The Replication Crises and the Trustworthiness of Empirical Evidence[*]

Aris Spanos
Department of Economics, Virginia Tech, USA

November 2018

Abstract

The primary aim of this paper is to call into question the presumption by the literature on the replication crises that replicability ensures the trustworthiness of empirical evidence. It is argued that the current focus of this literature on the abuse of significance testing as the main culprit is rather misplaced. This abuse is only a part of a much broader problem relating to the uninformed application of statistical methods without real understanding of their assumptions, limitations, proper implementation and interpretation of their inferential results. The abuse and misinterpretation of the p-value is just a symptom of the same uninformed implementation that can also render untrustworthy evidence replicable, when the same recipe-like implementation of statistical methods is followed by similarly uninformed practitioners. A case is made in this paper that the three most important sources of untrustworthy evidence in empirical modeling are: (i) statistical misspecification: invalid probabilistic assumptions imposed on the particular data, (ii) poor implementation of inferential procedures, and (iii) unwarranted evidential interpretations of their inferential results.

1 Introduction

It has been well-documented in several disciplines (Camerer et al., 2016, Hoffler, 2017, Johnson et al. 2016, Nosek and Lakens, 2013, National Academy of Science, 2016) that the majority of published empirical results are not replicable. That has been broadly interpreted as providing clear evidence that "most published research findings are false"; see Ioannidis (2005). In light of that, several contributors to this debate, as well as a few journal editors, pointed the finger at significance testing as a leading contributor to untrustworthy evidence and non-replicability problems. To understand the veracity of this charge leveled against significance testing, as well as the ramifications of different proposals to replace it, including Confidence Intervals (CIs), estimation-based effect sizes and redefining statistical significance (Mayo, 2018), there is an pressing need for a better understanding of the main sources of the

[*] An earlier version of this paper was presented as an invited keynote presentation at the 2018 Royal Statistical Society meeting in Cardiff, UK.

untrustworthy evidence problem, together with a balanced appraisal of the proposed methods to replace significance tests to meliorate the replicability problem.

The primary aim is to revisit the main argument of the replication crises literature (Begley and Ioannidis, 2015) with a view to make a case that replicability is neither necessary nor sufficient for the trustworthiness of empirical evidence. Indeed, the current focus on the abuse of significance testing as the main culprit (Ioannidis at al., 2017) is rather misplaced. The real problem is the untrustworthiness of evidence, and the leading cause of that is the recipe-like application of statistical methods without real understanding of their assumptions, limitations, proper implementation and warranted interpretations of their results; see also Stark and Salteli (2018). That is, the widespread abuse and misinterpretation of the p-value (p-hacking, multiple testing, cherry-picking) is only a symptom of the same uninformed implementation that contributes in many different ways to the problem of untrustworthy evidence. This calls into question the merit of securing replicability as a way to address the untrustworthy evidence problem, because the same mechanical implementation often ensures that untrustworthy evidence can be routinely replicated. This can easily happen when uninformed practitioners follow the same recipe-like implementation of statistical methods.

This perspective suggests a refocusing of the proposed strategies for securing the trustworthiness of published empirical evidence by appraising whether a particular study has circumvented or dealt with the potential errors and omissions that could have undermined the reliability of the particular inferences drawn. A case is made in this paper that the three most important sources of untrustworthy evidence in empirical modeling are: (i) statistical misspecification: invalid probabilistic assumptions imposed on the particular data, (ii) poor implementation of inferential procedures, and (iii) unwarranted evidential interpretations of their inferential results.

Ioannidis (2005) made his case by focusing exclusively on certain aspects of (ii), such as ad hoc rejection thresholds ($<.05$), cherry picking and p-hacking, to propose a Bayesian measure for an overall reliability of discipline-wide testing results, the Positive Predictive Value (PPV). Unfortunately, the PPV is riddled with egregious misinterpretations of frequentist error probabilities (section 3.3), rendering it completely inappropriate for the intended task.

The paper articulates a unifying framework for frequentist inference with a view to ensure *informed implementation* of statistical methods that is grounded on in-depth understanding of their assumptions, their limitations, their proper implementation and the warranted interpretation of their inferential results. For that, the following distinctions play important roles: (a) testing within vs. testing outside the prespecified statistical model, (b) pre-data vs. post-data error probabilities, (c) the modeling vs. the inference facet of statistical analysis, and (d) statistical vs. substantive information/model. Viewed in the context of this unifying framework, the different proposals to replace significance testing, such as observed CIs and estimation-based effects sizes, are equally susceptible to the above sources (i)-(iii) of untrustworthy evidence.

2 Revisiting the replicability problem

Reproducibility, viewed as the potential to reproduce particular inference results using the same data and methods, and *replicability*, viewed as the potential of such results to be independently confirmed by other researchers studying the same phenomenon of interest, are considered to be two crucial features of successful scientific research. The metaphor that motivates replicability is that of experimentation in a scientific lab where an experiment is repeated many times under controlled conditions and the empirical findings are replicated. Any systematic discrepancies indicate a problem that needs to be investigated; see Spanos (2010a). In that spirit, the founder of modern statistics declared: "In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to gives us a statistically significant result." (Fisher, 1935, p. 14)

The question that naturally arises is whether this *replication metaphor* is appropriate for appraising the trustworthiness of empirical evidence based on data generated outside the controlled conditions. The answer is not obvious when one refers to empirical modeling with *observational data* in social sciences.

First, empirical findings in hard sciences (physics, chemistry, biology) pertain to (a) laws of nature that are usually invariant with respect to the time and location of the investigation. Their experimental investigation is: (b) guided by reliable substantive knowledge pertaining to the phenomenon of interest, (c) framed in terms of tried and trusted procedural protocols, and (d) empirical knowledge has a high degree of cumulativeness. In contrast, empirical modeling in social sciences pertains to (e) fickle human behavior that is not invariant to time or location. The empirical modeling is (f) guided by tentative conjectures that are often treated as established knowledge, (h) by foisting a substantive model on the data and viewing empirical modeling as curve-fitting guided by goodness-of-fit. The end result is invariably (h) empirical models that are statistically and substantively misspecified.

Second, under the lab experimental controls it is relatively easy to generated data that can be viewed as realizations of an Independent and Identically Distributed (IID) process. In contrast, observable economic phenomena of interest give rise to data that often exhibit complicated forms of dependence and heterogeneity. These chance regularities need to be fully accounted for by the estimated statistical model in order to secure the reliability of any inference based on them. This renders modeling with observational data a lot more vulnerable to statistical misspecification.

Third, experimental investigation in the hard sciences is poking into nature itself using substantive models that provide accurate enough approximations of the reality they are probing. The substantive models in the social sciences are not accurate enough approximations of the phenomenon of interest for several reasons, including a huge gap between the theoretical concepts in terms of which a substantive (structural) model is framed and the available real-world data. For instance, there is a substantial difference between "demand" and "supply" and what the available data measure, usually "quantities transacted" and "the corresponding prices"; the former refer to intentions at a specific point in time and the latter to realizations over time. Hence, the need to distinguish between a statistical and a substantive model.

In light of the above, mimicking the empirical modeling practices of physicists and chemists, including replication and procedural protocols, might not be the best strategy for social scientists, when their primary aim is to secure the trustworthiness of empirical evidence.

2.1 Untrustworthy but replicable empirical evidence

What is insufficiently appreciated by the current literature on replicability is that 'the mechanical application of statistical methods' ensures that the reproducibility/replicability of inference results, by itself, will *not* address the untrustworthy evidence problem. For instance, dozens of MBA students continue to confirm a theory known as the *Efficient Market (EM) hypothesis*, by replicating and reproducing the original empirical results on a daily basis. To be more specific, the (weak) EM hypothesis asserts that 'changes in speculative prices are, in principle, *unpredictable* from their own past. Let P_t denote the speculative price of an asset (stock, bond, exchange rate, etc.) and define the log-returns associated with this asset to be: $y_t{=}\ln P_t - \ln P_{t-1}$. Fama (1970) framed this assertion in the form of a Normal, *Martingale Difference* model:

$$y_t{=}\varepsilon_t,\ (\varepsilon_t|\sigma(\mathbf{y}^0_{t-1})){\sim}\mathsf{NMD}(0,\sigma^2),\ t{\in}\mathbb{N}, \tag{1}$$

where $\sigma(\mathbf{y}^0_{t-1})$ denotes the σ-field generated by $\mathbf{y}^0_{t-1}{:=}(y_{t-1}, y_{t-2}, ..., y_1)$. The validity of the EM hypothesis is usually appraised by embedding (1) into an encompassing Autoregressive [AR(p)] model (table 1) and testing:

$$H_0\text{: } \alpha_i{=}0,\ H_1\text{: } \alpha_i{\neq}0,\ i{=}0, 1, ..., p. \tag{2}$$

Table 1: **Autoregressive [AR(**p**)] model**

$$y_t = \alpha_0 + \textstyle\sum_{i=1}^{p} \alpha_i y_{t-i} + u_t,\ t{\in}\mathbb{N},$$

{1}	**Normality:**	$\left(u_t|\sigma(\mathbf{y}^0_{t-1})\right) \backsim \mathsf{N}(.,.)$,
{2}	**Zero mean:**	$E\left(u_t|\sigma(\mathbf{y}^0_{t-1})\right){=}0$,
{3}	**Homoskedasticity:**	$E(u_t^2|\sigma(\mathbf{y}^0_{t-1})){=}\sigma_0^2$,
{4}	**Non-correlation:**	$E(u_t u_s|\sigma(\mathbf{y}^0_{t-1})){=}0,\ t{>}s,\ t, s{\in}\mathbb{N}$.

The main reason such untrustworthy results can be replicated is primarily the mechanical application of the same questionable methods to analyze them. One would be hard pressed to find a single published paper on the EM hypothesis in which the invoked probabilistic assumptions {1}-{4} have been validated.

Example 1. **Data**: weekly observations on the *US/Canadian* dollar exchange rate for the period July 1973 to December 1991, where y_t denotes log-returns. Estimation of an AR(2) model yielded:

$$y_t = \underset{(.018)}{.011} + \underset{(.035)}{.067}y_{t-1} + \underset{(.032)}{.002}y_{t-2} + \widehat{u}_t,\ s^2{=}.549,\ n{=}951. \tag{3}$$

The t-ratios $\tau(\alpha_0){=}\frac{.011}{.018}{=}.61[.282]$, $\tau(\alpha_1){=}1.914[.056]$, $\tau(\alpha_2){=}.063{=}[.95]$ confirm the EM hypothesis! Unfortunately, the invoked assumptions {1}, {3} and {4} are invalid, rendering such an inference unreliable. When a statistically adequate model that accounts for these departures is specified, in the form of a Student's t AR(2) model (table 2):

$$\widehat{y}_t = \underset{(.012)}{.000} + \underset{(.031)}{.104}y_{t-1} + \underset{(.035)}{.093}y_{t-2} + \widehat{v}_t,\ \widehat{\sigma}_0^2{=}.375,\ n{=}951,$$

$$\widehat{\omega}_t^2{=}\widehat{\sigma}_0^2(1{+}\textstyle\sum_{i=1}^2\{\underset{(.011)}{.255}\widetilde{y}_{t-1}{-}\underset{(.007)}{.021}[\widetilde{y}_{t-i}\widetilde{y}_{t-i-1}{+}\widetilde{y}_{t-i}\widetilde{y}_{t-i+1}]\}),$$

where $\widetilde{y}_{t-i}{:=}(\widetilde{y}_{t-i}{-}\overline{y})$, provides evidence *against* the EM hypothesis. Any attempt to account for the departures from {1}, {3} and {4} using an ARCH(p) or a GARCH(p,q) formulation, yields a misspecified model; Spanos (1995a).

Table 2- Student's t, AR (2) model

Statistical GM: $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + v_t,\ \ t{\in}\mathbb{N},$

for $\mathbf{y}^0_{t-1}{:=}(y_{t-1}, y_{t-2}, ..., y_1)$

[1] Student's t:	$D(y_t \vert\ \mathbf{y}^0_{t-1};\theta)$, is Student's t, $\nu{+}2$ d.f.
[2] Linearity:	$E(y_t\vert\sigma(\mathbf{y}^0_{t-1})){=}\beta_0 + \sum_{i=1}^{\ell}\beta_i y_{t-i},$
[3] Heteroskedasticity:	$Var(y_t\vert\sigma(\mathbf{y}^0_{t-1})){=}\omega^2_{\ell}(\mathbf{y}^0_{t-1}),$
[4] Markov:	$\{y_t,\ t{\in}\mathbb{N}\}$ is a Markov(2) process,
[5] t-invariance:	$\boldsymbol{\theta}{:=}(\beta_0, \beta_1, \beta_2, \sigma_0^2, \delta_1, \delta_2, \mu).$

Hence, the evidence for the EM hypothesis is a case of untrustworthy evidence that is easily reproducible and replicable due to *systematic errors* and *omissions* that would undermine the claim that replicability implies trustworthy empirical evidence.

The untrustworthiness of published empirical evidence is symptomatic of the uninformed implementation of statistical methods by practitioners, combined with the equally uninformed refereeing process that pays little attention to the three main sources of untrustworthiness: (i) statistical misspecification, (ii) poor implementation of inferential procedures, and (iii) unwarranted evidential interpretations of their inferential results. The current literature is focusing on some aspects of (ii) and (iii), but they totally ignore (i).

3 Frequentist testing: a coherent framework

It is no coincidence that the single most widely misused and abused procedure is that of significance testing, since, in addition to its neglected elements, the method can and is routinely applied in many different contexts (Cox and Hinkley, 1986, chs 3-4), but the limitations and interpretation of the results are different in each case.

3.1 Testing within vs. testing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$

Statistical model. The single most important concept in statistical modeling and inference is that of a (parametric) statistical model introduced by Fisher (1922). A statistical model defines the inductive premises of statistical inference and comprises the totality of probabilistic assumptions imposed (directly or indirectly) on the observed data. It is generically defined by:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}){=}\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}{\in}\Theta{\subset}\mathbb{R}^m\},\ \mathbf{x}{\in}\mathbb{R}^n_X,\ m{<}n, \tag{4}$$

where $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}{\in}\mathbb{R}^n_X$, denotes the *distribution of the sample* that encapsulates the assumed probabilistic structure. In what follows a statistical model is viewed as defining a stochastic Generating Mechanism (GM) assumed that could have generated the particular data $\mathbf{x}_0$; see Spanos (2006).

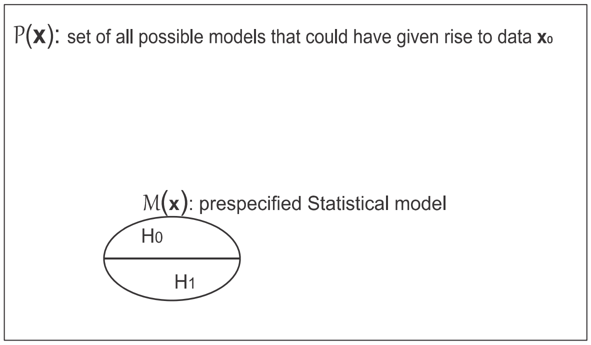

P(x): set of all possible models that could have given rise to data x₀
M(x): prespecified Statistical model
H0
H1

Fig. 3: Testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: N-P

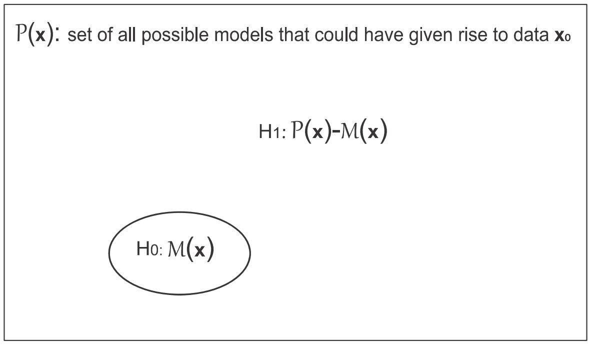

P(x): set of all possible models that could have given rise to data x₀
H1: P(x)-M(x)
H0: M(x)

Fig. 4: Testing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: M-S testing

Example 2. Consider the simple (IID) Normal model:

$$X_t \backsim \mathsf{NIID}(\mu,\sigma^2),\ \boldsymbol{\theta}{:=}(\mu,\sigma^2){\in}\mathbb{R}\times\mathbb{R}_+,\ t{=}1,2,...,n,..., \tag{5}$$

The specification (initial selection) of $\mathcal{M}_\theta(\mathbf{x})$ is based on imposing probabilistic assumptions on the stochastic process $\{X_t,\ t{\in}\mathbb{N}\}$ underlying the particular data that would render $\mathbf{x}_0$ a 'typical realization' thereof. This can be viewed as narrowing down the set $\mathcal{P}(\mathbf{x})$ of all possible statistical models that could have given rise to data $\mathbf{x}_0$ to a small subset $\mathcal{M}_\theta(\mathbf{x})$ (fig. 3).

Neyman-Pearson (N-P) testing is always within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ (fig. 3), and Mis-Specification (M-S) testing is always testing outside since it probes $[\mathcal{P}(\mathbf{x}){-}\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ (fig. 4) with a view to test the validity of the assumptions of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ vis-a-vis data $\mathbf{x}_0$.

3.2 Significance testing: probing within a statistical model

When frequentist testing is viewed as testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, Fisher's significance testing becomes a special case of Neyman-Pearson (N-P) testing, where H_0: $\theta{=}\theta_0$ is a point hypothesis. Contrary to Gigerenzer (1993), there is no intrinsic inconsistency between significance and N-P testing. This is because both methods, when viewed as testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, share the same framework.

(i) The *inductive premises* of inference are the same for both methods, i.e. $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

(ii) Their primary aim is to 'learn from data' about the 'true' $\boldsymbol{\theta}$ in Θ, denoted by $\boldsymbol{\theta}^*$, which is shorthand for saying that 'data $\mathbf{x}_0$ constitute a *typical realization* of the sample $\mathbf{X}$ from $\mathcal{M}^*(\mathbf{x}){=}\{f(\mathbf{x};\boldsymbol{\theta}^*)\}$, $\mathbf{x}{\in}\mathbb{R}^n_X$.

(iii) In light of (ii) the *whole* parameter space Θ is relevant for statistical purposes, irrespective of whether only a small subset is of interest from a substantive perspective; in principle, any single value in Θ could be $\boldsymbol{\theta}^*$. Their null and (implicit) alternative hypotheses are specified in terms of the parameters $\boldsymbol{\theta}\in\Theta\subset\mathbb{R}^m$ of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, and framed as a *partition* of the parameter space Θ:

$$H_0\text{: } \boldsymbol{\theta}\in\Theta_0 \text{ vs. } H_1\text{: } \boldsymbol{\theta}\in\Theta_1, \tag{6}$$

that corresponds to the partition of the sample space ($\mathbb{R}^n_X$) into an acceptance (C_0) and a rejection (C_1) region:

$$\mathbb{R}^n_X=\left\{\begin{array}{ccc}\boxed{C_0} & \leftrightarrow & \boxed{\Theta_0}\\ \boxed{C_1} & \leftrightarrow & \boxed{\Theta_1}\end{array}\right\}=\Theta$$

(iv) The underlying reasoning for both methods is *hypothetical,* 'what if' H_0 is true or false, which are not probabilistic events that one can condition upon. They represent *hypothetical scenarios* under which the sampling distribution of the test statistic $d(\mathbf{X})$ is evaluated.

(v) In light of (iii)-(iv), both methods rely on error probabilities (type I, II and the p-value) that: (a) are derived under these hypothetical scenarios, (b) are firmly attached to the testing procedure, and (c) relate to inferential claims about $\boldsymbol{\theta}$. Error probabilities are never assigned to $\boldsymbol{\theta}$ and they are *not* conditional probabilities.

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

$$H_0\text{: } \mu=\mu_0 \text{ vs. } H_1\text{: } \mu\neq\mu_0, \tag{7}$$

gives rise to the well-known *Student's t test* $T_\alpha:=\{\tau(\mathbf{X}), C_1(\alpha)\}$ defined by:

$$\tau(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s},\ C_1(\alpha)=\{\mathbf{x}\text{: } |\tau(\mathbf{x})|>c_\alpha\},\ s^2=\tfrac{1}{n-1}\textstyle\sum_{k=1}^n(X_k-\overline{X}_n)^2, \tag{8}$$

where c_α can be evaluated using the Student's t distribution. NOTE that $\overline{X}_n$ is the best estimator of θ^*. To evaluate the two types of error probabilities the distribution of $\tau(\mathbf{X})$ under both the null and alternatives are needed:

$$\begin{aligned}&[a]\ \tau(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s}\overset{\mu_0}{\backsim}\mathsf{St}(n-1),\\ &[b]\ \tau(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s}\overset{\mu_1}{\backsim}\mathsf{St}(\delta_1;n-1),\ \text{ for all } |\mu_1|>\mu_0,\end{aligned} \tag{9}$$

where $\delta_1=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$ is the *non-centrality parameter*. The sampling distributions in (iii) are used to specify the type I error probability and the power of the test T_α:

$$\begin{aligned}&\alpha=\mathbb{P}(|\tau(\mathbf{X})|>c_\alpha;\mu=\mu_0),\\ &p(\mu_1)=\mathbb{P}(|\tau(\mathbf{X})|>c_\alpha;\mu=\mu_1),\ \text{ for all } |\mu_1|>\mu_0.\end{aligned} \tag{10}$$

It is important to emphasize that the reasoning underlying frequentist testing is hypothetical: under $H_0(H_1)$ is shorthand for the scenario that the designated value $\mu_0(\mu_1)$ coincides with the 'true' value μ^*. (9)[a] should contrasted with the analogous pivotal result:

$$\tau(\mathbf{X};\mu)=\tfrac{\sqrt{n}(\overline{X}_n-\mu)}{s}\overset{\mu=\mu^*}{\backsim}\mathsf{St}(n-1), \tag{11}$$

that provides the basis for a $(1-\alpha)$ Confidence Intervals (CI) for μ:

$$\mathbb{P}(\overline{X}_n-\tfrac{s}{\sqrt{n}}c_{\frac{\alpha}{2}})\leq\mu<\overline{X}_n-\tfrac{s}{\sqrt{n}}c_{\frac{\alpha}{2}};\ \mu=\mu^*)=(1-\alpha). \tag{12}$$

The key difference between (9)[a] and (11) is that the latter is evaluated using *factual reasoning*; under μ^*, the true μ, whatever that value happens to be.

3.2.1 Pre-data vs. post-data error probabilities

The crucial difference between Fisher's significance testing and N-P testing stems from the fact that the former employs the p-value, a *post-data* ($d(\mathbf{x}_0)$ is known) error probability, and the latter uses *pre-data* error probabilities (type I and II) to define a rejection region, giving rise to the accept/reject H_0 rules. Both methods rely on thresholds for their inferences and the power of the test is relevant for significance testing, despite Fisher's (1955) claims to the contrary.

The real difference between pre-data and post data error probabilities is that the latter use *additional information* in the form of $d(\mathbf{x}_0)$. In particular, the sign of $d(\mathbf{x}_0)$ contains additional information pertaining to the direction of departure from H_0 indicated by data $\mathbf{x}_0$. This information renders one of the two tails in (9) irrelevant, and calls into question the concept of a *two-sided* p-value $p(\mathbf{x}_0)$.

Example 2 (continued). For (7) and test T_α:=$\{\tau(\mathbf{X}), C_1(\alpha)\}$, the p-value is:

$$p(\mathbf{x}_0)=\begin{cases} \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu{=}\mu_0), & \text{if } \tau(\mathbf{x}_0){>}0, \\ \mathbb{P}(\tau(\mathbf{X}) < \tau(\mathbf{x}_0); \mu{=}\mu_0), & \text{if } \tau(\mathbf{x}_0){<}0. \end{cases}.$$

3.3 Revisiting the Positive Predictive Value (PPV)

The claim by Ioannidis (2005) that 'most published research findings are false' has been particularly influential in blaming the use of significance testing for the such a state of affairs. His argument revolves around a posterior probability measure, the Positive Predictive Value (PPV), that has been accepted uncritically by both statisticians and other applied practitioners. Let us take a closer look at the PPV, which is defined in terms of 'events': F=H_0 is false, R=test rejects H_0, and takes the Bayesian probability formulation:

$$\text{PPV}{=}\Pr(F|R){=}\tfrac{\Pr(R|F)\Pr(F)}{\Pr(R|F)P(F)+\Pr(R|\overline{F})P(\overline{F})},$$

where $\Pr(R|F)$, $\Pr(F)$, $\Pr(\overline{R}|\overline{F})$ are referred to as 'sensitivity', 'prevalence' and 'specificity', respectively. Sensitivity aims to measure the proportion of actual rejections of H_0 that are correctly identified as such, specificity aims to measure the proportion of actual acceptances of H_0 that are correctly identified as such. Both depend crucially on prevalence $\Pr(F)$ that aims to measure the proportion of false H_0 in a certain population of nulls; whatever that could mean. PPV is a measure adapted from medical screening that aims to evaluate the reliability of medical diagnostic tests detecting an ailment in patients that revolves around the notions of 'false positive' and 'false negative'; see Fletcher and Fletcher (2005). To make sense of the PPV one needs to imagine testing of null hypotheses as a discipline wide activity undertaken by hundreds of practitioners contemplating thousands of null hypotheses, a proportion of which are false, say 20%! Unfortunately, this analogical reasoning behind the adaptation from medical screening gives the impression that $\Pr(R|F)$ and $\Pr(R|\overline{F})$ relate directly to the frequentist concepts of the 'power' and the 'significance level' of a test, respectively. This semblance, however, is utterly false.

First, frequentist error probabilities are never defined as *conditional* on the 'H_0 being true or false', because the latter do not constitute legitimate events in the context of a prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, upon which one can condition. Indeed, the notion of assigning probabilities to 'H_0 is true or false' is meaningless in the context of frequentist testing because they represent *hypothetical scenarios* under which the sampling distribution of the test statistic $(d(\mathbf{X}))$ is evaluated. Hence, the argument: "Whether the p value is stated as frequentists prefer $P(d(\mathbf{X}) > d(\mathbf{x}_0); H_0)$, or with Bayesian notation $P(d(\mathbf{X}) > d(\mathbf{x}_0)|H_0)$, for all practical purposes in my view, the p value, is indeed a probability conditional or conditioned on an assumption, the null hypothesis." (Schneider, 2018) bespeaks ignorance of basic probability theory; one can condition only on events and random variables, not assumptions in general. Moreover, pretending that it is a matter of notational choice shows further ignorance of the basic elements of frequentist testing. Just because the probabilities $\Pr(R|F)$, $\Pr(F)$, $\Pr(\overline{R}|\overline{F})$ are meaningful in the context of Bayesian inference, does not render them relevant for evaluating the reliability of frequentist testing results.

Second, in frequentist testing there is no such thing as discipline wide false positive/negative proportions that revolve around generic tests and generic null hypotheses analogous to medical screening devices. One can assert that the false positive of this screening device is 5%, but the analogical reasoning used to transfer such a notion to frequentist error probabilities is misplaced. Frequentist testing is local in the sense that it depend crucially on the particular the particular$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, the relevant test $(d(\mathbf{X}),\ C_1)$ and the particular data $(\mathbf{x}_0)$, including n; see Spanos (2013). Assuming that a certain proportion of the 'effects' tested in a particular field, say $\Pr(F)$=.2, are expected to be 'truly' non-null (or unreal relationships) relates to what Bayesians call the 'base rate', which is meaningless in the context of frequentist testing; see Spanos (2010b). Similarly, the power of a test is never a *point probability* chosen by cherry-picking the scenario 'H_0 is false', since it is defined for all $\theta_1 \in \Theta_1 = \Theta - \Theta_0$, see (10). Worse, using prior probabilities assigned to null hypotheses being true or false takes one down a Bayesian rabbit hole with numerous confusions awaiting the uninformed frequentist. For instance, engaging oneself in debates about the appropriateness of particular values assigned by Ioannidis (2005) to 'pretend' conditional probabilities $\Pr(R|F)$, $\Pr(\overline{R}|\overline{F})$, that would imply unreliable (low PPV) frequentist inference results, or protesting that $\Pr(F) < .5$ gives rise to misleading appraisals, amount to mindless war exercises on an imaginary map.

In summary, the PPV is nothing but a generic posterior measure of *untrustworthiness by association* based on a Bayesian meta-model of field-wide inferences. At best the PPV highlights certain well-known abuses of significance testing; see Mayo (2018). However, the price for that is that it reflects attention away from certain crucial sources of untrustworthy evidence, including statistical misspecification. It's pointless to argue about p-hacking and cherry-picking when the evaluated p-values are erroneous, in the first place.

3.4 Common confusions in frequentist testing

3.4.1 Significance testing and the role of power

A strong case can be made that the p-value and the accept/reject rules have no information relating to evidence for or against particular alternative hypotheses, since their evaluation is under H_0. That information calls for evaluating $\tau(\mathbf{X})$ under H_1.

The large n problem. This issue was initially raised by Berkson (1938) by arguing that Pearson's chi-square test statistic usually increases with n, and thus the p-value decreases as n increases, deducing that, there is always a large enough n to reject any null hypothesis however small the adopted threshold. Correcting Berkson's argument, the p-value decreases as n increases when $(\theta^{*}-\theta_0)\neq 0$ irrespective of the magnitude of the discrepancy. Note that in practice $(\theta^{*}-\theta_0)$ is estimated using $(\widehat{\theta}-\theta_0)$ where $\widehat{\theta}$ is an optimal estimator of θ^{*}. Hence, when $(\theta^{*}-\theta_0)\neq 0$ there is always a large enough n to reject any null hypothesis for any rejection threshold $c>0$. That means that a rejection of H_0 with $p(\mathbf{x}_0)=.03$ and $n=50$, does not have the same evidential weight for the falsity of H_0 as a rejection with $p(\mathbf{x}_0)=.03$ and $n=20000$. This questions the strategy of evaluating 'significance' using $p(\mathbf{x}_0)<.05$ and ignoring n.

This is known as the *fallacy of rejection*: (mis)interpreting reject H_0 [evidence against H_0] as evidence *for* a particular H_1; this can easily arise when a test has high enough power. An analogous fallacy can arise when there is not enough power, the *fallacy of acceptance*: (mis)interpreting a large p-value or accept H_0 [no evidence against H_0] as evidence for H_0; this can easily arise when a test has very low power.

Therefore, the problem arises when the p-value is detached from the particular test T_α and data $\mathbf{x}_0$, and is treated as providing the same evidence for a particular alternative H_1, regardless of the power of the test in question. For instance, in (9) the power increases with $\sqrt{n}$ since $\delta_1=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$, for all $\mu_1\in\Theta_1$, rendering the test more and more capable of detecting smaller and smaller discrepancies from H_0 with the same probability; see fig. 5a-b.

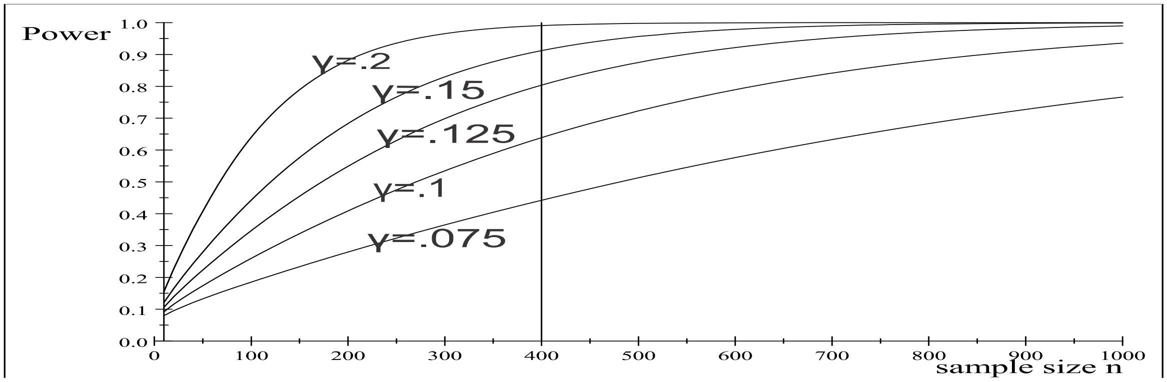


Fig. 5a: Power for different n and discrepancies $\gamma\sigma$ from the null

When viewed as *testing within* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, a significance test has a well-defined power function that becomes relevant when the rejection (acceptance) of H_0 with a small (large) enough p-value.

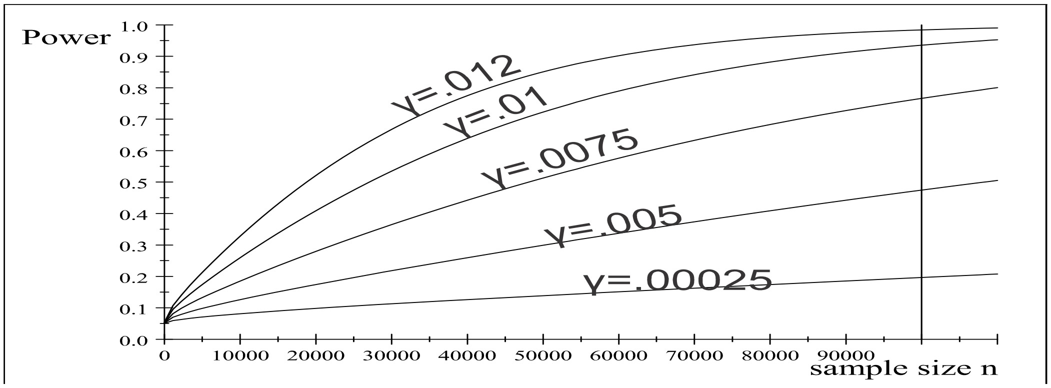


Fig. 5b: Power for different n and discrepancies $\gamma\sigma$ from the null

As Fisher (1935) argued: "By increasing the size of the experiment, we can render it more sensitive, meaning by this that it will allow of the detection of a lower degree of sensory discrimination, or ... quantitatively smaller departures from the null hypothesis." (pp. 21-22). This 'sensitivity' renders a rejection of H_0 with a large n (high power) very different in *evidential terms* from a rejection of H_0 with a small n (low power). That is, the p-value and the accept/reject rules were never meant to provide evidence for or against particular hypotheses beyond the coarse accept/reject H_0.

To counter the decrease in the p-value as n increases, some textbooks advise practitioners to use rules of thumb based on decreasing α for larger and larger sample sizes; see Lehmann and Romano (2005). Good (1988) proposes to standardize the p-value $p(\mathbf{x}_0)$ to the fixed sample size n=100 using the rule of thumb:

$$p_{100}(\mathbf{x}_0)=\min\left(.5,\ \left[p(\mathbf{x}_0)\cdot\sqrt{n/100}\right]\right),\ \ n>40.$$

Table 3: Actual $[p_n(\mathbf{x}_0)$=.01] vs. standardized p-value

n	50	100	500	1000	10000	100000	1000000
$p_{100}(\mathbf{x}_0)$	.007	.010	.022	.032	.100	.300	.5

3.5 Post-data Severity evaluation

The severity evaluation provides a more formal way to account for the sample size. Transforming the coarseness of the accept/reject testing results into *evidence for* or *against* a hypothesis or an inferential claim relating to H_0 or H_1 requires information about the generic capacity of the particular test in detecting discrepancies from H_0.This stems from the intuition that a small p-value or a rejection of H_0 based on a test with low power (e.g. a small n) for detecting a particular discrepancy γ provides *stronger* evidence for the presence of γ than using a test with much higher power (e.g. a large n).

Mayo and Spanos (2006) proposed a frequentist evidential account based on harnessing this intuition in the form of a post-data severity evaluation of the accept/reject results. This is based on using the generic capacity of the test to custom-tailor the discrepancy γ warranted by data $\mathbf{x}_0$. This evidential account can be used to circumvent the fallacies of rejection/acceptance.

This vulnerability of the p-value can be adequately addressed using the post-data severity assessment to provide an evidential account for frequentist inference; see Mayo (2018). The severity evaluation of the p-value $p(\mathbf{x}_0)$ and the accept/reject H_0 results is a *post-data* error probability that outputs an evidential interpretation based on inferential claims of the form $\theta \lesseqgtr \theta_0 + \gamma$, revolving around the warranted discrepancy γ^* from H_0 stemming from data $\mathbf{x}_0$ and test T_α. When the p-value is misinterpreted as providing evidence against H_0, its weakness is that a small $p(\mathbf{x}_0)$ indicates 'some' discrepancy from H_0, but it says nothing about its magnitude; the severity evaluation addresses this weakness; see Mayo and Spanos (2011).

Example. For the t-test when the null H_0: $\mu{=}\mu_0$ is rejected with $n{=}100$, $s{=}10$, $\tau(\mathbf{x}_0){=}2.7[p(\mathbf{x}_0){=}.004]$, the post-data severity evaluation will be based on:

$$SEV(T_\alpha;\mu{>}\mu_1){=}\mathbb{P}(\tau(\mathbf{X}){\leq}\tau(\mathbf{x}_0);\ \mu{=}\mu_1),\ \text{for}\ \mu_1{=}\mu_0{+}\gamma,\ \text{for}\ \gamma \geq 0,$$

where $\mu > \mu_1{=}\mu_0{+}\gamma$ is the relevant inferential claim. The warranted discrepancy from $\mu_1{=}\mu_0$ is determined by assuming a threshold, say $SEV(T_\alpha;\mu{>}\mu_1) \geq .90$ and calculating the associated biggest discrepancy $\gamma^*{\leq}1.41$. Note that the severity evaluation is always one-sided because it depends on the sign of $\tau(\mathbf{x}_0)$, like the p-value. To get some idea as to how the warranted discrepancy decreases as n increases, consider the overly simplistic case where $\tau(\mathbf{x}_0){=}2.7$ and $s{=}10$ remain the same, but $n{=}400$, then for $SEV(T_\alpha;\mu{>}\mu_1) \geq .90$, $\gamma^*{\leq}.705$.

More broadly, the severity evaluation provides an evidential interpretation of the coarse accept/reject rules that revolves around the warranted discrepancy from H_0. In this sense, the error statistical approach constitutes a refinement/extension of the Fisher-Neyman-Pearson frequentist inference when viewed as testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$; see Mayo and Spanos (2011).

Statistical vs. substantive significance. Establishing the magnitude of the discrepancy γ^* is as far as statistical information can take an inferential claim. Whether the statistically warranted discrepancy is substantively significant can only be established using substantive information to supplement the inferential claim. This raises the broader problem of relating the statistical with the substantive information.

4 Statistical vs. substantive adequacy

4.1 Statistical misspecification

Before any inferences are drawn, one needs to establish the statistical adequacy of the invoked statistical model $\mathcal{M}_\theta(\mathbf{x})$; the validity of its probabilistic assumptions vis-a-vis data $\mathbf{x}_0$. When any of the statistical model ($\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$) assumptions are invalid for data $\mathbf{x}_0$, both $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}{\in}\mathbb{R}^n_X$ and the likelihood function $L(\boldsymbol{\theta};\mathbf{x}_0) \propto f(\mathbf{x}_0;\boldsymbol{\theta})$, $\boldsymbol{\theta}{\in}\Theta$ are wrong. That, in turn, undermines the reliability of inference by distorting the sampling distribution $f(y_n;\boldsymbol{\theta})$ any statistic $Y_n{=}g(X_1, X_2, ..., X_n)$ (estimator, test statistic, predictor) derived via:

$$F(Y_n \leq y) = \underbrace{\int\int\cdots\int}_{\{\mathbf{x}:\ g(\mathbf{x})\leq y\}} f(\mathbf{x};\boldsymbol{\theta})d\mathbf{x},\ \forall y\in\mathbb{R}. \tag{13}$$

This can derail the optimality of these procedures (e.g. inconsistency) and induce sizeable discrepancies between the actual error probabilities (type I, II, p-values, coverage) and the nominal (assumed) ones – the ones derived by invoking these assumptions. Applying a .05 significance level test, when the actual type I error is closer to .9, will lead an inference astray.

It is important to emphasize that due to the reliance on the likelihood function, Bayesian inference is also equally vulnerable to statistical misspecification since the posterior distribution is: $\pi(\boldsymbol{\theta}|\mathbf{x}_0) \propto \pi(\boldsymbol{\theta})\cdot L(\boldsymbol{\theta};\mathbf{z}_0),\ \boldsymbol{\theta}\in\Theta$.
This is also the case for Akaike-type model selection procedures relying on $L(\boldsymbol{\theta};\mathbf{z}_0)$ (Spanos, 2010c), as well as nonparametric procedures whose statistical models impose potentially invalid dependence and heterogeneity assumptions.

4.2 Statistical inadequacy and the reliability of inference

The unreliability of inference stemming from imposing erroneous probabilistic assumptions on one's data (statistical misspecification) reveals itself in a variety of forms, including *inconsistent* estimators and *sizeable discrepancies* between actual and nominal (assumed) error probabilities (type I and II, coverage, etc.).

Table 4: Linear Regression (LR) and mean heterogeneity

	Adequate LR model				**Misspecified LR model**			
N=10000	True: Y_t=1.5+.5x_t+u_t, Estim: Y_t=β_0+$\beta_1 x_t$+u_t,				True: Y_t=1.5+$\underline{.13t}$+.5x_t+u_t Estim: Y_t=β_0+$\beta_1 x_t$+u_t,			
	n=50		n=100		n=50		n=100	
Parameters	Mean	Std	Mean	Std	Mean	Std	Mean	Std
$[\beta_0$=1.5$]\ \widehat{\beta}_0$	1.502	.122	1.500	.087	0.462	.450	0.228	.315
$[\beta_1$=.5$]\ \widehat{\beta}_1$	0.499	.015	0.500	.008	1.959	.040	1.989	.015
$[\sigma^2$=.75$]\ \widehat{\sigma}^2$	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
$\tau_{\beta_0}=\frac{\widehat{\beta}_0-\beta_0}{\widehat{\sigma}_{\beta_0}}$	0.004	$\boxed{.049}$	0.015	.050	-1.968	$\boxed{.774}$	-3.531	0.968
$\tau_{\beta_1}=\frac{\widehat{\beta}_1-\beta_1}{\widehat{\sigma}_{\beta_1}}$	-.013	$\boxed{.047}$	-.005	.049	35.406	$\boxed{1.00}$	100.2	1.000

Spanos and McGuirk (2001) illustrate this using simulation based on a LR model (N=10000 replications) under two different scenarios:

Scenario 1: the true and the estimated model coincide: Y_t=1.5+.5x_t+u_t,

Scenario 2: the true model is Y_t=1.5+$\underline{.13t}$+.5x_t+u_t but the estimated model model is Y_t=1.5+.5x_t+u_t.

The key conclusions under scenario 1 (statistically adequate model, table 4):

(i) the point estimates are *highly accurate* and the empirical 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 from n=50 to n=100.

The key conclusions under scenario 2 (statistically misspecified model):

(iii) the point estimates are *highly inaccurate* (symptomatic of *inconsistent* estimators) and the empirical type I error probabilities are *much larger* than the nominal (α=.05), and (iv) as n increases the inaccuracy of the estimates increases and the empirical type I error probabilities approach 1!

It is important to emphasize that the above simulation example is only indicative of the effects of departures from a single assumption. In practice, this is rather rare because more than one assumptions are often invalid; see Spanos and McGuirk (2001).

4.3 Substantive misspecification

Empirical modeling across different disciplines involves an intricate blending of *substantive* subject matter and *statistical information.* The substantive information stems form a theory or theories pertaining to the phenomenon of interest that could range from simple conjectures to intricate *substantive* (structural) models. Such information has an important and multifaceted role to play by demarcating the crucial aspects of the phenomenon of interest, e.g. suggesting the relevant variables and data, etc. In contrast, statistical information stems from the *chance regularities* exhibited by the chosen data $\mathbf{x}_0$. Scientific knowledge often begins with substantive conjectures based on subject matter information, but it becomes knowledge when its veracity is firmly grounded in real world data. In testing scientific (substantive) theories one needs to embed the substantive model into a statistical one. Only then can the data be brought to bear upon the adequacy of the scientific theory.

'**All models are wrong, but some are useful**'. When the trustworthiness of empirical results is questioned, practitioners often invoke the authority of George Box (1979) who coined this slogan, misinterpreting as rendering statistical misspecification inevitable. This is not what Box said: "Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model." (p. 202) which clearly suggests that the 'wrongness' he alludes to pertains to *substantive* adequacy, asserting that substantive models are crude approximations of the real world. Indeed, Box (1979) goes on to bring out the crucial role of testing assumptions using the residuals, by viewing empirical modeling as an iterative process driven by diagnostic checking (p. 204)"Suitable analysis of residuals can lead to our fixing up the model in other needed directions."

It is important to note that the use of the residuals to test the model assumptions was initially proposed and elaborated on in Box and Jenkins (1970). Hence, it is one thing to claim that all substantive models are not exact pictures of reality and completely another to claim that imposing invalid probabilistic assumptions on one's data is inevitable. In this sense, the misinterpretation of the Box slogan conflates two different questions (Spanos, 2010d):

[a] **statistical adequacy**: does $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ account for the chance regularities in $\mathbf{x}_0$? $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is built exclusively on the statistical information contained in data $\mathbf{x}_0$, and acts as a *mediator* between $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ and $\mathbf{x}_0$.

[b] **substantive adequacy:** does the model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ adequately captures (describes, explains, predicts) the phenomenon of interest? Substantive inadequacy arises, not from invalid probabilistic assumptions, but from highly unrealistic structural models, flawed *ceteris paribus* clauses, missing confounding factors, systematic approximation errors, etc. In this sense, probing for substantive adequacy is a considerably more complicated problem, which, at the very minimum, includes:

(i) securing statistical adequacy beforehand because without the reliability of any probing will be questionable, and

(ii) testing and confirming the validity of the overidentifying restrictions stemming from $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$, $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, $\boldsymbol{\varphi}\in\boldsymbol{\Phi}$.

4.4 The modeling vs. inference facets of statistical induction

How does one untangle the statistical from the substantive model? By separating the modeling from the inference facet of statistical analysis. The inference facet assumes the validity of $\mathcal{M}_{\theta}(\mathbf{x})$ with a view to secure the reliability and precision of inference, and the modeling facet aims to secure that validity. The modeling facet includes the cycle: (i) the specification of $\mathcal{M}_{\theta}(\mathbf{x})$, initial choice), (ii) Mis-Specification (M-S) testing and (iii) respecification when $\mathcal{M}_{\theta}(\mathbf{x})$ is found misspecified with a view to secure the statistical adequacy of the respecified model (table 5). Crudely put, conflating modeling with inference is analogous to mistaking the process of constructing a sailboat to preset specifications with sailing it in a competitive race; they are interrelated but separate facets. To borrow a phrase from Claeskens and Hjort (2008), p. xi, detaching the modeling from the inference facet is *not* 'the quiet scandal of statistics' as they claim. Imagine trying to construct a sailboat, using a pile of wooden planks, while sailing it! Instead, the quiet scandal of statistics is imposing invalid probabilistic assumptions on one's data and drawing unreliable inferences regardless.

Table 5: Empirical modeling and Inference

Modeling	1. Specification 2. Estimation 3. Mis-Specification (M-S) Testing 4. Respecification ∴ *Statistically adequate model*
Inference:	estimation, testing, prediction, simulation

The above framework also differs from traditional discussions by drawing a clear distinction between statistical and substantive information by proposing a purely probabilistic construal of the concept of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, rendering it very different from the relevant substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$. In testing substantive (scientific) hypotheses one needs to embed the substantive model into a statistical one. A statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is selected to meet two interrelated aims:

(i) to account for the chance regularities in data $\mathbf{x}_0$ by choosing a probabilistic structure for the generic stochastic process $\{X_t,\ t\in\mathbb{N}\}$ so as to render $\mathbf{z}_0$ a 'typical realization' thereof, and

(ii) to parameterize ($\boldsymbol{\theta}\in\Theta$) with a view to embed (parametrically) the substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ in its context via restrictions of the form $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$, $\boldsymbol{\theta}\in\Theta$, $\boldsymbol{\varphi}\in\Phi$, relating the statistical ($\boldsymbol{\theta}$) and substantive ($\boldsymbol{\varphi}$) parameters.

Only then can the data be brought to bear upon the adequacy of the scientific theory. Before substantive adequacy can be probed, however, one needs to establish the adequacy of the encompassing statistical model: the validity of its probabilistic assumptions vis-a-vis $\mathbf{x}_0$. Without it, there is no reason to presume that the inferences drawn are error reliable.

5 The reluctance to validate statistical models

Statistical adequacy is crucially important for inference because 'no trustworthy evidence for or against a substantive theory (or claim) can be secured on the basis of a statistically misspecified model'. In light of that, 'why are practitioners so reluctant to validate their statistical models?' There are several reasons for this neglect, including the following.

Theoretician vs. practitioner divide. An important contributor to the uninformed application of statistical tools that yields untrustworthy evidence, is a subtle disconnect between the theoretician (theoretical statistician), that leaves the practitioner unable to assess the appropriateness of different methods for particular data. The theoretician develops the statistical techniques associated with different statistical models for different types of data (time-series, cross-section, panel), and the practitioner implements these inferential tools using data, often observational. Each will do a much better job at their respective tasks if only they understood sufficiently well the work of the other. The theoretician will be more cognizant of the difficulties for the proper implementation of these tools, and make a conscious effort to elucidate their scope, applicability and limitations. Such knowledge will put the practitioner in a better position to produce trustworthy evidence by applying such tools only when appropriate. For instance, in proving that an estimator is Consistent and Asymptotically Normal (CAN), the theoretician could invoke *testable* assumptions. This will the practitioner a chance to appraise the appropriateness of different methods and do a much better job in producing trustworthy evidence by going the extra mile to test the validity of the invoked assumptions; see Spanos (2018).

Unfortunately, empirical modeling is currently dominated by a serious disconnect between these two since the theoretician is practicing *mathematical deduction* and the practitioner indulges in recipe-like *statistical induction* by transforming formulae into numbers using the data. The theoretician has no real motivation to elucidate the invoked deductive premises with a view to render them testable. If anything, the motivation is to invoke the mathematically weakest possible assumptions, irrespective of testability. Indeed, when challenged, theoreticians often argue (misleadingly) that the weaker the assumptions the less vulnerable the result to misspecification. The impression is more apparent than real. Weak probabilistic assumptions can be equally invalid as strong ones. What really ensures the reliability of inference is the *testability* of the inductive premises. As argued by Fisher (1922): "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).

2. Underestimation of the potential impact of statistical misspecification. The empirical literature appears to seriously underestimate the potentially devastating effects of statistical misspecification on the reliability of inference. This misplaced confidence in the reliability of inference stems from a number of different questionable arguments and claims often used in the traditional literature.

(a) The first is based on invoking *generic robustness results* whose generality and applicability is often greatly overvalued. Certain robustness results for particular departures from Normality are known, but are of limited value because they specify particular forms of non-Normality, e.g. retain symmetry. However, there are no robustness results for general departures from probabilistic assumptions pertaining to dependence and heterogeneity. For instance, in the case of the Linear Regression model (table 8), there are no robustness results for generic departures of the form:

$$E\left(u_t|X_t{=}x_t\right)\neq 0, Var\left(u_t|X_t{=}x_t\right)\neq\sigma^2,\ E\left(u_t u_s|X_t{=}x_t\right)\neq\sigma^2,\ t{>}s,\ t,s{\in}\mathbb{N}.$$

(b) The second questionable argument is that *asymptotic sampling distributions* render one's inferences less vulnerable to statistical misspecification. 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." (p. xiv)
The trustworthiness of inference results depend only on the approximate validity of the probabilistic model assumptions for the particular $\mathbf{Z}_0$ and nothing else.

3. Empirical modeling viewed as curve-fitting. In the empirical literature the statistical premises are misleadingly blended with the substantive premises of inference. This is primarily because empirical modeling is viewed as a *curve-fitting problem* guided by goodness-of-fit. The substantive (structural) model is foisted on the data in an attempt to quantify its unknown parameters. This treats the substantive information as established knowledge, and not as tentative conjectures to be tested against data. Unfortunately, excellent goodness-of-fit is neither necessary nor sufficient for statistical adequacy; see Spanos (2007). The end result of curve-fitting is often an estimated model that is misspecified, both statistically (invalid assumptions) and substantively; it doesn't shed trustworthy light on the phenomenon of interest.

4. Conflating M-S testing with N-P testing. This confusion stems primarily from the fact that the distinction between testing within vs. testing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is missing from the traditional literature. Moreover, the same procedures for constructing optimal tests, the Likelihood-ratio, the Lagrange Multiplier and the Wald procedures, are employed for both types of testing; see Godfrey (1988). This has led to a number of misleading claims and charges against M-S testing such as calling into question the legitimacy and value of the latter, including 'illegitimate double use of data', 'pre-test bias', 'infinite regress', etc. As argued in Spanos (2010c), these criticisms stem from insufficient understanding of the applicability, the primary objective and the underlying reasoning of M-S testing, combined with ignoring the crucial distinctions between modeling vs. inference and testing within vs. testing outside the boundaries of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

5. Non-testable probabilistic assumptions. Practitioners rarely have a complete list of testable probabilistic assumptions defining statistical models, even in cases where some of the probabilistic assumptions are made explicit. For instance, the incompleteness of the traditional specification of the Linear Regression (LR) model (table 8) becomes apparent when compared with a complete and testable specification in table 9. For instance, assumption [5] and the parameterization in table 9 are only implicit in table 8; see McGuirk and Spanos (2009).

Table 8: Linear Regression model: traditional specification

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

{1}	Normality:	$(u_t \vert X_t{=}x_t) \backsim \mathsf{N}(.,.)$,	$t{\in}\mathbb{N}$.
{2}	Zero mean:	$E\left(u_t \vert X_t{=}x_t\right){=}0$,	
{3}	Homoskedasticity:	$Var\left(u_t \vert X_t{=}x_t\right){=}\sigma^2$,	
{4}	Zero correlation:	$\{(u_t \vert X_t{=}x_t)\,,\ t{\in}\mathbb{N}\}$ is uncorrelated,	

Table 9: 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:	$(Y_t \vert X_t{=}x_t) \backsim \mathsf{N}(.,.)$,	$t{\in}\mathbb{N}$.
[2]	Linearity:	$E\left(Y_t \vert X_t{=}x_t\right){=}\beta_0+\beta_1 x_t$,	
[3]	Homoskedasticity:	$Var\left(Y_t \vert X_t{=}x_t\right){=}\sigma^2$,	
[4]	Independence:	$\{(Y_t \vert X_t{=}x_t)\,,\ t{\in}\mathbb{N}\}$ indep. process,	
[5]	t-invariance:	$(\beta_0,\beta_1,\sigma^2)$ are *not* changing with t,	

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

6 Misspecification testing: probing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$

In contrast to testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, Mis-Specification (M-S) testing constitute *testing outside* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ but within $[\mathcal{P}(\mathbf{x})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ for potential departures from the assumptions defining $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$; $\mathcal{P}(\mathbf{x})$ is the set of all possible statistical models that could have given rise to data $\mathbf{x}_0$. More formally, the key difference between M-S and N-P types of testing is:

Testing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: probing for validity of its assumptions

$$H_0\text{: } f(\mathbf{x};\boldsymbol{\theta}^*)\in\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}) \text{ vs. } \overline{H}_0\text{: } f(\mathbf{x};\boldsymbol{\theta}^*)\in\overline{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})}=[\mathcal{P}(\mathbf{x})-\mathcal{M}_{\theta}(\mathbf{x})]\,. \tag{14}$$

Testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: learning from data about $\mathcal{M}^*(\mathbf{x})$

$$H_0\text{: } f(\mathbf{x};\boldsymbol{\theta}^*)\in\mathcal{M}_0(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta_0\} \text{ vs. } H_1\text{: } f(\mathbf{x};\boldsymbol{\theta}^*)\in\mathcal{M}_{\mathbf{1}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta_1\}.$$

6.1 A coherent framework for M-S testing

In practice, however, $\overline{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})}$ cannot be fully operationalized, and thus M-S testing is more open-ended than N-P testing since it depends on how one renders probing $\overline{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})}=[\mathcal{P}(\mathbf{x})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ operational; see Spanos (2018).

The non-operational nature of $[\mathcal{P}(\mathbf{x})-\mathcal{M}_{\theta}(\mathbf{x})]$ raises a number of conceptual and technical issues, including the following.

(a) How to particularize $[\mathcal{P}(\mathbf{z})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})]$ to render it amenable to M-S testing.

(b) Securing the effectiveness/reliability of the diagnosis based on M-S tests.

It is important to emphasize the fact that M-S testing is a form of significance testing where null hypothesis is always defined by:

$$H_0\text{: all assumptions of } \mathcal{M}_{\theta}(\mathbf{x}) \text{ are valid for } \mathbf{x}_0, \tag{15}$$

and the default alternative is by default the negation of H_0. However, since $\overline{H}_0$: $\overline{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})}$ is non-operational, one needs to particularize it to some operational H_1, such that $H_1\subset[\mathcal{P}(\mathbf{z})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})]$, with the alternative in words being:

$$\begin{array}{l} H_1\text{: the stated departures from specific assumptions being tested,} \\ \textit{assuming} \text{ the rest of the assumption(s) of } \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}) \text{ hold for data } \mathbf{x}_0. \end{array} \tag{16}$$

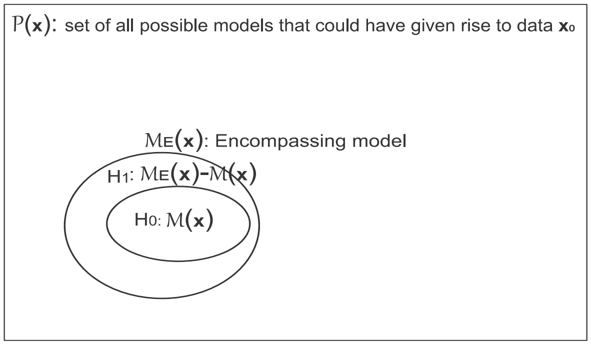

P(x): set of all possible models that could have given rise to data x₀
ME(x): Encompassing model
H1: ME(x)-M(x)
H0: M(x)

Fig. 6: M-S testing by encompassing

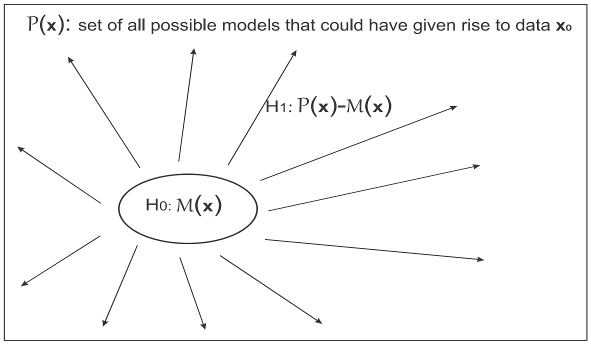

P(x): set of all possible models that could have given rise to data x₀
H1: P(x)-M(x)
H0: M(x)

Fig. 7: M-S testing: directions of departures

The particularization of $[\mathcal{P}(\mathbf{x}) - \mathcal{M}_\theta(\mathbf{x})]$ can take a number of different forms, including (i) parametric, (ii) nonparametric tests, as well as (iii) directions of departure. *Nonparametric* (omnibus) tests, such as the runs test, the Pearson chi-square and the Kolmogorov tests, are usually *non-directional* in the sense that the alternative hypothesis is defined as the negation of the null, rendering them particularly useful for M-S testing because that implies a broader local scope. The most serious weakness of nonparametric M-S tests is that they rarely provide information about the source of departure. For that information a practitioner needs to use parametric (directional) tests. Parametric M-S tests often take two forms. The first takes the form of a direction of departure from specific assumptions based on auxiliary regressions (fig. 7); see section 4.2. The second takes the form of encompassing $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ into a broader model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ (fig. 6) and testing the nesting restrictions; see Spanos (2018).

Mis-Specification (M-S) vs. N-P testing. (a) The fact that N-P testing is probing *within* the boundaries of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ and M-S testing is probing outside, i.e. $[\mathcal{P}(\mathbf{x}) - \mathcal{M}_\theta(\mathbf{x})]$, render the latter more vulnerable to the fallacy of rejection. Hence in practice one should *never* accept the particularized H_1 without further probing.

(b) In M-S testing the *type II error* [accepting the null when false] is often the more serious of the two errors. This because one will have another chance to correct for the type I error [rejecting the null when true] at the respecification stage, where a new model aims to account for the chance regularities the original model ignored. Hence, M-S testing is also more vulnerable to the fallacy of acceptance.

(c) The objective M-S testing is to probe as broadly beyond the null ($\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is valid) as possible, and thus tests with *low power* but broad (local) probing capacity have an important role to play. Curiously, the low local power is a blessing in M-S testing because when they indicate departures provides better evidence for its existence than parametric tests with very high power; see Spanos (2018).

6.2 Joint M-S testing vs. individual assumption tests

The form of null and alternative hypotheses for M-S testing in (15)-(16) call for assuming the validity of as fewer assumptions as possible under the alternative (16) to avoid misleading diagnosis. Moreover, model assumptions are usually interrelated, and thus testing them individually can give rise to misleading diagnoses.

A strong case can be made that the best strategy to avoid 'erroneous' diagnoses, minimize the number of maintained assumptions and enhance the scope of the tests is to use *joint M-S testing*. In particular, joint M-S testing based on auxiliary regressions has several distinct advantages over other procedures based on individual test statistics, such as Lagrange Multiplier tests for homoskedasticity, the Durbin-Watson and the Box-Pierce tests for no-autocorrelation, the Ramsey RESET test; see Godfrey (1988). In addition to minimizing the error of misdiagnoses, the explicit estimation of the auxiliary regressions enables the modeler to view the statistical significance of each individual term. For instance, a practitioner can easily cover up the presence of first order autocorrelation in the residuals by using a Box-Pierce test with a high order p of lags. Reporting the estimated auxiliary regressions associated with the performed M-S testing leaves no room for such mindless cover ups.

To simplify the discussion, the focus will be on the LR regression model in table 5, but the proposed auxiliary regressions can be easily extended to any statistical model of interest, including statistical models for cross-section and panel data.

The auxiliary regressions use the residuals $\{(\widehat{u}_t, \widehat{u}_t^2),\ t{=}1,2,...,n\}$ to probe for departures from model assumptions [2]-[5] (table 5), as they relate to:

$$H_0\text{: } E\left(Y_t \mid X_t{=}x_t\right){=}\beta_0{+}\beta_1 x_t,\ Var\left(Y_t \mid X_t{=}x_t\right){=}\sigma^2. \tag{17}$$

Any departures from assumptions [2]-[5] will change these two functions. To pick up such changes one uses additional terms, based on information already in the original data $\mathbf{Z}_0$, that indicate directions of potential departures from [2]-[5]:

$$\widehat{u}_t{=}\delta_0 + \delta_1 x_t + \overbrace{\delta_2 t}^{\overline{[5]}} + \overbrace{\delta_3 x_t^2}^{\overline{[2]}} + \overbrace{\delta_4 x_{t-1} + \delta_5 Y_{t-1}}^{\overline{[4]}} + v_{1t}, \tag{18}$$

H_0: $\delta_1{=}\delta_2{=}\delta_3{=}\delta_4{=}\delta_5{=}0$ vs. H_1: $\delta_1{\neq}0$ or $\delta_2{\neq}0$ or $\delta_3{\neq}0$ or $\delta_4{\neq}0$ or $\delta_5{\neq}0$.

$$\widehat{u}_t^2{=}\gamma_0 + \overbrace{\gamma_2 t}^{\overline{[5]}} + \overbrace{\gamma_1 x_t + \gamma_3 x_t^2}^{\overline{[3]}} + \overbrace{\gamma_4 x_{t-1}^2 + \gamma_5 Y_{t-1}^2}^{\overline{[4]}} + v_{2t}, \tag{19}$$

H_0: $\gamma_1{=}\gamma_2{=}\gamma_3{=}\gamma_4{=}\gamma_5{=}0$ vs. H_1: $\gamma_1{\neq}0$ or $\gamma_2{\neq}0$ or $\gamma_3{\neq}0$ or $\gamma_4{\neq}0$ or $\gamma_5{\neq}0$.

Intuitively, the above auxiliary regressions should be viewed as an attempt to probe the residuals $\{\widehat{u}_t,\ t{=}1,2,...,n\}$ for any remaining systematic information that might have been overlooked by the regression and skedastic functions in (17) in terms of assumptions [2]-[5].

A more formal justification/derivation of auxiliary regressions such as (18)-(19) can be based on the conditional expectation orthogonality theorem (Williams, 1991):

$$E\left([y{-}E(y|\sigma(\mathbf{X}))]{\cdot}h(\mathbf{X})\right){=}0, \text{ for any Borel-function } h(\mathbf{X}); \tag{20}$$

see Spanos (2018) for further details and extensions.

It is important to emphasize that the above F-type M-S tests based on (18)-(19) are viewed as significance tests whose only objective is to detect the presence of departures from the model assumptions. These tests are particularly vulnerable to the fallacies of acceptance and rejection. To circumvent the fallacy of rejection one should never adopt the particular alternative hypothesis (model) on the basis of a particular M-S test, without further testing. To deal with the fallacy of acceptance one should select a trenchant and comprehensive battery of M-S tests and ensure that the sample size n is large enough to enable one to detect any existing departures. Indeed, a rule of thumb for the minimal value of n is that: if n is not large enough to do a thorough job with M-S testing, it is not sufficient large for inference purposes. Note that the fallacy of acceptance is more serious for M-S testing because a false rejection could be remedied at the respecification stage, but there is no recourse in the case of a false acceptance.

Finally, in addition to applying joint M-S tests, one could render the probing more effective by using additional strategies, including: (a) judicious combinations of omnibus (non-parametric) and directional (parametric) tests, and (b) astute ordering of M-S tests so as to harness the interrelationship among the model assumptions. For instance, if the regression function is misspecified, the skedastic function-based M-S tests will give rise to misleading results, since the latter is defined with respect to a misspecified regression function.

The only assumption that (18)-(19) do not test for is [1] Normality. This is because the available M-S tests for [1] assume that the other assumptions are valid, rendering the results questionable when that is not the case. Hence, for a reliable test of Normality one should secure the validity of [2]-[5] beforehand.

6.3 Example. Capital Asset Pricing Model (CAPM)

Lai and Xing (2008), pp. 72-81, illustrate the CAPM using *monthly data* (n=64): y_t is excess (log) returns of Intel, x_t is the market excess (log) returns based on the SP500 index; the risk free returns is based on the 3-month Treasury bill rate. Estimation of the statistical (LR) model that nests the CAPM when the constant is zero yields:

$$y_t = \underset{(.009)}{.02} + \underset{(.237)}{1.996}x_t + \widehat{u}_t, \;\; R^2 = .536, \;\; s = .0498, \;\; n = 64, \tag{21}$$

where the standard errors are given in parentheses. Testing the significance of (β_0, β_1) using t-tests, the authors conclude that for α=.025 the coefficient β_0 is statistically insignificant but β_1 is significant, providing evidence for the CAPM.

The above inference results will be trustworthy when the LR probabilistic assumptions [1]-[5] (table 9) are valid for the particular data. The data $\{(y_t, x_t)\,,\; t=1, 2, ..., n\}$ (fig. 8-9) suggest that assumptions [4]-[5] are unlikely to be valid because the data exhibit very distinct time cycles and trends in the mean, and a shift in the variance after observation t=30. These are confirmed by the following auxiliary regressions:

$$\begin{gathered}\widehat{u}_t = \underset{(.011)}{.02} + \underset{(.311)}{.370}x_t + \underset{(.079)}{.091}t + \underset{(.094)}{.253}t^2 + \underset{(.075)}{.172}t^3 - \underset{(.122)}{.343}\,\widehat{u}_{t-1} + \widehat{v}_{1t}, \\ R^2 = .197, \;\; s = .0463, \;\; n = 63,\end{gathered} \tag{22}$$

$$\widehat{u}_t^2 = \underset{(.001)}{.002} - \underset{(.0005)}{.002}\,t + \underset{(.086)}{.141}\,\widehat{y}_t^2 + \widehat{v}_{2t}, \;\; R^2 = .2, \;\; s = .0037, \;\; n = 64 \tag{23}$$

The M-S testing results indicate that assumptions [3]-[5] are invalid (fig. 10).

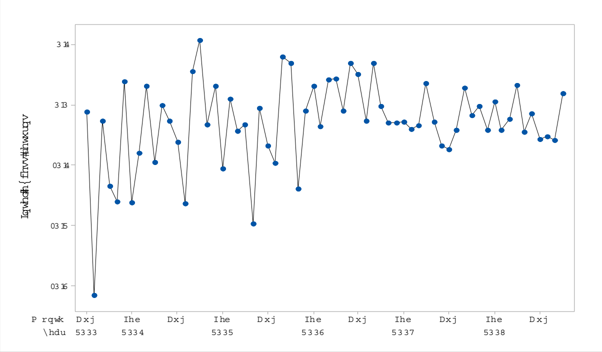


Fig. 8: Intel Corp. excess returns

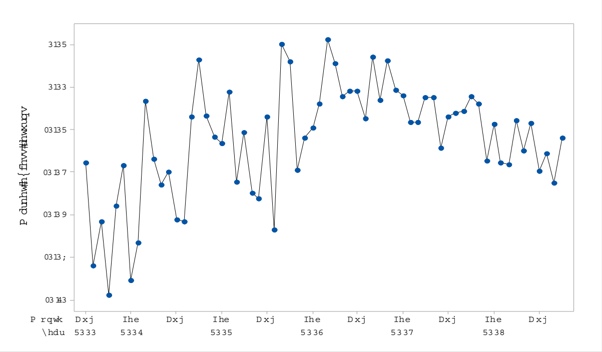


Fig. 9: Market excess returns

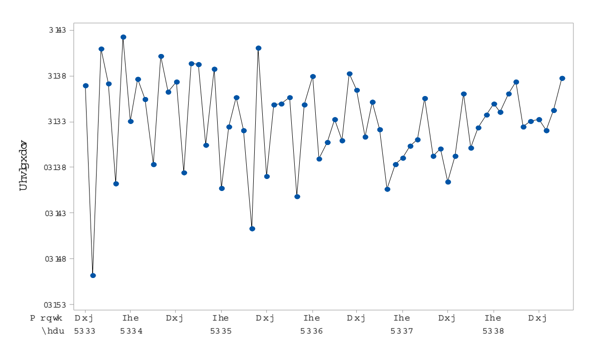


Fig. 10: Residuals from (22)

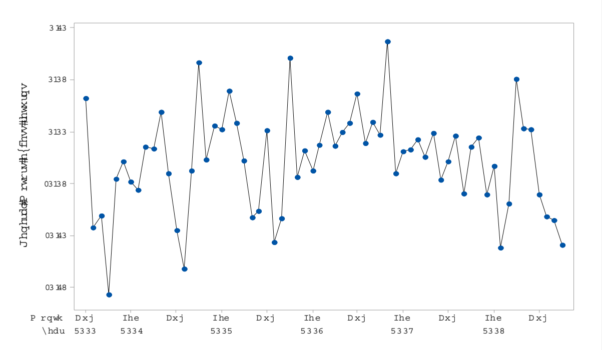


Fig. 11: General Motors excess returns

Probing for substantive adequacy (e.g. omitted variable). Consider posing the following substantive adequacy question to the statistically misspecified LR model in (21): Is z_{t-1}- last period's excess returns of General Motors an omitted variable in (21)? Adding z_{t-1} gives rise to:

$$y_t = \underset{(.008)}{.013} + \underset{(.233)}{2.086} x_t - \underset{(.129)}{.296} z_{t-1} + \widehat{\epsilon}_t, \;\; R^2 = .577, \;\; s = .0483, \;\; n = 64, \qquad (24)$$

which, when taken at face value, suggests that z_{t-1} is a relevant omitted variable. Really? Any variable that picks up the unmodeled trend (fig. 11), will appear to be statistically significant. Indeed, a simple respecification of the original model renders z_{t-1} insignificant:

$$y_t = \underset{(.015)}{.035} + \underset{(.315)}{2.319} x_t + \underset{(.081)}{.086} t + \underset{(.101)}{.190} t^2 + \underset{(.075)}{.158} t^3 - \underset{(.122)}{.316} y_{t-1} + \underset{(.498)}{.524} x_{t-1} - \underset{(.177)}{.164} z_{t-1} + \widehat{\varepsilon}_t. \qquad (25)$$

6.4 Example. Yule's 'nonsense-correlations'

Consider the empirical example in Yule (1926) using data on the ratio of Church of England marriages to all marriages (x_t) and the mortality rate (y_t) over the period 1866-1911, to demonstrate that their estimated correlation $\widehat{\rho}_{xy} = .9512$ appeared to be both very high and statistically significant. Yule made a genuine attempt to link 'nonsense-correlations' to the data in question *not* being 'random series'. He could not establish such a direct link because he was missing two key concepts that were yet to be fully integrated into statistics. The first was the concept of a prespecified 'parametric statistical model', introduced by Fisher (1922), comprising the totality of probabilistic assumptions imposed on the data. The second was the theory of stochastic processes founded by Kolmogorov (1933). The vague notion of a 'random series' was formalized into the concept of a realization of an IID stochastic process.

Yule's reverse engineering: "When we find that a theoretical formula applied to a particular case gives results which common sense judges to be incorrect, it is generally as well to examine the particular assumptions from which it was deduced, and see which of them are inapplicable to the case in point." (p. 4-5)

Emulating Yule's reverse engineering, let us elicit the implicit probabilistic assumptions that render the sample correlation coefficient:

$$\widehat{Corr(X_t, Y_t)} = \frac{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \overline{Y})(X_t - \overline{X})}{\sqrt{\left[\frac{1}{n}\sum_{t=1}^{n}(X_t - \overline{X})^2\right]\left[\frac{1}{n}\sum_{t=1}^{n}(Y_t - \overline{Y})^2\right]}}, \qquad (26)$$

$$\overline{X}=\tfrac{1}{n}\textstyle\sum_{t=1}^{n}X_t,\ \overline{Y}=\tfrac{1}{n}\sum_{t=1}^{n}Y_t,\ \widehat{Var(X_t)}=\tfrac{1}{n}\sum_{t=1}^{n}(X_t-\overline{X})^2,$$

$$\widehat{Var(Y_t)}=\tfrac{1}{n}\textstyle\sum_{t=1}^{n}(Y_t-\overline{Y})^2,\ \widehat{Cov(X_t,Y_t)}=\tfrac{1}{n}\sum_{t=1}^{n}(Y_t-\overline{Y})(X_t-\overline{X}),$$

a 'good' estimator of $Corr(X_t,Y_t)=\frac{Cov(X_t,Y_t)}{\sqrt{Var(X_t)Var(Y_t)}}$. The first implicit assumption is the *constancy* of the first two moments:

$$E(Y_t)=\mu_1,\ E(X_t)=\mu_2,\ Var(Y_t)=\sigma_{11},\ Var(X_t)=\sigma_{22},\ Cov(X_t,Y_t)=\sigma_{12},\ t\in\mathbb{N},$$

which corresponds to a form of the *ID assumption*. Second, the formulae for $\widehat{Var(X_t)}$ and $\widehat{Var(Y_t)}$, implicitly assume *non-correlation* over $t\in\mathbb{N}$, otherwise they should have included temporal covariance terms. Yule also sought to unveil the implicit distributional assumption "in order to reduce the formula to the very simple form given." (p. 5), knowing that the sample moments are not always 'optimal' estimators of the distribution moments; see Carlton (1946) for the case where the distribution is Uniform.

Given that under Normality the assumption of ID reduces to the constancy of the first two moments, and non-correlation coincides with *Independence* (I), one could make a case that the implicit parametric statistical model underlying the above formulae is *the simple bivariate Normal*:

$$\mathbf{Z}_t{:=}(Y_t,X_t)^\top \backsim \mathsf{NIID}(\boldsymbol{\mu}{:=}(\mu_1,\mu_2)^\top,\Sigma{:=}[\sigma_{ij}]_{ij=1}^2),\ t{=}1,2,...,n,... \qquad (27)$$

When any of the assumptions [1]-[4] (table 1) are invalid for the particular data $\mathbf{Z}_0$, the estimated correlation coefficient is likely to be 'spurious' (statistically untrustworthy). When any of the assumptions [1]-[4] are invalid for the particular data $\mathbf{Z}_0$, the estimated correlation coefficient is likely to be 'spurious' (statistically untrustworthy).

A glance at the t-plots of Yule's (1926) data (fig. 12-13) suggests that, to borrow his phrase: "Neither series, obviously, in the least resembles a random series" (aka IID); both data series exhibit mean t-heterogeneity (trending mean) and dependence (irregular cycles). To bring out the cycles in figures 12-13 more clearly one needs to subtract the trending means using a generic 3rd degree trend polynomial, as shown in figures 14-15, suggesting that assumptions [4]-[5] are likely to be invalid.

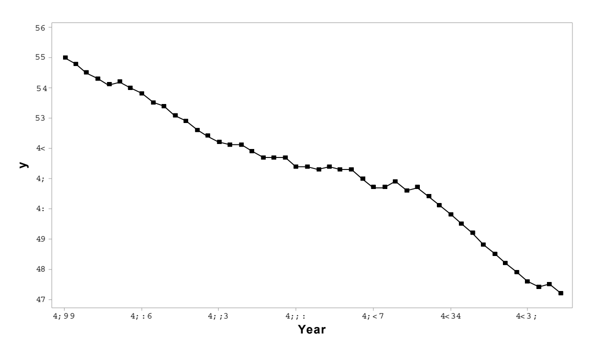


Fig. 12: t-plot of y_t-the mortality rate for the period 1866-1911

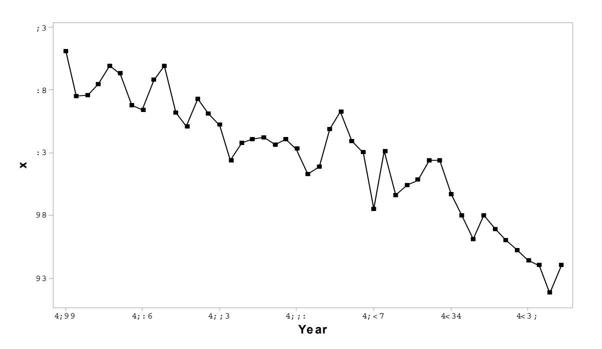
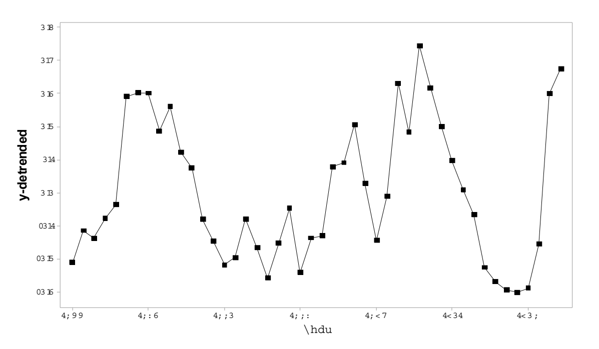


Fig. 13: t-plot of x_t-ratio of Church of England marriages to all marriages



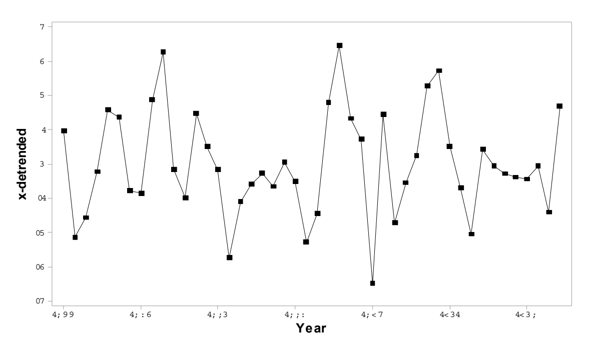


Fig. 14: t-plot of y_t-detrended

Fig. 15: t-plot of x_t-detrended

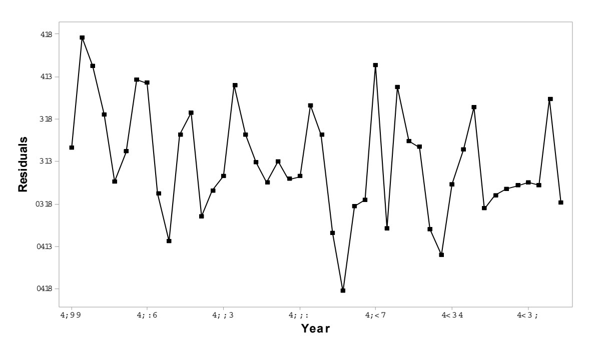
Given that the correlation (ρ_{12}) is directly related to the regression coefficient $(\alpha_1{=}(\sigma_{12}/\sigma_{22}))$ since $\rho_{12}{=}\alpha_1(\sqrt{\sigma_{22}}/\sqrt{\sigma_{11}})$, one can pose the question of statistical adequacy and inference in the context of the LR model, which yields:

$$y_t{=}\underset{(1.416)}{-10.847}+\underset{(.020)}{.419}x_t+\widehat{u}_t,\;\; R^2{=}.905,\;\; s{=}.664,\;\; n{=}46, \tag{28}$$

where $R^2{=}1{-}[\sum\widehat{u}_t^2/\sum_{t=1}^n(Y_t{-}\overline{Y})^2]$ denotes a goodness-of-fit measure, $s^2{=}(\frac{1}{n-2}\sum\widehat{u}_t^2)$ denotes the estimated variance of the regression, and in brackets below the estimates are reported the standard errors of the coefficients. Both coefficients (β_0,β_1) appear to be statistically significant since the t-ratios are:

$$\tau_0(\mathbf{z}_0){=}\tfrac{10.847}{1.416}{=}7.660[.000], \tau_1(\mathbf{z}_0){=}\tfrac{.419}{.020}{=}20.95[.000],$$

with the p-values given in square brackets; the implied correlation $\widehat{\rho}_{12}{=}\widehat{\alpha}_1(\sqrt{\widehat{\sigma}_{22}}/\sqrt{\widehat{\sigma}_{11}}))$ yields the values in Yule (1926): $\widehat{\rho}_{xy}{=}.419(\frac{4.854}{2.137}){=}.952[.000]$. However, the t-plot of the residuals (fig. 16) indicates that (28) is statistically misspecified since the residuals differs from that of a NIID realization (fig. 17) in so far as it exhibits distinct trends and cycles; see Spanos (1999), ch. 5. These results suggest that the estimator of β_1 in (28) is inconsistent and so is ρ_{xy}. *Hence,* and the t-tests for the significance of the coefficients are statistically untrustworthy.



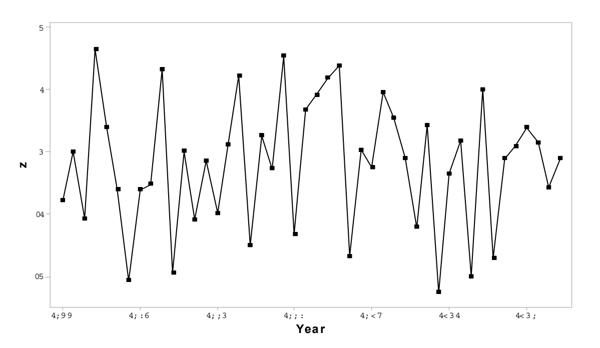


Fig. 16: Residuals from (28)

Fig. 17: t-plot of NIID data

6.5 Respecification, when a model is misspecified

The aim in respecification is to select probabilistic assumptions pertaining to the stochastic process $\{\mathbf{Z}_t{:=}(y_t,\mathbf{X}_t),\, t{\in}\mathbb{N}\}$ underlying the data that account for the chance regularities in $\mathbf{Z}_0$ not accounted for by the original model, as exemplified by the totality of M-S testing results. The traditional respecification usually takes the form of ad hoc 'error-fixing' moves, based on adopting the particularized alternative H_1 of the M-S test applied; a classic case of the fallacy of rejection.

Example: from OLS to GLS. Consider probing the validity of assumption [4] of the Linear Regression (LR) model (table 5) using the hypotheses:

$$H_0\text{: } \rho{=}0 \text{ vs. } H_1\text{: } \rho{\neq}0, \tag{29}$$

in the context of the encompassing model:

$$\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})\text{:} \quad Y_t{=}\beta_0+\beta_1 x_t+u_t,\ u_t{=}\rho u_{t-1}{+}\varepsilon_t,\ |\rho|{<}1,\ (\varepsilon_t|X_t{=}x_t)\backsim \mathsf{NIID}(0,\sigma_\varepsilon^2). \tag{30}$$

This was the first influential M-S test proposed by Durbin and Watson (1950), that became part of the LR output by the early software packages. The problem has been that the hypotheses in (29) looks suspiciously similar to a significance t-type test for β_1, and the literature misinterpreted it as another N-P test relating to the LR model. Worse, every econometrics textbook published after 1963 recommends that when the D-W test rejects H_0, one could correct the misspecification by adopting the alternative (encompassing) LR model (30). This strategy replaces the original Ordinary Least Squares (OLS) estimator of the parameters in $\boldsymbol{\theta}$ with the Generalized Least Squares (GLS) estimator of $\boldsymbol{\psi}$; see Greene (2012).

This respecification strategy, however, constitutes a classic example of the *fallacy of rejection*. When H_0 in (15) is rejected, the only inference that can be drawn is that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is misspecified and the data indicate a generic departure from assumption [4] (table 5), i.e. $E(u_t u_s|X_t{=}x_t){\neq}0$, but does *not* provide evidence *for* the particular alternative:

$$H_1\text{: } E(u_t u_s|X_t{=}x_t){=}(\tfrac{\rho^{|t-s|}}{1-\rho^2})\sigma_\varepsilon^2,\ t{>}s,\ t,s{=}1,2,...,n, \tag{31}$$

i.e. (30). This is because the D-W test would have rejected [4] for numerous particularized forms of departure from [4] within $[\mathcal{P}(\mathbf{z}){-}\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})]$, not just (31). For instance, the alternative encompassing model:

$$\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{z})\text{:} \quad Y_t{=}\alpha_0+\alpha_1 x_t+\alpha_2 Y_{t-1}+\alpha_3 x_{t-1}+v_t,\ (v_t|X_t{=}x_t)\backsim \mathsf{NIID}(0,\sigma_v^2), \tag{32}$$

would have elicited a similar rejection by the D-W test; see Spanos and McGuirk (2001). To establish the validity of (31) one needs to tests the probabilistic assumptions of $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ in (30) and secure their validity; evidence *against* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is *not* evidence *for* $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ or $\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{z})$. More often than not, error-fixing gives rise to respecified models with unnecessary and often implausible restrictions; see McGuirk and Spanos (2009).

Recall that in M-S testing the null is always H_0: $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is valid, but the alternative $\overline{H}_0$: $\mathcal{P}(\mathbf{z}){-}\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is non-operational. Hence, the modeler needs to select particularized alternatives or directions of departure that could never span the whole of $\mathcal{P}(\mathbf{z}){-}\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. This implies that when H_0 is rejected the particularized alternative H_1 is never an option for respecification purposes without further testing. In particular, the M-S testing based on auxiliary regressions, such as (18) and (19), could only provide information about departures from the original model assumptions. Significant coefficients indicate directions of departure from these assumptions. The auxiliary regressions do not provide a clear answer as to what the respecified model should look like, they only provide hints. That is decided by the statistical adequacy of the respecified model; its assumptions are tested anew and shown to be valid.

Model respecification. The above perspective offers a broader and more coherent vantage point from that stemming from the error process $\{(u_t|\mathbf{X}_t{=}\mathbf{x}_t),\ t{\in}\mathbb{N}\}$. It views the LR model as specified in terms of the regression and skedastic functions of $D(y_t \mid \mathbf{X}_t;\boldsymbol{\theta})$: $E(y_t \mid \mathbf{X}_t{=}\mathbf{x}_t){=}h(\mathbf{x}_t),\ Var(y_t \mid \mathbf{X}_t{=}\mathbf{x}_t){=}g(\mathbf{x}_t),\ \mathbf{x}_t{\in}\mathbb{R}_X^k$, where the functional forms $h(.)$ and $g(.)$, and the relevant parameterization $\boldsymbol{\theta}$ stem from the joint distribution $D(y, \mathbf{X}_t;\boldsymbol{\varphi})$. From this perspective, departures from particular assumptions might relate to both functions. For instance, the move of retaining the Linearity and Normality assumptions, but adopting an arbitrary form of Heteroskedasticity (Greene, 2012), can easily give rise to an internally inconsistent set of probabilistic assumptions; see Spanos (1995b).

Yule's example (continued). The misspecifications detected in Yule's LR model suggest that a way to get a statistically adequate model might be to respecify it by including trends and lags to account for the mean heterogeneity and dependence:

$$\begin{array}{c}y_t{=}\underset{(1.267)}{1.137}{-}\underset{(.016)}{.005}x_t{-}\underset{(.998)}{1.670}t{-}\underset{(.216)}{.406}t^2{+}\underset{(.076)}{.885}y_{t-1}{+}\underset{(.015)}{.006}x_{t-1}+\widehat{v}_t,\\ R^2{=}.996,\ s{=}.147,\ n{=}46\end{array}\qquad(33)$$

The estimated respecified model in (33) turns out to be statistically adequate, and hence it can be used to reliably confirm that $Corr\,(X_t, Y_t)=0$ (contemporaneously) and $Corr\,(X_{t-1}, Y_t)=0$ (temporally) since both coefficients of x_t and x_{t-2} are statistically insignificant. In contrast, both trends (t, t^2) and y_{t-1} are significant, confirming the above detected misspecifications. The results in (33) imply that a reliable estimate of $Corr\,(X_t, Y_t)$ is $\widehat{\rho}_{xy}{=}.038[.804]$, indicating that the original high linear association between X_t and Y_t, $\widehat{\rho}_{xy}{=}.952[.000]$, constitutes *statistically untrustworthy evidence.*

7 Summary and conclusions

The discussion in this paper calls into question the basic presumption underlying the replicability crises literature by making a case that replicability is neither necessary nor sufficient for trustworthy empirical evidence. It is argued that the PPV, as a posterior probability measure of untrustworthiness-by-association for discipline-wide testing, has no bearing on the trustworthiness of frequentist testing, since the latter is *local* in the sense that its results depend crucially on the particular $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, the relevant test $(d(\mathbf{X}),\ C_1)$ and the particular data $\mathbf{x}_0$; see Spanos (2013). Moreover, the abuse of significance testing is only a part of a much broader problem relating to the uninformed and mechanical implementation of statistical methods driven by software packages. The three most important sources of untrustworthy evidence in empirical modeling are: (i) statistical misspecification, (ii) poor understanding and implementation of frequentist inference procedures, and (iii) unwarranted evidential interpretations of their inferential results. Some of the alternative proposals to replace significance testing, such as using observed CIs and estimation-based effects sizes, are equally susceptible to the same sources (i)-(iii) of untrustworthy evidence.

To ensure *informed implementation* of statistical methods, the paper articulates a unifying framework for frequentist inference based on four key distinctions.

(a) Testing within vs. testing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. Testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ renders significance testing practically indistinguishable from N-P testing, and the latter is clearly different from M-S testing.

(b) Pre-data vs. post-data error probabilities. This brings out the differences between the significance level and the p-value, and motivates the post-data severity evaluation that provides a missing evidential interpretation of the p-value and the accept/reject rules. This can be used to address several foundations problems, including the fallacies of acceptance and rejection.

(c) Modeling vs. the inference facets of statistical analysis. This brings out the different questions posed to the data at different stages and the different procedures employed to answer these questions. It also helps to elucidate (d).

(d) Statistical vs. substantive information/model. This provides a bridge between the data and a scientific theory by showing that the data should be brought to bear upon substantive questions of interest, only when trenchant M-S testing has secured the adequacy of the statistical model.

The proposed modeling and inference framework offers suggestions for different ways the trustworthiness of published empirical results can be improved. A good starting point will be if journal referees and editors refuse to review papers that make no effort to demonstrate that their invoked statistical models are approximately valid for their data. A second step will be to ensure that the authors go the extra mile to ensure that their evidential interpretation of their p-values and accept/reject H_0 results is cogent by avoiding the well-known fallacies of acceptance and rejection. A third step is for editors and referees to demand that authors probe adequately for any potential substantive misspecifications after they secure the adequacy of the underlying statistical model. This includes testing whether any components of the substantive information belie the statistical information.

References

[1] Andreou, E. and A. Spanos (2003) "Statistical adequacy and the testing of trend versus difference stationarity", *Econometric Reviews*, **3**: 217-237.

[2] Begley, C.G. and J.P.A. Ioannidis (2015) "Reproducibility in Science: Improving the Standard for Basic and Preclinical Research," *Circulation Research*, **116**: 116–126.

[3] Box, G.E.P. (1979) "Robustness in the strategy of scientific model building", pp. 201–236 in Launer, R.L. and G.N. Wilkinson, *Robustness in Statistics*, Academic Press, London.

[4] Box, G.E.P. and G.M. Jenkins (1970) *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco.

[5] Carlton, A.G. (1946) "Estimating the Parameters of a Rectangular Distribution," *The Annals of Mathematical Statistics*, **17**: 355-358.

[6] Claeskens, G. and N. L. Hjort (2008) *Model Selection and Model Averaging,* Cambridge University Press, Cambridge.

[7] Cox, D.R. and D.V. Hinkley (1974) *Theoretical Statistics*, Chapman & Hall, London.

[8] Cubała, W.J., J. Landowski, J., and H.M. Wichowicz (2008), "Zolpidem abuse, dependence and withdrawal syndrome: sex as susceptibility factor for adverse effects", *British journal of clinical pharmacology*, **65**(3): 444-445.

[9] Fama, E. (1970) "Efficient capital markets: A review of theory and empirical work," *The Journal of Finance*, **25**: 383-417.

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

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

[12] Fisher, R. (1955) "Statistical methods and scientific induction", *Journal of the Royal Statistical Society, Series B,* **17**: 69-78.

[13] Fletcher, R.H. and Fletcher, S.W. (2005) *Clinical Epidemiology: the Essentials* (4th ed.), Lippincott Williams & Wilkins, Baltimore, ME.

[14] Godfrey, L.G. (1988) *Misspecification Tests in Econometrics*, Cambridge University Press, Cambridge.

[15] Greene, W.H. (2012), *Econometric Analysis*, 7th ed., New Jersey: Prentice Hall.

[16] Hacking, I. (1965) *Logic of Statistical Inference*, Cambridge University Press, Cambridge.

[17] Höffler, J.H. (2017) "Replication and economics journal policies", *American Economic Review*, 107(5): 52-55.

[18] Ioannidis, J.P.A., (2005) "Why most published research findings are false," *PLoS medicine*, 2: p.e124.

[19] Ioannidis, J.P.A., T.D. Stanley, and H. Doucouliagos (2017) "The Power of Bias in Economics Research," *Economic Journal*, 127: F236–F265.

[20] Kolmogorov, A.N. (1933) *Foundations of the theory of Probability*, 2nd English edition, Chelsea Publishing Co. NY.

[21] Lai, T.L., and H. Xing (2008), *Statistical models and methods for financial markets*, Springer, NY.

[22] Le Cam, L. (1986a) *Asymptotic Methods in Statistical Decision Theory*, Springer.

[23] McGuirk, A. and A. Spanos (2009) “Revisiting Error Autocorrelation Correction: Common Factor Restrictions and Granger Non-Causality”, *Oxford Bulletin of Economics and Statistics,* **71**: 273-294.

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

[25] Mayo, D.G. (2018) *Statistical Inference as Severe Testing: How to Get Beyond the Statistical Wars*, Cambridge University Press, Cambridge.

[26] Mayo, D.G. and A. Spanos (2004) “Methodology in Practice: Statistical Misspecification Testing”, *Philosophy of Science*, **71**: 1007-1025.

[27] 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.

[28] 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.

[29] National Academy of Sciences (2016) *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop.* Washington, DC: National Academies Press.

[30] Schneider, J.W. (2018) “Response to commentary on "Is NHST logically flawed",” *Scientometrics*, **116**: 2193–2194.

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

[32] Spanos, A. (1995b) “On Normality and the Linear Regression model”, *Econometric Reviews*, **14**: 195-203.

[33] Spanos, A. (1999), *Introduction to Probability Theory and Statistical Inference*, Cambridge University Press, Cambridge.

[34] Spanos, A. (2006) “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.

[35] Spanos, A. (2007) “Curve-Fitting, the Reliability of Inductive Inference and the Error-Statistical Approach”, *Philosophy of Science*, **74**: 1046–1066.

[36] Spanos, A. (2010a) “The Discovery of Argon: A Case for Learning from Data?” *Philosophy of Science*, **77**: 359-380.

[37] Spanos, A. (2010b) “Is Frequentist Testing Vulnerable to the Base-Rate Fallacy?” **Philosophy of Science**, **77**: 565-583

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

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

[40] Spanos, A. (2018) "Mis-Specification Testing in Retrospect", *Journal of Economic Surveys*, **32**: 541–577.

[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.

[42] Stark, P.B. and A. Saltelli, (2018) "Cargo-cult statistics and scientific crisis," *Significance*, **15**: 40-43.

[43] Williams, D. (1991) *Probability with Martingales*, Cambridge University Press, Cambridge.

[44] Yule, G.U. (1926) "Why do we sometimes get nonsense correlations between time series-a study in sampling and the nature of time series ", *Journal of the Royal Statistical Society*, **89**, 1-64.