Revisiting the 'spurious significance' associated with large n data: how the post-data severity can address that and related foundational problems

Aris Spanos
Dept. of Economics,
Virginia Tech

April 2023

Abstract

The recent explosion of big data with large sample sizes n has brought to the surface an old foundational problem known as the 'large n' relating to the misuse of frequentist testing to forge 'spurious' statistical significance. Large n data sets are universally considered a blessing because *they can* give rise to more accurate and trustworthy evidence. What is often ignored, however, is that these benefits take place *only when* certain preconditions are satisfied, the most crucial being the validity of the probabilistic assumptions invoked by the particular inference. The large n problem arises naturally in the Neyman-Pearson (N-P) testing due to the inherent trade-off between the type I and type II error probabilities around which the optimality of N-P tests revolves. It is well known since the 1930s that as n increases the p-value decreases, and the power increases. This calls into question the 'statistical significance' of parameters of interest using the conventional significance levels, .05, .025, .01, when n is very large. It is arged that the traditional 'rules of thumb' of decreasing α as n increases can alleviate but do not address the problem. A principled argument in the form of a post-data severity evaluation of the accept/reject H_0 results can address the large n and related problems, including the rigging of the significance level α, the arbitrariness of framing H_0 and H_1, the statistical vs. substantive significance, and estimation-based vs. testing-based effect sizes.

KEYWORDS: Large n problem; Student's t test; Neyman-Pearson testing; p-value; accept/reject H_0 results; non-central sampling distributions; post-data severity evaluation; warranted discrepancy; statistical vs. substantive significance; spurious statistical significance.

1 Introduction

Big data sets with large sample sizes (n) have become widely available in many scientific disciplines. Such data are universally considered a blessing because the additional sample information can potentially give rise to more *accurate* and *trustworthy evidence*. What is often neglected, however, is that the potential for such gains calls for certain preconditions to be met. The most important is that the invoked statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ – comprising the *probabilistic assumptions* imposed on data $\mathbf{x}_0$:=$(x_1, x_2, ..., x_n)$ – should be *validated* before any inferences are drawn.

The reason for the validation of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is that its *statistical adequacy* (approximate validity) secures the statistical *reliability* of inference by ensuring that (i) the nominal optimality (derived assuming $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is valid) of the inference procedures is in fact actual, as well as (ii) the *nominal* and the *actual* (based on data $\mathbf{x}_0$) error probabilities are approximately equal. In contrast, a *statistical misspecified* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ will give rise to non-optimal procedures and *sizeable discrepancies* between the actual and nominal error probabilities, rendering the inference results unreliable and the ensuing evidence untrustworthy. Applying a .05 significance level test when the actual type I error probability is closer to .97 (see Table 4), will yield spurious results and untrustworthy evidence; see Spanos (1989), Spanos and McGuirk (2001).

Modern frequentist statistics has been pioneered in the early 1920s by R.A. Fisher in the form of model-based statistical induction that revolves around a parametric statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. He also proposed a theory of optimal point estimation and a theory of significance testing driven by the p-value; see Fisher (1922, 1925a-b). J. Neyman and E. Pearson (N-P) supplemented that with an optimal theory of hypothesis testing, and J. Neyman (1937) proposed an optimal theory of interval estimation. Regrettably, these protagonists left largely unresolved several foundational problems that have bedeviled frequentist inference since the 1930s, including the following.

(1) *What is a statistical model* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ for data $\mathbf{x}_0$ and how is best selected?

(2) What is the *primary aim* of frequentist inference in learning from data $\mathbf{x}_0$?

(3) What is the *reasoning* underlying the derivation of the relevant sampling distributions employed to frame frequentist inference results?

(4) How can one bridge the gap between statistical *results* (estimation, testing) and *evidence* for or against an inferential claims?

(5) What are the respective roles of the '*substantive' subject matter* and the '*statistical' information* (chance regularities) in data $\mathbf{x}_0$, in empirical modeling?

(**6**) How can one establish the *statistical adequacy* of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ using effective *Mis-Specification* (M-S) *testing* and *respecification*?

(7) Foundational issues bedeviling frequentist testing since the 1930s:

(a) The claimed *incompatibility* between N-P and significance testing

(b) The arbitrariness in *framing* the N-P hypotheses H_0 and H_1.

(c) The *rigging* the significance level α to get a desired result.

(d) The *large/small n problems* yielding 'spurious' inference results.

(e) Distinguishing between *statistical* and *substantive significance*.

These foundational issues are interrelated since addressing one involves dealing with some of the others, but the main focus will be on the large n problem and its effects.

Section 2 summarizes Fisher's model-based frequentist statistics, with particular emphasis on Neyman-Pearson (N-P) testing, as a prelude to the discussion that follows. Section 3 revisits the large/small n problems in N-P testing and their implications for the trustworthiness of the ensuing evidence. Section 4 considers how the post-data severity evaluation of the accept/reject H_0 results (Mayo and Spanos, 2006, 2011) can address the large n problem. Section 5 considers briefly how the post-data severity evaluation can be used to deal with the other foundational problems.

2 Model-based frequentist statistics: an overview

All approaches to empirical modeling using statistics involve three basic components. [a] Questions of substantive interest (however vague or specific), [b] the relevant data $\mathbf{x}_0{:=}(x_1, x_2, ..., x_n)$ selected to shed light on these questions, and [c] a set of probabilistic assumptions comprising the (implicit) statistical model.

2.1 Fisher's model-based statistical induction

Model-based frequentist statistics was pioneered by Fisher (1922) as a form of statistical induction that revolves around a statistical model whose generic form is:

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

where $f(\mathbf{x};\boldsymbol{\theta})$ denotes the (joint) *distribution of the sample* $\mathbf{X}{:=}(X_1, X_2, ..., X_n)$, $\mathbb{R}^n_X$ is the sample space, and Θ the parameter space; see Spanos (2006b).

(1) What is a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ for data $\mathbf{x}_0$ and how is selected?

$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ constitutes a statistical mechanism framed in terms of probabilistic assumptions relating to the observable stochastic process $\{X_k,\ k{\in}\mathbb{N}\}$ underlying $\mathbf{x}_0$. It is selected on the basis that it *could have* given rise to data $\mathbf{x}_0$, and reconcile [a]-[b].

[b] $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is selected to account for all the chance regularity patterns exhibited by data $\mathbf{x}_0$ by choosing appropriate probabilistic assumptions relating to the underlying process $\{X_k,\ k{\in}\mathbb{N}\}$, from three broad categories, **Distribution, Dependence**, and **Heterogeneity.** Equivalently, $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is selected to render data $\mathbf{x}_0$ a 'truly typical realization' thereof; the 'typicality' can be tested using Mis-Specification (M-S) tests.

[a] Select a particular parametrization $\boldsymbol{\theta}{\in}\Theta$ for $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ that enables one to pose the substantive questions of interest to data $\mathbf{x}_0$; see Spanos (1986).

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

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

where $\boldsymbol{\theta}{:=}(\mu,\sigma^2)$, $\mathbb{R}{:=}(-\infty,\infty)$, $\mathbb{R}_+{:=}(0,\infty)$ and '$\mathsf{NIID}(\mu,\sigma^2)$' stands for 'Normal (N), Independent and Identically Distributed (IID), with mean μ and variance σ^2. Note that $f(\mathbf{x};\boldsymbol{\theta})$ encapsulates the probabilistic assumptions of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ since:

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

(2) What is the *primary aim* of frequentist inference in learning from data $\mathbf{x}_0$?

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

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

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

(3) What is the *reasoning* underlying the derivation of the relevant sampling distributions employed to frame frequentist inference results?

The derivation of the relevant sampling distributions are based on two different forms of reasoning:

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

(ii) *hypothetical* (hypothesis testing): H_0: $\boldsymbol{\theta}\in\Theta_0$ (presuming $\boldsymbol{\theta}\in\Theta_0$) vs. H_1: $\boldsymbol{\theta}\in\Theta_1$ (presuming $\boldsymbol{\theta}\in\Theta_1$); see Spanos (2019). Hence, $\boldsymbol{\theta}$ is always prespecified in (3).

The *sampling distribution,* $f(y_n;\boldsymbol{\theta}),\ \forall y\in\mathbb{R}_Y$, frames the uncertainty relating to the fact that data $\mathbf{x}_0$ constitutes a single realization (out of all $\mathbf{x}\in\mathbb{R}^n_X$) of the sample $\mathbf{X}$, and is used to calibrate the *capacity* (optimality) of the inference procedure in terms of the relevant error probabilities, coverage, type I, II and power.

The derivation of $f(y_n;\boldsymbol{\theta}),\ \forall y\in\mathbb{R}_Y$, in (3) presumes the validity of $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$, and thus Fisher (1922) emphasizes the importance of establishing the *statistical adequacy* of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: "For empirical as the specification of the hypothetical population [statistical model] 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." (p. 314). He goes on to emphasize the importance of model validation and the crucial role of Mis-Specification (M-S) testing 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." (p. 314) Fisher emphasized that statistical induction differs from other variants of induction in so far as its justification is empirical, stemming from the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, and not from any a priori stipulations; see Spanos (2022b).

Statistical adequacy plays a crucial role in securing the reliability of inference because it secures the approximate equality between the actual and the nominal error probabilities based on $\mathbf{x}_0$, ensuring the 'control' (keep track of) of these probabilities. In contrast, when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is *statistically misspecified*, (a) $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$, is erroneous, and that (b) distorts the sampling distribution $f(y_n;\boldsymbol{\theta})$ derived in (3), (c) giving rise to 'non-optimal' estimators and sizeable *discrepancies* between the actual and nominal error probabilities; see Spanos (2009). Hence, the way to 'control' the relevant

error probabilities is by establishing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. Regrettably, as Rao (2004) argues, the *statistical adequacy* of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is neglected in statistics courses: "They teach statistics as a deductive discipline of deriving consequences from given premises [$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$]. The need for examining the premises, which is important for practical applications of results of data analysis, is seldom emphasized." (p. 2)

2.2 Neyman-Pearson (N-P) testing

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

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

N-P testing yields the optimal (UMP) α-significance level test:

$$T_\alpha := [\tau(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s},\ C_1(\alpha)=\{\mathbf{x}\text{: } \tau(\mathbf{x})>c_\alpha\}], \tag{5}$$

where $\overline{X}_n=\frac{1}{n}\sum_{k=1}^{n} X_k,\ s_n^2=\frac{1}{(n-1)}\sum_{k=1}^{n}(X_k-\overline{X}_n)^2$, $C_1(\alpha)$ is the rejection region, and c_α is determined by the significance level α; see Lehmann and Romano (2005).

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

$$\tau(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s} \overset{\mu=\mu_0}{\backsim} \mathsf{St}(n-1), \tag{6}$$

is used to evaluate the type I error probability and the p-value:

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

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:

$$\tau(\mathbf{X})=\tfrac{\sqrt{n}\left(\overline{X}_n-\mu_0\right)}{\sigma} \overset{\mu=\mu_1}{\backsim} \mathsf{St}(\delta_1;n-1),\ \delta_1=\tfrac{\sqrt{n}(\mu_1-\mu_0)}{\sigma},\ \forall \mu_1>\mu_0, \tag{8}$$

where δ_1 is the noncentrality parameter of $\mathsf{St}(\delta_1;n-1)$, $\mu_1=\mu_0+\gamma_1$, $\gamma_1\geq 0$, with:

$$\mathcal{P}(\mu_1)=\mathbb{P}(\tau(\mathbf{X})>c_\alpha;\ \mu=\mu_1),\ \forall \mu_1>\mu_0, \tag{9}$$

defining the power of T_α evaluated based on (8). δ_1 indicates that the power increases monotonically with $\sqrt{n}$ and $(\mu_1-\mu_0)$ and decreases with σ.

The pre-data testing error probabilities (type I, II, and power) are:

(i) *hypothetical* and *unobservable* in principle since they revolve around θ^*,

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

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

The type I and II error probabilities are interrelated since there is a in-built trade-off between them. Neyman and Pearson (1933) addressed this trade-off by prespecifying α at a low value and maximizing $\mathcal{P}(\mu_1)$, $\forall \mu_1=\mu_0+\gamma_1\in\Theta_1$, $\gamma_1\pm 0$, to define a Uniformly Most Powerful (UMP) test; see Lehmann and Romano (2005).

The primary role of the pre-data testing error probabilities (type I, II, power) is to operationalize the notions of 'statistically significant/insignificant' in terms of the

sampling distribution of $\tau(\mathbf{X})$. The testing results 'accept/reject H_0' depend crucially on the particular statistical context (Spanos, 2019, ch. 13):

(i) $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, (ii) H_0: $\theta\in\Theta_0$ vs. H_1: $\theta\in\Theta_1$, (iii) T_α:=$\{d(\mathbf{X}), C_1(\alpha)\}$, (iv) data $\mathbf{x}_0$. (10)

The most egregious *misinterpretation* of the accept/reject H_0 results is to detach them from their statistical context in (10) and use stars, say α=.1 [*], α=.05 [**], α=.01 [***], to indicate statistical significance at different α.

2.3 Inference 'results' vs. inferential claims about evidence

(4) How can one relate statistical *results* (estimation, testing, prediction) and *evidence* for or against an inferential claim in frequentist inference?

In statistical inference, it is important to distinguish between *statistical results*, such as a point estimate, say $\widehat{\theta}(\mathbf{x}_0)$, an observed $(1-\alpha)$ CI, say $[L(\mathbf{x}_0), U(\mathbf{x}_0)]$, and an accept or reject H_0 outcome, and what *inferential claims* such results can justify.

Example 1 (continued). It is often presumed that the optimality of point estimators of (μ, σ^2): $\widehat{\mu}(\mathbf{X})$=$\overline{X}_n$=$\frac{1}{n}\sum_{t=1}^{n} X_t$ and $s^2(\mathbf{X})$=$\frac{1}{(n-1)}\sum_{t=1}^{n}(X_t-\overline{X}_n)^2$, justify the following inferential claims for a *large enough* n.

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

$$\widehat{\mu}(\mathbf{x}_0)\simeq\mu^*, \text{ and } s^2(\mathbf{x}_0)\simeq\sigma_*^2, \text{ when } n \text{ is large enough.} \tag{11}$$

The inferential claim in (11) is unwarranted since $\widehat{\mu}(\mathbf{x}_0)$ represents a single point $\mathbf{X}$=$\mathbf{x}_0$ of the estimator's $\widehat{\mu}(\mathbf{X})$ sampling distribution $f(\widehat{\mu}(\mathbf{x});\boldsymbol{\theta})$, $\forall\mathbf{x}\in\mathbb{R}^n$. That is the reason for reporting point estimates by attaching their standard errors (SE), say $\widehat{\mu}(\mathbf{x}_0)\pm 2SE(\widehat{\mu}(\mathbf{x}_0))$. This is formalized using interval estimation.

(b) The inferential claim associated with an $(1-\alpha)$ optimal CI for μ:

$$\text{CI}(\mathbf{X})=\mathbb{P}(\overline{X}_n-c_{\frac{\alpha}{2}}(\tfrac{s_n}{\sqrt{n}})\leq\mu<\overline{X}_n+c_{\frac{\alpha}{2}}(\tfrac{s_n}{\sqrt{n}});\ \mu=\mu^*)=(1-\alpha), \tag{12}$$

relates to CI($\mathbf{X}$) overlaying μ^* with probability $(1-\alpha)$. This does not justify the inferential claim that the observed CI:

$$\text{CI}(\mathbf{x}_0)=[\overline{x}_n-c_{\frac{\alpha}{2}}(\tfrac{s_n}{\sqrt{n}})\leq\mu<\overline{x}_n+c_{\frac{\alpha}{2}}(\tfrac{s_n}{\sqrt{n}})],$$

overlays μ^* with probability $(1-\alpha)$, or that values of μ in the middle of CI($\mathbf{x}_0$) are more probable than the ones at the end points. CI($\mathbf{x}_0$) might or might not include μ^* for a particular $\mathbf{x}_0$. Post data, the factual reasoning (what if μ=μ^*) has transpired, and no probability can be assigned to CI($\mathbf{x}_0$).

(c) In N-P testing the 'accept or reject H_0' results do not imply that there is evidence for H_0 or H_1; see Spanos (2014, 2021a-b).

2.4 Substantive vs. statistical models

(5) What are the respective roles of the '*substantive' subject matter information* and the '*statistical' systematic information* (chance regularities) in data $\mathbf{x}_0$?

Behind every substantive (structural) model estimable with data $\mathbf{z}_0$:

$$\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})=\{f(\mathbf{x};\boldsymbol{\varphi}),\ \boldsymbol{\varphi}\in\Phi\},\ \mathbf{x}\in\mathbb{R}^n_X,\quad \Phi\subset\mathbb{R}^p,\ p\leq m,$$

there is always an implicit statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ that comprises solely the probabilistic assumptions imposed on $\{(Y_t|\ \mathbf{X}_t=\mathbf{x}_t)\,,\ t\in\mathbb{N}\}$ without the statistically redundant substantive restrictions.

Example 4. The statistical model implicit in a Capital Asset Pricing Model (CAPM) is a Linear Regression (LR) model (Table 3):

$$\begin{aligned}
&\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z}): && (Y_t-x_{2t})=\alpha_1(x_{1t}-x_{2t})+\varepsilon_t,\ (\varepsilon_t|\mathbf{X}_t=\mathbf{x}_t)\backsim\mathsf{NIID}(0,\ \sigma^2_\varepsilon),\ t\in\mathbb{N},\\
&\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}): && Y_t=\beta_0+\beta_1x_{1t}+\beta_2x_{2t}+u_t,\ (u_t|\mathbf{X}_t=\mathbf{x}_t)\backsim\mathsf{NIID}(0,\ \sigma^2_u),\ t\in\mathbb{N},\\
&\mathbf{g}(\boldsymbol{\varphi},\boldsymbol{\theta})=\mathbf{0}: && \beta_0=0,\ \beta_1+\beta_2-1=0,\ \text{where } \boldsymbol{\varphi}=(\alpha_1,\sigma^2_\varepsilon),\ \boldsymbol{\theta}=(\beta_0,\beta_1,\beta_2,\sigma^2_u).,
\end{aligned}$$

where $\mathbf{Z}_t$:=(Y_t, x_{1t}, x_{2t}), denote 'returns' for: Y_t-a particular asset, x_{1t}-market, and x_{2t}-safe asset, respectively. The two models, $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ and $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, are related via restrictions $\mathbf{g}(\boldsymbol{\varphi},\boldsymbol{\theta})=\mathbf{0}$ whose *validity* for data $\mathbf{z}_0$, can be used to gauge whether $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ belies the data. Testing $\mathbf{g}(\boldsymbol{\varphi},\boldsymbol{\theta})=\mathbf{0}$ should always be based on a statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ to secure the reliability of the test; see Spanos (1986, 1989, 2006).

3 The large n problem in N-P testing

The large n problem arises naturally in N-P testing due to the in-built trade-off between the type I and type II error probabilities around which the optimality of N-P tests revolves. Increasing n raises the power of a test and thus to avoid rejecting H_0 for smaller and smaller discrepancies one needs to reduce α to counter-balance the increase in power, but how? That has been a key question since the 1930s.

3.1 How the large n problem affects the p-value

Berkson (1938), in applying a chi-square test observed that "... when the numbers in the data are quite large, the P's [the 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." He went on: "If, then, we know in advance the P that will result..., it is no test at all." (p. 527)

Empirical example 1 (continued). Consider the hypotheses:

$$H_0\text{: } \mu\leq\mu_0 \text{ vs. } H_1\text{: } \mu>\mu_0,\ \mu_0=2, \tag{13}$$

in the context of (2) using the following information:

$$n=100,\ \alpha=.05,\ c_\alpha=1.66,\ \overline{x}_n=2.317,\ \text{and } s^2=3.7675\ (s=1.941). \tag{14}$$

The test statistic $\tau(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{s}$ yields: $\tau_n(\mathbf{x}_0)=\frac{\sqrt{100}(2.317-2)}{1.941}=1.6533[.0528]$,

where the p-value is $p_n(\mathbf{x}_0)=.0528$, indicating 'accept H_0' at $\alpha=.05$.

The key question of interest is how increasing n beyond $n=100$ will affect the result of N-P testing. There are two possible scenarios to contemplate.

Scenario 1 assumes that all different values of $n\geq 100$ yield the same observed $\tau(\mathbf{x}_0)$. This scenario has been explored in Mayo and Spanos (2006, 2011).

Scenario 2 assumes that the change in the estimates $\overline{x}_n$ and s_n^2 are 'relatively small' to render $\tau(\mathbf{x}_0)=\frac{\sqrt{n}(\overline{x}_n-\mu_0)}{s_n}$ approximately constant for all $n\geq 100$. How appropriate is this scenario in elucidating the large n problem? The *first* issue to consider is how an increase in n will affect the estimates:

$$\overline{x}_n=\frac{1}{n}\sum_{t=1}^{n} x_t,\ \ s_n^2=\frac{1}{n-1}\sum_{t=1}^{n}(x_t-\overline{x}_n)^2. \tag{15}$$

The answer is: 'if the NIID assumptions are valid' for data $\mathbf{x}_n$, the changes in $\overline{x}_n$ 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 likely to bring small changes in $\overline{x}_n$ and s_n^2. In addition, since $\tau(\mathbf{x}_0)$ involves both $\overline{x}_n$ and s_n in a ratio, any changes from increasing n will affect both, and that could potentially minimize the changes on $\tau_n(\mathbf{x}_0)$ since an increase/decrease in $\overline{x}$, is likely to change s_n in tandem.

The *second* issue to consider is that the *strong consistency* of the estimators, $\overline{X}_n,\ s_n^2$, does imply a gradual *increase* in the 'precision' of the estimates as n increases, IF the invoked $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is statistically adequate. The problem, however, is that this increase in precision cannot be quantified with any accuracy for a given n. As argued by Le Cam (1986): "... limit theorems "as n tends to infinity" are logically devoid of content about what happens at any particular n. ... Unfortunately, the approximation bounds we could get are too often too crude and cumbersome to be of any practical use." (p. xiv). Hence, the unwarranted claim in (11).

The *third* issue is that $\frac{(\overline{x}_n-\mu_0)}{s_n}$ for $\mu_0=0$ is known in psychology as the 'effect size' for μ (Cohen, 1988) widely used to infer the magnitude of the 'scientific' effect.

The above claims for scenario 2 can be verified using carefully designed simulations that ensure that the replicated samples $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ are statistically adequate.

Empirical example 1 (continued). Let us pursue scenario 2 assuming that any marginal changes in $\overline{x}_n=2.317$ and $s_n=1.941$ for $n=100$ will keep $\frac{(\overline{x}_n-\mu_0)}{s_n}$ constant.

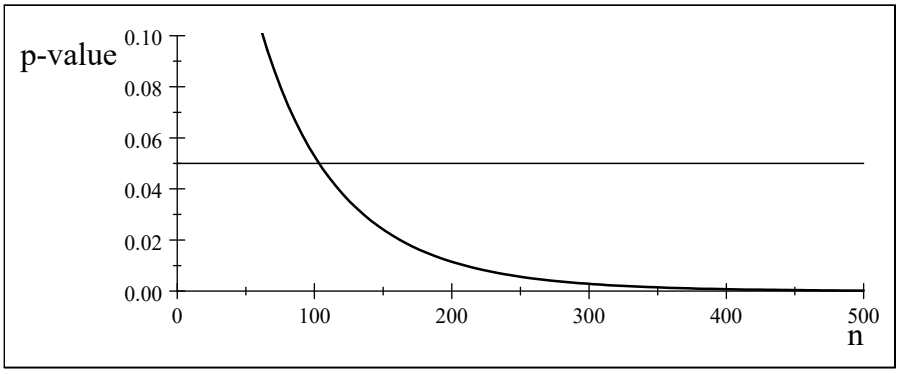

Fig. 1: the p-value curve for different sample sizes n

Figure 1 depicts the p-value curve for $100<n\leq 1000$ indicating that one can easily manipulate n to get the desired result since:

(a) For $n < 105$ the p-value will yield $p_n(\mathbf{x}_0) > \alpha=.05$, 'accept H_0',

(b) For $n > 105$ the p-value will yield $p_n(\mathbf{x}_0) < \alpha=.05$, 'reject H_0'.

Table 1 reports the values of $\tau_n(\mathbf{x}_0)$ and $p_n(\mathbf{x}_0)$ as n increases, showing that $p_n(\mathbf{x}_0)$ decreases rapidly down to tiny values for $n\geq$10000.

Table 1: the p-value with increasing n (constant estimates $\overline{x}_n, s_n$)								
n	100	120	150	300	500	1000	2000	10000
$\tau_n(\mathbf{x}_0)$	1.633	1.789	2.0	2.829	3.652	5.165	7.304	16.332
$p_n(\mathbf{x}_0)$	.0528	.038	.024	.0025	.00014	.00000015	$.2 \times 10^{-12}$	.0000...

The small n problem is equally pernicious as the large n since it undermines, not only the primary inferences that revolve around $\boldsymbol{\theta}$, but also the *model validation* using Mis-Specification (M-S) testing. This is because neither form of testing will be effective enough; their power will be too low to detect existing discrepancies. Hence, a rule of thumb for an adequate sample size is that if n is not large enough for comprehensive model validation to be effective (high enough power for detecting existing departures from the probabilistic assumptions of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$), is not large enough for reliable inferences; see Spanos (2022a).

3.2 The large n problem and the power of a test

Neyman and Pearson (1933) understood the difference between inference results and evidence for or against an evidential claim by arguing against (mis)interpreting 'accept H_0' as evidence for H_0' and 'reject H_0' as evidence for H_1. Mayo and Spanos (2006) framed these false interpretations in terms of two fallacies:

Fallacy of acceptance: misinterpreting 'accept H_0' (no evidence against H_0) as evidence *for* H_0'. This could arise when n is small enough.

Fallacy of rejection: misinterpreting 'reject H_0' (evidence against H_0) as providing evidence *for* H_1'. This could arise when n is large enough.

These fallacies point the finger at the power of a test as the primary culprit since, for a fixed α, the power increases monotonically with $\sqrt{n}$, and thus the test will detect smaller and smaller discrepancies γ from μ=μ_0, as n increases. To demonstrate that, let us evaluate the discrepancy that test T_α can detect with power .8 ($P(\mu_1)$=.8) as n increases. Specific examples of n holding the power constant at .8 are shown in table 2, indicating that as n increases the test detects smaller and smaller discrepancies from μ=μ_0 with high enough power, say $\mathcal{P}(\mu_1)$=.8.

Table 2: Discrepancy γ detected with $P(\mu_1)$=.8 as n increases								
n	100	200	500	1000	10000	100000	1000000	20000000
γ	.486	.344	.217	.154	.0485	.01535	.00486	.0034
$\mathcal{P}(\mu_1)$	.8	.8	.8	.8	.8	.8	.8	.8

Fisher (1935) was the first to raise the large n problem: "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." (pp. 21-22).

The power curves in figure 2 reflect the increases in n from figure 1 and provide a more complete picture of how such increases affect the sensitivity of the test.

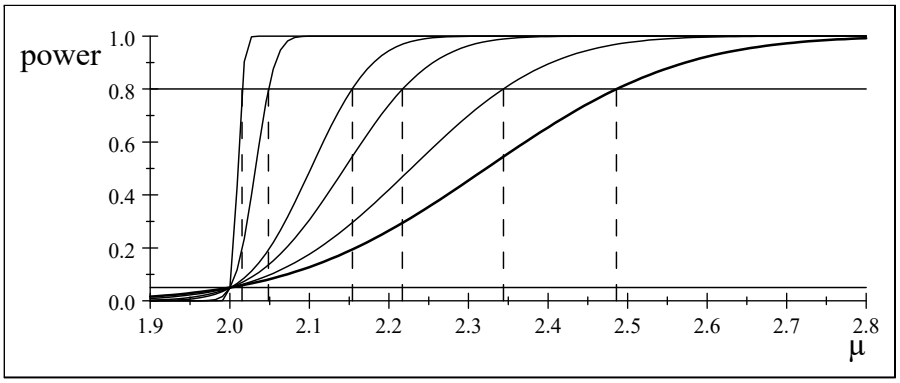

Fig. 2: the power curve for different sample sizes n

3.3 The empirical literature and the large n problem

How does the large n problem relate to the current empirical literature? Examples of extreme misuse of frequentist testing abound in all applied fields. In econometrics data sets with a very large n are considered a blessing because they can yield more accurate and trustworthy evidence, blithely ignoring the *necessary preconditions* for these gains to materialize.

(i) One needs to establish the *statistical adequacy* of the invoked statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ *before* any inferences are drawn. Also, a statistically misspecified $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ renders discussions about the large n problem irrelevant due to sizeable discrepancies between the nominal and actual error probabilities!

(ii) One needs to implement statistical inference by selecting the *most effective* (optimal) inference procedures in the context of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

(iii) One needs to interpret the ensuing *inference results pertinently*, e.g. avoid unwarranted inferential claims or conflate 'inference results' with 'evidence' for or against hypotheses.

Example 2A (Abouk et al. 2022, Appendix J: Table 1. p. 99) On the basis of an estimated Linear Regression (LR) model with n=24732966, it is claimed that the estimates $\widehat{\beta}_k$=.004, SE($\widehat{\beta}_k$)=.002, render the coefficient β_k of a crucial variable x_k statistically significant at α=.05!

(i) Statistical misspecification. The authors of this claim brushed aside any questions relating to the *validity* of the probabilistic assumptions invoked when the LR model is estimated for inference purposes (e.g. t-tests). For inference purposes the probabilistic assumptions that matter are not the ones for the error process $\{(u_t|X_t{=}x_t),\ t{\in}\mathbb{N}\}$, but those of the observable process $\{(Y_t|X_t{=}x_t),\ t{\in}\mathbb{N}\}$ underlying data $\mathbf{z}_0{:=}\{(x_t,y_t),\ t{=}1,2,...,n)$; see Spanos (1986, 2006a-b, 2010).

A complete specification of the LR model in terms of internally consistent and testable set of probabilistic assumptions relating to the observable process $\{(Y_t|X_t{=}x_t),\ t{\in}\mathbb{N}\}$ is given in Table 3 comprising assumptions [1]-[5]. The statistical parametrization of

$\boldsymbol{\theta}:=(\beta_0,\beta_1,\sigma^2)$ constitutes an integral part of the specification that addresses numerous confusions in textbook econometrics, including the omitted variables bias, endogeneity/exogeneity, Instrumental Variables, GMM, etc., etc.; see Spanos (1986, 2006a-c, 2009, 2010a).

Table 3: Normal, Linear Regression model

Statistical GM: $Y_t=\beta_0+\beta_1 x_t+u_t,\ \ t\in\mathbb{N}:=(1,2,...,n,...)$

[1]	Normality:	$(Y_t\|X_t=x_t)\backsim \mathsf{N}(.,.)$,	$t\in\mathbb{N}$.
[2]	Linearity:	$E(Y_t\|X_t=x_t)=\beta_0+\beta_1 x_t$,	
[3]	Homoskedasticity:	$Var(Y_t\|X_t=x_t)=\sigma^2$,	
[4]	Independence:	$\{(Y_t\|X_t=x_t),\ t\in\mathbb{N}\}$ indep. process,	
[5]	t-invariance:	$\boldsymbol{\theta}:=(\beta_0,\beta_1,\sigma^2)$ are *not* changing with t,	

$\beta_0=E(y_t)-\beta_1E(X_t),\ \beta_1=(\frac{Cov(X_t,y_t)}{Cov(X_t)}),\ \sigma^2=Var(y_t)-\beta_1Cov(X_t,y_t).$

(**6**) How can one establish the *statistical adequacy* of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ using effective *Mis-Specification (M-S) testing* and *respecification*?

The most effective form of Mis-Specification (M-S) testing for evaluating the statistical adequacy of the LR model comes in the form of joint tests based on auxiliary regressions for probing the assumptions [2]-[5] (Table 1.

The first auxiliary regression specifies how departures from different assumptions might affect the regression function $E(Y_t\mid X_t=x_t)=\beta_0+\beta_1x_t$:

$$\widehat{u}_t=\delta_0+\delta_1x_t+\overbrace{\delta_2t}^{\neg[5]}+\overbrace{\delta_3x_t^2}^{\neg[2]}+\overbrace{\delta_4x_{t-1}+\delta_5Y_{t-1}}^{\neg[4]}+v_{1t}, \tag{16}$$

H_0: $\delta_1=\delta_2=\delta_3=\delta_4=\delta_5=0$ vs. H_1: $\delta_1\neq0$ or $\delta_2\neq0$ or $\delta_3\neq0$ or $\delta_4\neq0$ or $\delta_5\neq0$.
where $\neg[k]$ denotes the negation of $[k]$. The second auxiliary regression specifies how departures from different assumptions might affect $Var(Y_t\mid X_t=x_t)=\sigma^2$:

$$\widehat{u}_t^2=\gamma_0+\overbrace{\gamma_2t}^{\neg[5]}+\overbrace{\gamma_1x_t+\gamma_3x_t^2}^{\neg[3]}+\overbrace{\gamma_4x_{t-1}^2+\gamma_5Y_{t-1}^2}^{\neg[4]}+v_{2t}, \tag{17}$$

H_0: $\gamma_1=\gamma_2=\gamma_3=\gamma_4=\gamma_5=0$ vs. H_1: $\gamma_1\neq0$ or $\gamma_2\neq0$ or $\gamma_3\neq0$ or $\gamma_4\neq0$ or $\gamma_5\neq0$.
Intuitively, the above auxiliary regressions can be viewed as attempts to probe the residuals $\{\widehat{u}_t,\ t=1,2,...,n\}$ for any *remaining systematic statistical information* that might have been missed by the LR model. More formally, the additional terms in (16) and (17) will be *zero* when assumptions [2]-[5] are valid for data $\mathbf{Z}_0$. It is no accident that M-S tests are often specified in terms of the residuals since $\{\widehat{u}_t,\ t=1,2,...,n\}$ constitute a maximum ancillary statistic; see Spanos (2010b).

3.3.1 Statistical misspecification and its consequences

To illustrate the effects of invalid assumptions on the reliability of inference, consider the following **simulation example** (Spanos and McGuirk, 2001) based on $N=10,000$ replications with $n=50$ and $n=100$ using the following two scenarios.

S1. The **estimated LR model is statistically adequate**. That is, the probabilistic assumptions of [1]-[5] are valid for data $\mathbf{z}_0$; the true and estimated models coincide: $Y_t=\beta_0+\beta_1 x_t+u_t$.

When the estimated LR model is statistically adequate: (i) the empirical means of the N point estimates are *highly accurate* and the empirical (actual) type I error probabilities associated of the t-tests are very close to the nominal (α=.05) even for a sample size n=50, , and (ii) their accuracy improves as n increases.

S2. $Y_t=\beta_0+\beta_1 x_t+u_t$ is **estimated**, but the **true model** is $Y_t=\delta_0+\delta_1 t+\beta_1 x_t+u_t$. This renders invalid assumption [5] since $\beta_0(t)=\delta_0+\delta_1 t$.

Table 4: Linear Regression (LR) and misspecification

	S1: Adequate LR model				**S2: Misspecified LR model**			
Replications: N=10000	True: Y_t=1.5+0.5x_t+u_t, Estim: $Y_t=\beta_0+\beta_1 x_t+u_t$,				True: Y_t=1.5+.13t+.5x_t+u_t Estim: $Y_t=\beta_0+\beta_1 x_t+u_t$,			
	n=50		n=100		n=50		n=100	
Parameters	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$[\beta_0$=1.5] $\hat{\beta}_0$	1.502	.122	1.500	.087	0.462	.450	0.228	.315
$[\beta_1$=.5] $\hat{\beta}_1$	0.499	.015	0.500	.008	1.959	.040	1.989	.015
$[\sigma^2$=.75] $\hat{\sigma}^2$	0.751	.021	0.750	.010	2.945	.384	2.985	.266
$[\mathcal{R}^2$=.25]R^2	0.253	.090	0.251	.065	0.979	.003	0.995	.001
t-statistics	Mean	α=.05	Mean	α=.05	Mean	α=.05	Mean	α=.05
$\tau_{\beta_0}=\frac{\hat{\beta}_0-\beta_0}{\hat{\sigma}_{\beta_0}}$	0.004	.049	0.015	.050	-1.968	0.774	-3.531	0.968
$\tau_{\beta_1}=\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}_{\beta_1}}$	-.013	.047	-.005	.049	35.406	1.000	100.2	1.000

When the estimated model is misspecified: (i)* the empirical overall means based on the N estimates are *highly inaccurate* (inconsistency) and the actual type I error probabilities are *much larger* (> .77) than the nominal (α=.05), and (ii)* as n increases the inaccuracy of the estimates increases and the actual type I error probabilities approach 1! This undermines completely the claim that the combination of (a) a very large n and (b) invoking asymptotic inferences can sidestep the problem of establishing statistical adequacy is fallacious. The only criterion for evaluating the *reliability* of frequentist inferences is: **actual error probabilities≃ nominal ones**. This pertains to a wide variety of robustness claims currently invoked by textbook econometrics. There are *no real robustness results* for generic departures from IID; see Spanos (2002). The only credible robustness claim relates to certain forms of non-Normality, assuming that all the other assumptions [2]-[5] are valid! It should also be emphasized that in table 3 only assumption [5] is invalid, but in practice there are often several such invalid assumptions. This simulation also illustrates the inanity of addressing the large n problem when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is misspecified; one cannot keep track of the error probabilities to be able to adjust them.

(ii) Large n problem. The statistical significance (with n=24732966) at α=.05 of β_k is taken at face value, oblivious to the large n problem in N-P testing.

(iii) Conflating 'testing results' with 'evidence'. The authors claim *evidence for* $\beta_k \neq 0$, and proceeds to draw conclusions about its substantive implications for the effectiveness of different economic policies.

Let us elaborate on **(ii)** and **(iii)** using the reported results in Abouk et al. (2022), using scenario 2. For the LR model the sampling distribution of $\widehat{\boldsymbol{\beta}}:=(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}_1)$ is:

$$(\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})|\mathbf{X}) \backsim \mathsf{N}(\mathbf{0}, \sigma^2\mathbf{Q}_X^{-1}),\ \lim_{n\to\infty}(\tfrac{\mathbf{X}^\top\mathbf{X}}{n})=\mathbf{Q}_X=[q_{ij}]_{i,j=1}^m.$$

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

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

Example 2A (Abouk et al. 2022). Report $\widehat{\beta}_k$=.004, $\mathrm{SE}(\widehat{\beta}_k)=\sqrt{\frac{s}{n\sqrt{q_{kk}}}}$=.002, and $p(\mathbf{z}_0)$<.05, implying $\frac{\sqrt{n}(\widehat{\beta}_k-0)}{s\sqrt{q_{kk}}}=\frac{\sqrt{24732966}(.004)}{\sqrt{98.932}}$=2, for $c_{.025}$=1.96 and $p(\mathbf{z}_0)$=.045.

Using this information, we can reconstruct what would $p_n(\mathbf{z}_0)$ have been for different n using scenario 2 in Table 5, which indicate that the claim of statistical significance ($\beta_k \neq 0$) will be unwarranted for any $n<24,000,000$.

Table 5: the p-value with increasing n (constant estimates)

n	100	500	1000	2000	10000	10×10^4	10×10^5	20×10^5	20×10^6	24×10^6
$\tau_n(\mathbf{x}_0)$	.004	.009	.0127	.018	.040	.127	.402	.569	1.798	1.970
$p_n(\mathbf{x}_0)$	.997	.993	.990	.986	.967	.899	.688	.570	.072	.049

The large n problem is particularly pernicious in applied micro studies with a huge n since spurious statistical significance results are often used to frame policy decisions.

Example 2B (Abouk et al. 2022). The authors evaluate the difference between the two means using the t-test (Lehmann and Romano, 2005):

$$T_\alpha^d:\ \ \tau(\mathbf{Z})=[\sqrt{\tfrac{n_1n_2}{n_1+n_2}}(\overline{X}_n-\overline{Y}_n)/s_n] \overset{\mu_1=\mu_2}{\backsim} \mathsf{St}(n_1+n_2-2),\ C_1=\{\mathbf{z}:\ |\tau(\mathbf{z})|>c_{\frac{\alpha}{2}}\}. \quad (18)$$

The authors (Table 2, p. 43) report 50 ANOVA results for the difference between two means $(\overline{x}_n-\overline{y}_n)$, 48 of which are very small, $(\overline{x}_n-\overline{y}_n)<.1$, but their p-values are tiny, $<.0000$, due to n=24,732,966. Surprising, however, for the remaining 2, whose $(\overline{x}_n-\overline{y}_n)$=0 ($\overline{x}_n$=13.2, $\overline{y}_n$=13.2) and ($\overline{x}_n$=2.51, $\overline{y}_n$=2.51), the reported p-values are .8867 and .0056, respectively. This is surprising since for $(\overline{x}_n-\overline{y}_n)$=0 one expects $\tau(\mathbf{z}_0)$=0 and $p(\mathbf{z}_0)$=1. What did go wrong? Blame the statistical software for using (at least) 12 digit decimal precision and not 2.

For example, when $(\overline{x}_n$=2.51+.000001$-\overline{y}_n$=2.51) =.000001$\neq$0, yields $p(\mathbf{z}_0)$=.0056, which will occur with n_1=11,004,078, n_2=13,728,888, s_n=.000032 $\rightarrow \tau(\mathbf{z}_0)$=2.77.

3.4 '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 and Romano (2005).

1. **Naive rule of adjustment** to α:

n	100	200	500	1000	10000	20000	200000	$\cdots$
α	.05	.025	.01	.001	.000001	.000000001	.000000000001	$\cdots$

2. Good (1988) proposes to **standardize the p-value** $p(\mathbf{x}_0)$ to the relative sample size n=100 and α=.05 using the following rule of thumb:

$$p_{100}(\mathbf{x}_0)=\min\left(.5,\left[p_n(\mathbf{x}_0)\cdot\sqrt{n/100}\right]\right),\ n>10. \tag{19}$$

Empirical example (continued). The above formula yields:

n=120, $p_{120}(\mathbf{x}_0)$=.038: $p_{100}(\mathbf{x}_0)$=.0416
n=150, $p_{150}(\mathbf{x}_0)$=.024: $p_{100}(\mathbf{x}_0)$=.0294
n=300, $p_{300}(\mathbf{x}_0)$=.002: $p_{100}(\mathbf{x}_0)$=.0035
n=500, $p_{500}(\mathbf{x}_0)$=.0002: $p_{100}(\mathbf{x}_0)$=.0045
n=1000, $p_{1000}(\mathbf{x}_0)$=.000001: $p_{100}(\mathbf{x}_0)$=.000003
n=10000, $p_{10000}(\mathbf{x}_0)$=.000...0: $p_{100}(\mathbf{x}_0)$=.000...0

The above results suggest that rules of thumb can be useful in helping to select α to take into account the inherent trade-off between the type I and II error probabilities, and avoid spurious results based on over-sensitive tests. They do not address the large n problem, however, since they are ad hoc and the p-value will be very close to zero beyond n=100000 in practice. Instead, what is needed is to replace the rules of thumb with a principled argument that provides an evidential interpretation of the accept/reject H_0 results by taking fully into account the relevant statistical context: (i) $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, (ii) H_0: $\theta\in\Theta_0$ vs. H_1: $\theta\in\Theta_1$, (iii) T_α:=$\{d(\mathbf{X}),C_1(\alpha)\}$, (iv) data $\mathbf{x}_0$.

4 The post-data severity evaluation

The key idea underlying the general concept of 'severe testing' is that an inferential claim H is warranted only when the different ways it can be false have been adequately probed and forfended; see Mayo (1996). Applying that general idea to the N-P testing 'accept/reject H_0' results takes the form of a post-data severity evaluation of such results with a view to use $d(\mathbf{x}_0)$ in the context of test T_α to evaluated the warranted discrepancy γ from the null value θ=θ_0 with high enough probability.

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

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

(C-2) with very high probability, test T_α would have produced a result that 'accords less well' with H than $\mathbf{x}_0$ does, if H were false (Mayo and Spanos, 2006, 2011).

4.1 Case 1: accept H_0

Let us illustrate the post-data severity evaluation based on C-1 and C-2.

Empirical example 1 (continued). Recall that for α=.05, c_α=1.66, $\overline{x}_n$=2.317, s=1.941, and n=100, test T_α in (5) yields: $\frac{\sqrt{100}(2.317-2)}{1.941}$=1.6533[.0507], *'accept* H_0'.

(C-1) indicates that $\mathbf{x}_0$ accords with H_0, and thus the relevant inferential claim is:

$$\mu\leq\mu_1=\mu_0+\gamma,\ \text{for}\ \gamma\geq 0. \tag{20}$$

(C-2) suggests that the post-data severity evaluation relating to (20) should be based on evaluating the probability of the event: "outcomes $\mathbf{x}$ that accord less well with $\mu \leq \mu_1$ than $\mathbf{x}_0$ does", i.e. event $[\mathbf{x}\text{: } \tau(\mathbf{x}) > \tau(\mathbf{x}_0)]$, defining $\text{SEV}(T_\alpha; \mu \leq \mu_1)$ by:

$$SEV(T_\alpha; \mu \leq \mu_1) = \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0);\ \mu = \mu_1),\ \forall \mu \in \Theta_1 = (2, 3),$$

for different values of $\gamma = \mu_1 - \mu_0$, with the evaluations being based on:

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{s} \overset{\mu=\mu_1}{\backsim} \mathsf{St}(\delta_1; n-1),\ \delta_1 = \frac{\sqrt{n}(\mu_1 - \mu_0)}{s},\ \forall \mu_1 > \mu_0.$$

In the case of 'accept H_0' one is seeking the *'smallest' discrepancy* from $\mu_0 = 2$ with high enough probability.

The post-data severity evaluated for typical values is reported in table 6.

Table 6: Post-data Severity evaluation for $\mu_1 \leq 2.0 + \gamma$

γ	.05	.1	.15	.20	.30	.317	.40	.481	.60	.70
μ_1	2.05	2.1	2.15	2.2	2.3	2.317	2.4	2.481	2.6	2.7
$\text{SEV}(\mu \leq \mu_1)$	.086	.133	.196	.274	.465	.500	.665	.800	.926	.974

The evaluation of $SEV(T_\alpha; \mu \leq \mu_1)$ itself takes the form:

for $\gamma = .05$, $1.6332 - \frac{\sqrt{100}(2.05-2)}{1.941} = 1.3756$ yields $SEV(T_\alpha; \mu \leq \mu_1) = .086$,

for $\gamma = .1$, $1.6332 - \frac{\sqrt{100}(2.1-2)}{1.941} = 1.118$ yields $SEV(T_\alpha; \mu \leq \mu_1) = .133$,

for $\gamma = .481$, $1.6332 - \frac{\sqrt{100}(2.481-2)}{1.941} = -.845$ yields $SEV(T_\alpha; \mu \leq \mu_1) = .8$.

Figure 3, depicting the post-data severity curve for all $\mu \in [1.8, 3.0]$, indicates that the discrepancy warranted by data $\mathbf{x}_0$ and test T_α with probability (SEV) .8 is $\gamma^\ddagger \leq .481$ ($\mu_1 \leq 2.481$). Also, the results in table 5 bring out two important cases.

(i) The warranted discrepancy $\mu_1 \leq 2.481$ is bigger than $\overline{x}_n = 2.317$.

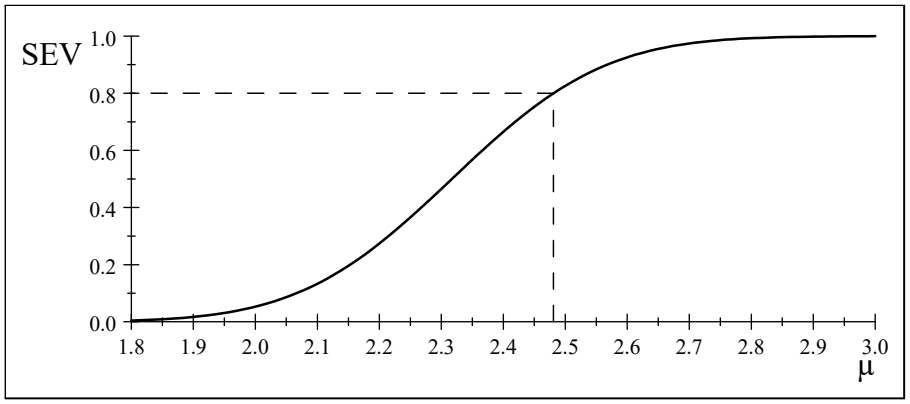

Fig. 3: the post-data severity curve (accept H_0)

(ii) The probability associated with the relevant inferential claim $\mu \leq \mu_1 = \mu_0 + \gamma$, for $\mu_1 = 2.317 = \overline{x}_n$, is always .5; not high enough for $\gamma = .317$ to be the discrepancy warranted by $\mathbf{x}_0$.

4.2 The post-data severity and the large n problem

4.2.1 Case 1: accept H_0

Using the fact that the point estimates for $\overline{x}_n$ and s do not usually change significantly as n increases for values beyond n=100 when the data $\mathbf{x}_0$ stem from an IID sample, we will consider the following counter-factual scenario.

Counter-factual scenario: what if the estimates $\overline{x}_n$=2.317 and s_n=1.941 remain constant as n increases? What would the warranted discrepancy be at SEV(γ)=.8?

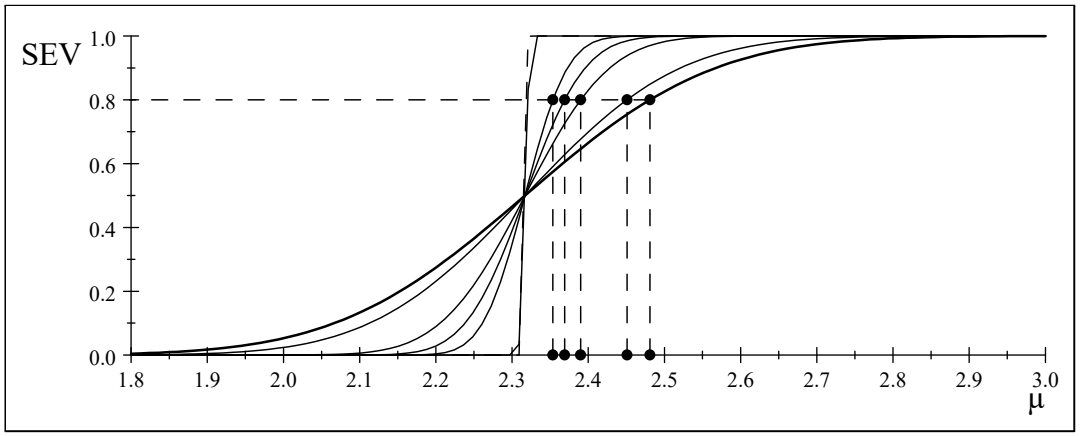

Fig. 4: the severity curve (accept H_0) for different n (same estimates)

Empirical example 1 (continued). Figure 4 depicts the post-data severity curves as the sample size n (keeping the estimates $\overline{x}_n$=2.317 and s_n=1.941 constant), indicating that increasing n renders the severity curves steeper and steeper, reducing the warranted discrepancy $\gamma^{\ddagger}_n$ monotonically (Table 7) until they reach the lower bound at $\gamma^{\ddagger}\leq$.3217 ($\overline{x}_n$=2.317).

Table 7: Post-data severity ($SEV(\gamma)$=.8, $\mu_1\leq 2+\gamma_1$, $\overline{x}_n$=2.317, s=1.941)

n:	100	120	150	200	300	500	1000	2000	20000	200000
$\tau_n(\mathbf{x}_0)$:	1.633	1.789	2.0	2.310	2.829	3.652	5.165	7.304	23.097	73.038
μ_1=2+γ:	2.481	2.467	2.451	2.434	2.412	2.390	2.369	2.3535	2.329	2.321

This makes sense in practice since $\overline{X}_n$ and s_n are strongly consistent estimators of μ^* and σ_*, i.e. $\mathbb{P}(\lim_{n\to\infty}\widehat{\theta}_n(\mathbf{X})=\theta^*)$=1, and thus their accuracy (precision) improves as n increases beyond a certain threshold N. Recall that in the case of 'accept H_0' one is seeking the 'smallest discrepancy' from μ_0=2. Hence, as n increases SEV renders the warranted discrepancy $\gamma^{\ddagger}_n$ 'more accurate' by reducing it, until it reaches the lower bound around $\gamma^{\ddagger}\leq$.317 ($\mu_1=\overline{x}_n$).

4.2.2 Case 2: reject H_0

Empirical example 3. For the hypotheses in (13), let the estimate of μ be $\overline{x}_n$=2.726 retaining s=1.941, yields:

$$\tau_n(\mathbf{x}_0)=\frac{\sqrt{100}(2.726-2)}{1.941}=3.740[.00009],$$

which indicates a strong 'reject H_0'. In contrast to the case 'accept H_0' for 'reject H_0' one is seeking the 'largest' discrepancy from μ_0=2 with high severity. The relevant inferential claim takes the form:

$$\mu \geq \mu_1 = \mu_0 + \gamma, \text{ for } \gamma \geq 0, \tag{21}$$

and its post-data probabilistic evaluation is based on:

$$SEV(T_\alpha; \mu \geq \mu_1) = \mathbb{P}(\tau(\mathbf{X}) < \tau(\mathbf{x}_0);\ \mu = \mu_1),\ \forall \mu \in \Theta_1 = (2, 3).$$

Table 8 indicates that the discrepancy γ warranted with $SEV(T_\alpha; \mu \geq \mu_1)$=.8 is $\gamma\ddagger \leq .56$ ($\mu_1 \leq 2.562$), and the discrepancy for $\mu_1 \leq \overline{x}_n$=2.726 has $SEV(T_\alpha; \mu \geq \mu_1)$=.5.

Table 8: Post-data Severity evaluation for $\mu_1 \geq 2 + \gamma$

γ	.1	.2	.3	.4	.5	.562	.6	.726	.8	.9
μ_1	2.1	2.2	2.3	2.4	2.5	2.562	2.6	2.562	2.6	2.7
SEV($\mu \geq \mu_1$)	.999	.996	.984	.951	.876	.800	.741	.500	.352	.186

Counter-factual: what if the original (n=100) estimates $\overline{x}_n$=2.726 and s=1.941 remained the same but the sample size n increases. How would the warranted discrepancy at SEV($\mu_0+\gamma$))=.8 change?

Figure 5 depicts all the severity curves for n=100, 200, 500, 1000, 10000, indicating that increasing the sample size n (keeping $\overline{x}_n$=2.726 and s=1.941 constant) renders the curves steeper and steeper, 'increasing' the warranted discrepancy monotonically (reject H_0) up to the lower bound $\gamma\ddagger \leq .75$ ($\mu_1 = \overline{x}_n$=2.726).

As in the case of accept H_0, the strong consistency of $\overline{X}_n$ and s_n ensures that increasing n in practice will improve the precision of the warranted discrepancy $\gamma^\ddagger_n$.

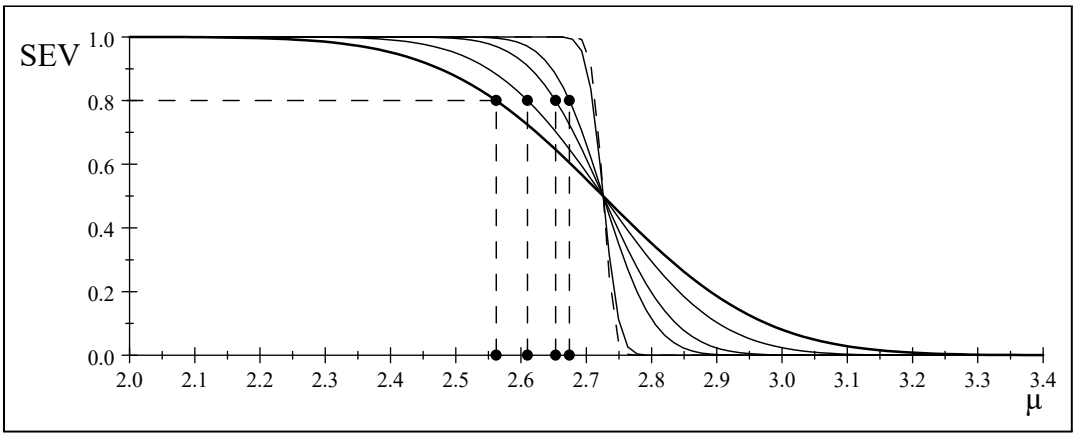

Fig. 5: the severity curve (reject H_0) for different n (same estimates)

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

$$\frac{\sqrt{n}(\widehat{\beta}_k - 0)}{s\sqrt{q_{kk}}} = \frac{\sqrt{24732966}(.004)}{\sqrt{98.932}} = 2.0[.045].$$

What is the warranted discrepancy γ from β_k=0 with high severity, say .977, in light of n=24732966? The answer is $\gamma\ddagger \leq .0000001$, and not $\widehat{\beta}_k$=.004 or Cohen's $d(\mathbf{z}_0)$=.0004.

4.2.3 Key features of the post-data SEV evaluation

(1). It provides a *principled argument* to replace the ad hoc rules of thumb by an evidential interpretation of the accept/reject H_0 results based on the discrepancy $\gamma \gtrless 0$ from $\mu=\mu_0$ *warranted* by data $\mathbf{x}_0$ and test T_α with high enough probability.

(2). The $SEV(\mu_0+\gamma)$ evaluation is a statistical context-specific *error probability* grounded in (10), with the sign and magnitude of $\tau_n(\mathbf{x}_0)$ suggesting the direction of relevant discrepancies γ.

(3). What distinguishes the $SEV(\mu_0+\gamma)$ evaluation from other attempts to deal with the large n problem is that its outputting of the discrepancy $\mu_1=\mu_1-\mu_0$ is based on the *non-central distribution*:

$$\left(\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{s}-\boxed{\frac{\sqrt{n}(\mu_1-\mu_0)}{s}}\right)\overset{\mu=\mu_1}{\backsim}\mathsf{St}(n-1),\ \text{for all } \mu_1\in\Theta_1. \tag{22}$$

This ensures that the warranted discrepancy $\gamma^\ddagger$ is evaluated using the 'same' sample size n, counter-balancing the effect of n on $\tau_n(\mathbf{x}_0)$. In contrast, the p-value is evaluated under $\mu=\mu_0$ for which $\delta_0=0$, and the power replaces $\tau_n(\mathbf{x}_0)$ in (22) with a constant c_α; see Spanos (2013a-b, 2014).

(4). The evaluation of the warranted $\gamma_n^\ddagger$ accounts for the increase in n by enhancing its precision. In the case of accept (reject) H_0 this comes in the form of reducing (increasing) $\gamma_n^\ddagger$ as n increases since one is seeking the smallest (largest) discrepancy from μ_0. For a very large n the warranted discrepancy $\gamma^\ddagger$ will approach the value $\widehat{\theta}_n(\mathbf{x}_0)$ since for a statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, $\widehat{\theta}_n(\mathbf{x}_0)$ should be accurate enough.

(5). The post-data severity perspective can explain the flaws of the p-value since (i) $p(\mathbf{x}_0)<\alpha$ indicates the presence of 'some' discrepancy γ, but provides no information about its magnitude, (ii) the distribution $\tau_n(\mathbf{X})$ is evaluated only under H_0, and thus (iii) $p(\mathbf{x}_0)$ is vulnerable to the large/small n problems; see Spanos (2021b).

5 Additional foundational problems in N-P testing

5.1 N-P vs. significance testing: the claimed incompatibility

Fisher's significance testing and the Neyman-Pearson (N-P) testing constitute two variants of frequentist testing which are often presented as essentially 'incompatible' and any attempt to blend them will result in an "inconsistent hybrid" which is "burdened with conceptual confusion" (Gigerenzer, 1993, p. 324).

This claim is questionable since both approaches employ *the same*:

(a) underlying *statistical set up*, in the form of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$,

(b) underlying *'hypothetical' reasoning*, and

(c) probing $\theta=\theta_0$ within the same statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

The key difference is that the *pre-data* error probabilities (type I, II, power) for the N-P testing aim to *calibrate* the generic capacity of a test, and the *post-data* p-value for Fisher's significance testing aims to provide a *measure of discordance* between $\theta=\theta_0$ and $\mathbf{x}_0$. Once the pre-data vs. post-data distinction is made explicit, the two approaches can be harmoniously blended; see also Lehmann (1993).

The confusion in the current literature stems from burdening Fisher's post-data p-value with a definition adapted to fit the N-P testing framing of H_0 and H_1.

Pre-data definition. The p-value is the probability of obtaining a result 'equal to or more extreme' than the one observed $\mathbf{x}_0$, when H_0 is true'.

In the context of N-P testing, the clause 'equal to or more extreme' is invariably interpreted in light of the *framing* of H_1. This has led to the p-value being viewed as *the smallest significance level* $\alpha_{\min}$ at which H_0 would have been rejected when H_0 is true. There is nothing wrong with this definition when used in N-P testing. In the context of Fisher's significance testing with a point null, say $\theta{=}\theta_0$, however, this definition can be misleading since the additional information relating to the sign and magnitude of $d(\mathbf{x}_0)$, pinpoints the direction of relevant departures from $\theta{=}\theta_0$, which can be at odds with the direction indicated by the particular H_1. To rectify that, the severity evaluation suggests an alternative *post-data* definition of the p-value.

Post-data definition. The p-value is the probability of all possible outcomes $\mathbf{x}{\in}\mathbb{R}^n_X$ that accord less well with H_0 than $\mathbf{x}_0$ does $[d(x_0)]$, when H_0 is true.

The key difference is that the post-data definition ensures that the sign and magnitude of $d(\mathbf{x}_0)$, and *not* the particular N-P framing of H_1, determine the direction of the relevant departures. To illustrate that let us revisit the Berger example where $d(\mathbf{x}_0){=}-2.6833$ pointing at departures of the form $\theta_1{=}\theta_0{+}\gamma,\ \gamma{<}0$. Hence, the post-data p-value is $p(\mathbf{x}_0){=}\mathbb{P}(d(\mathbf{X}){<}d(\mathbf{x}_0);\ \theta{=}\theta_0){=}.036$, which is at odds with the direction indicated by H_1: $\theta{>}.5$ due to Berger's rigged framing.

When the post-data p-value is viewed from the severity perspective, $p(\mathbf{x}_0){<}\alpha$ indicates the presence of 'some' discrepancy γ, but provides no information about its magnitude (Mayo and Spanos, 2006) since (i) the sampling distribution underlying $p(\mathbf{x}_0)$ is evaluated only under H_0, (ii) $p(\mathbf{x}_0)$ is vulnerable to the large/small n problems (e.g. high/low power), and (iii) the pre-data N-P framing; see Spanos (2021b).

5.2 Statistical *results* vs. *evidence* for an inferential claim

Given that the primary aim of frequentist testing is to learn from data $\mathbf{x}_0$ about θ^*, the testing results accept/reject H_0 are too coarse to provide genuine evidence about θ^*. For instance, in testing the hypotheses, H_0: $\mu{\leq}\mu_0$ vs. H_1: $\mu{>}\mu_0$, using the UMP test T_α (section 2.4), a rejection of H_0 with data $\mathbf{x}_0$ warrants the coarse claim $\theta^*{\in}(\mu_0,\infty)$, which is not informative enough for μ^*.

Example 4. For the simple Bernoulli model in (26), the relevant data refer to newborns during a year where $(X{=}1)$ stands for a 'boy (B)' and $(X{=}0)$ stands for a 'girl (G)'. Let the hypotheses of interest by H_0: $\theta \leq \theta_0$ vs. H_1: $\theta > \theta_0$, for $\theta_0{=}.5$, where $\theta{=}E(X){=}\mathbb{P}(X{=}1)$. The data come from two different locations and more than 3 centuries apart, $\mathbf{x}_1$ (Cyprus, 1993) and $\mathbf{x}_2$ (London, 1687).

Applying the optimal test $T^{>}_\alpha$ to both data sets yields the results in Table 9, with p-values in square brackets. The post-data severity curves (fig. 6) depict the $SEV(T_\alpha;\mathbf{x}_i;\theta{>}\theta_1)$ for $\theta_1{=}\theta_0{+}\gamma_1$, for data $\mathbf{x}_i,\ i{=}1,2$, relating to the relevant inferential claim $\theta{>}\theta_1{=}\theta_0{+}\gamma_1$, for $\gamma_1{>}0$; the underlying distribution of $d(\mathbf{X})$ used for these

evaluations is given in (28).

Table 9: N-P testing of H_0: $\theta \leq .5$ vs. H_1: $\theta > .5$		
Data	H_0: $\theta \leq .5$ vs. H_1: $\theta > .5$	SEV=.85
Cyprus 1993: $\mathbf{x}_1$: 5442 (B), 5072 (G):	$d(\mathbf{x}_1)=\frac{\sqrt{10514}(\frac{5442}{10514}-.5)}{\sqrt{.5(1-.5)}}=3.600[.00016]$	$\gamma^{\ddagger} \leq .0125$,
London 1687: $\mathbf{x}_2$: 7737 (B), 7214 (G):	$d(\mathbf{x}_2)=\frac{\sqrt{14951}(\frac{7737}{14951}-.5)}{\sqrt{.5(1-.5)}}=4.277[.000009]$	$\gamma^{\ddagger} \leq .0132$,

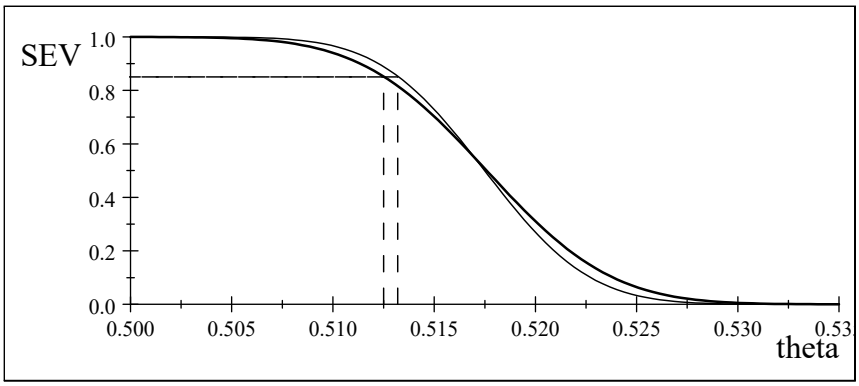

Fig. 6: Severity curve for data $\mathbf{x}_1$ and $\mathbf{x}_2$

Although the SEV is always attached to the relevant inferential claim $\theta>\theta_1$, it can be intuitively viewed as indicating that the relevant neighborhood of θ^* is around θ_1=.513 $\pm\,\varepsilon$, for $0<\varepsilon \leq .001$, with probability .85.

5.3 Statistical vs. substantive significance

The small p-values stemming from a large enough n is often misconstrued as conflating statistical with substantive significance, which is erroneous since the apparent statistical significance is spurious. To address the problem of statistical vs. substantive significance one needs to consider two different but interrelated dimensions of learning from data about phenomena of interest.

The first dimension relates to the precondition is that there is reliable enough knowledge relating to the sign and magnitude of a parameter, say $\varphi^{\blacklozenge}$, stemming from substantive subject matter information; ideally from a substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$.

The second dimension relates to having trustworthy evidence for the sign and magnitude of a statistical parameter θ stemming from a statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

The post-data severity evaluation of the accept/reject H_0 results can provide the link between the two parameters to address the problem. This is achieved by relating the discrepancy $\gamma^{\ddagger}$ from θ_0 (θ_1=$\theta_0 \pm \gamma^{\ddagger}$) warranted by test T_α and data $\mathbf{x}_0$ with high probability to the substantively determined $\varphi^{\blacklozenge}$.

Example 4 (continued). In human biology (Hardy, 2002) is known that the *substantive value* for the ratio of boys to all newborns is $\varphi^{\blacklozenge}\simeq$.5122. In the above example it is shown that for the two different data sets, $\mathbf{x}_1$ and $\mathbf{x}_2$, the severity-based warranted discrepancies are $\gamma_1^{\ddagger}\leq$.0125 ($\theta_1\leq$.5125) and $\gamma_2^{\ddagger} \leq$ .0132 ($\theta_2\leq$.5132),

respectively. Hence, it is clear that both statistically determined values also entail substantive significance since $\varphi^{\blacklozenge}\simeq.5122 \leq .5125 < .5132$. That is, the testing-based warranted discrepancy from the θ_0 provides a reliable evaluation of the 'scientific effect'; see Spanos (2021a-b).

5.4 Post-data severity vs. estimation-based effect sizes

Related to the above statistical vs. substantive significance is the concerted effort in psychology (Cohen, 1969/1988) to replace p-values with estimation-based effect sizes based on reformulating test statistics to get rid of its reliance on n, as a way to circumvent the large n problem. As argued by Abelson (1995) : "One advantage of the raw size effect as a measure is that its expected value is independent of the size of the sample used to perform the significance test." (p. 46)

Although there are many different formulae for effect sizes in the context of different statistical models (Ellis, 2010), the discussion will focus on Cohen's d in an attempt to relate it the post-data severity; see Spanos (2021a) for broader discussion. Cohen (1969) proposed $d(\mathbf{z}_0)=\frac{(\overline{x}_n-\overline{y}_n)}{s_n}$ in the context of a simple bivariate Normal model underlying the testing of the difference between two means; see section 3.3 above. Given that the relevant test statistic is $\tau(\mathbf{Z})=[\sqrt{N}(\overline{X}_n-\overline{Y}_n)/s_n]$, where:

$$N=\tfrac{n_1n_2}{n_1+n_2},\ s_n^2=\tfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)},\ s_1^2=\tfrac{1}{(n_1-1)}\textstyle\sum_{i=1}^{n_1}(X_i-\overline{X}_n)^2,\ s_2^2=\tfrac{1}{(n_2-1)}\sum_{i=1}^{n_2}(Y_i-\overline{Y}_n)^2,$$

Cohen's $d(\mathbf{z}_0)$ constitutes a point estimate of the parameter $\varphi=\frac{(\mu_1-\mu_2)}{\sigma}$, $\sigma^2=E(s_n^2)$, with $d(\mathbf{Z})=[(\overline{X}_n-\overline{Y}_n)/s_n]$ the corresponding Maximum Likelihood (ML) estimator which is reparametrization invariant. This renders $d(\mathbf{z}_0)$ questionable and as as a way to circumvent the large N problem for two reasons.

First, $d(\mathbf{z}_0)=[(\overline{x}_n-\overline{y}_n)/s_n]$ is vulnerable to the small N problem. Abelson (1995) points out: "... it could be argued that with a smallish sample, one might obtain a big apparent effect without being able to reject the null hypothesis. In other words, an effect size ought not to be judged in totally in isolation, but in conjunction with the p-value."

Second, the inferential claim $d(\mathbf{z}_0)\simeq\varphi^*$ for a large enough N, is *unwarranted* for Cohen's $d(\mathbf{z}_0)$, like all point estimates. This stems from the fact that $d(\mathbf{z}_0)$ represents a single point $\mathbf{Z}=\mathbf{z}_0$ of the sampling distribution of the pivot $\tau(\mathbf{Z};\varphi)$:

$$\tau(\mathbf{Z};\varphi)=[\sqrt{N}(\overline{X}_n-\overline{Y}_n)/s_n]\overset{\varphi=\varphi^*}{\backsim}\mathsf{St}(n_1+n_2-2),\ \forall\mathbf{z}\in\mathbb{R}^{n_1+n_2}, \tag{23}$$

which is derived using *factual* (what if $\varphi=\varphi^*$) reasoning. Legitimate inferential claims relating to φ need to be framed in terms of (23). Hence, eliminating N from $\tau(\mathbf{Z};\varphi)$ does not address the large N problem – it conceals it – since $(\overline{X}_n-\overline{Y}_n)/s_n$ does not have a well-defined sampling distribution without $\sqrt{N}$. To address the problem one needs a legitimate sampling distribution relating to $(\overline{X}_n-\overline{Y}_n)/s_n$; see Spanos (2021a).

A more effective way to deal with the large N problem is to use the post-data severity evaluation of the accept/reject H_0 results based on:

$$\tau(\mathbf{Z})=\tfrac{\sqrt{N}(\overline{X}_n-\overline{Y}_n)}{s}\overset{\mu_1\neq\mu_2}{\backsim}\mathsf{St}(\delta_1;n_1+n_2-2),\ \delta_1=\tfrac{\sqrt{N}(\mu_1-\mu_2)}{\sigma}. \tag{24}$$

The post-data severity of the discrepancy $\gamma=(\mu_1-\mu_2)\neq 0$ warranted by test T^d_α and data $\mathbf{z}_0$ with high enough severity. In contrast to Cohen's $d(\mathbf{z}_0)$ estimation-based effect size, the warranted discrepancy $\gamma^\ddagger$, provides a 'test-based effect size' which is calibrated in terms of post-data error probabilities. This evaluates the 'scientific effect' a lot more effectively than any point estimate could.

5.5 The framing of the N-P hypotheses H_0 and H_1

How arbitrary is the N-P framing? It is not as arbitrary as the wide-spread misuse/abuse of N-P testing might suggest! Consider the following example proposed by Berger (2019) to demonstrate the fatuity of the p-value.

Example 5: **Berger.** Consider testing the hypotheses:

$$H_0\text{: } \theta=.5 \text{ vs. } H_1\text{: } \theta>.5. \tag{25}$$

in the context of the simple Bernoulli (Ber) model:

$$X_k \backsim \mathsf{BerIID}(\theta,\theta(1-\theta)),\quad x_k=0,1,\ 0<\theta<1,\ k\in\mathbb{N}, \tag{26}$$

assuming $\alpha=.05,\ c_\alpha=1.645,\ n=20,\ \overline{x}_n=.2$. An optimal (UMP) test for (25) is:

$T^{>}_\alpha:=\{d(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}},\ \mathcal{C}_1(\alpha)=\{\mathbf{x}\text{: } d(\mathbf{x})>c_\alpha\}\}$, yielding: $d(\mathbf{x}_0)=-2.683$,

with a p-value $p(\mathbf{x}_0)=.996$, indicating 'accept H_0'; ha, ha,...,ha!

What is the real culprit behind this apparently absurd result? A hint about the real culprit can be gleaned by evaluating the discrepancy γ_1 for $\theta_1=.5+\gamma_1$ that ensures high power, say $\mathcal{P}(\theta_1)=.8$, which is $\gamma^\ddagger_1=.2636$ ($\theta_1=.7636$); see fig. 7. This discrepancy, however, is utterly uninteresting from a statistical perspective since the data $\mathbf{x}_0$ ($\overline{x}_n=.2$) indicate that $\theta^*\in(0,.5)$, which is excluded by the framing in (25). This suggests that the real source of the absurd result 'accept H_0' is likely to be the *framing* of H_0 and H_1 as it relates to the power of this test.

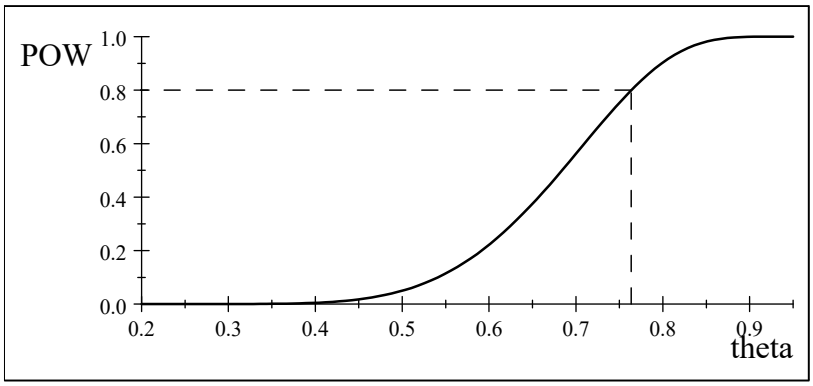

Fig. 7: Power curve for Berger's example

In particular, the (implicit) power of test T_α to detect discrepancies around $\theta_1=.2$ ($\overline{x}_n=.2$) is $\mathcal{P}(\theta_1=.2)=.0000003$ (figure 9), hence the absurd result 'accept $\theta=.5$'. In that sense Berger's example does not expose the fatuity of the p-value, but the absurdity of the original framing in (25) that excludes $\theta^*\in(0,5)$, which turns the probing

with a UMP test into 'a wild goose chase', which gets worse as n increases since for n=100, $\mathcal{P}(\theta_1=.2)\simeq 0$.

A careful reading of Neyman and Pearson (1933) reveals two crucial *preconditions* for the framing of H_0: $\theta\in\Theta_0$ vs. H_1: $\theta\in\Theta_1$ that secure the effectiveness of N-P testing and the informativeness of its results in learning from data about θ^*:

[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 to ensure that the type I error is *the more serious* of the two (using the analogy with a criminal trial for which H_0: not guilty; p. 296).

Given that an optimal N-P test is chosen by fixing α to a low value and selecting a test that minimizes the type II error (accepting H_0 when false) probability, [2] implies that low type II error probability (or equivalently high power) is needed around the potential neighborhood of θ^*- the true value of θ.

Berger's framing H_0: θ=.5 vs. H_1: θ>.5, however, ensures that the test T_α has high power for discrepancies $\theta_1\in(.5,1)\subset\Theta$ which is the 'wrong' subset as a result of flouting both preconditions [1]-[2].

What would a proper framing be in this example? When no reliable pre-data information about the potential neighborhood of θ^* is available, one should always use a two-sided test to avoid negating [1], i.e. $\theta^*\notin[\Theta_0\cup\Theta_1]$. What happens when the framing is H_0: θ=.5 vs. H_1: $\theta\neq$.5? One would 'reject H_0' with $p(\mathbf{x}_0)$=.0073, since $\mathcal{P}(\theta_1=.2\pm\epsilon)\geq.94$ for ϵ=.1. An even better framing that satisfies [1]-[2] is:

$$H_0\text{: }\theta\geq.5\text{ vs. }H_1\text{: }\theta<.5.$$

Applying the UMP test $T_\alpha^<:=\{d(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}},\ \mathcal{C}_1(\alpha)=\{\mathbf{x}\text{: }d(\mathbf{x})<c_\alpha\}\}$.

$$d(\mathbf{x}_0)=\frac{\sqrt{20}(.2-.5)}{\sqrt{.5(1-.5)}}=-2.6833[.0036],\text{ rejects }H_0.$$

Hence, the relevant severity inferential claim is $\theta_1\leq\theta_0-\gamma,\ \gamma>0$:

$$SEV(T_\alpha;\theta\leq\theta_1)=\mathbb{P}(d(\mathbf{X})\leq d(\mathbf{x}_0);\ \theta=\theta_1),\ \forall\theta\in(0,.5),\tag{27}$$

for different values of $\gamma=\theta_1-\theta_0$, with the SEV evaluations based on:

$$\frac{[d(\mathbf{X})-\delta(\theta_1)]}{\sqrt{V(\theta_1)}}\overset{\theta=\theta_1}{\backsim}\mathsf{Bin}\,(0,1;n)\simeq\mathsf{N}(0,1),\ \delta(\theta_1)=\frac{\sqrt{n}(\theta_1-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}},\ V(\theta_1)=\frac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)}.\tag{28}$$

It is important to note that the inequality for the evaluation in (27) stems from $\gamma\leq 0$.

The post-data severity curve in (27) (figure 8), indicates that the warranted discrepancy with high enough probability, say .7 (n=20), associated with the relevant inferential claim:

$$\theta_1<\theta_0+\gamma,\ \gamma<0\text{ is }\theta_1\leq.157.$$

5.6 Rigging the significance level α to get a desired result

Another problem with N-P testing is the dichotomous nature of the accept/reject H_0 rules which render the choice of α vulnerable to rigging by selecting it after the p-value has been evaluated. Indeed, this issue has been raised numerous times in the literature as a major flaw because intuitively $p(\mathbf{x}_0)$=.049 and $p(\mathbf{x}_0)$=.051 should give rise to similar evidence despite the reversal of the 'accept/reject H_0'.

To illustrate this, let us return to empirical example 1, where the UMP test yields $\tau(\mathbf{x}_0)=\frac{\sqrt{100}(2.317-2)}{1.941}=1.6533[.0507]$, indicating 'accept H_0' at $\alpha=.05$. What if, one wants to rig the result by selecting $\alpha=.051$, which will reverse the original result. Although this might seem too obvious as a rigging attempt, one can contrive a much less obvious examples with a larger n.

The post-data severity evaluation addresses this issue by:

(a) transforming these inference results into evidence pertaining to 'the discrepancy $\gamma^{\ddagger}$ warranted by test T_α and data $\mathbf{x}_0$', and

(b) avoiding the dichotomy created by α using the information relating to the sign and magnitude of $\tau(\mathbf{x}_0)$.

For instance, in the above example $\tau(\mathbf{x}_0)=1.6533$ suggests that the discrepancy from $\mu_0=2$ indicates by $\tau(\mathbf{x}_0)$ is clearly $\gamma>0$ and that calls for retaining the severity curve in figure 3, despite the rigged reject H_0 at $\alpha=.051$ result. Probing for warranted discrepancies $\gamma<0$ makes no sense post-data, and the severity curve remains the same since it revolves around $\tau(\mathbf{x}_0)$ and not c_α. One might object to this argument by contrasting it to empirical example 3 where $\overline{x}_n=2.726$ gives rise to reject H_0. Looking at $\tau_n(\mathbf{x}_0)=\frac{\sqrt{100}(2.726-2)}{1.941}=3.740[.00009]$, however, reveals that the change is anything but marginal; see Spanos (2013a-b).

This issue relates to the widely accepted argument that in Fisher's significance testing the p-value is invariably ambiguous since it does not designate a direction of departure analogous to H_1 of N-P testing. This claim, however, ignores the fact that the p-value is a *post-data error probability*, and thus the sign and magnitude of $\tau(\mathbf{x}_0)$ provide additional information relating to the direction of departure which usually eliminates one of the two tails, rendering the post-data p-value invariably one-sided.

6 Summary and conclusions

The accept/reject H_0 results of N-P testing and the p-value are highly vulnerable to *the large n problem* since the optimality of N-P tests revolves around the inherent trade-off between the type I and II error probabilities. Hence, the detection of statistical significance based on conventional significance levels, $\alpha=$ .1, .05, .025, .01, is likely to be spurious when a large enough n, say $n>1000$,. Such spurious results arise since the observed test statistic $d_n(\mathbf{x}_0)$ (the p-value $p_n(\mathbf{x}_0)$) increases (decreases) monotonically with$\sqrt{n}$ as n increases. This reflects the fact that the power of a test increases monotonically with n, rendering the particular test more and more capable of detecting smaller and smaller discrepancies $\pm\gamma$ from the null value $\theta=\theta_0$.

The post-data severity evaluation can address the large n and related problems (1)-(7) by outputting the warranted discrepancy $\gamma^{\ddagger}$ warranted by data $\mathbf{x}_0$ and test T_α with high probability.

(a) The SEV transforms the accept/reject H_0 'results' into 'evidence' for particular inferential claims of the form $\theta \gtrless \theta_0+\gamma_1,\ \gamma_1\neq 0$.

(b) The key feature of SEV in addressing the large n problem is that the evaluation

of $\gamma^{\ddagger}$ is always relative to the same n as $\tau_n(\mathbf{x}_0)$ that gives rise to the accept/reject H_0 results. That is, in evaluating $\gamma^{\ddagger}$ the SEV counter-balances the effect of n on $\tau_n(\mathbf{x}_0)=\frac{\sqrt{n}(\overline{x}_n-\mu_0)}{s}$ by the non-centrality parameter $\delta_1=\frac{\sqrt{n}(\mu_1-\mu_0)}{s}$.

(c) SEV harnesses the additional precision of the estimates from a larger n by decreasing (increasing) the warranted discrepancy $\gamma^{\ddagger}$ depending on whether $\tau_n(\mathbf{x}_0)$ gave rise to a accept (reject) H_0.

(d) The SEV outputting the warranted discrepancy $\gamma^{\ddagger}$, can be used to address other foundational problems, including the rigging the significance level α, the framing of H_0 and H_1, distinguishing between statistical and substantive significance, as well as a testing-based effect size for the magnitude of the 'substantive' effect.

In conclusion, it is important to reiterate that securing genuine 'learning from data $\mathbf{x}_0$' about phenomena of interest begins with establishing the *statistical adequacy* of the invoked statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. The latter is the toll one is called to pay for the statistical reliability and the trustworthiness of the ensuing evidence. Without that the statistical analysis and the ensuing results degenerate into tinkering with meaningless numbers and making up stories without any real evidence stemming from one's data.

References

[1] Abelson, R.P. (1995) *Statistics as Principled Argument*, Lawrence Erlbaum, NJ.

[2] Abouk, R, S. Adams, B. Feng, J.C. Maclean, M. Pesko (2022) "The Effects of e-cigarette taxes on pre-pregnancy and prenatal smoking," NBER Working Paper 26126.

[3] Berger, J. (2019) "Four types of frequentism and their interplay with Bayesianism", Duke University.

[4] Berkson, J. (1938) "Some difficulties of interpretation encountered in the application of the chi-square test," *Journal of the American Statistical Association*, 33: 526-536.

[5] Cohen, J. (1969/1988) *Statistical power analysis for the behavioral sciences*, (2nd ed.), Lawrence Erlbaum, NJ.

[6] Ellis, P.D. (2010) *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*, Cambridge University Press, Cambridge.

[7] Fisher, R.A. (1922) "On the mathematical foundations of theoretical statistics", *Philosophical Transactions of the Royal Society A*, 222: 309-368.

[8] Fisher, R.A. (1925a) *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh.

[9] Fisher, R.A. (1925b) "Theory of statistical estimation", *Proceedings of the Cambridge Philosophical Society*, **22**: 700–725.

[10] Fisher, R.A. (1935) *The Design of Experiments*, Oliver and Boyd, Edinburgh.

[11] Gigerenzer, G. (1993) "The superego, the ego, and the id in statistical reasoning," *A handbook for data analysis in the behavioral sciences: Methodological issues*, pp. 311-339.

[12] Good, I.J. (1988), "The interface between statistics and philosophy of science," *Statistical Science*, **3**: 386-397.

[13] Lehmann, E.L. (1993) "The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two?" *Journal of the American statistical Association*, 88(424): 1242-1249.

[14] Lehmann, E.L. and J.P. Romano (2005) *Testing Statistical Hypotheses*, Springer, NY.

[15] Mayo, D.G. (1996) *Error and the Growth of Experimental Knowledge*, The University of Chicago Press, Chicago.

[16] Mayo, D.G. and A. Spanos. (2006) "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction", *The British Journal for the Philosophy of Science,* 57: 323-357.

[17] Mayo, D.G. and A. Spanos (2011), "Error Statistics", pp. 151-196 in the Handbook of Philosophy of Science, vol. 7: Philosophy of Statistics, D. Gabbay, P. Thagard, and J. Woods (editors), Elsevier.

[18] McCullagh, P. (2002) "What is a statistical model?" *Annals of Statistics*, **30**: 1225-1267.

[19] Neyman, J. (1937) "Outline of a Theory of Statistical Estimation based on the Classical Theory of Probability", *Philosophical Transactions of the Royal Statistical Society of London*, A, 236: 333–380.

[20] Neyman, J. and E.S. Pearson (1933) "On the problem of the most efficient tests of statistical hypotheses", *Philosophical Transactions of the Royal Society, A,* 231, 289-337.

[21] Rao, C.R. (2004), "Statistics: Reflections on the Past and Visions for the Future," *Amstat News*, 327: 2-3.

[22] Spanos, A. (1986) *Statistical Foundations of Econometric Modelling*, Cambridge University Press, Cambridge.

[23] Spanos, A. (1989) "Early Empirical Findings on the Consumption Function, Stylized Facts or Fiction: a Retrospective View," *Oxford Economic Papers*, 41: 150-169.

[24] Spanos, A. (1995), "On theory testing in Econometrics: modeling with nonexperimental data", *Journal of Econometrics,* **67**: 189-226.

[25] Spanos, A. (2002) "Parametric versus Non-parametric Inference: Statistical Models and Simplicity," pp. 181-206 in *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple*, edited by A. Zellner, H. A. Keuzenkamp and M. McAleer, Cambridge University Press, Cambridge.

[26] Spanos, A. (2006a) "Econometrics in retrospect and prospect," In *New Palgrave Handbook of Econometrics*, vol. 1, ed. T. C. Mills and K. Patterson, Macmillan, London.

[27] Spanos, A. (2006b) "Where Do Statistical Models Come From? Revisiting the Problem of Specification", pp. 98-119 in *Optimality: The Second Erich L. Lehmann Symposium*, edited by J. Rojo, Lecture Notes-Monograph Series, vol. 49, Institute of Mathematical Statistics.

[28] Spanos, A. (2006c) "Revisiting the Omitted Variables Argument: substantive vs. statistical adequacy", *Journal of Economic Methodology*, **13**: 179–218.

[29] Spanos, A. (2009) "Statistical Misspecification and the Reliability of Inference: the simple t-test in the presence of Markov dependence", *The Korean Economic Review*, 25: 165-213.

[30] Spanos, A. (2010a) "Statistical Adequacy and the Trustworthiness of Empirical Evidence: Statistical vs. Substantive Information", *Economic Modelling*, 27: 1436–1452.

[31] Spanos, A. (2010b) "Akaike-type Criteria and the Reliability of Inference: Model Selection vs. Statistical Model Specification", *Journal of Econometrics,* **158**: 204-220.

[32] Spanos, A. (2013a) "Revisiting the Likelihoodist Evidential Account", *Journal of Statistical Theory and Practice*, **7**: 187-195.

[33] Spanos, A. (2013b) "Who Should Be Afraid of the Jeffreys-Lindley Paradox?", *Philosophy of Science*, 80: 73-93.

[34] Spanos, A. (2014) "Recurring Controversies about P values and Confidence Intervals Revisited," *Ecology*, 95(3): 645—651.

[35] Spanos, A. (2018) "Mis-Specification Testing in Retrospect", *Journal of Economic Surveys*, Vol. 32, No. 2: 541–577.

[36] Spanos, A. (2019) *Introduction to Probability Theory and Statistical Inference: Empirical Modeling with Observational Data*, 2nd edition, Cambridge University Press, Cambridge.

[37] Spanos, A. (2021a) "Revisiting noncentrality-based confidence intervals, error probabilities and estimation-based effect sizes", *Journal of Mathematical Psychology*, 104, p. 102580.

[38] Spanos, A. (2021b) "Bernoulli's golden theorem in retrospect: error probabilities and trustworthy evidence", *Synthese*, 199: 13949-13976..

[39] Spanos, A. (2022a) "Severity and Trustworthy Evidence: Foundational Problems versus Misuses of Frequentist Testing", *Philosophy of Science*, 89(2): 378-397.

[40] Spanos, A. (2022b) "Frequentist Model-based Statistical Induction and the Replication crisis", *Journal of Quantitative Economics*, volume in honor of C.R. Rao; 10.1007/s40953-022-00312-z.

[41] Spanos, A. and A. McGuirk (2001) "The Model Specification Problem from a Probabilistic Reduction Perspective," *Journal of the American Agricultural Association*, 83: 1168-1176.