.

A closer look at the RLR approach to evidence reveals that $LR(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\mathbf{x}_0)$ provides nothing more than a ranking of all the different pairs of values of $\boldsymbol{\theta}$ in $(0,1)$ relative to $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_0)$, regardless of the 'true' value $\boldsymbol{\theta}^*$. This can be easily demonstrated using simulation with a known $\boldsymbol{\theta}^*$ to show that the RLR search using different replications $\mathbf{x}_i$, $i=1,2,...,N$, and the associated ML estimates $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_i)$, $i=1,2,...,N$, is unlikely that any one of them will be equal or very close to the true value. To get anything close to $\boldsymbol{\theta}^*$ one should use all the replications for a sufficiently large N to approximate closely the sampling distribution, $f(\widehat{\boldsymbol{\theta}}_n(\mathbf{x});\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}_X^n$, of $\widehat{\boldsymbol{\theta}}_n(\mathbf{X})$, and use the overall average $\frac{1}{N}\sum_{i=1}^N\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_k)\simeq\boldsymbol{\theta}^*$. This, however, is based on N data sets of sample size n not just the one, $\mathbf{x}_0$. Hence, the ratio $LR(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\mathbf{x}_0)$ contains no information relating to $f(\widehat{\boldsymbol{\theta}}_n(\mathbf{x});\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}_X^n$, beyond being a single observation. In this sense, the RLR is as unduly data-dependent as the point estimate $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_0)$ in (16) and the associated observed CI in (17) since it also ignores the uncertainty stemming from $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_0)$ being a single observation from $f(\widehat{\boldsymbol{\theta}}_n(\mathbf{x});\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}_X^n$.

More importantly, comparing the results of the SEV evaluation in Table 1 with those based on $LR(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\mathbf{x}_0)$ in Table 5 indicates clearly that the two approaches to evidence are incompatible. The SEV evaluation relating to the ML estimate $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_0)$ will always yield 0.5, rendering it unwarranted in principle! Worse, the SEV evaluation of the discrepancy associated with $\boldsymbol{\theta}_1=0.51557$ is warranted with probability 0.9, but the value $\boldsymbol{\theta}_2=0.5285$, that the RLR approach assigns the same 'strength of evidence' relative to $\boldsymbol{\theta}_0=0.52204$ is warranted with probability 0.099! Which of the two approaches gives rise to pertinent evidence stemming from constituting a sound inductive generalization of the relevant statistical results, the accept/reject for the SEV, and the ML estimate for the RLR?

One can make a credible case that the ML estimate $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_0)$, would lie in some broad neighborhood of $\boldsymbol{\theta}^*$, but narrowing that down sufficiently to learn about $\boldsymbol{\theta}^*$ cannot be attained using $LR(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\mathbf{x}_0)$. This stems from the fact that the RLR approach revolves around the ML estimate which is based on an *estimation perspective* that cannot be deployed post-data since $\mathbf{X}=\mathbf{x}_0$ has occurred, and thus $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ has transpired. In contrast, the SEV

evaluation is based on a *testing perspective*, which is equally pertinent for evaluating error probabilities pre-data and post-data.

The most crucial feature of the SEV evaluation is that it converts the accept/reject H_0 results into evidence by taking into account the statistical context in (15) including the power of the test. This is important since detecting a particular discrepancy, say γ_1, provides stronger evidence for its presence when the power is lower than higher; see Spanos [18]. The underlying intuition can be illustrated by imagining two people searching for metallic objects on the same beach, one is using a highly sensitive metal detector that can detect small nails, and the other a much less sensitive one. If both detectors begin to buzz simultaneously, the one more likely to have found something substantial is the less sensitive one! The SEV evaluation harnesses this intuition by custom-tailoring the power of the test, replacing the original c_α with the post-data $d(\mathbf{x}_0)$, to establish the discrepancy γ_1 warranted by data $\mathbf{x}_0$ with high enough probability. As argued above, this enables the SEV evidential account to circumvent several foundational problems, including the large n problem. In contrast, the RLR account will yield the same strength of evidence for any two values of $\boldsymbol{\theta}$, irrespective of the size of n.

6. Summary and Conclusions

The replication crisis has exposed the apparent untrustworthiness of published empirical evidence, but its narrow attribution to certain abuses of frequentist testing can be called into question as 'missing the forest for the trees'. A stronger case can be made that the real culprit is the much broader problem of the *uninformed and recipe-like, implementation* of statistical methods, which contributes to the untrustworthiness in many different ways, including [a] imposing invalid probabilistic assumptions on one's data, and [b] conflating unduly data-specific inference results' with 'evidence for or against inferential claims about $\boldsymbol{\theta}^*$', which represent inductive generalizations of such results.

The above discussion makes a case that the post-data severity (SEV) evaluation provides an evidential account of the accept/reject H_0 results, in the form of a discrepancy $\gamma \neq 0$ from $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, warranted with high enough probability by data $\mathbf{x}_0$ and test T_α. The SEV evaluation is framed in terms of a post-data error probability that accounts for the statistical context in (15), as well as the uncertainty stemming from inference results relying on a single realization of the sample $\mathbf{X} = \mathbf{x}_0$. The SEV evaluation perspective is used to call into question Royall's [11] LR approach to evidence as another rendering of the fallacious claim $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_0) \simeq \boldsymbol{\theta}^*$ for a large enough n.

The SEV evaluation is also shown to elucidate/address several foundational issues confounding frequentist testing since the 1930s, including (i) statistical vs. substantive significance, (ii) the large n problem, and (iii) the alleged arbitrariness of the N-P framing H_0 and H_1, used to undermine the coherence of frequentist testing. The SEV also oppugns the proposed alternatives to replace or modify frequentist testing, using statistical results, such as observed CIs, effects sizes, and redefining significance, all of which are equally vulnerable to [a]–[c] undermining the trustworthiness of empirical evidence. In conclusion, it is important to reiterate that unless one has already established the statistical adequacy of the invoked $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, any discussions relating to reliable inference results and trustworthy evidence based on tail area probabilities are unwarranted.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: All data used are available in published sources.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

N-P	Neyman–Pearson
MDS	Medical diagnostic screening
M-S	Misspecification
CI	Confidence interval
SEV	Post-data severity

References

1. National Academy of Sciences. *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop*; NA Press: Washington, DC, USA, 2016.
2. Wasserstein, R.L.; Lazar, N.A. ASA's statement on p-values: Context, process, and purpose. *Am. Stat.* **2016**, *70*, 129–133. [CrossRef]
3. Baker, M. Reproducibility crisis. *Nature* **2016**, *533*, 353–366.
4. Hoffler, J.H. Replication and Economics Journal Policies. *Am. Econ. Rev.* **2017**, 107, 52–55. [CrossRef]
5. Ioannidis, J.P.A. Why most published research findings are false. *PLoS Med.* **2005**, *2*, e124. [CrossRef] [PubMed]
6. Benjamin, D.J.; Berger, J.O.; Johannesson, M.; Nosek, B.A.; Wagenmakers, E.J.; Berk, R.; Bollen, K.A.; Brembs, B.; Brown, L.; Camerer, C.; et al. Redefine statistical significance. *Nat. Hum. Behav.* **2017**, *33*, 6–10. [CrossRef] [PubMed]
7. Spanos, A. Revisiting noncentrality-based confidence intervals, error probabilities and estimation-based effect sizes. *J. Mathematical Stat. Psychol.* **2021**, *104*, 102580. [CrossRef]
8. Spanos, A. Curve-Fitting, the Reliability of Inductive Inference and the Error-Statistical Approach. *Philos. Sci.* **2007**, *74*, 1046–1066. [CrossRef]
9. Leek, J.T.; Peng, R.D. Statistics: P values are just the tip of the iceberg. *Nature* **2015**, *520*, 520–612. [CrossRef]
10. Spanos, A. On theory testing in Econometrics: modeling with nonexperimental data. *J. Econom.* **1995**, *67*, 189–226. [CrossRef]
11. Royall, R. *Statistical Evidence: A Likelihood Paradigm*; Chapman & Hall: New York, NY, USA, 1997.
12. Fisher, R.A. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. A* **1922**, *222*, 309–368.
13. Spanos, A. Mis-Specification Testing in Retrospect. *J. Econ. Surv.* **2018**, *32*, 541–577. [CrossRef]
14. Spanos, A. Where Do Statistical Models Come From? Revisiting the Problem of Specification. In *Optimality: The Second Erich L. Lehmann Symposium*; Rojo, J., Ed.; Lecture Notes-Monograph Series; Institute of Mathematical Statistics: Beachwood, OH, USA, 2006; Volume 49, pp. 98–119.
15. Spanos, A. *Probability Theory and Statistical Inference: Empirical Modeling with Observational Data*; Cambridge University Press: Cambridge, UK, 2019.
16. Spanos, A. Akaike-type Criteria and the Reliability of Inference: Model Selection vs. Statistical Model Specification. *J. Econom.* **2010**, *158*, 204–220. [CrossRef]
17. Spanos, A. Frequentist Model-based Statistical Induction and the Replication crisis. *J. Quant. Econ.* **2022**, *20* (Suppl. 1), 133–159. [CrossRef]
18. Spanos, A. Severity and Trustworthy Evidence: Foundational Problems versus Misuses of Frequentist Testing. *Philos. Sci.* **2022**, *89*, 378–397. [CrossRef]
19. Casella, G.; Berger, R.L. *Statistical Inference*, 2nd ed.; Duxbury: Pacific Grove, CA, USA, 2002.
20. Schweder, T.; Hjort, N.L. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*; Cambridge University Press: Cambridge, UK, 2016.
21. Lehmann, E.L.; Romano, J.P. *Testing Statistical Hypotheses*; Springer: New York, NY, USA, 2005.
22. Owen, D.B. Survey of Properties and Applications of the Noncentral t-Distribution. *Technometrics* **1968**, *10*, 445–478. [CrossRef]
23. Neyman, J.; Pearson, E.S. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. A* **1933**, *231*, 289–337.
24. Fisher, R.A. *The Design of Experiments*; Oliver and Boyd: Edinburgh, UK, 1935.
25. Spanos, A. Revisiting the Large n (Sample Size) Problem: How to Avert Spurious Significance Results. *Stats* **2023**, *6*, 1323–1338. [CrossRef]
26. Spanos, A.; McGuirk, A. The Model Specification Problem from a Probabilistic Reduction Perspective. *J. Am. Agric. Assoc.* **2001**, *83*, 1168–1176. [CrossRef]
27. Le Cam, L. *Asymptotic Methods in Statistical Decision Theory*; Springer: New York, NY, USA, 1986.
28. Neyman, J. Note on an article by Sir Ronald Fisher. *J. R. Stat. Ser. B* **1956**, *18*, 288–294. [CrossRef]
29. Spanos, A. Recurring Controversies about P values and Confidence Intervals Revisited. *Ecology* **2014**, *95*, 645–651. [CrossRef]
30. Ellis, P.D. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*; Cambridge University Press: Cambridge, UK, 2010.
31. Mayo, D.G.; Spanos, A. Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction. *Br. J. Philos. Sci.* **2006**, *57*, 323–357. [CrossRef]
32. Spanos, A. Who Should Be Afraid of the Jeffreys-Lindley Paradox? *Philos. Sci.* **2013**, *80*, 73–93. [CrossRef]
33. Mayo, D.G. *Error and the Growth of Experimental Knowledge*; The University of Chicago Press: Chicago, IL, USA, 1996.

34. Mayo, D.G.; Spanos, A. Error Statistics. In *Handbook of Philosophy of Science, Volume 7: Philosophy of Statistics*; Gabbay, D., Thagard, P., Woods, J., Eds.; Elsevier: Amsterdam, The Netherlands, 2011; pp. 151–196.
35. Arbuthnot, J. An argument for Divine Providence, taken from the constant regularity observed in the birth of both sexes. *Philos. Trans.* **1710**, *27*, 186–190.
36. Hardy, I.C.W. (Ed.) *Sex Ratios: Concepts and Research Methods*; Cambridge University Press: Cambridge, UK, 2002.
37. Good, I.J. Standardized tail-area probabilities. *J. Stat. Comput. Simul.* **1982**, *16*, 65–66. [CrossRef]
38. Berger, J. Four Types of Frequentism and their Interplay with Bayesianism. *N. Engl. J. Stat. Data Sci.* **2022**, 1–12. [CrossRef]
39. Cohen, J. The Earth is round ($p < 0.05$). *Am. Psychol.* **1994**, *49*, 997–1003.
40. Abelson, R.P. *Statistics as Principled Argument*; Lawrence Erlbaum: Mahwah, NJ, USA, 1995.
41. Hacking, I. *Logic of Statistical Inference*; Cambridge University Press: Cambridge, UK, 1965.
42. Fisher, R.A. *Statistical Methods for Research Workers*; Oliver and Boyd: Edinburgh, UK, 1925.
43. Spanos, A. Revisiting the Likelihoodist Evidential Account. *J. Stat. Pract.* **2013**, *7*, 187–195. [CrossRef]
44. Reid, N. Likelihood. In *Statistics in the 21st Century*; Raftery, A.E., Tanner, M.A., Wells, M.T., Eds.; Chapman & Hall: London, UK, 2002; pp. 419–430.
45. Barnard, G.A. The logic of statistical inference. *Br. J. Philos. Sci.* **1972**, *23*, 123–132. [CrossRef]
46. Mayo, D.G. *Statistical Inference as Severe Testing: How to Get Beyond the Statistical Wars*; Cambridge University Press: Cambridge, UK, 2018.
47. Hacking, I. Review: Likelihood. *Br. J. Philos. Sci.* **1972**, *23*, 132–137. [CrossRef]
48. Edwards, A.W.F. *Likelihood*; Cambridge University Press: Cambridge, UK, 1972.
49. Hacking, I. The Theory of Probable Inference: Neyman, Peirce and Braithwaite. In *Science, Belief and Behavior: Essays in Honour of R. B. Braithwaite*; Mellor, D., Ed.; Cambridge University Press: Cambridge, UK, 1980; pp. 141–60 .
50. Pearson, E.S.; Neyman, J. On the problem of two samples. *Bull. Acad. Pol. Sci.* **1930**, 73–96.
51. Williams, D. *Weighing the Odds: A Course in Probability and Statistics*; Cambridge University Press: Cambridge, UK, 2001.
52. Lele, S.R. Evidence functions and the optimality of the law of likelihood. In *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*; Taper, M.L., Lele, S.R., Eds.; University of Chicago Press: Chicago, IL, USA, 2004; pp. 191–216.
53. Taper, M.L.; Lele, S.R. *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*; University of Chicago Press: Chicago, IL, USA, 2004.