

Article

Revisiting the Replication Crisis and the Untrustworthiness of Empirical Evidence

Aris Spanos 

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

Abstract: The current replication crisis relating to the non-replicability and the untrustworthiness of published empirical evidence is often viewed through the lens of the Positive Predictive Value (PPV) in the context of the Medical Diagnostic Screening (MDS) model. The PPV is misconstrued as a measure that evaluates ‘the probability of rejecting H_0 when false’, after being metamorphosed by replacing its false positive/negative probabilities with the type I/II error probabilities. This perspective gave rise to a widely accepted diagnosis that the untrustworthiness of published empirical evidence stems primarily from abuses of frequentist testing, including p-hacking, data-dredging, and cherry-picking. It is argued that the metamorphosed PPV misrepresents frequentist testing and misdiagnoses the replication crisis, promoting ill-chosen reforms. The primary source of untrustworthiness is *statistical misspecification*: invalid probabilistic assumptions imposed on one’s data. This is symptomatic of the much broader problem of the *uninformed and recipe-like implementation* of frequentist statistics without proper understanding of (a) the invoked probabilistic assumptions and their validity for the data used, (b) the reasoned implementation and interpretation of the inference procedures and their error probabilities, and (c) warranted evidential interpretations of inference results. A case is made that Fisher’s model-based statistics offers a more pertinent and incisive diagnosis of the replication crisis, and provides a well-grounded framework for addressing the issues (a)–(c), which would unriddle the non-replicability/untrustworthiness problems.

Keywords: replication crisis; untrustworthy evidence; non-replicability; false positive/negative rates; medical diagnostic testing; Positive Predictive Value; type I/II error probabilities; Neyman-Pearson testing; *p*-value; statistical adequacy; post-data severity evaluation



Academic Editor: Wei Zhu

Received: 1 April 2025

Revised: 28 April 2025

Accepted: 15 May 2025

Published: 20 May 2025

Citation: Spanos, A. Revisiting the Replication Crisis and the Untrustworthiness of Empirical Evidence. *Stats* **2025**, *8*, 41. <https://doi.org/10.3390/stats8020041>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

All approaches to statistical inference revolve around three common elements:

- (A) *Substantive subject matter information* (however vague, specific, or formal) which specifies the questions of interest. This often comes in the form of an *aPriori Postulated (aPP) model*, say $\mathcal{M}_\varphi(\mathbf{z})$, $\varphi \in \Phi$, based on an equation with an error term (usually white-noise) attached. The quintessential example of an aPP model is the linear model:

$$Y_t = \alpha_0 + \alpha_1 x_t + \varepsilon_t, \quad \varepsilon_t \sim \text{NIID}(0, \sigma^2), \quad t \in \mathbb{N} = (1, 2, \dots, n, \dots),$$

where ‘NIID’ stands for ‘Normal (N), Independent (I) and Identically Distributed (ID).

- (B) *Appropriate data* $\mathbf{x}_0 := (x_1, x_2, \dots, x_n)$ that could shed light on the pertinence of the substantive information framed by $\mathcal{M}_\varphi(\mathbf{x})$, $\varphi \in \Phi$.
- (C) A set of *probabilistic assumptions* imposed (directly or indirectly via unobservable error terms) on data \mathbf{x}_0 , comprising the invoked statistical model, say $\mathcal{M}_\theta(\mathbf{z})$, $\theta \in \Theta$,

whose validity underwrites the reliability of inference and the trustworthiness of the ensuing evidence.

For the discussion that follows, it is important to define the key concepts incisively. An empirical study is said to be *replicable* if its statistical results: (i) can be independently confirmed with very similar or consistent results by other researchers, (ii) using the same or akin data, and (iii) studying the same phenomenon of interest. Statistical evidence is said to be *untrustworthy* when: (a) any of probabilistic assumptions $\mathcal{M}_\theta(\mathbf{x})$ is invalid for data \mathbf{x}_0 , which (b) undermines the optimality and reliability of the ensuing inference procedures, and/or (c) misinterpreting the inference results by eliciting unwarranted evidence relating to the parameters θ .

In a highly influential paper, Ioannidis [1] made a case that “most published research findings are false” by proposing a widely accepted explanation based on:

- [S1] Viewing the untrustworthiness of evidence problem at a discipline-wide level as an instantiation of the Positive Predictive Value (PPV) measure of the Medical Diagnostic screening (MDS) model, after replacing its false positive/negative probabilities with the type I/II error probabilities, and adding a prior distribution for the unknown parameter θ .
- [S2] Inferring that the non-replicability stems primarily from abuses of frequentist testing inflating the nominal $\alpha = 0.05$ into a higher actual type I error probability, yielding high rates of rejections of true null hypotheses which lower the replication rates.

During the last two decades, the replication crisis literature has taken presumptions [S1]–[S2] at face value and focused primarily on amplifying how the non-replication of published empirical results provides *prima facie* evidence of their untrustworthiness, pointing the finger at several abuses of frequentist testing, including p-hacking, data-dredging, optional stopping, double dipping, and HARKing; see Baker [2], Höffler [3], Wasserstein et al. [4]. The apparent non-replicability has been affirmed in several disciplines (Simmons et al. [5]; Camerer et al. [6] inter alia). As a result of that, several leading statisticians and journal editors called for reforms (Benjamin et al. [7]; Shrout and Rodgers [8] inter alia), which include:

- [i] replacing *p*-values with Confidence Intervals (CIs) and effect sizes, and
- [ii] redefining statistical significance by reducing the conventional threshold of $\alpha = 0.05$ to a smaller value, say $\alpha = 0.005$; see Johnson et al. [9].

The **primary objective** of this paper is twofold. *First*, to argue that the untrustworthiness of published empirical evidence has been endemic in several scientific disciplines since the 1930s, and the leading contributor to the untrustworthiness has been **statistical misspecification**: invalid probabilistic assumptions imposed on the data in (C) above. An effective way to corroborate that is to replicate several influential papers in a particular research area with the same or akin data, and affirm or deny their trustworthiness or/and replicability by first testing the validity of their invoked probabilistic assumptions; see Andreou and Spanos [10], Do and Spanos [11]. In contrast, Ioannidis [1] makes no such attempt to corroborate his claims, but instead employs a metaphor, based on the PPV, that invokes stipulations [S1]–[S2] grounded on several questionable presuppositions.

Second, to put forward an alternative explanation about the main sources of the untrustworthiness of published empirical evidence and propose ways to ameliorate the problem. It is argued that the untrustworthiness stems primarily from a much broader problem relating to the uninformed and recipe-like implementation of frequentist statistics (a) without proper understanding of the invoked statistical model $\mathcal{M}_\theta(\mathbf{x})$, and its validity for data \mathbf{x}_0 , (b) using incongruous implementations of frequentist inference procedures that misconstrue their error probabilities, as well as (c) invoking unwarranted evidential

interpretations of their results, such as misinterpreting ‘accept H_0 ’ (no evidence against H_0) as evidence *for* H_0 ’ or misinterpreting ‘reject H_0 ’ (evidence against H_0) as evidence *for* a particular H_1 ’. As incisively argued by Stark and Saltelli [12], p. 41:

“... Practitioners go through the motions of fitting models, computing p -values or confidence intervals, or simulating posterior distributions. They invoke statistical terms and procedures as incantations, with scant understanding of the assumptions or relevance of the calculations, or even the meaning of the terminology. This demotes statistics from a way of thinking about evidence and avoiding self-deception to a formal “blessing” of claims”.

It is argued that Fisher’s [13] model-based frequentist inference framework provides effective ways to address both the non-replicability and the untrustworthiness problems by dealing with the issues (a)–(c) raised above. Special emphasis is placed on distinguishing between ‘inference results’, such as a point estimate, an observed Confidence Interval (CI), an effect size, and accept/reject H_0 , which are too coarse and unduly ‘data-specific, and their inductive generalizations framed as evidence for germane inferential claims relating to the ‘true’ θ , say θ^* . The latter are outputted by the **post-data severity evaluation** (SEV) of testing results. This distinction is crucial since inference results are not often replicable with ‘akin’ data, but the SEV inferential claims would be replicable when the invoked statistical model is statistically adequate (valid assumptions) for the particular data.

Section 2 provides a summary of Fisher’s [13] model-based frequentist statistics, with particular emphasis on the statistical adequacy of $\mathcal{M}_\theta(\mathbf{x})$ (implicitly) invoked by all model-based inferences (frequentist and Bayesian), as a prelude to the discussion. Section 3 revisits the MDS model (a bivariate simple Bernoulli) underlying the PPV to compare and contrast its false positive/negative rates with the Neyman-Pearson (N-P) type I/II error probabilities. Section 4 elaborates on the uninformed and recipe-like implementation of frequentist statistics by highlighting several unwarranted evidential interpretations of their results. Particular emphasis is placed on the role of establishing the statistical adequacy of $\mathcal{M}_\theta(\mathbf{x})$ to ensure that the actual error probabilities approximate closely the nominal ones, rendering them tractable. Section 5 discusses how the post-data severity (SEV) evaluation of the accept/reject H_0 results (Mayo and Spanos [14]) transmutes them into replicable evidence which could give rise to learning from \mathbf{x}_0 about the phenomena of interest.

2. Model-Based Frequentist Statistics

2.1. Fisher’s Model-Based Statistical Framing

This section provides a bird’s eye view of frequentist inference, in general, and N-P testing in particular, in an attempt to preempt needless confusion relating to what is traditionally known as the Null Hypothesis Significance Testing (NHST) practice; see Nickerson [15], Spanos [16].

Fisher [13] founded model-based statistical induction that revolves around the concept of a prespecified parametric statistical model, whose generic form is:

$$\mathcal{M}_\theta(\mathbf{x}) = \{f(\mathbf{x}; \theta), \theta \in \Theta\}, \mathbf{x} \in \mathbb{R}_X^n, \text{ for } \Theta \subset \mathbb{R}^m, m < n, \quad (1)$$

where $f(\mathbf{x}; \theta)$ denotes the distribution of the sample $\mathbf{X} := (X_1, X_2, \dots, X_n)$, \mathbb{R}_X^n the sample space, and Θ the parameter space. $\mathcal{M}_\theta(\mathbf{x})$ can be viewed as a particular parameterization ($\theta \in \Theta$) of the observable stochastic process $\{X_k, k \in \mathbb{N} := (1, 2, \dots, n, \dots)\}$ underlying data \mathbf{x}_0 . $\mathcal{M}_\theta(\mathbf{x})$ provides an ‘idealized’ description of a statistical mechanism that could have given rise to \mathbf{x}_0 . The *main objective* of model-based frequentist inference is to give rise to learning from data \mathbf{x}_0 by *narrowing down* Θ to a small neighborhood around the ‘true’ value of θ in Θ , say θ^* , whatever value that happens to be; see Spanos [17] for further discussion.

The *statistical adequacy* of $\mathcal{M}_\theta(\mathbf{x})$ for data \mathbf{x}_0 plays a crucial role in securing the reliability of inference and the trustworthiness of evidence because it ensures that the actual error probabilities approximate closely the nominal ones, enabling the ‘control’ of these unobservable probabilities. As a result, when $\mathcal{M}_\theta(\mathbf{x})$ is *misspecified*:

- (a) the distribution of the sample $f(\mathbf{x}; \theta)$, $\mathbf{x} \in \mathbb{R}_X^n$, in (1) is erroneous,
- (b) rendering the likelihood function $L(\theta; \mathbf{x}_0) \propto f(\mathbf{x}_0; \theta)$, $\theta \in \Theta$, invalid,
- (c) distorting the sampling distribution $f(y_n; \theta)$ of any relevant statistic $Y_n = h(X_1, X_2, \dots, X_n)$ (estimator, test, predictor).

In turn, (a)–(c) give rise to (i) ‘non-optimal’ inference procedures, and (ii) induce sizeable *discrepancies* between the *actual* and *nominal* error probabilities; arguably the most crucial contributor to the untrustworthiness of empirical evidence. Applying a 0.05 significance level test, when the actual type I error probability is 0.97 (Spanos [18] Table 15.5) will give rise to untrustworthy evidence. Hence, the only way to keep track of the relevant error probabilities is to establish the statistical adequacy of $\mathcal{M}_\theta(\mathbf{x})$ to forefend the unreliability of inference stemming from (a)–(c), using thorough Mis-Specification (M-S) testing; see Spanos [19]. This will secure the optimality and reliability of the ensuing inferences, giving rise to trustworthy evidence.

2.2. Frequentist Inference: Estimation

Example 1. Consider the simple Normal model:

$$\mathcal{M}_\theta(\mathbf{x}): X_t \sim \text{NIID}(\mu, \sigma^2), \theta := (\mu, \sigma^2) \in [\mathbb{R} \times \mathbb{R}_+] = (0, \infty), x_t \in \mathbb{R}, t \in \mathbb{N}, \tag{2}$$

The sampling distributions of the optimal estimators of $\theta = (\mu, \sigma^2)$, $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$, and $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$, evaluated under factual reasoning ($\theta = \theta^*$) are:

$$\text{(i) } \bar{X}_n \stackrel{\theta = \theta^*}{\rightsquigarrow} N(\mu^*, \frac{\sigma_*^2}{n}), \text{ (ii) } \frac{(n-1)s^2}{\sigma^2} \stackrel{\theta = \theta^*}{\rightsquigarrow} \chi^2(n-1), \tag{3}$$

and (iii) \bar{X}_n and s^2 are independent, where $\chi^2(n-1)$ denotes the chi-square distribution with $(n-1)$ degrees of freedom (d.f.). Assumption (iii) implies:

$$E(h_1(\bar{X}_n) \cdot h_2(s^2)) = E[h_1(\bar{X}_n)] \cdot E[h_2(s^2)],$$

for any well-behaved (Borel) functions $h_i(\cdot)$, $i = 1, 2$; see Spanos [16]. The *factual* ($\theta = \theta^*$) reasoning ensures that $E(\bar{X}_n - \mu) = 0$ and $E(\frac{(n-1)s^2}{\sigma^2}) = (n-1)$, yielding:

$$E(\frac{\sqrt{n}(\bar{X}_n - \mu)}{s}) \stackrel{\mu = \mu^*}{=} E(\bar{X}_n - \mu^*) \cdot E(\sqrt{n}/s) = 0, \text{ for any } E(\sqrt{n}/s) > 0.$$

This gives rise to the sampling distribution of the pivot:

$$\tau(\mathbf{X}; \mu) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{s} \stackrel{\theta = \theta^*}{\rightsquigarrow} \text{St}(n-1), \tag{4}$$

where $\text{St}(n-1)$ denotes the Student’s t distribution with $(n-1)$ d.f.

(4) underlies the $(1-\alpha)$ Uniformly Most Accurate (UMA) CI for μ :

$$\mathbb{P}(\bar{X}_n - c_{\frac{\alpha}{2}}(\frac{s}{\sqrt{n}}) \leq \mu < \bar{X}_n + c_{\frac{\alpha}{2}}(\frac{s}{\sqrt{n}})); \theta = \theta^* = 1 - \alpha, \tag{5}$$

where $c_{\frac{\alpha}{2}}$ denotes the distribution threshold of the relevant tail area with $(\alpha/2)$ probability. The optimal property of UMA denotes a CI with the shortest expected length.

2.3. Neyman-Pearson (N-P) Testing

Example 1 (continued). In the context of (2), testing the hypotheses:

$$H_0: \mu \leq \mu_0 \text{ vs. } H_1: \mu > \mu_0, \tag{6}$$

gives rise to the Uniformly Most Powerful (UMP) α -significance level N-P test:

$$T_\alpha := [\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s}, C_1(\alpha) = \{\mathbf{x}: \tau(\mathbf{x}) > c_\alpha\}], \tag{7}$$

(Lehmann and Romano [20]) where $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$, $s^2 = \frac{1}{(n-1)} \sum_{k=1}^n (X_k - \bar{X}_n)^2$, $C_1(\alpha)$ is the rejection region, and c_α is determined by the prespecified α .

The distribution of $\tau(\mathbf{X})$ evaluated using hypothetical reasoning (what if $\mu = \mu_0$):

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s} \stackrel{\mu = \mu_0}{\sim} \text{St}(n-1), \tag{8}$$

ensures that $E(\bar{X}_n - \mu_0) \stackrel{\mu = \mu_0}{=} \mu_0 - \mu_0 = 0$, underlying the evaluations of:

$$\alpha = \mathbb{P}(\tau(\mathbf{X}) > c_\alpha; \mu = \mu_0), \quad p(\mathbf{x}_0) = \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu = \mu_0). \tag{9}$$

The sampling distribution of $\tau(\mathbf{X})$ evaluated under H_1 (what if $\mu = \mu_1$) is:

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s} \stackrel{\mu = \mu_1}{\sim} \text{St}(\delta_1; n-1), \text{ for all } \mu_1 > \mu_0, \tag{10}$$

where $\delta_1 = E\left[\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s}\right] \stackrel{\mu = \mu_1}{=} \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}$ is the noncentrality parameter, and (10) is used to evaluate the type II error probability $\beta(\mu_1)$ and the power for a given α :

$$\mathcal{P}(\mu_1) = (1 - \beta(\mu_1)) = \mathbb{P}(\tau(\mathbf{X}) > c_\alpha; \mu = \mu_1), \text{ for } \mu_1 = \mu_0 + \gamma_1, \gamma_1 \geq 0. \tag{11}$$

It should be emphasized that these error probabilities are assigned to a particular N-P test, e.g., T_α , to calibrate its *generic* (for any $\mathbf{x} \in \mathbb{R}^n$) *capacity* to detect different discrepancies $\gamma_1 = (\mu_1 - \mu_0)$ from μ_0 ; see Neyman and Pearson [21].

The primary role of these error probabilities is to operationalize the notions of ‘statistically significant/insignificant’ in the form of ‘accept/reject H_0 results’. The optimality of N-P tests revolves around an in-built trade-off between the type I and II error probabilities, and an optimal N-P test is derived by prespecifying α at a low value and minimizing the type II error $\beta(\mu_1)$, or maximizing the power $\mathcal{P}(\mu_1) = (1 - \beta(\mu_1))$, $\forall \mu_1 = \mu_0 + \gamma_1, \gamma_1 \geq 0$. In summary, the error probabilities have several *key* attributes Spanos [16]:

- [i] They are assigned to the test procedure T_α to ‘calibrate’ its *generic* (for any $\mathbf{x} \in \mathbb{R}_X^n$) *capacity* to detect different *discrepancies* γ from $\mu = \mu_0$.
- [ii] They cannot be conditional on θ , an unknown constant (not a random variable).
- [iii] There is a built-in trade-off between the type I and II error probabilities.
- [iv] They frame the accept/reject H_0 rules in terms of ‘statistical approximations’ based on the distribution of $d(\mathbf{X})$ evaluated using hypothetical reasoning.
- [v] They are *unobservable* since they revolve around θ^* -true value of θ .

This was clearly explained in Fisher’s 1955 reply to a letter from John Tukey: “A level of significance is a probability derived from a hypothesis [hypothetical reasoning], not one asserted in the real world” (Bennett [22] p. 221).

2.4. An Inconsistent Hybrid Logic Burdened with Confusion?

In an insightful discussion, Gigerenzer [23] describes the traditional narrative of statistical testing created by textbook writers in psychology during the 1950s and 1960s, as based on ‘a hybrid logic’: “Neither Fisher nor Neyman and Pearson would have accepted this hybrid as a theory of statistical inference. The hybrid logic is inconsistent from both perspectives and burdened with conceptual confusion” (p. 324)

It is argued that Fisher’s model-based statistical induction could provide a unifying ‘reasoning’ that elucidates the similarities and differences between Fisher’s inductive inference and Neyman’s inductive behavior; see Halpin and Stam [24].

Using the t-test in (7), the two perspectives have several *common components*:

- (i) a prespecified statistical model $\mathcal{M}_\theta(\mathbf{x})$,
- (ii) the framing of hypotheses in terms of $\theta := (\mu, \sigma^2)$,
- (iii) a test statistic $\tau(\mathbf{X})$,
- (iv) a null hypothesis $H_0: \mu = \mu_0$,
- (v) the sampling distribution of $\tau(\mathbf{X})$ evaluated under H_0 , and
- (vi) a probability threshold α (0.1, 0.01, 0.025, 0.05, etc.) to decide when H_0 is discordant/rejected.

- The N-P perspective adds to the common components (i)–(vi),
- (vii) re-interpreting α as a prespecified(pre-datat) type I error probability,
- (viii) an alternative hypothesis $H_1: \mu \neq \mu_0$ to supplement $H_0: \mu = \mu_0$,
- (ix) the sampling distribution of $\tau(\mathbf{X})$ evaluated under H_1 ,
- (x) the type II error probability and the power of a test.

These added components frame an optimal theory of N-P testing based on constructing an optimal test statistic $\tau(\mathbf{X})$ and framing a rejection region $C_1(\alpha)$ to maximize the power.

The key to fusing the two perspectives is the ‘hypothetical reasoning’ underlying the derivation of both sampling distributions in (v) and (ix). The crucial difference is that the type I and II (power) error probabilities are *pre-data* because they calibrate the test’s generic (for all $\mathbf{x} \in \mathbb{R}^n$) capacity to detect discrepancies from H_0 , but the *p*-value is a *post-data* error probability since its evaluation is based on $\tau(\mathbf{x}_0)$.

The traditional textbook narrative considers Fisher’s significance testing, based on $H_0: \mu = \mu_0$ and guided by the *p*-value, problematic since the absence of an explicit alternative H_1 renders the power and the *p*-value ambivalent, e.g., one-sided or two-sided? This, however, misconstrues Fisher’s significance testing since his *p*-value is invariably one-sided because, post-data, the *sign* of $\tau(\mathbf{x}_0)$ designates the relevant tail. In fact, the ambivalence originates in the N-P-laden definition of the *p*-value: ‘the probability of obtaining a result ‘equal to or more extreme’ than the one observed \mathbf{x}_0 when H_0 is true’. The clause ‘equal or more extreme’ is invariably (mis)interpreted in terms of H_1 . Indeed, the *p*-value is related to α by interpreting $p(\mathbf{x}_0)$ as the smallest significance level for which a true H_0 would have been rejected; see Lehmann and Romano [20]. A more pertinent *post-data* definition of the *p*-value that averts this ambivalence is: ‘the probability of all sample realizations $\mathbf{x} \in \mathbb{R}_X^n$ that accord less well (in terms of $\tau(\mathbf{x})$) with H_0 than \mathbf{x}_0 does, when H_0 is true’; see Spanos [18].

Similarly, the power of a test comes into play with just the common components (i)–(vi) since the probability threshold α defines the distribution threshold c_α for the probability of detecting discrepancies of the form $\gamma_1 = \pm (\mu_1 - \mu_0)$ using the non-central Student’s *t* in (10), originally derived by Fisher [25]. Indeed, Fisher [26] pp. 21–22, was the first to recognize the effect of increasing n on the power (he called *sensitivity*): “By increasing the size of the experiment, we can render it more sensitive, meaning by this that it will allow of the detection of ... quantitatively smaller departures from the null hypothesis”, which is particularly useful in experimental design; see Box [27].

The above arguments suggest that when the underlying hypothetical reasoning and the pre-data vs. post-data error probabilities are delineated, there is no substantial conflict or conceptual confusion between the Fisher and N-P perspectives. What remains problematic, however, is that neither the p -value nor the accept/reject H_0 results provide cogent evidence because they are too coarse to designate a small neighborhood containing μ^* , to engender any genuine learning from data via $\tau(\mathbf{x}_0)$. The post-data severity (SEV) evaluation offers such an evidential interpretation in the form of a discrepancy from μ_0 warranted by $\tau(\mathbf{X})$ and data \mathbf{x}_0 with high enough probability; see Section 5.

3. The Medical Diagnostic Screening Perspective

3.1. Revisiting the MDS Statistical Model

The statistical analysis of MDS was pioneered by Yerushalmy [28] and Neyman [29]. The concept of the false-positive (negative) rate relates to the proportion of results of a medical diagnostic test indicating falsely that a medical condition exists (does not exist). ‘Sensitivity’ denotes the proportion of positives that are correctly identified, and ‘specificity’ is the proportion of negatives that are correctly identified.

Example 2. Viewing the MDS in the context of model-based inference, the invoked $\mathcal{M}_\theta(\mathbf{z})$ is a simple (bivariate) Bernoulli (Ber) model:

$$\mathcal{M}_\theta(\mathbf{z}): \begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim \text{BerIID} \left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} \theta_1(1-\theta_1) & \theta_3 \\ \theta_3 & \theta_2(1-\theta_2) \end{pmatrix} \right), i \in \mathbb{N}, \quad (12)$$

where $\theta_i \in (0, 1)$, $i = 1, 2, 3$, $(X = 0)$ -test positive, $(X = 1)$ -test negative, $(Y = 0)$ -disease, and $(Y = 1)$ -no disease; see Spanos [30]. $\mathcal{M}_\theta(\mathbf{z})$ is often presented in the form of the contingency in Table 1, where (Bishop et al. [31]):

$$\theta_1 = p(1,0) + p(1,1), \theta_2 = p(0,1) + p(1,1), \theta_3 = p(1,1) - p_1(1)p_2(1).$$

Table 1. 2×2 contingency table.

$x \setminus y$	0	1	Total
0	$p(0,0)$	$p(0,1)$	$p_1(0)$
1	$p(1,0)$	$p(1,1)$	$p_1(1)$
total	$p_2(0)$	$p_2(1)$	1

The MDS revolves around several measures relating to the ‘effectiveness’ of the screening, as shown in Table 2, representing the parameters of substantive interest ($\varphi_i, i = 1, 2, \dots, 6$) in terms of which inferences are often framed.

Table 2. Medical Diagnostic Screening Measures.

Prevalence: $\varphi_1 := \mathbb{P}(Y = 0) = p_2(0)$	Sensitivity: $\varphi_2 := \mathbb{P}(X = 0 Y = 0) = \frac{p(0,0)}{p_2(0)}$
False-positive: $\varphi_3 := \mathbb{P}(X = 0 Y = 1) = \frac{p(0,1)}{p_2(1)}$	False-negative: $\varphi_4 := \mathbb{P}(X = 1 Y = 0) = \frac{p(1,0)}{p_2(0)}$
Specificity: $\varphi_5 := \mathbb{P}(X = 1 Y = 1) = \frac{p(1,1)}{p_2(1)}$	PPV: $\varphi_6 := \mathbb{P}(Y = 0 X = 0) = \frac{p(0,0)}{p_1(0)}$

The relevant data $\mathbf{z}_0 := \{(x_i, y_i), i = 1, 2, \dots, n\}$ for the MDS are usually generated under ‘controlled conditions’ in order to secure the validity of the IID assumptions, and the results are based on a large number of medical tests, say 2000, carried out with specimens prepared in a lab that is *known* to be positive or negative; see Senn [32]. This renders $(X = i, Y = j), i, j = 0, 1$, *observable events* whose frequencies can be used to estimate, not

only the probabilities in Table 1, but also the parameters of interest $\varphi_i, i = 1, 2, \dots, 6$, in Table 2. The unknown parameters $p(i, j), i, j = 0, 1$, are estimated using Maximum Likelihood (ML):

$$\hat{p}_{ML}(i, j) = \frac{1}{n} \sum_{k=1}^n \mathbf{1}(X=i, Y = j) = \frac{f_{ij}}{n}, i, j = 0, 1, \tag{13}$$

where $\mathbf{1}(\cdot)$ is the indicator function and f_{ij} denotes the frequencies in each cell. ML estimators are used since they are invariant to reparametrizations needed to estimate the parameters of interest $\varphi := (\varphi_1, \varphi_2, \dots, \varphi_6)$ (Table 2); see Greenhouse and Mantel [33], Nissen-Meyer [34].

Example 2 (empirical). The ML estimates in (13), in the form of the observed relative frequencies based on $n = 2000$, are given in Table 3. The point estimates of the MDS measures in Table 2 are shown in Table 4, where the specificity is reasonable, but the PPV = 0.412 indicates that the screening is not reliable enough.

Table 3. 2 × 2 table of relative frequencies.

$x \setminus y$	0	1	Total
0	$\frac{131}{2000} = 0.065$	$\frac{187}{2000} = 0.094$	$\frac{318}{2000} = 0.159$
1	$\frac{11}{2000} = 0.006$	$\frac{1671}{2000} = 0.835$	$\frac{1682}{2000} = 0.841$
total	$\frac{142}{2000} = 0.071$	$\frac{1858}{2000} = 0.929$	1

Table 4. Estimates of MDS measures.

$\hat{\varphi}_1 = \hat{p}_2(0) = 0.071$	$\hat{\varphi}_2 = \frac{\hat{p}(0,0)}{\hat{p}_2(0)} = \frac{0.065}{0.071} = 0.915$
$\hat{\varphi}_3 = \frac{\hat{p}(0,1)}{\hat{p}_2(1)} = \frac{0.094}{0.929} = 0.101$	$\hat{\varphi}_4 = \frac{\hat{p}(1,0)}{\hat{p}_2(0)} = \frac{0.006}{0.071} = 0.085$
$\hat{\varphi}_5 = \frac{\hat{p}(1,1)}{\hat{p}_2(1)} = \frac{0.835}{0.929} = 0.899$	$\hat{\varphi}_6 = \frac{\hat{p}(0,0)}{\hat{p}_1(0)} = \frac{0.065}{0.094} = 0.691$

Note that all the above probabilities are estimable since X and Y are observable (Senn [32], in contrast to the N-P error probabilities that revolve around θ^*).

3.2. The PPV and Its Impertinent Metamorphosis

The PPV is a well-known measure of the effectiveness of a MDS evaluating: ‘the probability of a patient testing positive, given that the patient has the disease’:

$$\text{PPV: } \mathbb{P}(Y = 0|X = 0) = \frac{\mathbb{P}(Y = 0, X = 0)}{\mathbb{P}(X = 0)} = \frac{\mathbb{P}(X = 0|Y = 0) \cdot \mathbb{P}(Y = 0)}{\mathbb{P}(X = 0|Y = 0) \cdot \mathbb{P}(Y = 0) + \mathbb{P}(X = 0|Y = 1) \cdot \mathbb{P}(Y = 1)}, \tag{14}$$

derived from the probabilities of the binary observable events in Table 1.

Ioannidis [1] metamorphosed the PPV by (a) replacing the false positive/negative with the type I/II error probabilities:

$$\mathbb{P}(X = 0|Y = 0) = \text{Pr}(R|F) := (1 - \beta(\theta_1)), \mathbb{P}(X = 0|Y = 1) = \text{Pr}(R|\bar{F}) := \alpha, \tag{15}$$

where $\mathbb{P}(\theta_1) = (1 - \beta(\theta_1)), \theta_1 \in \Theta_1$, is the power, and (b) adding a prior:

$$\pi(\theta) = \begin{cases} \omega & \text{for } \theta \neq \theta^*, \\ (1 - \omega) & \text{for } \theta = \theta^*, \end{cases} \forall \theta \in \Theta := (0, 1), \tag{16}$$

where θ^* denotes the ‘true’ θ , to define the Metamorphosed PPV:

$$\text{M-PPV: } \text{Pr}(F|R) = \frac{\text{Pr}(R|F) \text{Pr}(F)}{\text{Pr}(R|F) \text{Pr}(F) + \text{Pr}(R|\bar{F}) \text{Pr}(\bar{F})} := \frac{(1 - \beta(\theta_1))\omega}{(1 - \beta(\theta_1))\pi + \alpha(1 - \omega)}, \tag{17}$$

aiming to evaluate ‘the probability of rightful rejections of H_0 (when false)’.

Unfortunately, the substitutions used in (17) are incongruous since the two sides in (15) have nothing in common. In particular, the type I/II error probabilities are testing-based, grounded on hypothetical reasoning ($\theta = \theta_0$ or $\theta = \theta_1$), and satisfying features [i]–[v] (Section 2.3). In contrast, the false positive/negative probabilities are estimation-based, grounded on factual reasoning ($\theta = \theta^*$), and belying [i]–[v] since:

- [i]* They are assigned to the observable events ($X = 0|Y = 1$) and ($X = 1|Y = 0$).
- [ii]* They are observable conditional $\mathbb{P}(X = 0|Y = 1)$ and $\mathbb{P}(X = 1|Y = 0)$ probabilities.
- [iii]* There is *no* trade-off between $\mathbb{P}(X = 0|Y = 1)$ and $\mathbb{P}(X = 1|Y = 0)$.
- [iv]* They relate only to the bivariate Bernoulli model in (12) (Table 2).
- [v]* They are *observable* probabilities that can be estimated using data \mathbf{z}_0 .

Example 3. In the field of psychology, assume $\omega = \Pr(F) = 0.1$ (10% false nulls), $(1-\beta) =: \Pr(R|F) = 0.8$ (high enough), and $\alpha =: \Pr(R|\bar{F}) = 0.15$ (higher than 0.05 due to abuses of the p -value) yields $M\text{-PPV} = 0.372 < 0.5$, which Ioannidis [1] would interpret as evidence that “most published research findings [in psychology] are false”.

This evaluation raises serious questions of pertinence. ‘Why would the power $\mathcal{P}(\theta_1) =: \Pr(R|F) = 0.8$, denoting the generic capacity of a N-P test to detect all discrepancies $\gamma_1 = (\theta_1 - \theta_0)$, for all $\theta_1 > \theta_0$ relating to H_1 , detecting an arbitrary discrepancy $\gamma_1^{\dagger} = (\theta_1^{\dagger} - \theta_0)$ with $\mathcal{P}(\theta_1) = 0.8$ be of any interest in practice? Why 10% false nulls? Given that $\theta \in (0, 1)$, $\Pr(\theta_0 = \theta^*) = 0$, for any $\theta_0 \in (0, 1)$, even when one assumes (erroneously) that such an assignment makes sense.

To reveal the problems with the evaluation of the M-PPV at a more practical level, consider the task of *collecting* (IID) *data*, analogous to the MDS data in Table 3, from the ‘population’ of all published empirical studies in psychology. Such a task is instantly rendered impossible by the fact that ‘ H_0 is true/false’ is *unobservable* since they revolve around θ^* . This is why Ioannidis makes no such attempt and instead evaluates the M-PPV by plucking numbers from thin air.

As mentioned above, a more effective procedure to make a case for the untrustworthiness of published empirical evidence would be to replicate the most cited empirical papers in a particular research field with the same or akin data, and affirm or deny their untrustworthiness/non-replicability, by probing their statistical adequacy first. A recent example of such replication is discussed in Do and Spanos [11], which relates to the most famous empirical relationship in macroeconomics, the Phillips curve. Some of the most highly cited/influential published papers in major economics journals since the late 1950s have been replicated using the original data. Their replication confirmed their numerical results with minor differences in precision. However, when the statistical adequacy is probed, all these papers were found to be statistically misspecified, rendering their empirical findings untrustworthy.

By focusing on the untrustworthiness at a discipline-wide level, the M-PPV has nothing to say about the virtues or flaws of the individual papers. Therefore, an empirical study in psychology that does an excellent job of forefending statistical misspecification and erroneous interpretations of its inference results, should not be dismissed as *untrustworthy by association*, just because one’s conjectural evaluation yields $M\text{-PPV} < 0.5$. Worse, the Bonferroni-type corrections for the questionable practices blamed by the replication literature make sense only at the level of an individual study. Indeed, such adjustments will be pointless when that study is statistically misspecified, since the latter would render the error probabilities intractable by the induced sizeable discrepancies between the actual and nominal ones (Section 4.1).

In summary, Ioannidis [1] makes his case using an incongruous metaphor free of any empirical evidence in conjunction with highly questionable claims, including:

- [a] The accept/reject H_0 results are in essence misconstrued as evidence for H_0/H_1 .
- [b] The observable/conditional false positive/negative probabilities of the PPV are viewed as equivalent to the unobservable/unconditional type I/II error probabilities.
- [c] Imputing a contrived ‘prior’ in (16) into N-P testing, by viewing the falsity of H_0 and H_1 at a discipline-wide level as a ‘bag of nulls, a proportion of which is false’.
- [d] Invoking a direct causal link between the replicability and the trustworthiness of empirical evidence.

3.3. Could the M-PPV Shed Any Light on Untrustworthiness?

The question that naturally arises is why the case by Ioannidis [1], grounded on the M-PPV, has been so widely accepted. A plausible explanation could be that its apparent credibility stems from the same uninformed and recipe-like implementation of statistics, which cursorily uses the terms ‘type I/II error probabilities’ and ‘false positive/negative probabilities’ interchangeably, ignoring their fundamental differences in their nature and crucial attributes [i]–[v] vs. [i]*–[v]*.

To make the impertinence of this claim more transparent, consider using the same Bernoulli model in (12) for the Berkeley admissions data (Freedman et al. [35] pp. 17–20), which would alter the BDS terminology of the (X, Y) random variables (Table 5).

Table 5. MDS vs. Berkeley admissions.

MDS:	$(X = 0)$ -test positive	$(X = 1)$ -test negative	$(Y = 0)$ -disease	$(Y = 1)$ -no disease
Admissions:	$(X = 0)$ -deny	$(X = 1)$ -admit	$(Y = 0)$ -female	$(Y = 1)$ -male

This would leave unchanged the statistical analysis in Tables 1, 3 and 4, but the MDS terminology in Table 2 would seem absurd for the data on potential gender discrimination in the admissions. In fact, any attempt to relate $\mathbb{P}(X = 0|Y = 0) = \text{Pr}(\text{denying admission}|female)$, $\mathbb{P}(X = 0|Y = 1) = \text{Pr}(\text{denying admission}|male)$ to the power of a test and the type I error probability, respectively, would be considered ludicrous.

Regrettably, the M-PPV has created a sizeable literature, including in philosophy of science, based on selecting different values for (α, β, π) in (17), and plotting the M-PPV as a function of the odds ratio, $(1 - \pi) / \pi$, or the power $(1 - \beta)$ to propose various impertinent diagnoses and remedies for untrustworthiness; see Nosek et al. [36] and Munafò et al. [37]. For instance, Bird [38] argues: “If most of the hypotheses under test are false, then there will be many false hypotheses that are apparently supported by the outcomes of well conducted experiments and null-hypothesis significance tests with a type-I error rate (α) of 5%. Failure to recognize this is to commit the fallacy of ignoring the base rate” (p. 965) is ill-thought-out. First, the charge that frequentist testing is vulnerable to the base-rate fallacy is erroneous since frequentist error probabilities cannot be conditional on a constant θ ; see Spanos [30]. Second, statistical hypotheses share no similitude to ‘a bag of lottery tickets one of which will win’ since parameters are unknown constants and take values over a continuum, hence $\theta_0 \neq \theta^*$ will invariably be the case in practice with $\theta_0 = \theta^*$ only by happenstance. Worse, Bird’s [38] recommendation that the trustworthiness of evidence will improve by: “Seek(ing) means of generating more likely research hypotheses to be true” (p. 968) reveals an arrant misunderstanding of how N-P testing gives rise to learning about θ^* from data \mathbf{z}_0 . Indeed, interrogatory N-P probing using optimal tests will yield more reliable inferences, irrespective of how many null values $\theta_0 \neq \theta^*$ have been probed since learning is attained irrespective of whether H_0 is accepted or rejected; see Section 5.

4. Uninformed Implementation of Statistical Modeling and Inference

A strong case can be made that the widespread abuse of frequentist testing is only symptomatic of a much broader problem relating to the *uninformed, and recipe-like, implementation* of statistical methods that contributes to untrustworthy evidence, in several different ways, the most important of which are the following:

- (a) **Statistical misspecification:** invalid probabilistic assumptions are imposed on the particular data \mathbf{x}_0 , by ignoring the approximate validity (statistical adequacy) of the probabilistic assumptions comprising the statistical model $\mathcal{M}_\theta(\mathbf{x})$. Statistical misspecification is endemic in disciplines like economics where probabilistic assumptions are often assigned to unobservable error term(s), but what matters for the reliability of inference and the trustworthiness of the ensuing evidence is whether the probabilistic assumptions imposed (indirectly) on the observable process $\{X_t, t \in \mathbb{N}\}$ underlying the data \mathbf{x}_0 are valid or not; see Spanos [18].
- (b) **Unwarranted evidential interpretations** of inference results, including: (i) an optimal point estimate $\hat{\theta}(\mathbf{x}_0)$ implies that $\hat{\theta}(\mathbf{x}_0) \simeq \theta^*$ for n large enough, (ii) attaching the coverage probability to observed CIs, and (iii) misinterpreting testing accept/reject H_0 results by detaching them from their particular statistical context (Spanos [17]), and (iv) conflating nominal and actual error probabilities by ignoring Bonferroni-type adjustments stemming from abuses of N-P testing. Focusing on (iv), however, overlooks the forest for the trees, since: “*p* values are just the tip of the iceberg.” (Leek and Peng [39]).
- (c) **Questionable statistical modeling practices** which include (i) foisting an aPriori Postulated (aPP) substantive model $\mathcal{M}_\varphi(\mathbf{z}, \varphi \in \Phi)$ on data \mathbf{Z}_0 using a curve-fitting procedure, and (ii) evaluating its ‘appropriateness’ using goodness-of-fit/prediction measures, without recognizing that the latter is neither necessary nor sufficient for statistical adequacy, which is invariably ignored; see Spanos [18].

In practice, an aPriori Postulated (aPP) (substantive) model should be nested within its implicit statistical model $\mathcal{M}_\theta(\mathbf{z})$ comprising (a) the probabilistic assumptions imposed (often implicitly) on the observable process $\{\mathbf{Z}_t, t \in \mathbb{N}\}$ underlying \mathbf{Z}_0 and (b) the ensuing parametrization $\theta \in \Theta$, providing (c) the crucial link between $\mathcal{M}_\varphi(\mathbf{z})$ and the real-world mechanism that generated \mathbf{Z}_0 . The statistical parameters $\theta \in \Theta$ are framed to relate them to the substantive parameters $\varphi \in \Phi$, relating the two via restrictions, say $\mathbf{g}(\varphi, \theta) = \mathbf{0}$, which need to be tested to ensure that $\mathcal{M}_\varphi(\mathbf{z})$ is substantively adequate vis-a-vis data \mathbf{Z}_0 .

Example 4. Consider the $\mathcal{M}_\varphi(\mathbf{z})$ -Capital Asset Pricing Model (CAPM) where:

$$\begin{aligned} \mathcal{M}_\varphi(\mathbf{z}): (Y_t - x_{2t}) &= \alpha_1(x_{1t} - x_{2t}) + \varepsilon_t, \varphi := (\alpha_1, \sigma_\varepsilon^2), \varepsilon_t \sim \text{NIID}(0, \sigma_\varepsilon^2), t \in \mathbb{N}, \\ \mathcal{M}_\theta(\mathbf{z}): Y_t &= \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t, \theta := (\beta_0, \beta_1, \beta_2, \sigma_u^2), u_t \sim \text{NIID}(0, \sigma_u^2), t \in \mathbb{N}. \end{aligned} \tag{18}$$

where $\mathbf{Z}_t := (Y_t, X_{1t}, X_{2t})$, Y_t -portfolio returns, X_{1t} -market returns, and X_{2t} - returns of a risk free asset, with $\mathcal{M}_\theta(\mathbf{z})$ being the underlying statistical (linear regression) model.

The parameters of the two models are related via the substantive restrictions:

$$\mathbf{g}(\varphi, \theta) = \mathbf{0}: \beta_0 = 0, \beta_1 + \beta_2 = 1. \tag{19}$$

For the estimated $\mathcal{M}_\varphi(\mathbf{z})$ to yield trustworthy evidence, (a) $\mathcal{M}_\theta(\mathbf{z})$ should be statistically adequate for data \mathbf{Z}_0 , and (b) the restrictions in (19) should *not* belie the data \mathbf{Z}_0 , which is seldom the case in published empirical papers; see Spanos [18].

4.1. Statistical Adequacy and Replication in Practice

A particularly important contributor to untrustworthy evidence is the *statistical misspecification* of the invoked $\mathcal{M}_\theta(\mathbf{x})$. Statistically adequate models are (approximately) replicable

with akin data because they are unique. In contrast, there are numerous ways $\mathcal{M}_\theta(\mathbf{z})$ can be statistically misspecified; see Spanos [18]. Indeed, untrustworthy evidence is easy to replicate when practitioners employ the same uninformed and recipe-like, implementation of statistical methods, ignoring their statistical adequacy “... an analysis can be fully reproducible and still be wrong.” (Leek and Peng [39]). This calls into question the Ioannidis [1] stipulations [S1]–[S2], since statistical misspecification is the primary source of untrustworthiness giving rise to (i) ‘non-optimal’ inference procedures, and (ii) inducing sizeable *discrepancies* between the *actual* and *nominal* error probabilities (Section 2.1). These discrepancies render any Bonferroni-type adjustments for p-hacking, data-dredging, multiple testing and cherry-picking, irrelevant. Hence, a crucial precondition for relating replication/non-replication to the trustworthiness/untrustworthiness is to establish the statistical adequacy of the invoked $\mathcal{M}_\theta(\mathbf{x})$ using akin data.

Example 5. Consider two studies based on akin data sets $\mathbf{x}_1, \mathbf{x}_2, n = 100$, invoking the same the simple Normal in (2) $\mathcal{M}_\theta(\mathbf{x})$, giving rise to the following results:

$$X_{1t} = \underset{(0.103)}{3.208} + \hat{u}_{1t}, s_1 = 1.029, \quad X_{2t} = \underset{(0.106)}{3.195} + \hat{u}_{2t}, s_2 = 1.134, \quad (20)$$

with the standard errors in brackets, and $\hat{u}_{it}, i = 1, 2, t = 1, \dots, n$, denoting the residuals.

Taking the results in (20) at face value, Table 6 compares them and their ensuing CIs and N-P tests whose *p*-values are given in square brackets. An informal test of the difference between the two means, $H_0: (\mu_1 - \mu_2) = 0$ vs. $H_1: (\mu_1 - \mu_2) \neq 0$, yields $\tau(\mathbf{x}_1, \mathbf{x}_2) = 0.085[0.932]$, rendering the results of the two studies almost identical. Despite that, it’s not obvious how an umpire could opine whether the results with data \mathbf{x}_2 constitutes a successful replication with trustworthy evidence of the results with \mathbf{x}_1 .

Table 6. Statistical Inference Results.

	Point Estimates	Observed CIs	$H_0: \mu_i = 3.2 \ i = 1, 2$
\mathbf{x}_1 :	$(\bar{x}_1 = 3.208 \ s_1 = 1.029)$	$(3.006, 3.410)$	$\tau(\mathbf{x}_1) = 0.078[0.938]$
\mathbf{x}_2 :	$(\bar{x}_2 = 3.195 \ s_2 = 1.134)$	$(2.973, 3.417),$	$\tau(\mathbf{x}_2) = -0.044[0.965]$

To answer that one needs to evaluate the statistical adequacy of both estimated models in (20) using informal t-plots of the data combined with formal M-S testing (Spanos [19]) to probe their statistical adequacy. The t-plot for data \mathbf{x}_1 (Figure 1) exhibits no obvious departures from the NIID assumptions, but the ‘irregular cycles’ exhibited by \mathbf{x}_2 (Figure 2) indicate a likely departures from ‘Independence’.

The validity of the IID assumptions can be tested using the runs test (Spanos, 2019) [18]:

$$d_R(\mathbf{X}) = \frac{[E(R) - R]}{\sqrt{Var(R)}} \stackrel{\text{IID}}{\approx} N(0, 1), d_R(\mathbf{X}) \stackrel{\text{non-IID}}{\sim} N(\delta_n, \tau_n^2), \delta_n \neq 0, \tau_n^2 > 0, \quad (21)$$

where *R*-number of runs, $E(R) = \frac{2n-1}{3}$, $Var(R) = \frac{16n-29}{90}$, δ_n and τ_n^2 depend only on *n*. The Anderson-Darling (A-D) tests for Normality are also given in Table 7.

Table 7. Simple M-S tests for IID.

Data	Runs Test for IID	Normality Test
\mathbf{x}_1 :	$d_R(\mathbf{x}_1) = \frac{(50-51)}{4.178} = 0.239[0.918]$	$A-D(\mathbf{x}_1) = 0.414[0.330]$
\mathbf{x}_2 :	$d_R(\mathbf{x}_2) = \frac{(32-51)}{4.178} = 4.548[0.000]$	$A-D(\mathbf{x}_2) = 0.177[0.918]$

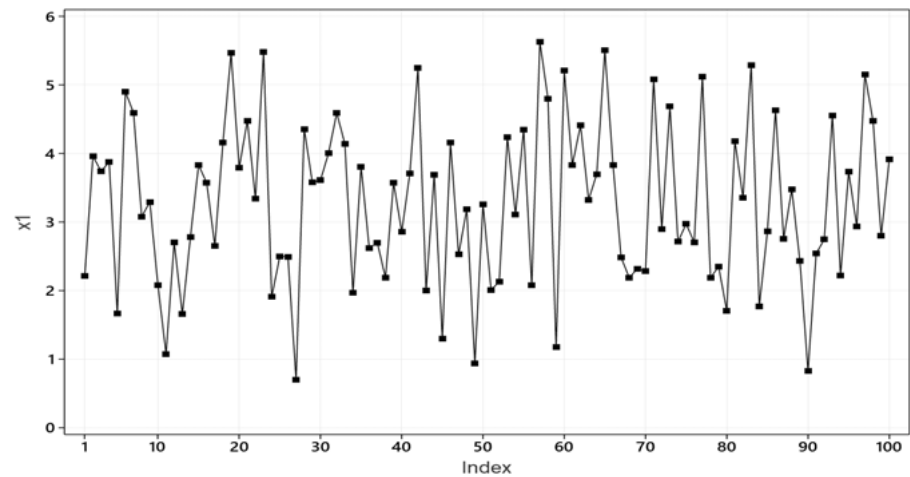


Figure 1. t-plot of data x_1 .

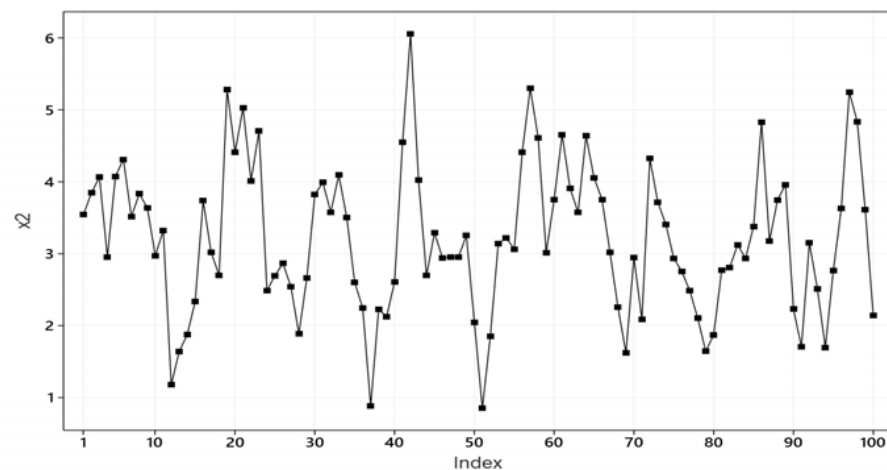


Figure 2. t-plot of data x_2 .

The M-S testing results in Table 7 confirm that the simple Normal model in (2) is statistically adequate for data x_1 , but misspecified with data x_2 since the ‘Independence’ assumption is invalid. This is corroborated by the t-plot of the residuals $\{\hat{u}_{2t}, t = 1, 2, \dots, n\}$ in Figure 3, which exhibits similar irregular cycles as Figure 2. This implies that the inference results in Table 6 based on x_2 will be *unreliable* since the non-independence induces sizeable discrepancies between the actual and nominal error probabilities; see Spanos [18].

Thus, the results based on data x_2 do not represent a successful replication (with trustworthy evidence) of those based on data x_1 . To secure trustworthy evidence for any inferential claim using data x_2 one needs to respecify the original $\mathcal{M}_\theta(x)$ in (2) by selecting an alternative statistical model $\mathcal{M}_\theta(x_2)$ aiming to account for the dependence mirrored by the irregular cycles in Figures 2 and 3.

An obvious choice in this case is to replace the assumptions of IID with Markov dependence and stationarity, which give rise to an Autoregressive [AR(1)] model; see Spanos [18]. Estimating the AR(1) model with data x_2 yields:

$$X_{2t} = 1.331 + 0.583X_{2t-1} + \hat{v}_{2t}, \quad s_2 = 0.874. \quad (22)$$

(0.278) (0.083)

The t-plot of the residuals $\{\hat{v}_{2t}, t = 1, 2, \dots, n\}$ from (22) in Figure 4 indicates that the Markov dependence assumption accounts for the irregular cycles exhibited by the residuals in Figures 2 and 3. The statistical adequacy of (22) is confirmed by the M-S tests:

$$d_R(\mathbf{v}_2) = \frac{(51-50.1)}{4.157} = 0.217[0.853], A-D(\mathbf{v}_2) = 0.438[0.290].$$

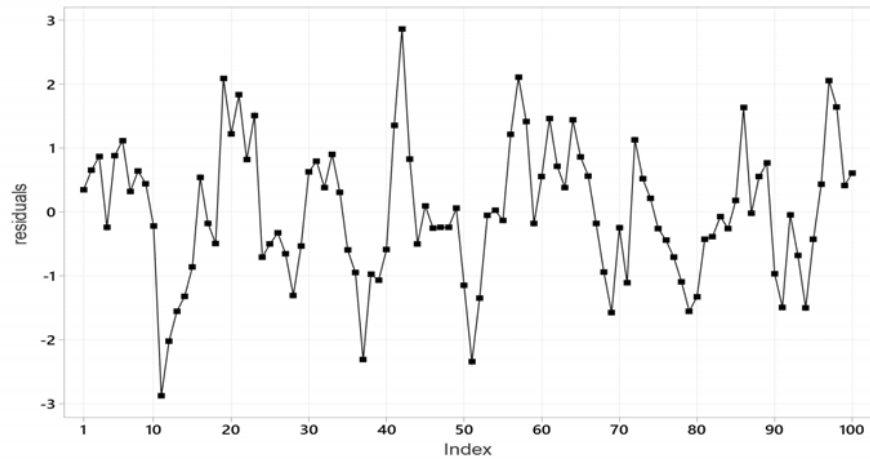


Figure 3. t-plot of data x_2 .

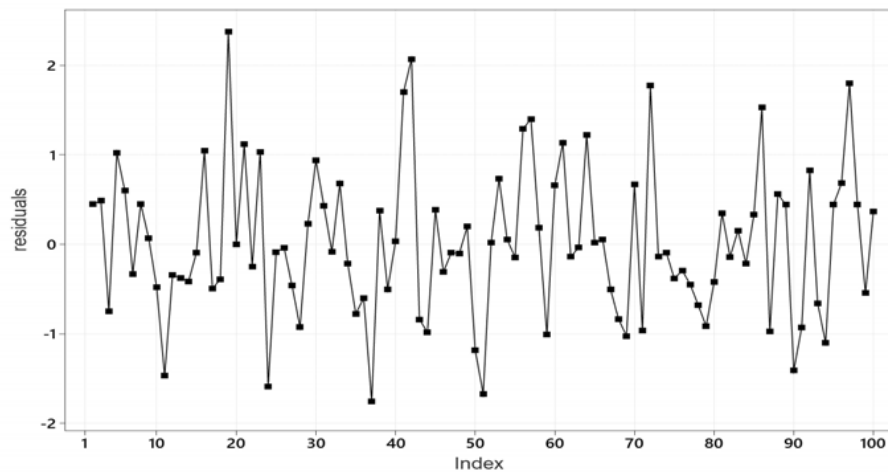


Figure 4. t-plot of data x_2 .

4.2. Fallacious Evidential Interpretations of Inference Results

Another crucial source of the untrustworthiness of evidence is unwarranted and fallacious evidential interpretations of unduly data-specific inference results as evidence for particular inferential claims.

Example 1 (continued). The optimality of the point estimators, $\hat{\mu}(\mathbf{X}) = \bar{X}_n = \frac{1}{n} \sum_{t=1}^n X_t$, $s^2(\mathbf{X}) = \frac{1}{(n-1)} \sum_{t=1}^n (X_t - \bar{X}_n)^2$, does not entail:

$$\hat{\mu}(\mathbf{x}_0) \simeq \mu^*, \text{ and } s^2(\mathbf{x}_0) \simeq \sigma_*^2, \text{ for } n \text{ large enough,} \tag{23}$$

where \simeq indicates approximate equality. The claims in (23) are unwarranted since $\hat{\mu}(\mathbf{x}_0)$ and $s^2(\mathbf{x}_0)$ ignore the relevant uncertainty associated with $\hat{\mu}(\mathbf{x}_0)$ and $s^2(\mathbf{x}_0)$ rely exclusively on single point, $\mathbf{X} = \mathbf{x}_0$ from the sampling distributions, $f(\hat{\mu}(\mathbf{x}); \theta)$ and $f(s^2(\mathbf{x}); \theta)$, $\forall \mathbf{x} \in \mathbb{R}^n$; see Spanos [16]. Invoking asymptotic properties for $\hat{\mu}(\mathbf{X})$ and $s^2(\mathbf{X})$, such as consistency, will not justify the claims in (23). As argued by Le Cam [40] p. xiv: "... limit theorems 'as n tends to infinity' are logically devoid of content about what happens at any particular n ".

(b) It is important to emphasize that the problem with the claim $\hat{\theta}(\mathbf{x}_0) \simeq \theta^*$ also extends to estimation-based effect sizes; see Ellis [41]. For instance, in the case of testing the differences between two means, Cohen's $d = \frac{(\bar{x}_n - \bar{y}_n)}{s}$ is nothing more than

an estimate $h(\mathbf{z}_0) = \frac{(\bar{x}_n - \bar{y}_n)}{s}$ relating to the estimator $h(\mathbf{Z}) = \frac{(\bar{X}_n - \bar{Y}_n)}{s}$ of $h(\boldsymbol{\theta}) = \frac{(\mu_1 - \mu_2)}{\sigma}$

In that sense, the claim that $\frac{(\bar{x}_n - \bar{y}_n)}{s} \simeq \frac{(\mu_1^* - \mu_2^*)}{\sigma^*}$ is a variation on the same unwarranted claim: $\hat{\theta}(\mathbf{x}_0) \simeq \theta^*$ for n large enough; see Spanos [16]

(c) The inferential claim associated with (5) relates to the random CI($\mathbf{X}; \mu$) overlaying μ^* with probability $(1 - \alpha)$, but does not extend to $CI(\mathbf{x}_0)$ Spanos [16]:

$$CI(\mathbf{x}_0) = [\bar{x}_n - c_{\frac{\alpha}{2}}(s(\mathbf{x}_0)/\sqrt{n}), \bar{x}_n + c_{\frac{\alpha}{2}}(s(\mathbf{x}_0))]. \tag{24}$$

As Neyman [42], p. 288, argued: "... valid probability statements about random variables usually cease to be valid if the random variables are replaced by their particular values". i.e., the factual reasoning makes no sense post-data [\mathbf{x}_0 has occurred].

(d) It is also well-known that interpreting the 'accept/reject H_0 results' as evidence for H_0/H_1 is unsound giving rise to two related fallacies (Mayo and Spanos [14]):

Fallacy of acceptance: misinterpreting 'accept H_0 ' (no evidence against H_0) as evidence for H_0 '. This could easily arise in cases where n is too small and the test has no sufficient power to detect a actual discrepancy $\gamma_1 = \theta_0 - \theta_1 \pm 0$.

Fallacy of rejection: misinterpreting 'reject H_0 ' (evidence against H_0) as evidence for a particular H_1 '. This could easily arise when n is very large and the test is sensitive enough to detect tiny discrepancies; see Spanos [43].

The above arguments call into question the M-PPV-inspired reforms relating to replacing p -values with CIs and effect sizes, and redefining statistical significance by reducing the conventional thresholds for α , as ill-thought-out.

5. Testing Results vs. the Post-Data Severity (SEV) Evaluation

The reason why the accept/reject H_0 results are not routinely replicable with akin data is that they are sensitive to (i) the framing of H_0 and H_1 , (ii) the prespecified α and (iii) the sample size n . A principled procedure transmuted the unduly data-dependent accept/reject H_0 results into evidence relating to θ^* is the post-data severity (SEV) evaluation, guided by the sign and magnitude of $d(\mathbf{x}_0)$. A hypothesis H (H_0 or H_1) passes a severe test T_α with data \mathbf{x}_0 if (Mayo and Spanos [14]):

- (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.

5.1. Case 1: Reject H_0

Example 6. Consider the simple Bernoulli model:

$$X_k \sim \text{BerIID}(\theta, \theta(1-\theta)), x_k=0, 1, 0 < \theta < 1, k \in \mathbb{N}, \tag{25}$$

where $\theta = E(X_k) = \mathbb{P}(X_k = 1)$, and the hypotheses:

$$H_0: \theta \leq \theta_0 \text{ vs. } H_1: \theta > \theta_0, \text{ for } \theta_0 = 0.5 \tag{26}$$

It can be shown that the t-type test (Lehmann and Romano, [20]):

$$T_\alpha^> := \{d(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}}, C_1(\alpha) = \{\mathbf{x}: d(\mathbf{x}) > c_\alpha\}, \text{ for } \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \tag{27}$$

is optimal in the sense of Uniformly Most Powerful (UMP). The sampling distribution of $d(\mathbf{X})$ evaluated under H_0 (hypothetical: what if $\theta = \theta_0$) is:

$$d(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \stackrel{\theta=\theta_0}{\simeq} \text{Bin}(0, 1; n) \simeq \text{N}(0, 1), \tag{28}$$

where the ‘scaled’ Binomial distribution, $\text{Bin}(0, 1; n)$, can be approximated (\simeq) by the $N(0, 1)$, which is used to evaluate α and the p -value:

$$\alpha = \mathbb{P}(d(\mathbf{X}) > c_\alpha; \theta = \theta_0), \quad p(\mathbf{x}_0) = \mathbb{P}(d(\mathbf{X}) > d(\mathbf{x}_0); \theta = \theta_0). \tag{29}$$

The sampling distribution of $d(\mathbf{X})$ under H_1 (hypothetical: what if $\theta = \theta_1$) is:

$$d(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \stackrel{\theta = \theta_1}{\simeq} \text{Bin}\left(\delta(\theta_1), \sqrt{V(\theta_1)}; n\right), \text{ for } \theta_1 = \theta_0 + \gamma_1, \gamma_1 \geq 0, \tag{30}$$

$$\delta(\theta_1) = \frac{\sqrt{n}(\theta_1 - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}}, \quad V(\theta_1) = \frac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)},$$

whose probabilities can be approximated using:

$$(\sqrt{V(\theta_1)})^{-1}[d(\mathbf{X}) - \delta(\theta_1)] \stackrel{\theta = \theta_1}{\simeq} \text{Bin}(0, 1; n) \simeq N(0, 1). \tag{31}$$

(31) is used to derive the type II error and the power of the test $T_\alpha^>$ in (27) which increases monotonically with \sqrt{n} and $(\mu_1 - \mu_0)$ and decreases with $V(\theta_1)$.

Example 6 (continued). For the simple Bernoulli model in (25), the relevant data \mathbf{x}_0 refer to newborns for 2020 in Cyprus, 5190 male ($X = 1$) and 4740 female ($X = 0$), i.e., $n = 9930$. The optimal test $T_\alpha^>$ in (27), based on $\alpha = 0.001$ and $\hat{\theta}_n(\mathbf{x}_0) = 0.523$, yields:

$$d(\mathbf{x}_0) = 4.5158, \text{ indicating ‘reject } H_0 \text{’ with } p(\mathbf{x}_0) = 0.000003.$$

This, combined with $d(\mathbf{x}_0) > 0$, suggest that condition C-1 implies that \mathbf{x}_0 accords with H_1 , and condition C-2 indicates that the relevant event relates to: “outcomes \mathbf{x} that accord less well with $\theta > \theta_1$ than \mathbf{x}_0 does”, i.e., event $\{\mathbf{x}: d(\mathbf{x}) \leq d(\mathbf{x}_0)\}, \forall \mathbf{x} \in \{0, 1\}^n$, and its probability relates to the inferential claim $\theta > \theta_1 = \theta_0 + \gamma_1, \gamma_1 > 0$:

$$SEV_R(T_\alpha^>; \theta > \theta_1) = \sup_{\theta_1 \in \Theta_1} \mathbb{P}(d(\mathbf{X}) \leq d(\mathbf{x}_0); \theta = \theta_1) = p_1, \tag{32}$$

where $\Theta_1 = (0.5, 1), p_1 > 0.5$ is a prespecified (high enough) probability. The evaluation of (32), based on (31) gives rise to the severity curve depicted in Figure 5, assigning probabilities relating to the warrant with data \mathbf{x}_0 of different discrepancies $\gamma_1 = (\theta_1 - \theta_0) > 0$.

The largest discrepancy γ warranted with severity 0.9 by test $T_\alpha^>$ and data \mathbf{x}_0 is $\gamma^\ddagger \leq 0.01623$ ($\theta_1^\ddagger = 0.51623$); see Table 8. How does the post-data SEV evaluation convert the unduly data-specific accept/reject H_0 results into evidence, giving rise to learning from data \mathbf{x}_0 about θ^* ? While the SEV evaluation is always attached to $T_\alpha^>$ in (27), as it relates to the relevant inferential claim $\theta > \theta_1, \gamma^\ddagger \leq 0.01623$ ($\theta_1^\ddagger = 0.51623$) warranted with probability 0.9, could be viewed as narrowing down the coarse accept/reject H_0 result indicating that $\theta^* \in (0.5, 1)$ to the much narrower $\theta^* \in (0.516 \pm \varepsilon)$ for $0 < \varepsilon \leq 0.0002$.

Table 8. $SEV(T_\alpha^>; \theta := \theta_1)$ for ‘reject $\theta_0 = 0.5$ ’ and ‘accept $\theta_0 = (18/35)$ ’ with $(T_\alpha^>; \mathbf{x}_0)$.

$\theta_0 = 0.5: \gamma_1 =$	0.01	0.013	0.015	0.0155	0.01623	0.017	0.018	0.02265	0.025	0.03
$\theta_0 = (\frac{18}{35}): \gamma_1 =$	***	***	***	0.0012	0.0019	0.0027	0.0037	0.0084	0.011	0.016
$\theta_1 =$	0.51	0.513	0.515	0.5155	0.51623	0.517	0.518	0.52265	0.525	0.53
$Sev(\theta := \theta_1) =$	0.994	0.973	0.937	0.923	0.900	0.870	0.824	0.500	0.320	0.071

*** indicate “no entry”.

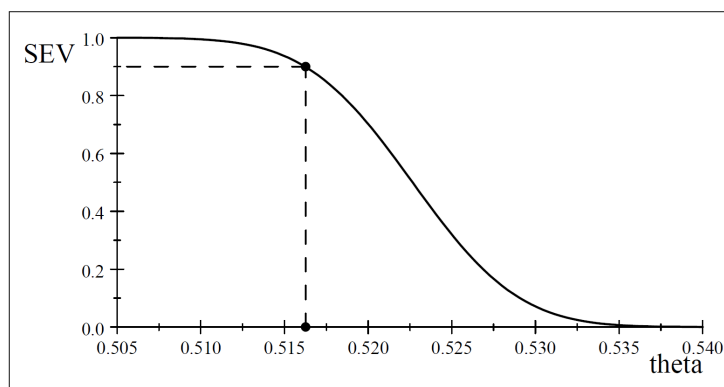


Figure 5. Severity curve for data \mathbf{x}_0 .

5.2. Case 2: Accept H_0

Example 6 (continued). Consider replacing $\theta_0 = 0.5$ in (26) with the Nicholas Bernoulli value $\theta_0 = (18/35) \simeq 0.5143$. Applying the UMP test $T_\alpha^>$ yields:

$$d_B(\mathbf{x}_0) = 1.669, \text{ with a } p\text{-value } p(\mathbf{x}_0) = 0.0476,$$

indicating ‘accept H_0 ’ at $\alpha = 0.001$, but given $d_B(\mathbf{x}_0) > 0$, the relevant $SEV_A(T_\alpha^>; \theta > \theta_1)$ for the inferential claim, $\theta > \theta_1 = \theta_0 + \gamma_1, \gamma_1 > 0$, is:

$$SEV_A(T_\alpha^>; \theta > \theta_1) = \inf_{\theta_1 \in \Theta_1} \mathbb{P}(d(\mathbf{X}) \leq d(\mathbf{x}_0); \theta = \theta_1) = p_1 > 0.5, \tag{33}$$

whose evaluation is based on (31), which is identical to that of rejecting H_0 based on $\theta_0 = 0.5$, in Figure 5, despite the change of the result from reject to accept $\theta_0 = (18/35)$! Indeed, comparing the results for $SEV_R(T_\alpha^>; \theta > \theta_1)$ and $SEV_A(T_\alpha^>; \theta := \theta_1)$ in Table 8, is clear that even though the discrepancies γ_1 are different ($\theta_0 = 0.5, \theta_0 = (18/35)$), the two curves are identical $\forall \theta_1 \in (0.5, 1)$. As a result, for the same inferential claim, $\theta := \theta_1 = \theta_0 + \gamma_1, \gamma_1 := 0$, the largest discrepancy $\gamma_1 := 0$ at 0.9 probability from $\theta_0 = 0.5$ (reject H_0) coincides with the smallest discrepancy γ_1 from $\theta_0 = (18/35)$ (accept H_0). Also, the low severity (0.5) for the point estimate $\hat{\theta}_n(\mathbf{x}_0) = 0.52266$ ensures that $\gamma = 0.02266$ will never be a warranted discrepancy, i.e., there is strong evidence against the claim $\hat{\theta}_n(\mathbf{x}_0) \simeq 0.52266$.

What renders the SEV evaluation different is that it calibrates the warranted γ_1^\ddagger , with high enough probability. In that sense, it can be viewed as a **testing-based effect size** that provides a more reliable evaluation of the ‘scientific effect’; see Spanos [16]. This should be contrasted with the estimation-based effect size measure, $(\hat{\theta}_n(\mathbf{x}_0) - 0.5) / \sqrt{0.5(1 - 0.5)} = 0.045$ (Ellis, 2010) [41], which is equally fallacious as that claim that $\hat{\theta}_n(\mathbf{x}_0) \simeq \theta^*$ for n large enough.

Statistical vs. substantive significance. The SEV evaluation would also address this problem by relating the discrepancy γ^\ddagger from θ_0 ($\theta_1 = \theta_0 \pm \gamma^\ddagger$) warranted by test T_α and data \mathbf{x}_0 with high probability to the substantively determined value φ^\blacklozenge . For example, in human biology (Hardy, [44]) it is known that the *substantive value* for the ratio of boys to all newborns is $\varphi^\blacklozenge \simeq 0.5122$. Comparing φ^\blacklozenge with the severity-based warranted discrepancy, $\gamma_1^\ddagger \leq 0.01556$ ($\theta_1 \leq 0.51623$) suggests that the statistically determined γ_1^\ddagger includes the substantive value since $\varphi^\blacklozenge \simeq 0.5122 < 0.51623$. In fact, the SEV evaluation of $\theta_1 \leq 0.5122$ is 0.981.

5.3. Post-Data Severity and the Trustworthiness of Evidence

The above discussion suggests that similar point estimates, observed CIs, and testing results do not guarantee a successful replication or/and trustworthy evidence. Although the statistical adequacy of the underlying statistical models is necessary, it is not sufficient, especially when the sample size n of such studies is different.

An obvious way to address this issue is to use the distinction between statistical results and evidence and compare the warranted discrepancies γ from the null value, $\theta = \theta_0$ by the tests in question with high enough severity. To illustrate, consider an example that utilizes similar data from two different countries, three centuries apart.

Example 6 (continued). To bring out the replicability of the SEV evidential claims, consider data \mathbf{x}_1 that refer to newborns during 1668 in London (England), 6073 boys, 5560 girls, $n = 11,633$ (Arbuthnot [45]), 352 years before the Cyprus 2020 data. The optimal test in (27) yields $d(\mathbf{x}_1) = 4.756$, with $p(\mathbf{x}_1) = 0.0000001$, rejecting $H_0: \theta \leq 0.5$. This result is close to $d(\mathbf{x}_0) = 4.5158$, with $p(\mathbf{x}_0) = 0.000003$ from Cyprus for 2020, but the question of interest is whether the latter constitutes a successful replication of the 1668 data. Using $SEV(T_\alpha^>; \mathbf{x}_1; \theta > \theta_1) = 0.9$, the warranted discrepancy from $\theta_0 = 0.5$ by test $T_\alpha^>$ and London data \mathbf{x}_1 is $\gamma_1^\ddagger \leq 0.0161$ which is almost identical to that with the Cyprus data $\gamma_0^\ddagger \leq 0.0162$, affirming the replicability of the evidence that $\theta^* \in (0.5, 0.516]$.

The two severity curves in Figure 6 for data \mathbf{x}_0 and \mathbf{x}_1 are almost identical for the range of values of θ that matters ($SEV(T_\alpha^>; \mathbf{x}_0; \theta > \theta_1) > 0.8$), despite the fact that $n = 9930$ vs. $n = 11,633$. Hence, for a statistically adequate $\mathcal{M}_\theta(\mathbf{x})$, the SEV could provide a more robust measure of replicability of trustworthy evidence, than point estimates, effect sizes, observed CIs, or p -values; see Spanos [16].

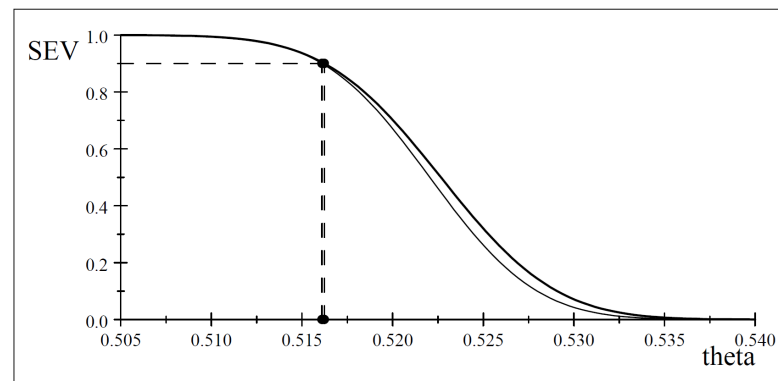


Figure 6. Severity curves for data \mathbf{x}_0 and \mathbf{x}_1 .

The key to enabling the $SEV(T_\alpha^>; \mathbf{x}_0; \theta > \theta_1)$ evaluation to circumvent the problem with different sample sizes n is that the two components in $[d(\mathbf{x}_0) - \delta(\theta_1)]$ use the same n to evaluate the discrepancy $\gamma_1 = \theta_1 - \theta_0 \geq 0$, addressing the above-mentioned fallacies of acceptance and rejection. In contrast, the power of $T_\alpha^>$ is evaluated using $[c_\alpha - \delta(\theta_1)]$, where c_α is based on the prespecified α , and the p -value is evaluated only under $\theta = \theta_0$, $d(\mathbf{X}) \stackrel{\theta = \theta_0}{\sim} N(0,1)$, rendering both evaluations vulnerable to the large/small n problems.

The SEV can be used to address other foundational problems, including distinguishing between statistical and substantive significance, as well as providing a testing-based effect size for the magnitude of the ‘substantive’ effect. An example of that is illustrated in Spanos (2023) [17], where a published paper reports a t-test for an estimated regression coefficient $\hat{\beta} = 0.004$ claimed to be significant at $\alpha = 0.05$ with $n = 24,732,966$ and a p -value $p = 0.045$. The SEV evaluation of that claim indicates that the warranted magnitude of the coefficient would be considerably smaller at $\beta \leq 0.00000001$ with high enough severity.

6. Summary and Conclusions

The case made by Ioannidis [1], based on M-PPV misrepresents frequentist testing and misdiagnoses the replication crisis, due to its incongruous stipulations [S1]–[S2] based on highly questionable presuppositions [a]–[d]. The most ill-chosen incongruousness is the colligating of the false positive/negative with the type I/II error probabilities. This led to

promoting ill-informed reforms, including replacing p -values with CIs and effect sizes and lowering α .

The above discussion has made a case that the non-replicability and the untrustworthiness of empirical evidence stem primarily from the uninformed and recipe-like implementation of statistical modeling and inference. This ignores fundamental issues relating to (a) establishing the statistical adequacy of the invoked statistical model $\mathcal{M}_\theta(\mathbf{x})$ for data \mathbf{x}_0 , (b) the reasoned implementation of frequentist inference and the pertinent interpretation of their error probabilities, as well as (c) warranted evidential interpretations of inference results. Fisher's model-based statistics provides an incisive explanation for the non-replication/untrustworthiness of empirical evidence and an appropriate framework for addressing (a)–(c). The statistical adequacy of $\mathcal{M}_\theta(\mathbf{x})$ is *necessary* for securing the trustworthiness of empirical evidence and ensuring the crucial link between trustworthiness and replicability, as both are key aspects firmly attached to individual studies. The inveterate problem with frequentist testing is that the accept/reject H_0 results are too coarse and unduly data-specific to provide cogent evidence for θ^* . This problem can be addressed using their post-data SEV evaluation to output the discrepancy $\gamma_1 = (\theta_1 - \theta_0)$ warranted by data \mathbf{x}_0 with high probability, which narrows down the coarseness of these results. From this perspective, abuses of frequentist testing represents the tip of the untrustworthy evidence iceberg, with the remainder stemming mainly from statistical/substantive misspecification.

As Stark and Saltelli [12] aptly argue: “The problem is one of cargo-cult statistics – the ritualistic miming of statistics rather than conscientious practice. This has become the norm in many disciplines, reinforced and abetted by statistical education, statistical software, and editorial policies” (p. 40).

A redeeming value of the Ioannidis [1] case might be that its provocative title raised awareness of the endemic untrustworthiness of published empirical evidence. This could potentially initiate a fertile dialogue leading to ameliorating this thorny problem. This could happen when the focus is redirected to the reasoned implementation of frequentist modeling and inference guided by statistical adequacy. That is, replicating influential empirical papers in a particular research area with the same or akin data, and evaluating their statistical and substantive adequacy, represents the most effective way to establish the untrustworthiness of published empirical evidence.

The endemic untrustworthiness of published empirical evidence in most disciplines could be redressed but would require a lot of difficult changes, including: (i) a substantial overhaul of the teaching of probability theory and statistics, (ii) crucial changes in the current statistical software to include M-S testing, and (iii) changes in the editorial policies for papers that rely on statistical inference requiring corroboration of the statistical adequacy for the invoked statistical models.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are publicly available.

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

Abbreviations

The following abbreviations are used in this manuscript:

PPV	Positive Predictive Value
M-PPV	Metamorphosed PPV
MDS	Medical Diagnostic Screening

NIID	Normal, Independent and Identically Distributed
M-S	Mis-Specification
N-P	Neyman–Pearson
UMP	Uniformly Most Powerful
CAPM	Capital Asset Pricing Model
SEV	post-data severity evaluation

References

- Ioannidis, J.P.A. Why most published research findings are false. *PLoS Med.* **2005**, *2*, e124. [[CrossRef](#)] [[PubMed](#)]
- Baker, M. Reproducibility crisis. *Nature* **2016**, *533*, 353–366.
- Hoffler, J.H. Replication and Economics Journal Policies. *Am. Econ. Rev.* **2017**, *107*, 52–55. [[CrossRef](#)]
- Wasserstein, R.L.; Schirm, A.L.; Lazar, N.A. Moving to a world beyond ‘ $p < 0.05$ ’. *Am. Stat.* **2019**, *73*, 1–19.
- Simmons, J.P.; Nelson, L.D.; Simonsohn, U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allow Presenting Anything as Significant. *Psychol. Sci.* **2011**, *22*, 1359–1366. [[CrossRef](#)]
- Camerer, C.F.; Dreber, A.; Forsell, E.; Ho, T.H.; Huber, J.; Johannesson, M.; Kirchler, M.; Almenberg, J.; Altmejd, A.; Chan, T.; et al. Evaluating replicability of laboratory experiments in economics. *Science* **2016**, *351*, 1433–1436. [[CrossRef](#)]
- 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](#)]
- Shrout, P.E.; Rodgers, J.L. Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annu. Rev. Psychol.* **2018**, *69*, 487–510. [[CrossRef](#)]
- Johnson, V.E.; Payne, R.D.; Wang, T.; Asher, A.; Mandal, S. On the Reproducibility of Psychological Science. *J. Am. Stat. Assoc.* **2017**, *112*, 1–10. [[CrossRef](#)]
- Andreou, E.; Spanos, A. Statistical adequacy and the testing of trend versus difference stationarity. *Econom. Rev.* **2003**, *22*, 217–237. [[CrossRef](#)]
- Do, H.P.; Spanos, A. Revisiting the Phillips curve: The empirical relationship yet to be validated. *Oxf. Bull. Econ. Stat.* **2024**, *86*, 761–794. [[CrossRef](#)]
- Stark, P.B.; Saltelli, A. Cargo-cult statistics and scientific crisis. *Significance* **2018**, *15*, 40–43. [[CrossRef](#)]
- Fisher, R.A. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. A* **1922**, *222*, 309–368.
- 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](#)]
- Nickerson, R.S. Null Hypothesis Significance Testing: A review of an old and continuing controversy. *Psychol. Methods* **2000**, *5*, 241–301. [[CrossRef](#)] [[PubMed](#)]
- Spanos, A. Revisiting noncentrality-based confidence intervals, error probabilities, and estimation-based effect sizes. *J. Math. Stat. Psychol.* **2021**, *104*, 102580. [[CrossRef](#)]
- Spanos, A. Revisiting the Large n (Sample Size) Problem: How to Avert Spurious Significance Results. *Stats* **2023**, *6*, 1323–1338. [[CrossRef](#)]
- Spanos, A. *Probability Theory and Statistical Inference: Empirical Modeling with Observational Data*; Cambridge University Press: Cambridge, UK, 2019.
- Spanos, A. Mis-Specification Testing in Retrospect. *J. Econ. Surv.* **2018**, *32*, 541–577. [[CrossRef](#)]
- Lehmann, E.L.; Romano, J.P. *Testing Statistical Hypotheses*; Springer: New York, NY, USA, 2005.
- Neyman, J.; Pearson, E.S. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. A* **1933**, *231*, 289–337.
- Bennett, J.H. *Statistical Inference and Analysis: Selected Correspondence of RA Fisher*; Clarendon Press: Oxford, UK, 1990.
- Gigerenzer, G. The superego, the ego, and the id in statistical reasoning. In *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*; Psychology Press: London, UK, 1993; pp. 311–339.
- Halpin, P.F.; Stam, J.H. Inductive Inference or Inductive Behavior: Fisher and Neyman: Pearson Approaches to Statistical Testing in Psychological Research (1940–1960). *Am. J. Psychol.* **2006**, *119*, 625–653. [[CrossRef](#)]
- Fisher, R.A. Properties of Hh functions. In *Introduction to the British Association of Mathematical Tables*; British Association: London, UK, 1931; Volume 1, p. 26.
- Fisher, R.A. *The Design of Experiments*; Oliver and Boyd: Edinburgh, UK, 1935.
- Box, J.F. R.A. Fisher, *The Life of a Scientist*; Wiley: New York, NY, USA, 1978.
- Yerushalmy, J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Rep. (1896–1970)* **1947**, *62*, 1432–1449. [[CrossRef](#)]
- Neyman, J. Outline of statistical treatment of the problem of diagnosis. *Public Health Rep. (1896–1970)* **1947**, *62*, 1449–1456. [[CrossRef](#)]

30. Spanos, A. Is Frequentist Testing Vulnerable to the Base-Rate Fallacy? *Philos. Sci.* **2010**, *77*, 565–583. [[CrossRef](#)]
31. Bishop, Y.V.; Fienberg, S.E.; Holland, P.W. *Discrete Multivariate Analysis*; MIT Press: Cambridge, MA, USA, 1975.
32. Senn, S. *Statistical Issues in Drug Development*, 3rd ed.; Wiley: New York, NY, USA, 2021.
33. Greenhouse, S.W.; Mantel, N. The evaluation of diagnostic tests. *Biometrics* **1950**, *6*, 399–412. [[CrossRef](#)]
34. Nissen-Meyer, S. Evaluation of screening tests in medical diagnosis. *Biometrics* **1964**, *20*, 730–755. [[CrossRef](#)]
35. Freedman, D.; Pisani, R.; Purves, R. *Statistics*, 3rd ed.; Norton: New York, NY, USA, 1998.
36. Nosek, B.A.; Hardwicke, T.E.; Moshontz, H.; Allard, A.; Corker, K.S.; Dreber, A.; Fidler, F.; Hilgard, J.; Struhl, M.K.; Nuijten, M.B.; et al. Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.* **2022**, *73*, 719–748. [[CrossRef](#)]
37. Munafò, M.R.; Nosek, B.A.; Bishop, D.V.M.; Button, K.S.; Chambers, C.D.; Percie du Sert, N.; Simonsohn, U.; Wagenmakers, E.J.; Ware, J.J.; Ioannidis, J.P.A. A manifesto for reproducible science. *Nat. Hum. Behav.* **2017**, *1*, 1–9. [[CrossRef](#)]
38. Bird, A. Understanding the Replication Crisis as a Base Rate Fallacy. *Br. J. Philos. Sci.* **2021**, *72*, 965–993. [[CrossRef](#)]
39. Leek, J.T.; Peng, R.D. Statistics: P values are just the tip of the iceberg. *Nature* **2015**, *520*, 520–612. [[CrossRef](#)]
40. Le Cam, L. *Asymptotic Methods in Statistical Decision Theory*; Springer: New York, NY, USA, 1986.
41. Ellis, P.D. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*; CUP: Cambridge, UK, 2010.
42. Neyman, J. Note on an article by Sir Ronald Fisher. *J. R. Stat. Ser. B* **1956**, *18*, 288–294. [[CrossRef](#)]
43. Spanos, A. How the Post-Data Severity Converts Testing Results into Evidence for or against Pertinent Inferential Claims. *Entropy* **2024**, *26*, 95. [[CrossRef](#)] [[PubMed](#)]
44. Hardy, I.C.W. (Ed.) *Sex Ratios: Concepts and Research Methods*; Cambridge University Press: Cambridge, UK, 2002.
45. 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.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.