

Article

Revisiting the Large n (Sample Size) Problem: How to Avert Spurious Significance Results

Aris Spanos 

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

Abstract: Although large data sets are generally viewed as advantageous for their ability to provide more precise and reliable evidence, it is often overlooked that these benefits are contingent upon certain conditions being met. The primary condition is the approximate validity (statistical adequacy) of the probabilistic assumptions comprising the statistical model $\mathcal{M}_\theta(\mathbf{x})$ applied to the data. In the case of a statistically adequate $\mathcal{M}_\theta(\mathbf{x})$ and a given significance level α , as n increases, the power of a test increases, and the p -value decreases due to the inherent trade-off between type I and type II error probabilities in frequentist testing. This trade-off raises concerns about the reliability of declaring ‘statistical significance’ based on conventional significance levels when n is exceptionally large. To address this issue, the author proposes that a principled approach, in the form of post-data severity (SEV) evaluation, be employed. The SEV evaluation represents a post-data error probability that converts unduly data-specific ‘accept/reject H_0 results’ into evidence either supporting or contradicting inferential claims regarding the parameters of interest. This approach offers a more nuanced and robust perspective in navigating the challenges posed by the large n problem.

Keywords: large n problem; Neyman–Pearson testing; p -value; post-data severity evaluation; spurious statistical significance

JEL Classification: C12; C18; C51; C52; C55



Citation: Spanos, A. Revisiting the Large n (Sample Size) Problem: How to Avert Spurious Significance Results. *Stats* **2023**, *6*, 1323–1338. <https://doi.org/10.3390/stats6040081>

Academic Editor: Wei Zhu

Received: 9 November 2023

Revised: 26 November 2023

Accepted: 30 November 2023

Published: 5 December 2023



Copyright: © 2023 by the author. 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

The recent availability of big data with large sample sizes (n) in many scientific disciplines brought to the surface an old and largely forgotten foundational issue in frequentist testing known as the ‘large n problem’. It relates to inferring that an estimated parameter is statistically significant based on conventional thresholds, $\alpha = 0.1, 0.05, 0.025, 0.01$, when n is very large, say $n > 100,000$.

1.1. A Brief History of the Large n Problem

As early as 1938, Berkson [1] brought up the large n problem by pointing out its effect on the p -value:

“... when the numbers in the data are quite large, the P’s [p -values] tend to come out small... if the number of observations is extremely large—for instance, on the order of 200,000—the chi-square P will be small beyond any usual limit of significance. ... If, then, we know in advance the P that will result..., it is no test at all.” (p. 527)

In 1935 Fisher [2] pinpointed the effect of the large n problem on the power (he used the term *sensitivity*) of the test:

“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)

In 1942 Berkson [3] returned to the large n problem by arguing that it also affects the power of a test, rendering it relevant for a sound evidential interpretation:

“In terms of the Neyman–Pearson (N-P) formulation they have different powers for any particular alternative, and hence are likely to give different results in any particular case.” (p. 334)

One of the examples he uses to make his case is from Fisher’s book [4] which is about testing the linearity assumption of a Linear Regression (LR) model.

Unfortunately, the example allowed Fisher [5] to brush aside the broader issues of misinterpreting frequentist tests as evidence for or against hypotheses. In his response, he focuses narrowly on the particular example and recasts their different perspectives on frequentist testing as a choice between ‘objective tests of significance’ and ‘subjective impressions based on eyeballing’ a scatterplot:

“He has drawn the graph. He has applied his statistical insight and his biological experience to its interpretation. He enunciates his conclusion that ‘on inspection it appears as straight a line as one can expect to find in biological material.’ The fact that an objective test had demonstrated that the departure from linearity was most decidedly significant is, in view of the confidence which Dr. Berkson places upon subjective impressions, taken to be evidence that the test of significance was misleading, and therefore worthless.” (p. 692).

In his reply, Berkson [6], (p. 243), reiterated the role of the power:

“When with such specific tests one has sufficient numbers, they become sensitive tests; in the terminology of Neyman and Pearson they become ‘powerful’.”

As a result of the exchange between Berkson and Fisher, the broader issue of ‘inference results’ vs. ‘evidence’ was put aside by the statistics literature until the late 1950s. Berkson’s example from Fisher [4] was an unfortunate choice since the test in question is not a N-P type test, which probes within the boundaries of the invoked statistical model, $\mathcal{M}_\theta(\mathbf{x})$, by framing the hypotheses of interest in terms of its unknown parameters θ . It is a misspecification test, which probes outside the boundaries of $\mathcal{M}_\theta(\mathbf{x})$ to evaluate the validity of its assumptions—the linearity in this case; see Spanos [7].

In 1957 Lindley [8] presented the large n problem as a conflict between frequentist and Bayesian testing by arguing that the p -value will reject $H_0: \theta = \theta_0$ as n increases, whereas the Bayes factor will favor H_0 . Subsequently, this became known as the Jeffreys–Lindley paradox (Spanos [9]) for further discussion.

In 1958 Lehmann [10] raised the issue of the sample size influencing the power of a test, and proposed decreasing α as n increases to counter-balance the increase in its power:

“By habit, and because of the convenience of standardization in providing a common frame of reference, these values [$\alpha = 0.1, 0.05, 0.025, 0.01$] became entrenched as conventional levels to use. This is unfortunate since the choice of significance level should also take into consideration the power that the test will achieve against alternatives of interest. There is little point in carrying out an experiment that has only a small chance of detecting the effect being sought when it exists. Surveys by Cohen [11] and Freiman et al. [12] suggest that this is, in fact, the case for many studies.” (Lehmann [13], pp. 69–70).

Interestingly, the papers cited by Lehmann [13] have been published in psychology and medical journals, respectively, indicating that practitioners in certain disciplines became aware of the small/large n problems and began exploring solutions. Cohen [11,14], was particularly influential in convincing psychologists to supplement, or even replace the p -value, with a statistic that is free of n , known as the ‘effect size’, to address the large n problem. For instance, Cohen’s $d = \frac{(\bar{x}_n - \bar{y}_n)}{s}$ is the recommended effect size when one is testing the difference between two means based on the test statistic $\frac{\sqrt{N}[(\bar{X}_n - \bar{Y}_n) - \gamma]}{s}$ in testing $H_0: \gamma = 0$ vs. $H_1: \gamma \neq 0$; see Lehmann [13].

In 1982 Good [15], p. 66, went a step further than Lehmann [10] to propose a rule of thumb: “*p*-values ... should be standardized to a fixed sample size, say $N = 100$, by replacing $P[p_n(\mathbf{x}_0)]$ with $\min(0.5, [p_n(\mathbf{x}_0) \cdot \sqrt{n/100}])$, $n > 10$.”

1.2. Large n Data and the Preconditions for More Accurate and Trustworthy Evidence

Large n data are universally considered a blessing since they can *potentially* give rise to more accurate and trustworthy evidence. What is often neglected, however, is that the potential for such gains requires certain ‘modeling’ and ‘inference’ stipulations to be met before any such accurate and trustworthy evidence can materialize.

[a] The most crucial *modeling* stipulation is for the practitioner to establish the *statistical adequacy* (approximate validity) of the invoked probabilistic assumptions imposed on one’s data, say $\mathbf{x}_0 := (x_1, x_2, \dots, x_n)$, comprising the relevant statistical model $\mathcal{M}_\theta(\mathbf{x})$; see Spanos [16]. Invalid probabilistic assumptions induce sizeable discrepancies between the nominal error probabilities—derived assuming the validity of $\mathcal{M}_\theta(\mathbf{x})$ —and the actual ones based on $\mathbf{x}_0$, rendering the inferences unreliable and the ensuing evidence untrustworthy. Applying a 0.05 significance level test when the actual type I error probability is closer to 0.97, due to invalid probabilistic assumptions (Spanos [17], p. 691) will yield spurious inference results and untrustworthy evidence; see Spanos [18] for additional examples.

[b] The most crucial *inference* stipulation for the practitioner is to distinguish between raw ‘inference results’, such as point estimates, effect sizes, observed CIs, ‘accept/reject H_0 ’, and *p*-values, and ‘evidence for or against inferential claims’ relating to unknown parameters θ . Conflating the two often gives rise to fallacious evidential interpretations of such inference results, as well as unwarranted claims, including spurious statistical significance with a large n . The essential difference between ‘inference results’ and ‘evidence’ is that the former rely unduly on the particular data $\mathbf{x}_0 := (x_1, x_2, \dots, x_n)$, which constitutes a single realization $\mathbf{X} = \mathbf{x}_0$ of the sample $\mathbf{X} := (X_1, \dots, X_n)$. In contrast, sound evidence for or against inferential claims relating to θ needs to account for that uncertainty.

The **main objective** of the paper is to make a case that the large n problem can be addressed using a principled argument based on the post-data severity (SEV) evaluation of the unduly data-specific accept/reject results. This is achieved by accounting for the inference-related uncertainty to provide an evidential interpretation of such results that revolves around the discrepancy $\gamma \neq 0$ from the null value warranted by the particular test and data $\mathbf{x}_0$ with high enough post-data error probability. This provides an inductive generalization of the accept/reject results that enhances learning from data.

As a prelude to the discussion that follows, Section 2 summarizes Fisher’s [19] model-based frequentist statistics, with particular emphasis on Neyman–Pearson (N-P) testing. Section 3 considers the large n problem and its implications for the *p*-value and the power of an N-P test. Section 4 discusses examples from the empirical literature in microeconometrics where the large n problem is largely ignored. Section 5 explains how the post-data severity evaluation of the accept/reject H_0 results can address the large n problem and is illustrated using hypothetical and actual data examples.

2. Model-Based Frequentist Statistics: An Overview

2.1. Fisher’s Model-Based Statistical Induction

Model-based frequentist statistics was pioneered by Fisher [19] in the form of statistical induction that revolves around a statistical model whose generic form is:

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

and revolves around the *distribution of the sample* $\mathbf{X} := (X_1, X_2, \dots, X_n)$, $f(\mathbf{x}; \theta)$, $\mathbf{x} \in \mathbb{R}_X^n$, $\mathbb{R} := (-\infty, \infty)$, which encapsulates its probabilistic assumptions. $\mathbb{R}_X^n$ denotes the sample space, and Θ the parameter space; see Spanos [16].

Unfortunately, the term ‘model’ is used to describe many different constructs across different disciplines. In the context of empirical modeling using statistics, however, the relevant models can be grouped into two broad categories: ‘substantive’ (structural, a priori

postulated) and ‘statistical’ models. Although these two categories of models are often conflated, a statistical model $\mathcal{M}_\theta(\mathbf{x})$ comprises solely the probabilistic assumptions imposed (explicitly or implicitly) on the particular data $\mathbf{x}_0$; see McCullagh [20]. Formally, $\mathcal{M}_\theta(\mathbf{x})$ is a stochastic mechanism framed in terms of probabilistic assumptions from three broad categories: Distribution (D), Dependence (M), and Heterogeneity (H), assigned to the observable stochastic process $\{X_t, t \in \mathbb{N} := (1, 2, \dots, n, \dots)\}$ underlying data $\mathbf{x}_0$.

The specification (initial selection) of $\mathcal{M}_\theta(\mathbf{x})$ has a twofold objective:

- (a) $\mathcal{M}_\theta(\mathbf{x})$ is selected to account for all the chance regularity patterns—the systematic statistical information—in data $\mathbf{x}_0$ by choosing appropriate probabilistic assumptions relating to $\{X_k, k \in \mathbb{N}\}$. Equivalently, $\mathcal{M}_\theta(\mathbf{x})$ is selected to render data $\mathbf{x}_0$ a ‘typical realization’ therefrom, and the ‘typicality’ can be evaluated using *Mis-Specification (M-S) testing*, which evaluates the approximate validity of its probabilistic assumptions.
- (b) $\mathcal{M}_\theta(\mathbf{x})$ is parametrized $[\theta \in \Theta]$ to enable one to shed light on the substantive questions of interest using data $\mathbf{x}_0$. When these questions are framed in terms of a *substantive model*, say $\mathcal{M}_\varphi(\mathbf{x})$, $\varphi \in \Phi$, one needs to bring out the implicit statistical model in a way that ensures that the two sets of parameters are related via a set of restrictions $g(\varphi, \theta) = 0$ colligating φ to the data $\mathbf{x}_0$ via θ ; see Spanos [16].

Example 1. A widely used example in practice is the simple Normal model:

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

where ‘NIID’ stands for ‘Normal (D), Independent (M), and Identically Distributed (H)’.

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

The *cornerstone* of frequentist inference is the concept of a *sampling distribution*, $f(y_n; \theta) = dF_n(y)/dy$, for all $(\forall) y \in \mathbb{R}_Y$, of a statistic $Y_n = g(X_1, X_2, \dots, X_n)$ (estimator, test, predictor), derived via:

$$F_n(y) = \mathbb{P}(Y_n \leq y) = \underbrace{\int \int \cdots \int}_{\{\mathbf{x}: g(\mathbf{x}) \leq y\}} f(\mathbf{x}; \theta) d\mathbf{x}, \quad \forall y \in \mathbb{R}_Y. \quad (3)$$

The derivation of $f(y_n; \theta)$, $\forall y \in \mathbb{R}_Y$, in (3) presumes the validity of $f(\mathbf{x}; \theta)$, $\mathbf{x} \in \mathbb{R}_X^n$,

which in the case of (2) is: $f(\mathbf{x}; \theta) \stackrel{\text{NIID}}{=} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2} \sum_{k=1}^n (x_k - \mu)^2\right\}$, $\mathbf{x} \in \mathbb{R}^n$.

In light of the crucial role of the distribution of the sample, Fisher [19], p. 314, emphasized the importance of establishing the *statistical adequacy* (approximate validity) of the invoked statistical model $\mathcal{M}_\theta(\mathbf{x})$:

“For empirical as the specification of the hypothetical population $[\mathcal{M}_\theta(\mathbf{x})]$ may be, this empiricism is cleared of its dangers if we can apply a rigorous and objective test of the adequacy with which the proposed population represents the whole of the available facts.”

He went on to underscore the crucial importance of Mis-Specification (M-S) testing (testing the approximate validity of the probabilistic assumptions comprising $\mathcal{M}_\theta(\mathbf{x})$) as the way to provide an empirical justification for statistical induction:

“The possibility of developing complete and self-contained tests of goodness of fit deserves very careful consideration, since therein lies our justification for the free use which is made of empirical frequency formulae.” (Fisher [19], (p. 314).)

Statistical adequacy plays a crucial role in securing the reliability of inference because it ensures the approximate equality between the actual and the nominal error probabilities based on $\mathbf{x}_0$, assuring that one can keep track of the relevant error probabilities. In contrast, when $\mathcal{M}_\theta(\mathbf{x})$ is *statistically misspecified* (Spanos, [21]),

- (a) $f(\mathbf{x}; \theta)$, $\mathbf{x} \in \mathbb{R}_X^n$, and the likelihood function $L(\mathbf{x}_0; \theta) \propto f(\mathbf{x}_0; \theta)$, $\forall \theta \in \Theta$ are erroneous,
- (b) distorting the sampling distribution $f(y_n; \theta)$ derived via (3), as well as
- (c) giving rise to ‘non-optimal’ estimators and sizeable *discrepancies* between the actual error probabilities and nominal—derived assuming $\mathcal{M}_\theta(\mathbf{x})$ is valid.

In light of that, the practical way to keep track of the relevant error probabilities is to establish the statistical adequacy of $\mathcal{M}_\theta(\mathbf{x})$. When $\mathcal{M}_\theta(\mathbf{x})$ is misspecified, any attempt to adjust the relevant error probabilities is ill-fated because the actual error probabilities are unknown due to being sizeably different from the nominal ones.

Regrettably, as Rao [22], p. 2, points out, validating $\mathcal{M}_\theta(\mathbf{x})$ using comprehensive M-S testing is neglected in statistics courses:

“They teach statistics as a deductive discipline of deriving consequences from given premises [$\mathcal{M}_\theta(\mathbf{x})$] The need for examining the premises, which is important for practical applications of results of data analysis is seldom emphasized. ... The current statistical methodology is mostly model-based, without any specific rules for model selection or validating a specified model. ” (p. 2)

See Spanos [23] for further discussion.

2.2. Neyman–Pearson (N-P) Testing

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

$$H_0: \mu \leq \mu_0 \text{ vs. } H_1: \mu > \mu_0, \quad (4)$$

an optimal (UMP) α -significance level test (Neyman and Pearson, 1933 [24]) is:

$$T_\alpha := [\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s}, C_1(\alpha) = \{\mathbf{x}: \tau(\mathbf{x}) > c_\alpha\}], \quad (5)$$

where $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$, $s_n^2 = \frac{1}{(n-1)} \sum_{k=1}^n (X_k - \bar{X}_n)^2$, $C_1(\alpha)$ denotes the rejection region, and c_α is determined by the significance level α ; see Lehmann [13].

The sampling distribution of $\tau(\mathbf{X})$ evaluated under H_0 (hypothetical) is:

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s} \stackrel{\mu = \mu_0}{\sim} \text{St}(n-1), \quad (6)$$

where ‘ $\text{St}(n-1)$ ’ denotes the Student’s t distribution with $(n-1)$ degrees of freedom, which provides the basis for evaluating the type I error probability and the p -value:

$$\alpha = \mathbb{P}(\tau(\mathbf{X}) > c_\alpha; \mu = \mu_0), \quad p(\mathbf{x}_0) = \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu = \mu_0). \quad (7)$$

That is, both the type I error probability and the p -value in (7) are evaluated using *hypothetical reasoning*, that interprets ‘ $\mu = \mu_0$ is true’ as ‘what if’ $\mu_0 = \mu^*$.

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

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma} \stackrel{\mu = \mu_1}{\sim} \text{St}(\delta_1; n-1), \quad \delta_1 = \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}, \quad \forall \mu_1 = \mu_0 + \gamma_1, \gamma_1 \geq 0, \quad (8)$$

where δ_1 is the noncentrality parameter of $\text{St}(\delta_1; n-1)$, which provides the basis for evaluating the power of test T_α :

$$\mathcal{P}(\mu_1) = \mathbb{P}(\tau(\mathbf{X}) > c_\alpha; \mu = \mu_1), \quad \forall \mu_1 = \mu_0 + \gamma_1, \gamma_1 \geq 0. \quad (9)$$

where $\delta_1 = \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}$ in (8) indicates that the power increases monotonically with $\sqrt{n}$ and $(\mu_1 - \mu_0)$ and decreases with σ .

$$\mathcal{P}(\mu_1) = \mathbb{P}(\tau(\mathbf{X}) > c_\alpha; \mu = \mu_1), \quad \forall \mu_1 = \mu_0 + \gamma_1, \gamma_1 \geq 0. \quad (10)$$

Equation (8) shows that the power increases with $\sqrt{n}$ and $(\mu_1 - \mu_0)$ and decreases with σ .

The optimality of N-P tests revolves around an inherent trade-off between the type I and II error probabilities. To address that trade-off, Neyman and Pearson [24] proposed to construct an optimal test by prespecifying α at a low value and minimizing the type II error $\beta(\mu_1)$, or maximizing the power $\mathcal{P}(\mu_1) = (1 - \beta(\mu_1))$, $\forall \mu_1 = \mu_0 + \gamma_1, \gamma_1 \geq 0$.

A question that is often overlooked in traditional expositions of N-P testing is:

Where does prespecifying the type I error probability at a low threshold come from?

A careful reading of Neyman and Pearson [24] reveals the answer in the form of two crucial stipulations relating to the framing of $H_0: \theta \in \Theta_0$ and $H_1: \theta \in \Theta_1$, to ensure the effectiveness of N-P testing and the informativeness of the ensuing results:

1. Θ_0 and Θ_1 should form a *partition* of Θ (p. 293) to avoid $\theta^* \notin [\Theta_0 \cup \Theta_1]$.
2. Θ_0 and Θ_1 should be framed in such a way to ensure that the type I error is the more serious of the two, using the analogy with a criminal trial, p. 296, with H_0 : no guilty, to secure a low probability for sending an innocent person to jail.

Unveiling the intended objective of stipulation 2, suggests that the requirement for a small (prespecified) α is to ensure that the test has a low probability of rejecting a true null hypothesis, i.e., when $\theta_0 = \theta^*$. Minimizing the type II error probability implies that an optimal test should have the lowest possible probability (or equivalently, the highest power) for accepting (rejecting) $\theta = \theta_0$, when false, i.e., when $\theta_0 \neq \theta^*$. That is, an optimal test should have a high power around the potential neighborhood of θ^* . This implies that when no reliable information about this potential neighborhood is available, one should use a two-sided test to avert the case where the test has no or very low power around θ^* ; see Spanos [25].

For the reason that the power increases with n , it is important to take that into account in selecting an appropriate α to avoid both an under-powered and an over-powered test, forefending the ‘small n ’ and ‘large n problems’, respectively. *First*, for a given α , one needs to calculate the value of n needed for T_α to have sufficient power to detect parameter discrepancies $\gamma \neq 0$ of interest; see Spanos [26]. *Second*, for a large n one needs to adjust α to avoid an ultra-sensitive test that could detect tiny discrepancies, say $\gamma = 0.0000001$, and misleadingly declare them statistically significant.

The primary role of the *pre-data* testing error probabilities (type I, II, and power) is to operationalize the notions of ‘statistically significant/insignificant’ in terms of statistical approximations framed in terms of a test statistic $\tau(\mathbf{X})$ and its sampling distribution. These error probabilities calibrate the capacity of the test to shed sufficient light on θ^* , giving rise to learning from data. In this sense, the reliability of the testing results ‘accept/reject H_0 ’ depends crucially on the particular testing statistical context (Spanos, [17], ch. 13):

$$(i) \mathcal{M}_\theta(\mathbf{x}), (ii) H_0: \theta \in \Theta_0 \text{ vs. } H_1: \theta \in \Theta_1, (iii) T_\alpha := \{d(\mathbf{X}), C_1(\alpha)\}, (iv) \text{ data } \mathbf{x}_0, \quad (11)$$

which includes, not only the adequacy of $\mathcal{M}_\theta(\mathbf{x})$ vis-a-vis data $\mathbf{x}_0$, the framing of H_0 and H_1 , and as well as n ; see Spanos ([16,17]). For instance, when $n > 10,000$, *detaching* the accept/reject H_0 results from their statistical context in (11), and claiming statistical significance at conventional thresholds will often be an unwarranted claim. Let us elaborate on this assertion.

3. The Large n Problem in N-P Testing

3.1. How Could One Operationalize as n Increases?

As mentioned above, for a given α , increasing n increases the test’s power and decreases the p -value. What is not so obvious is how to operationalize the clause ‘as n increases’ since data $\mathbf{x}_0$ usually come with a specific sample size n . Assuming that one begins with large enough n to ensure that the Mis-Specification (M-S) tests have sufficient power to detect existing departures from the probabilistic assumptions of the invoked $\mathcal{M}_\theta(\mathbf{x})$, say $n = 100$, there are two potential scenarios one could contemplate.

Scenario 1 assumes that all different values of $n \geq 100$ give rise to the same observed $\tau(\mathbf{x}_0)$. This scenario has been explored by Mayo and Spanos [27,28].

Scenario 2 assumes that as n increases beyond $n = 100$ the change in the estimates $\bar{x}_n$ and s_n^2 are ‘relatively small’ to render the ratio $(\bar{x}_n - \mu_0)/s_n$ approximately constant.

Scenario 2 seems realistic enough to shed light on the large n problem for two reasons. *First*, when the NIID assumptions are valid for $\mathbf{x}_n$, the changes in $\bar{x}_n$ and s_n^2 from increasing n are likely to be ‘relatively small’ since $n = 100$ is sufficiently large to provide a reliable initial estimate, and thus, increasing n is unlikely to change the ratio drastically. *Second*, the estimate $(\bar{x}_n - \mu_0)/s_n$ where μ_0 is a value of interest, is known as the ‘effect size’ for μ in psychology (Ellis, [29]), and is often used to infer the magnitude of the ‘scientific’ effect, irrespective of n . Let us explore the effects of increasing n using scenario 2.

3.2. The Large n Problem and the p -Value

Empirical example 1 (continued). Consider the hypotheses in (4) for $\mu_0 = 2$ in the context of (2) using the following information:

$$n = 100, \alpha = 0.05, c_\alpha = 1.66, \bar{x}_n = 2.317, \text{ and } s^2 = 3.7675 \text{ (} s = 1.941 \text{)}. \quad (12)$$

The test statistic $\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s}$ yields: $\tau_n(\mathbf{x}_0) = \frac{\sqrt{100}(2.317 - 2)}{1.941} = 1.6533$, with a p -value, $p_n(\mathbf{x}_0) = 0.0528$, indicating ‘accept H_0 ’. The question of interest is how increasing n beyond $n = 100$ will affect the result of N-P testing.

Consider the issue of how the p -value changes as n increases when $\frac{(\bar{x}_n - \mu_0)}{s_n}$ is constant (scenario 2). The p -value curve in Figure 1 for $50 < n \leq 500$ indicates that one can manipulate n to obtain the desired result since (a) for $n < 105$ the p -value will yield $p_n(\mathbf{x}_0) > \alpha = 0.05$, ‘accept H_0 ’, (b) for $n > 105$ the p -value will yield $p_n(\mathbf{x}_0) < \alpha = 0.05$, ‘reject H_0 ’.

Table 1 reports particular values of $\tau_n(\mathbf{x}_0)$ and $p_n(\mathbf{x}_0)$ from Figure 1 as n increases, showing that $p_n(\mathbf{x}_0)$ decreases rapidly down to tiny values for $n \geq 10,000$, confirming Berkson’s [1], (p. 334), observation in the introduction that the p -value “... will be small beyond any usual limit of significance.”

Table 1. The p -value as n increases (keeping $(\bar{x}_n - \mu_0)/s_n$ constant).

n	100	120	150	300	500	1000	2000	10,000
$\tau_n(\mathbf{x}_0)$	1.633	1.789	2.0	2.829	3.652	5.165	7.304	16.332
$p_n(\mathbf{x}_0)$	0.0528	0.038	0.024	0.0025	0.00014	0.00000015	0.2×10^{-12}	0.0000...

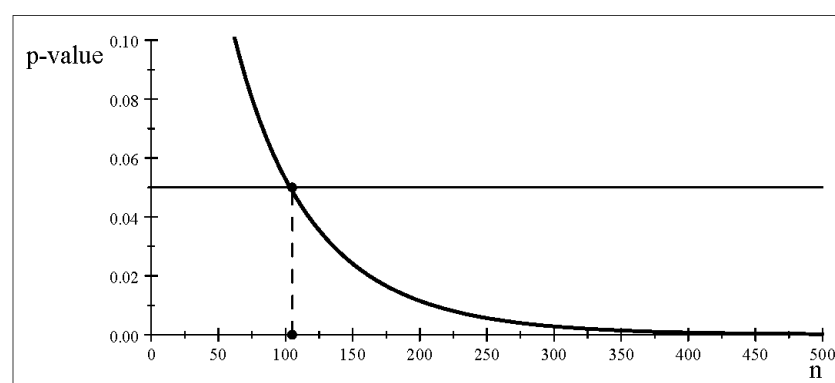


Figure 1. The p -value curve for different sample sizes n .

3.3. The Large n Problem and the Power of a Test

Let us consider the power of $T_\alpha^>$ at 0.8 ($\mathbb{P}(\mu_1) = 0.8$) as n increases under scenario 2 ($(\bar{x}_n - \mu_0)/s_n$ held constant), and evaluate its effect on the size of the detected discrepancies $\gamma_1 = \mu_1 - \mu_0$.

Table 2 reports several values of n showing how the test detects smaller and smaller discrepancies from $\mu = 2$, confirming Fisher's [2] quotation in the introduction. This can be seen more clearly in Figure 2, where the power curves become steeper and steeper as n increases.

Table 2. Discrepancy γ_1 detected with $P(\mu_1) = 0.8$ as n increases.

n	100	200	500	1000	10,000	100,000	1,000,000	20,000,000
γ_1	0.486	0.344	0.217	0.154	0.0485	0.01535	0.00486	0.0034

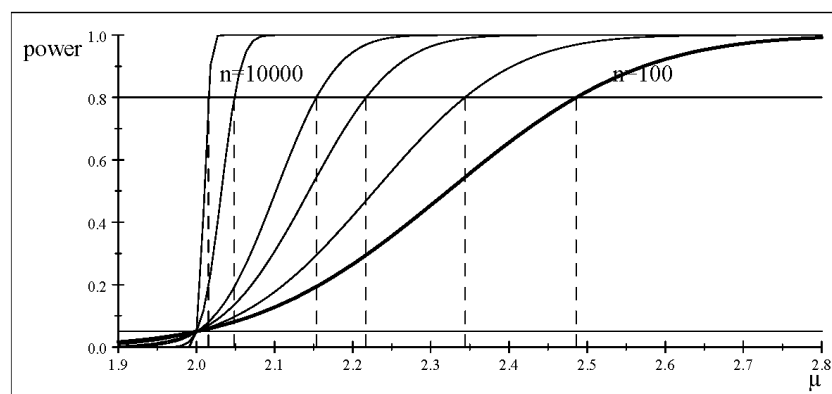


Figure 2. The power curve for different sample sizes n .

One might object to this quote as anachronistic since Fisher [30] was clearly against the use of the type II error probability and the power of a test due to his falsificationist stance where ‘accept H_0 ’ would not be an option. The response is that Fisher acknowledged the role of power using the term ‘sensitivity’ in the quotation (Section 1.1), as well as explicitly in Fisher [31], p. 295. Indeed, the presence of a rejection threshold α in Fisher’s significance testing brings into play the pre-data type I and II error probabilities, while his p -value is a post-data error probability; see Spanos [17].

4. The Empirical Literature and the Large n Problem

Examples of misuse of frequentist testing abound in all applied fields, but the large n problem is particularly serious in applied microeconometrics where very large sample sizes n are often employed to yield spurious significance results, which are then used to propose policy recommendations relating to relevant legislation; see Pesko and Warman [32].

4.1. Empirical Examples in Microeconometrics

Empirical example 2A (Abouk et al. [33], Appendix J: Table 1. p. 99). Based on an estimated LR model in (Table 7). with $n = 24,730,930$, it is inferred that the estimates $\hat{\beta}_k = 0.004$, $SE(\hat{\beta}_k) = 0.002$, rendering the coefficient β_k of a key variable x_k statistically significant at $\alpha = 0.05$.

Such empirical results and the claimed evidence based on the estimated models are vulnerable to three potential problems.

- (i) **Statistical misspecification** and the ensuing untrustworthy evidence since the discussion in the paper ignores the statistical adequacy of the estimated statistical models. Given that to evaluate the statistical adequacy of any published empirical study, one needs the original data to apply thorough misspecification testing (Spanos, [7]), that issue will be sidestepped in the discussion that follows.
- (ii) **Large n problem.** One of the many examples in the paper relates to the claimed statistical significance (with $n = 24,730,930$) at $\alpha = 0.05$ of β_k , ignoring its potential ‘spurious’ statistical significance results stemming from the large n problem in N-P testing.

- (iii) **Conflating ‘testing results’ with ‘evidence’.** The authors claim *evidence* for $\beta_k \neq 0$, and proceed to infer its implications for the effectiveness of different economic policies.

Empirical example 2A (continued). Abouk et al. [33] report $\hat{\beta}_k = 0.004$, $SE(\hat{\beta}_k) = \sqrt{\frac{s}{n\sqrt{q_{kk}}}} = 0.002$, and $p(\mathbf{z}_0) < 0.05$, implying:

$$\frac{\sqrt{n}(\hat{\beta}_k - 0)}{s\sqrt{q_{kk}}} = \frac{\sqrt{24,730,930}(0.004)}{\sqrt{98.932}} = 2, \text{ for } c_{0.025} = 1.96 \text{ and } p(\mathbf{z}_0) = 0.045.$$

The relevant sampling distribution of $\hat{\beta} := (\hat{\beta}_0, \hat{\beta}_1)$ for the LR model is:

$$(\sqrt{n}(\hat{\beta} - \beta)|\mathbf{X}) \sim N(\mathbf{0}, \sigma^2 \mathbf{Q}_X^{-1}), \lim_{n \rightarrow \infty} (\frac{\mathbf{X}^T \mathbf{X}}{n}) = \mathbf{Q}_X = [q_{ij}]_{i,j=1}^m > 0.$$

Focusing on just one coefficient β_k the t -test for its significance is:

$$\tau(\mathbf{y}) = \frac{\sqrt{n}(\hat{\beta}_k - 0)}{s/\sqrt{q_{kk}}} \underset{n \rightarrow \infty}{\beta_k = 0} N(0, 1), C_1(\alpha) = \{\mathbf{y}: |\tau(\mathbf{y})| > c_{\frac{\alpha}{2}}\}.$$

Using the information in example 2A, we can reconstruct what would $p_n(\mathbf{z}_0)$ have been for different n using scenario 2. The results in Table 3 indicate that the authors’ claim of statistical significance ($\beta_k \neq 0$) at $\alpha = 0.05$ will be unwarranted for any $n < 24,000,000$.

Table 3. The p -value with increasing n (constant estimates).

n	100	500	1000	2000	10,000	10×10^4	10×10^5	20×10^5	20×10^6	24×10^6
$\tau_n(\mathbf{x}_0)$	0.004	0.009	0.0127	0.018	0.040	0.127	0.402	0.569	10.798	10.970
$p_n(\mathbf{x}_0)$	0.997	0.993	0.990	0.986	0.967	0.899	0.688	0.570	0.072	0.049

Example 2B. Abouk et al. [33], Table 2, p. 57, report 42 ANOVA results for the difference between two means $(\bar{x}_{n_1} - \bar{y}_{n_2})$, 40 of these tests have tiny p -values, less than 0.0000, even though the differences between the two means, $(\bar{x}_{n_1} - \bar{y}_{n_2})$ appear to be very small. Surprisingly, two of the estimated differences are zero, $(\bar{x}_{n_1} - \bar{y}_{n_2}) = 0$ ($\bar{x}_{n_1} = 13.2$, $\bar{y}_{n_2} = 13.2$ and $\bar{x}_{n_1} = 2.51$, $\bar{y}_{n_2} = 2.51$), but their reported p -values are 0.8867 and 0.0056, respectively; one would have expected $p(\mathbf{z}_0) = 1$ when $\tau(\mathbf{z}_0) = 0$. Looking at these results, one wonders what went wrong with the reported p -values.

The hypotheses of interest take the form (Lehmann, [13]):

$$H_0: (\mu_1 - \mu_2) = 0 \text{ vs. } H_1: (\mu_1 - \mu_2) \neq 0$$

with the optimal test being the t -test $T_\alpha^\neq = \{\tau(\mathbf{Z}), C_1 = \{\mathbf{z}: |\tau(\mathbf{z})| > c_{\frac{\alpha}{2}}\}\}$ where:

$$\tau(\mathbf{Z}) = [\sqrt{N}(\bar{X}_{n_1} - \bar{Y}_{n_2} - \gamma)/s_N] \overset{\mu_1 - \mu_2 = 0}{\rightsquigarrow} \text{St}(n_1 + n_2 - 2), N = \frac{n_1 n_2}{n_1 + n_2}, \quad (13)$$

$$s_N^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}, s_1^2 = \frac{1}{(n_1 - 1)} \sum_{i=1}^{n_1} (X_i - \bar{X}_{n_1})^2, s_2^2 = \frac{1}{(n_2 - 1)} \sum_{i=1}^{n_2} (Y_i - \bar{Y}_{n_2})^2.$$

A close look at this t -test suggests that the most plausible explanation for the above ‘strange results’ is the statistical software using (at least) 12-digit decimal precision. Given that the authors report mostly 3-digit estimates, the software is picking up tiny discrepancies, say < 0.0000001 , which, when magnified by $\sqrt{N}$, could yield the reported p -values. That is, the reported results have (inadvertently) exposed the effect of the large n on the p -value.

4.2. Meliorating the Large n Problem Using Rules of Thumb

In light of the inherent trade-off between the type I and type II error probabilities, some statistics textbooks advise practitioners to use ‘rules of thumb’ based on decreasing α as n increases; see Lehmann [13].

1. Ad hoc rules for adjusting α as n increases

$n =$	100	200	500	1000	10,000	20,000	200,000	...
$\alpha =$	0.05	0.025	0.01	0.001	0.0001	0.00001	0.00000001	...

2. Good [15], p. 66, proposed to standardize the p -value $p_n(\mathbf{x}_0)$ relative to $N = 100$.

Applying his rule of thumb, $\min(0.5, [p_n(\mathbf{x}_0) \cdot \sqrt{n/100}])$, $n > 100$, to the p -value curve in Figure 1, based on $\tau_n(\mathbf{x}_0) = 1.6533$, yields:

$n =$	100	120	150	300	500	1000	5000	10,000
$p_n(\mathbf{x}_0) =$	0.0528	0.038	0.024	0.002	0.0002	0.000001	2.4×10^{-20}	0.000...0
$p_{100}(\mathbf{x}_0) =$	0.0528	0.042	0.029	0.0035	0.0045	0.000003	10.7×10^{-19}	0.000...0

The above numerical examples suggest that such rules of thumb for selecting α as n increases can meliorate the problem; they do not, however, address the large n problem since they are ad hoc, and their suggested thresholds decrease to nearly zero beyond $n = 10,000$.

4.3. The Large n Problem and Effect Sizes

Although there is no authoritative definition of the notion of an effect size, the one that comes closest to its motivating objective is Thompson's [34]:

"An effect size is a statistic quantifying the extent to which the sample statistics diverge from the null hypothesis." (p. 172)

The notion is invariably related to a certain frequentist test and has a distinct resemblance to its test statistic. To shed light on its effectiveness in addressing the large n problem, consider a simpler form of the test in (13) where the sample size n is the same for both $\mathbf{X}$ and $\mathbf{Y}$ and $\text{Cov}(X_i, Y_i) = 0$ for $i = 1, 2, \dots, n$. The optimal t -test is based on a simple bivariate Normal distribution with parameters (μ_1, μ_2, σ^2) for $H_0: \gamma = \gamma_0$ vs. $H_1: \gamma \neq \gamma_0$, $\gamma = \mu_1 - \mu_2$, $\gamma_0 = 0$, takes the form:

$$\tau(\mathbf{Z}) = \frac{\sqrt{N}[(\bar{X}_n - \bar{Y}_n) - \gamma]}{s} \stackrel{\gamma=\gamma_0}{\rightsquigarrow} \text{St}(2n-2), N = (n/2), \quad (14)$$

$$\tau(\mathbf{Z}) \stackrel{\gamma=\gamma_1}{\rightsquigarrow} \text{St}(\delta_1; 2n-2), \delta_1 = \frac{\sqrt{N}(\gamma_1 - \gamma_0)}{\sigma}, \text{ for } \gamma_1 \neq \gamma_0,$$

The recommended effect size for this particular test is the widely used Cohen's $d = [(\bar{x}_n - \bar{y}_n)/s]$. As argued by Abelson [35], the motivation underlying the choice of the effect size statistic:

"... is that its expected value is independent of the size of the sample used to perform the significance test." (p. 46)

That is, if one were to view Cohen's $d(\mathbf{z}_0) = [(\bar{x}_n - \bar{y}_n)/s]$ as a point estimate of the unknown parameter $\psi = [(\mu_1 - \mu_2)/\sigma]$, the statistic referred to by Thompson above is $\hat{\psi}(\mathbf{Z}) = [(\bar{X}_n - \bar{Y}_n)/s]$, confirming Abelson's claim that $E[\hat{\psi}(\mathbf{Z})] = \psi$ is free of n .

The question that naturally arises at this point is to what extent deleting $\sqrt{N}$ and using the point estimate $\hat{\psi}(\mathbf{z}_0) = [(\bar{x}_n - \bar{y}_n)/s]$ of ψ as the effect size result associated with $\gamma = (\mu_1 - \mu_2)$ addresses the large n problem. The demonstrable answer is that it does not since the claim that $\hat{\psi}(\mathbf{z}_0)$ approximates closely ψ^* for a large enough n is unwarranted; see Spanos [36]. The reason is that a point estimate $\hat{\psi}(\mathbf{z}_0)$ represents a single realization relating to the relevant sampling distribution, i.e., it represents a 'statistical result' that ignores the relevant uncertainty. Indeed, the $\hat{\psi}(\mathbf{Z})$, by itself, does not have a sampling distribution without $\sqrt{N}$, since its relevant sampling distribution relates to:

$$\tau(\mathbf{Z}; \gamma) = \frac{\sqrt{N}[(\bar{X}_n - \bar{Y}_n) - \gamma]}{s} \stackrel{\gamma=\gamma^*}{\rightsquigarrow} \text{St}(2n-2),$$

where γ^* denotes the true value of $(\mu_1 - \mu_2)$; see Spanos [37] for more details.

5. The Post-Data Severity Evaluation (SEV) and the Large n Problem

The post-data severity (SEV) evaluation of the accept/reject H_0 results is a principled argument that provides an evidential account for these results. Its main objective is to transform the unduly data-specific accept/reject H_0 results into evidence. This takes the form of an inferential claim that revolves around the discrepancy $\gamma_1 = \mu_1 - \mu_0$ warranted by data $\mathbf{x}_0$ and test T_α with a high enough probability; see Spanos [9].

A hypothesis H (H_0 or H_1) passes a severe test T_α with $\mathbf{x}_0$ 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 [27].

5.1. Case 1: Accept H_0

In the case of ‘accept H_0 ’, the SEV evaluation is seeking the smallest ‘discrepancy from $\mu_0 = 2$ with a high enough probability.

Empirical example 1 (continued). For $\alpha = 0.05$, $c_\alpha = 1.66$, $\bar{x}_n = 2.317$, $s = 1.941$, and $n = 100$, test T_α in (5) for the hypotheses in (4) yields: $\tau(\mathbf{x}_0) = 1.6533$, with $p(\mathbf{x}_0) = 0.0507$, indicating ‘accept H_0 ’. (C-1) indicates that $\mathbf{x}_0$ accords with H_0 , and since $\tau(\mathbf{x}_0) > 0$ the relevant inferential claim is $\mu \leq \mu_1 = \mu_0 + \gamma$, for $\gamma > 0$. Hence, (C-2) calls for evaluating the probability of the event: “outcomes $\mathbf{x} \in \mathbb{R}^n$ that accord less well with $\mu \leq \mu_1$ than $\mathbf{x}_0$ does”, i.e., $[\mathbf{x}: \tau(\mathbf{x}) > \tau(\mathbf{x}_0)]$, $\forall \mathbf{x} \in \mathbb{R}^n$:

$$SEV(T_\alpha; \mu \leq \mu_1) = \min_{\mu \in \Theta_1} \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu = \mu_1) > p_1,$$

where $\Theta_1 := (2, \infty)$, $\gamma = (\mu_1 - \mu_0) > 0$, for a large enough $p_1 > 0.5$, and evaluated based on:

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s} \stackrel{\mu = \mu_1}{\sim} \text{St}(\delta_1; n - 1), \delta_1 = \frac{\sqrt{n}(\mu_1 - \mu_0)}{s}, \forall \mu_1 > \mu_0, \quad (15)$$

$$E(\tau(\mathbf{X})) = \left(\frac{\sqrt{(n-1)/2} \Gamma[(n-2)/2]}{\Gamma[(n-1)/2]} \right) \delta_1 := m \text{ and } Var(\tau(\mathbf{X})) = \frac{n-1}{n-3} (1 + \delta^2) - m^2; \text{ see Owen [38].}$$

That is, the central and non-central Student t sampling distributions differ not only with respect to their mean but also in terms of the variance as well as the higher moments since for a non-zero δ_1 the non-central Student’s t is non-symmetric. The post-data severity curve for all $\mu \in [1.8, 3]$ in Figure 3 indicates that the discrepancy warranted by data $\mathbf{x}_0$ and test T_α with probability 0.8 is $\gamma^\dagger \leq 0.481$ ($\mu_1 \leq 2.481$).

The post-data severity evaluated for typical values reported in Table 4 reveals that the probability associated with the discrepancy relating to the estimate $\bar{x}_n = 2.317$ ($\gamma_1 = 0.317$) is *never* high enough to be the discrepancy warranted by $\mathbf{x}_0$ and test T_α since $SEV(\gamma_1 = 0.317) = 0.5$. This calls into question any claims relating to point estimates and observed CIs more generally since they represent an uncalibrated (it ignores the relevant uncertainty) single realization ($\mathbf{x}_0$) of the sample $\mathbf{X}$ relating to the relevant sampling distribution.

Table 4. Post-data severity evaluation (SEV) for $\mu_1 \leq 2.0 + \gamma$.

γ	0.05	0.1	0.15	0.20	0.30	0.317	0.40	0.481	0.60	0.70
μ_1	2.05	2.1	2.15	2.2	2.3	2.317	2.4	2.481	2.6	2.7
$SEV(\mu \leq \mu_1)$	0.086	0.133	0.196	0.274	0.465	0.500	0.665	0.800	0.926	0.974

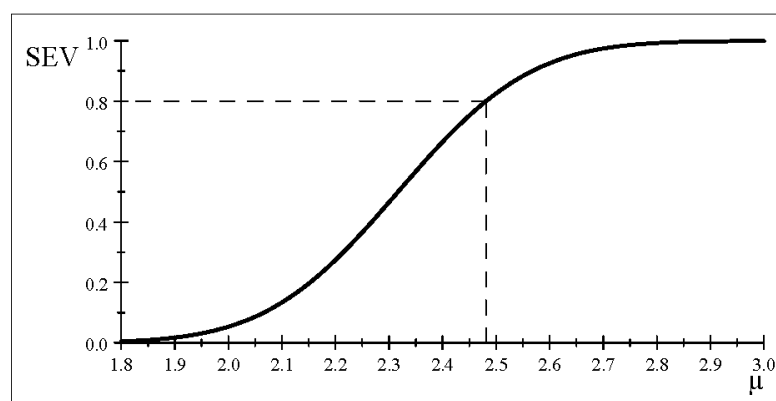


Figure 3. The post-data severity curve (accept H_0).

5.2. The Post-Data Severity and the Large n Problem

5.2.1. Case 1: Accept H_0

Empirical example 1 (continued). Consider scenario 2 where $[(\bar{x}_n - \mu_0)/s_n]$ remains constant as n increases. Table 5 gives particular examples of n showing how $\tau_n(\mathbf{x}_0)$ increases and μ_1 decreases for a fixed $SEV(\gamma) = 0.8$.

Table 5. Post-data severity ($SEV(\gamma) = 0.8, \mu_1 \leq 2 + \gamma_1, \bar{x}_n = 2.317, s = 1.941$).

n :	100	120	150	200	300	500	1000	2000	20,000	200,000
$\tau_n(\mathbf{x}_0)$:	1.633	1.789	2.0	2.310	2.829	3.652	5.165	7.304	23.097	73.038
$\mu_1 = 2 + \gamma$:	2.481	2.467	2.451	2.434	2.412	2.390	2.369	2.3535	2.329	2.321

Figure 4 presents the different SEV curves as n increases ($[(\bar{x}_n - \mu_0)/s]$ held constant). It confirms the results of Table 5, by showing that the curves become steeper and steeper as n increases. This reduces the warranted discrepancy $\gamma_n^\dagger$ monotonically (Table 5), converging to the lower bound at $\gamma^\dagger \leq 0.3217$ ($\bar{x}_n = 2.317$) as $n \rightarrow \infty$. That is, the warranted discrepancies indicated on the x-axis, at a constant probability 0.8 (y-axis), become smaller and smaller as n increases, with the lower bound being the value of μ where its optimal estimator converges to μ^* with probability one as $n \rightarrow \infty$.

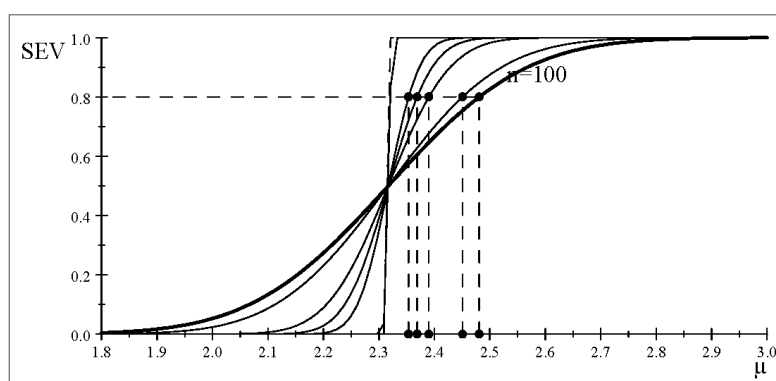


Figure 4. The severity curve (accept H_0) for different n (same estimates).

This makes statistical sense in practice since $\bar{X}_n$ and s_n are *strongly consistent* estimators of μ^* and σ_* , i.e., $\mathbb{P}(\lim_{n \rightarrow \infty} \hat{\theta}_n(\mathbf{X}) = \theta^*) = 1$, and thus their accuracy (precision) improves as n increases beyond a certain threshold $n > N$. Recall that in the case of ‘accept H_0 ’ one is seeking the ‘smallest discrepancy’ from $\mu_0 = 2$. Hence, as n increases SEV renders the warranted discrepancy $\gamma_n^\dagger$ ‘more accurate’ by reducing it, until it reaches the lower bound around $\gamma^\dagger \leq (\bar{x}_n - \mu_0) [\mu_1 = \bar{x}_n]$ as $n \rightarrow \infty$; note that the latter is the most accurate value for μ^* , only when $\mathcal{M}_\theta(\mathbf{x})$ is statistically adequate!

5.2.2. Case 2: Reject H_0

Empirical example 3. Consider changing the estimate of μ to $\bar{x}_n = 2.726$ retaining $s = 1.941$, which yields: $\tau_n(\mathbf{x}_0) = 3.740[0.00009]$, indicating ‘reject H_0 ’ in (4). In contrast to the case ‘accept H_0 ’, when the test ‘rejects H_0 ’ one is seeking the ‘largest’ discrepancy from $\mu_0 = 2$ with high probability. In light of $\tau_n(\mathbf{x}_0) > 0$, the relevant inferential claim is: $\mu \geq \mu_1 = \mu_0 + \gamma$, for $\gamma > 0$, and its post-data probabilistic evaluation is based on:

$$SEV(T_\alpha; \mu \geq \mu_1) = \max_{\forall \mu \in \Theta_1} \mathbb{P}(\tau(\mathbf{X}) < \tau(\mathbf{x}_0); \mu = \mu_1) > p_1, \quad (16)$$

where $\Theta_1 := (2, \infty)$, $p_1 > 0.5$, and the evaluation of (16) is based on (15).

Table 6 indicates that the discrepancy γ warranted with $SEV(T_\alpha; \mu \geq \mu_1) = 0.8$ is $\gamma^\dagger \leq 0.562$ ($\mu_1 \leq 2.562$), and the discrepancy for $\mu_1 \leq 2.726$ has $SEV(T_\alpha; \mu \geq \mu_1) = 0.5$.

Table 6. Post-data severity evaluation (SEV) for $\mu_1 \geq 2 + \gamma$.

γ	0.1	0.2	0.3	0.4	0.5	0.562	0.6	0.726	0.8	0.9
μ_1	2.1	2.2	2.3	2.4	2.5	2.562	2.6	2.562	2.6	2.7
$SEV(\mu \geq \mu_1)$	0.999	0.996	0.984	0.951	0.876	0.800	0.741	0.500	0.352	0.186

Figure 5 depicts the severity curves for $n = 100, 200, 500, 1000, 10,000$, indicating that keeping the probability constant at 0.8 (y-axis), as n increases (keeping $[(\bar{x}_n - \mu_0)/s_n]$ constant) renders the curves steeper and steeper, thus ‘increasing’ the warranted discrepancy monotonically (x-axis) toward approaching the upper bound, which is the value of μ where its optimal estimator converges to μ^* with probability one as $n \rightarrow \infty$. This is analogous to the case of ‘accept H_0 ’, with the lower bound replaced by an upper bound, which is equal to μ^* in both cases. That is, the SEV evaluation increases the precision of the discrepancy warranted with high (but constant) probability as n increases.

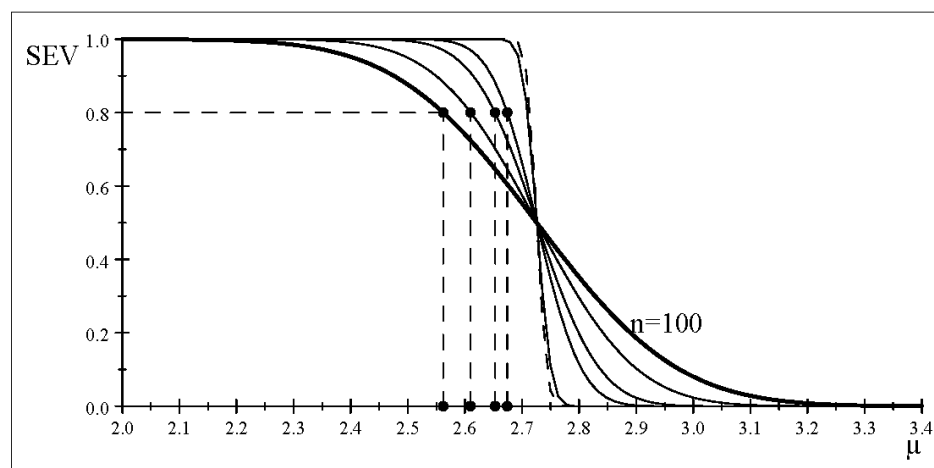


Figure 5. The severity curve (reject H_0) for different n (same estimates)

Example 2A (Abouk et al., 2022) continued. The reported t -test result is:

$$\frac{\sqrt{n}(\hat{\beta}_k - 0)}{s\sqrt{q_{kk}}} = \frac{\sqrt{24,730,930}(0.004)}{\sqrt{98.932}} = 2.0[0.045].$$

What is the warranted discrepancy $\gamma^\dagger$ from $\beta_k = 0$ with high enough severity, say 0.977 (in light of $n = 24,730,930$)? The answer is $\beta_k = \gamma^\dagger \leq 0.0000001$, which calls into question $\hat{\beta}_k = 0.004$, as well as Cohen’s $d(\mathbf{z}_0) = 0.0004$; see Ellis [29]. This confirms the tiny discrepancies from zero due to the large n problem conjectured in Section 4.1.

One might object to the above conclusion relating to the issue of spurious significance by countering that a tiny discrepancy is still different from zero, and thus, the statistical

significance is well-grounded. Regrettably, such a counter-argument is based on a serious misconstrual of statistical inference in general, where learning from data takes the form of ‘statistical approximations’ framed in terms of a statistic or a pivot and its sampling distribution. This distinguishes statistics from other sub-fields of applied mathematics. No statistical inference is evaluated in terms of a binary choice of right and wrong! Indeed, this confusion lies at the heart of misinterpreting the binary accept/reject results as evidence for or against hypotheses or claims; see Mayo and Spanos [28].

Example 2B (Abouk et al. [33]) continued. As argued in Section 4.1, the reported results for the difference between two means exemplify the effect of the large n problem on the p -value. To be able to quantify that effect, however, one needs the exact p -values, but only two of the 42 reported p -values are given exactly, 0.8867 and 0.0056; the other 40 are reported as less than 0.0000. Also, the particular sample sizes for the different reported tests are not given, and thus the overall n_1 and n_1 will be used. Focusing on the one with the p -value = 0.0056, one can retrieve the observed t -statistic, which can then be used for the SEV evaluation. For $\bar{x}_n = 2.51$, $\bar{y}_n = 2.51$, the observed t -statistic is $\tau(\mathbf{z}_0) = 2.77$, and using the SEV evaluation one can show that the warranted discrepancy γ from $(\mu_1 - \mu_2) = 0$ with $SEV = 0.97$ ($N = 6, 108, 194$) is $\gamma^\dagger \leq 0.00000068$, which confirms that the t -test detects tiny discrepancies between the two estimated means, as conjectured in Section 4.1. That is, all but one (p -value = 0.8867) of the reported 42 test results in Table 2 (Abouk et al. [33] p. 43) constitute cases of spurious rejection of the null due to the large n problem. Even if one were to reduce N by a factor of 40, the warranted discrepancies would still be tiny.

5.2.3. Key Features of the Post-Data SEV Evaluation

- (a) The $SEV(\mu_0 + \gamma)$ is a post-data *error probability*, evaluated using *hypothetical reasoning*, that takes fully into account the testing statistical context in (11) and is guided by the sign and magnitude of $\tau_n(\mathbf{x}_0)$ as indicators of the direction of the relevant discrepancies γ from $\mu = \mu_0$. This is particularly important because the *factual reasoning*, what if $\theta = \theta^*$, underlying point estimation and Confidence Intervals (CIs), does not apply post-data.
- (b) The $SEV(\mu_0 + \gamma)$ evaluation differs from other attempts to deal with the large n problem in so far as its outputting of the discrepancy $\gamma_1 = \mu_1 - \mu_0$ is always based on the non-central distribution (Kraemer and Paik, [39]):

$$\left(\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s} - \frac{\sqrt{n}(\mu_1 - \mu_0)}{s} \right) \stackrel{\mu=\mu_1}{\approx} \text{St}(n-1), \text{ for all } \mu_1 \in \Theta_1. \quad (17)$$

This ensures that the warranted discrepancy $\gamma^\dagger$ is evaluated using the ‘same’ sample size n , counter-balancing the effect of n on $\tau_n(\mathbf{x}_0)$. Note that ‘ $\approx$ ’ denotes an approximation.

- (c) The evaluation of the warranted $\gamma_n^\dagger$ with high probability accounts for the increase in n by enhancing its precision. As n increases $\gamma^\dagger$ will approach the value $\hat{\theta}_n(\mathbf{x}_0)$, since for a statistically adequate $\mathcal{M}_\theta(\mathbf{x})$, $\hat{\theta}_n(\mathbf{x}_0)$ approaches θ^* due to its strong consistency.
- (d) The SEV can be used to address other foundational problems, including distinguishing between ‘statistical’ and ‘substantive’ significance. It also provides a testing-based effect size for the magnitude of the ‘scientific’ effect by addressing the problem with estimation-based effect sizes raised in Section 4.3. Also, the SEV evaluation can shed light on several proposed alternatives to (or modifications of) N-P testing by the replication crisis literature (Wasserstein et al. [40]), including replacing the p -value with effect sizes and observed CIs and redefining statistical significance (Benjamin et al. [41]) irrespective of the sample size n ; see Spanos [25,37].

6. Summary and Conclusions

The large n problem arises naturally in the context of N-P testing due to the in-built trade-off between the type I and II error probabilities, around which the optimality of N-P tests revolves. This renders the accept/reject H_0 results and the p -value highly vulnerable to the large n problem. Hence, for $n > 10,000$, the detection of statistical significance based

on conventional significance levels will often be spurious since a consistent N-P test will detect smaller and smaller discrepancies as n increases; see Spanos [9].

The post-data severity (SEV) evaluation can address the large n problem by converting the unduly data-specific accept/reject H_0 ‘results’ into ‘evidence’ for a particular inferential claim of the form $\theta \geq \theta_0 + \gamma_1$, $\gamma_1 \neq 0$. This is framed in the form of the discrepancy $\gamma_1^\dagger$ warranted by data $\mathbf{x}_0$ and test T_α with high enough probability. The SEV evidential account is couched in terms of a *post-data* error probability that accounts for the uncertainty arising from the undue data-specificity of accept/reject H_0 results. The SEV differs from other attempts to address the large n problem in so far as its evaluation is invariably based on a non-central distribution whose non-centrality parameter uses the same n as the observed test statistic $\tau(\mathbf{x}_0)$ to counter-balance the effect induced by n in outputting γ_1 .

The SEV evaluation was illustrated above using two empirical results from Abouk et al. [33]. Example 2A concerns an estimated coefficient, $\hat{\beta}_k = 0.004$, $SE(\hat{\beta}_k) = 0.002$, in a LR model, which is declared statistically significant at $\alpha = 0.05$ with $n = 24,730,930$. The SEV evaluation yields a discrepancy $\gamma^\dagger \leq 0.0000001$ from $\beta_k = 0$ warranted by data $\mathbf{z}_0$ and the t -test with probability 0.977. Example 2B concerns the difference between two means whose estimates are $\bar{x}_{n_1} = 2.51$, $\bar{y}_{n_2} = 2.51$, but the t -test outputted $\tau(\mathbf{z}_0) = 2.77$ with $N = 6,108,194$. The SEV evaluation yields a discrepancy $\gamma^\dagger \leq 0.00000068$ from $(\mu_1 - \mu_2) = 0$ warranted by data $\mathbf{z}_0$ and the t -test with probability 0.97. Both empirical examples represent cases of spurious statistical significance results stemming from exceptionally large sample sizes, $n = 24,730,930$, and $N = 6,108,194$, respectively.

Funding: This research received no external funding.

Institutional Review Board Statement: Non-applicable.

Informed Consent Statement: Non-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:

M-S	Mis-Specification
N-P	Neyman–Pearson
UMP	Uniformly Most Powerful
SE	Standaard Error
SEV	post-data severity evaluation

References

1. Berkson, J. Some difficulties of interpretation encountered in the application of the chi-square test. *J. Am. Stat.* **1938**, *33*, 526–536. [CrossRef]
2. Fisher, R.A. *The Design of Experiments*; Oliver and Boyd: Edinburgh, UK, 1935.
3. Berkson, J. Tests of significance considered as evidence. *J. Am. Assoc.* **1942**, *37*, 325–335. [CrossRef]
4. Fisher, R.A. *Statistical Methods for Research Workers*; Oliver and Boyd: Edinburgh, UK, 1925.
5. Fisher, R. A. Note on Dr. Berkson’s criticism of tests of significance. *J. Am. Stat. Assoc.* **1943**, *38*, 103–104. [CrossRef]
6. Berkson, J. Experience with Tests of Significance: A Reply to Professor R. A. Fisher. *J. Am. Assoc.* **1943**, *38*, 242–246. [CrossRef]
7. Spanos, A. Mis-Specification Testing in Retrospect. *J. Econ. Surv.* **2018**, *32*, 541–577. [CrossRef]
8. Lindley, D.V. A statistical paradox. *Biometrika* **1957**, *44*, 187–192. [CrossRef]
9. Spanos, A. Who Should Be Afraid of the Jeffreys-Lindley Paradox? *Philos. Sci.* **2013**, *80*, 73–93. [CrossRef]
10. Lehmann, E.L. Significance level and power. *Ann. Math. Stat.* **1958**, *29*, 1167–1176. [CrossRef]
11. Cohen, J. The statistical power of abnormal-social psychological research: A review. *J. Abnorm. Soc. Psychol.* **1962**, *65*, 145–153. [CrossRef]
12. Freiman, J.A.; Chalmers, T.C.; Smith, H.; Kuebler, R.R. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *N. Engl. J. Med.* **1978**, *299*, 690–694. [CrossRef]
13. Lehmann, E.L. *Testing Statistical Hypotheses*, 2nd ed.; Wiley: New York, NY, USA, 1986.
14. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum: Hoboken, NJ, USA, 1988.

15. Good, I.J. Standardized tail-area probabilities. *J. Stat. Comput. Simul.* **1982**, *16*, 65–66. [\[CrossRef\]](#)
16. 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.
17. Spanos, A. *Introduction to Probability Theory and Statistical Inference: Empirical Modeling with Observational Data*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2019.
18. Spanos, A. Statistical Misspecification and the Reliability of Inference: The simple *t*-test in the presence of Markov dependence. *Korean Econ. Rev.* **2009**, *25*, 165–213.
19. Fisher, R.A. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc.* **1922**, *222*, 309–368.
20. McCullagh, P. What is a statistical model? *Ann. Stat.* **2002**, *30*, 1225–1267. [\[CrossRef\]](#)
21. Spanos, A. Statistical Adequacy and the Trustworthiness of Empirical Evidence: Statistical vs. Substantive Information. *Econ. Model.* **2010**, *27*, 1436–1452. [\[CrossRef\]](#)
22. Rao, C.R. Statistics: Reflections on the Past and Visions for the Future. *Amstat. News* **2004**, *327*, 2–3.
23. Spanos, A. Frequentist Model-based Statistical Induction and the Replication crisis. *J. Quant. Econ.* **2022**, *20*, 133–159. [\[CrossRef\]](#)
24. Neyman, J.; Pearson, E.S. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc.* **1933**, *231*, 289–337.
25. Spanos, A. How the Post-data Severity Converts Testing Results into Evidence for or Against Pertinent Inferential Claims. *Entropy* **2023**, under review.
26. Spanos, A. Severity and Trustworthy Evidence: Foundational Problems versus Misuses of Frequentist Testing. *Philos. Sci.* **2022**, *89*, 378–397. [\[CrossRef\]](#)
27. Mayo G.D.; Spanos, A. Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction. *Br. J. Philos. Sci.* **2006**, *57*, 323–357. [\[CrossRef\]](#)
28. Mayo, D.G.; Spanos, A. Error Statistics. In *The Handbook of Philosophy of Science*; Gabbay, D., Thagard, P., Woods, J., Eds.; Elsevier: Amsterdam, The Netherlands, 2011; Volume 7: Philosophy of Statistics, pp. 151–196.
29. 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.
30. Fisher, R.A. Statistical methods and scientific induction. *J. R. Soc. Ser. Stat. Methodol.* **1955**, *17*, 69–78. [\[CrossRef\]](#)
31. Fisher, R.A. Two new properties of mathematical likelihood. *Proc. R. Soc. Lond. Ser.* **1934**, *144*, 285–307.
32. Pesko, M.F. and C. Warman Re-exploring the early relationship between teenage cigarette and e-cigarette use using price and tax changes. *Health Econ.* **2022**, *31*, 137–153. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Abouk, R.; Adams, S.; Feng, B.; Maclean, J.C.; Pesko, M. The Effects of e-cigarette taxes on pre-pregnancy and prenatal smoking. *NBER Work. Pap.* **2022**, 26126, Revised June 2022. Available online: <https://www.nber.org/system/files/workingpapers/w26126/w26126.pdf> (accessed on 5 October 2023).
34. Thompson, B. *Foundations of Behavioral Statistics: An Insight-Based Approach*; Guilford Press: New York, NY, USA, 2006.
35. Abelson, R.P. *Statistics as Principled Argument*; Lawrence Erlbaum: Hoboken, NJ, USA, 1995.
36. Spanos, A. Bernoulli's golden theorem in retrospect: Error probabilities and trustworthy evidence. *Synthese* **2021**, *199*, 13949–13976. [\[CrossRef\]](#)
37. Spanos, A. Revisiting noncentrality-based confidence intervals, error probabilities and estimation-based effect sizes. *J. Math.* **2021**, *104*, 102580. [\[CrossRef\]](#)
38. Owen, D.B. Survey of Properties and Applications of the Noncentral t-Distribution. *Technometrics* **1968**, *10*, 445–478. [\[CrossRef\]](#)
39. Kraemer, H.C.; Paik, M. A central t approximation to the noncentral t distribution. *Technometrics* **1979**, *21*, 357–360. [\[CrossRef\]](#)
40. Wasserstein, R.L.; Schirm, A.L.; Lazar, N.A. Moving to a world beyond " $p < 0.05$ ". *Am. Stat.* **2019**, *73*, 1–19.
41. 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. [\[CrossRef\]](#)

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.