The Replication Crises and the Trustworthiness of Empirical Evidence in Economics*

Aris Spanos
Department of Economics, Virginia Tech, USA

January 2019

Abstract

The paper questions the presumption by the replication crises literature that replicability ensures the trustworthiness of empirical evidence. It is argued that the abuse of significance testing is only a symptom 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. Indeed, in certain cases the same uninformed implementation can also render untrustworthy evidence replicable. It is argued that the trustworthiness of empirical evidence should be assessed at the individual study level, and not at a fictitious discipline-wide level evaluating untrustworthiness by association. 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 inference procedures, and (iii) unwarranted evidential interpretations of their inferential results. Moreover, alternative methods to replace significance testing, such as observed CIs and estimation-based effects sizes, are equally vulnerable to (i)-(iii).

1 Introduction

It has been well-documented in several disciplines that the majority of published empirical results are not replicable, and a sizeable percentage is not even reproducible; see Camerer et al. (2016); Hoffler (2017); Ioannidis et. al. (2017); Johnson et al. (2016); Nosek and Lakens (2014); National Academy of Science (2016). *Reproducibility* refers to being able to reproduce particular inference results using the same data and inferential methods. *Replicability* is broader and pertains to the potential of such results to be independently confirmed by other researchers studying the same phenomenon of interest. The non-replicability of published empirical results has been

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

broadly interpreted as providing clear evidence that "most published research findings are false"; see Ioannidis (2005). In light of that, several leading statisticians in different applied fields, as well as a few journal editors, pointed the finger at significance testing as a leading contributor to untrustworthy evidence and called for reforms; see Benjamin et. al (2017). The proposed reforms include replacing p-values with Confidence Intervals (CIs), using estimation-based effect sizes and redefining statistical significance (Mayo, 2018; Haig, 2016). To understand the merits of the charges against significance testing, as well as the ramifications of different proposals to replace it, there is a pressing need for a better understanding of the main sources of the untrustworthy evidence, together with a balanced appraisal of the proposed reforms to ameliorate the replicability problem.

Ioannidis (2005) made his case by focusing almost exclusively on well-known abuses of significance testing, such as ad hoc rejection thresholds ($<.05$), cherry picking, data-dredging, multiple testing and p-hacking, that introduce 'biases' into the empirical findings. These abuses are arbitrarily bundled into a Bayesian measure for an overall reliability of discipline-wide testing results, the Positive Predictive Value (PPV), borrowed from medical screening diagnostic devices. Unfortunately, the PPV has two major flaws, rendering it completely inappropriate for appraising the reliability of frequentist testing results. First, it is a discipline-wide Bayesian measure of *untrustworthiness by association* based on posterior probabilities framed in terms of false positive/negative rates. Second, the claimed analogies between the 'false positive' and 'false negative' probabilities for medical diagnostic devices and the type I and II frequentist error probabilities constitute egregious misinterpretations of the latter (section 3.3).

The primary aim of this paper is to revisit a key argument of the replication crises literature (Begley and Ioannidis, 2015) with a view to call into question the presumption that replicability secures trustworthy evidence. It is argued that, broadly speaking, *replicability* is *neither necessary nor sufficient* for *trustworthy empirical evidence*. Moreover, the current focus on the abuse of significance testing as the main culprit (Ioannidis at al., 2017) is much too narrow to explain the extent of the untrustworthiness of evidence problem. This is because the widespread abuse and misinterpretation of the p-value (p-hacking, multiple testing, cherry-picking, low power studies) is only a symptom of a much broader problem relating to *uninformed implementation* of statistical methods that contributes in *many different ways* to the problem of untrustworthy evidence. The main argument of this paper is (a) a call to separate replicability from trustworthiness of evidence and focus on the latter at the individual study level, and (b) make a case that the leading cause of the observed untrustworthiness is the recipe-like application of statistical methods without real understanding of their assumptions, limitations, proper implementation and warranted interpretations of their results; see Stark and Salteli (2018). This questions the credibility of the argument that securing replicability addresses the untrustworthy evidence problem, since the same mechanical implementation often

ensures that untrustworthy evidence can be routinely replicated, when uninformed practitioners follow the same recipe-like application of statistical methods. The same problem also questions the evaluation of the trustworthiness of evidence at a discipline wide level. The trustworthiness of the empirical results from a heedful implementation of statistical methods in a particular study cannot be undermined by the uniformed implementation of such methods by other practitioners in the same discipline.

The above comments suggest a refocusing of the proposed strategies to secure 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 that the three main sources of untrustworthy evidence in empirical modeling are: (i) statistical misspecification: invalid probabilistic assumptions imposed on the particular data, (ii) poor implementation of inference procedures, and (iii) unwarranted evidential interpretations of their inferential results.

The paper articulates a unifying framework for frequentist inference with a view to ensure an *informed implementation* of statistical methods that is grounded on in-depth understanding of their assumptions, their limitations, their proper implementation and the warranted interpretations 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 framework, proposals to replace significance testing, such as observed CIs, redefining statistical significance and estimation-based effects sizes, are equally vulnerable to the above main sources (i)-(iii) of untrustworthy evidence.

2 Revisiting the replicability problem

The metaphor that motivates invoking replicability as a way to secure trustworthy empirical evidence is that of experimentation in a scientific lab where an experiment is repeated many times under controlled conditions and the empirical findings are confirmed. 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)

2.1 Replicability and the trustworthiness of evidence

The question that should be asked is whether this *replication metaphor* is appropriate for appraising the trustworthiness of empirical evidence based on data generated beyond a controlled experimental setup. The answer is not obvious when one refers to empirical modeling with *observational data* in the social sciences. For instance,

would a macroeconometric model for the Italian economy replicate for any other European economy, assuming we use exactly the same data series and time period?

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. It is often insufficiently appreciated that substantive conjectures becomes knowledge only when their veracity is validated by the relevant data. The end result is invariably (h) empirical models that are statistically and substantively misspecified because excellent fit is neither necessary nor sufficient for statistical adequacy; Spanos (2007).

Second, viewing replicability as pertaining to the potential of empirical results to be independently confirmed by other researchers studying the same phenomenon of interest, raises serious issues when the latter refer to economic or social phenomena. This is because due to their invariance too time and location, physical phenomena can be viewed as isolated systems or can be isolated sufficiently in the lab under controlled conditions. No such isolation is possible for economic phenomena giving rise to observational data. This is why experimental results in economics are always vulnerable to external validity problems.

Third, 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 intricate 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 such data. This renders modeling with observational data a lot more vulnerable to statistical misspecification.

Fourth, 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. There are no such substantive models in the social sciences for several reasons, including the 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; see Spanos (1995a). Hence, the need to distinguish between a statistical and a substantive model.

Fifth, when viewed from a purely statistical perspective, it is important to note

that even in the best case scenario where each study: (i) uses the same sample size n, (ii) the data measure the same variables, (iii) the data constitute realizations of the same generating mechanism, and (iv) the estimated model is statistically adequate, the inference results for each study, including point and interval estimates, accept/reject testing results and p-values, will *not* coincide; there is always some sampling variability. The best one can expect is that they provide approximately the same broad empirical evidence; see Spanos and Mayo (2015). This is particularly important for the p-values for the different studies since $\mathbf{x}_0$ will be different, but when the same rule, $p(\mathbf{x}_0) \leq \alpha$, is used for a significant result, the statistic $p(\mathbf{X})$ is Uniformly distributed over [0,1] (Cox and Hinkley, 1974), and thus $\mathbb{P}(p(\mathbf{X}) \leq \alpha;\ H_0)=\alpha$.

Returning the question pertaining to a macroeconometric model for Italy, the answer is that it is highly unlikely that it will replicate for any other European country due to the fact that condition (iii) is likely to be false, since economic phenomena are often not invariant to location. Indeed, it is highly likely that the relevant explanatory variables might differ between different economies. Does that imply that the model for Italy constitutes untrustworthy evidence? Assuming that condition (iv) holds, the model will give rise to statistically trustworthy empirical evidence for Italy.

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

2.2 Replicable but untrustworthy 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.

The (weak) EM hypothesis asserts that 'changes in speculative prices are, in principle, unpredictable from their own past'. Letting P_t denote the speculative price of an asset (stock, bond, exchange rate, etc.), Fama (1970) framed this claim in terms of its log-returns, $y_t=\ln P_t-\ln P_{t-1}$, being a Normal, Martingale Difference:

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

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

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

The main reason such untrustworthy results can be replicated is primarily due to 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]-[5] are validated.

Table 1: Normal, AutoRegressive (AR(1)) Model

Statistical GM:		$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_1 y_{t-1} + u_t,\ t\in\mathbb{N}.$	
[1]	Normality:	$(y_t\|\sigma(\mathbf{y}_{t-1})) \backsim \mathsf{N}(.,.),$	$\left.\begin{array}{l}\\\\\\\\\\\end{array}\right\} t\in\mathbb{N}.$
[2]	Linearity:	$E\left(y_t\|\sigma(\mathbf{y}_{t-1})\right) = \alpha_0 + \sum_{i=1}^{p} \alpha_1 y_{t-1},$	
[3]	Homoskedasticity:	$Var\left(y_t\|\sigma(\mathbf{y}_{t-1})\right) = \sigma_0^2,$	
[4]	Markov:	$\{y_t,\ t\in\mathbb{N}\}$ is a Markov(p) process,	
[5]	t-invariance:	$(\alpha_i,\ i=0,1,...,p,\sigma_0^2)$ are *constant* over t,	

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:

$$y_t = \underset{(.018)}{.012} + \underset{(.035)}{.086}y_{t-1} + \underset{(.032)}{.002}y_{t-2} + \widehat{u}_t,\ s^2{=}.549,\ n{=}952. \tag{3}$$

Choosing a significance level of $\alpha{=}.01$, in light of the fact that the sample size is large, $n{=}954$, the t-ratios $\tau(\alpha_0){=}\frac{.011}{.018}{=}.61[.514]$, $\tau(\alpha_1){=}2.486[.013]$, $\tau(\alpha_2){=}.063{=}[.944]$, where the p-values are in square brackets, indicate that the coefficients are insignificant, confirming the EM hypothesis restrictions (2). However, it turns out that the invoked assumptions [1] and [3] are invalid, rendering such an inferences potentially unreliable. When a statistically adequate model that accounts for these departures is obtained after respecification, the estimated autoregressive function becomes:

$$y_t = \underset{(.012)}{.000} + \underset{(.031)}{.104}y_{t-1} + \underset{(.035)}{.093}y_{t-2} + \widehat{v}_t,\ \widehat{\sigma}_0^2{=}.375,\ n{=}952, \tag{4}$$

whose t-ratios $\tau(\alpha_1){=}3.355[.0004]$, $\tau(\alpha_2){=}2.657[.0039]$ provide clear evidence *against* the EM hypothesis; see section 6.4-5. That is, the evidence for the EM hypothesis is easily reproducible/replicable but untrustworthy due to *misspecification errors*, a major source of untrustworthiness largely ignored by the replication crises literature.

It is important to emphasize that untrustworthy evidence can be replicable even for experimental data. For instance, replicating an empirical regularity, an exhibit in experimental economics (Sugden, 2005), could be the result of uninformed practitioners following the same recipe-like application of statistical procedures.

3 Frequentist testing: a coherent framework

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:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}){=}\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}{\in}\Theta{\subset}\mathbb{R}^m\},\ \mathbf{x}{\in}\mathbb{R}_X^n,\ m{<}n, \tag{5}$$

where $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}{\in}\mathbb{R}_X^n$, 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).

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

$$X_t \backsim \mathsf{NIID}(\mu, \sigma^2),\ \boldsymbol{\theta}{:=}(\mu, \sigma^2){\in}\mathbb{R} \times \mathbb{R}_+,\ t{=}1, 2, ..., n, ..., \tag{6}$$

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. 1).

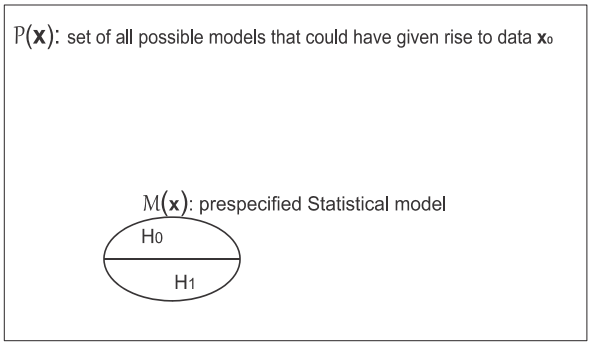

Fig. 1: Testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: N-P

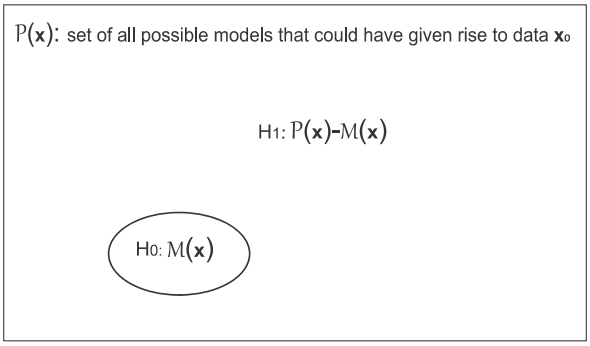

Fig. 2: Testing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: M-S testing

Neyman-Pearson (N-P) testing is always within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ (fig. 1), and Mis-Specification (M-S) testing is always testing outside since it probes $[\mathcal{P}(\mathbf{x}){-}\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ (fig. 2) with a view to test the validity 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 share the same setup. (i) The same *inductive premises* of inference, i.e. $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

(ii) Their common 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) The null and alternative hypothesis always framed in terms of the parameters $\boldsymbol{\theta}{\in}\Theta{\subset}\mathbb{R}^m$ of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, and for every null hypothesis there is a default alternative with which they form a *partition* of the parameter space Θ:

$$H_0\text{: } \boldsymbol{\theta}{\in}\Theta_0 \text{ vs. } H_1\text{: } \boldsymbol{\theta}{\in}\Theta_1. \tag{7}$$

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

$$\mathbb{R}^n_X{=}\left\{\begin{array}{|c|}\hline C_0 \\ \hline C_1 \\ \hline \end{array}\leftrightarrow\begin{array}{|c|}\hline \Theta_0 \\ \hline \Theta_1 \\ \hline \end{array}\right\}{=}\Theta$$

That is, 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 value of $\boldsymbol{\theta}$ in Θ could be $\boldsymbol{\theta}^*$.

(iv) The underlying reasoning is *hypothetical* in the sense that 'H_0 is true or false', represent *hypothetical scenarios* under which the sampling distribution of the test statistic $d(\mathbf{X})$ is evaluated.

(v) In light of (iii)-(iv), frequentist testing relies on error probabilities (type I, II and the p-value) that are (a) firmly attached to the testing procedure, and (b) relate to inferential claims about $\boldsymbol{\theta}$. That is, 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:

$$H_0\text{: } \mu=\mu_0 \text{ vs. } H_1\text{: } \mu\neq\mu_0, \tag{8}$$

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

$$\tau(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s},\ C_1(\alpha)=\{\mathbf{x}\text{: } |\tau(\mathbf{x})|>c_\alpha\},\ s^2=\tfrac{1}{n-1}\textstyle\sum_{k=1}^n(X_k-\overline{X}_n)^2, \tag{9}$$

where c_α can be evaluated using the Student's t distribution. To evaluate the type I and II error probabilities the distribution of $\tau(\mathbf{X})$ under both H_0 and H_1 are needed:

$$\begin{aligned}&\text{[a] } \tau(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s} \overset{\mu=\mu_0}{\backsim} \mathsf{St}(n-1),\\ &\text{[b] } \tau(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s} \overset{\mu=\mu_1}{\backsim} \mathsf{St}(\delta_1;n-1),\ \text{for all } |\mu_1|>\mu_0,\end{aligned} \tag{10}$$

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

$$\begin{aligned}&\alpha=\mathbb{P}(|\tau(\mathbf{X})|>c_\alpha;\ \mu=\mu_0),\\ &p(\mu_1)=\mathbb{P}(|\tau(\mathbf{X})|>c_\alpha;\ \mu=\mu_1),\ \text{for all } |\mu_1|>\mu_0.\end{aligned} \tag{11}$$

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 (10) irrelevant, and calls into question the concept of a *two-sided* p-value $p(\mathbf{x}_0)$.

Example 2 (continued). For (8) and test $T_\alpha:=\{\tau(\mathbf{X}), C_1(\alpha)\}$, the p-value is:

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

As a measure of discordance, the p-value is the probability of all outcomes $\mathbf{x}\in\mathbb{R}_X^n$ that are more discordant with H_0 than $\mathbf{x}_0$ is, when H_0 is true. The p-value is a legitimate post-data error probability because it denotes the smallest threshold (significance level) at which H_0 would have been rejected.

Confidence Intervals (CIs). The sampling distribution in (10)[a] should contrasted with the analogous pivotal (distributional) result:

$$\tau(\mathbf{X};\mu)=\tfrac{\sqrt{n}(\overline{X}_n-\mu)}{s}\overset{\mu=\mu^*}{\backsim}\mathsf{St}(n-1), \tag{12}$$

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

$$\mathbb{P}(\overline{X}_n-\tfrac{s}{\sqrt{n}}c_{\frac{\alpha}{2}})\leq\mu<\overline{X}_n+\tfrac{s}{\sqrt{n}}c_{\frac{\alpha}{2}};\ \mu=\mu^*)=(1-\alpha)\,, \tag{13}$$

where $(1-\alpha)$ denotes the *coverage probability*. The key difference between (10)[a] and (12) is that the coverage probability is evaluated using *factual reasoning*, under μ^*, the true μ. Note that the coverage error probability α is firmly attached to $\overline{X}_n$, a function of the sample $\mathbf{X}$, and the relevant inferential claim is that the random interval $[\overline{X}_n+\frac{s}{\sqrt{n}}c_{\frac{\alpha}{2}}),\ \overline{X}_n-\frac{s}{\sqrt{n}}c_{\frac{\alpha}{2}}]$ covers μ^* with probability $(1-\alpha)$.

Poor implementation. When applying the above t-test one should be aware that α and the power function $p(\mu_1)$, for $|\mu_1|>\mu_0$, calibrate the generic capacity of this test in detecting discrepancies from H_0. Pre-data the practitioner should be mindful of the potential errors that could undermine the reliability of the ensuring inferences. For instance, testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ assumes its statistical adequacy, and so does the CI. In practice, this needs to be established using trenchant Mis-Specification (M-S) testing to probe the NIID assumptions in (6) beforehand; see section 7. A precondition for effective M-S testing is to ensure that the sample size n is large enough for trenchant probing. The rule of thumb is that if n is not large enough to test effectively the validity of the model assumptions, it is not large enough for inference. For simple (IID) statistical models $n\geq 40$ should be large enough, but for models that involve complicated forms of dependence and heterogeneity $n\geq 100$ might be necessary. Having established the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, n should also be evaluated in relation to the substantive questions of interest to ensure that the test in question has sufficient power to detect discrepancies of interest, when present.

3.3 Revisiting replicability: the ideal case

To shed some light on what replicability amounts to in practice, let us consider an ideal scenario where the data are generated by simulating a Linear Regression (LR) model (table 7) whose statistical Generating Mechanism (GM) is completely known:

$$Y_t=\beta_0+\beta_1x_t+\sigma\epsilon_t,\ \epsilon_t\backsim\mathsf{N}(0,1),\ t=1,2,...,n, \tag{14}$$

where $(\beta_0=1.5,\ \beta_1=.5,\ \sigma^2=.755)$, $\epsilon_k\backsim\mathsf{N}(0,1)$ denotes *pseudo-random* numbers from $\mathsf{N}(0,1)$. Using the LR assumptions [1]-[5] in table 7, one can derive the optimal estimators:

$$\widehat{\beta}_0=\overline{Y}-\widehat{\beta}_1\overline{x},\ \widehat{\beta}_1=\tfrac{\sum_{t=1}^n(Y_t-\overline{Y})(x_t-\overline{x})}{\sum_{t=1}^n(x_t-\overline{x})^2},\ s^2=\tfrac{1}{n-2}\sum_{t=1}^n\widehat{u}_t^2, \tag{15}$$

where $\overline{Y}=\frac{1}{n}\sum_{t=1}^{n}Y_t$, $\overline{x}=\frac{1}{n}\sum_{t=1}^{n}x_t$, and the residuals $\{\widehat{u}_t=(Y_t-\widehat{\beta}_0-\widehat{\beta}_1x_t),\ t=1,2,...,n\}$. The sampling distributions of $(\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\sigma}^2_{ML})$, as well as the t-ratios and the R^2, are:

$$\widehat{\beta}_0\backsim\mathsf{N}(\beta_0,\sigma^2\left(\tfrac{1}{n}+\varphi_x\overline{x}^2\right)),\ \varphi_x=(\tfrac{1}{\sum_{t=1}^{n}(x_t-\overline{x})^2}),\ \widehat{\beta}_1\backsim\mathsf{N}(\beta_1,\sigma^2\varphi_x),\ \tfrac{(n-2)s^2}{\sigma^2}\backsim\chi^2(n-2),$$
$$\tau_0(\mathbf{z})=\tfrac{\widehat{\beta}_0-1.5}{\sqrt{s^2\left(\frac{1}{n}+\varphi_x\overline{x}^2\right)}}\overset{H_0}{\backsim}\mathsf{St}\left(n-2\right),\ \tau_1(\mathbf{z})=\tfrac{\widehat{\beta}_1-.5}{\sqrt{s^2\varphi_x}}\overset{H_0}{\backsim}\mathsf{St}\left(n-2\right),\ R^2=1-\tfrac{\sum_{t=1}^{n}\widehat{u}_t^2}{\sum_{t=1}^{n}(Y_t-\overline{Y})^2}. \tag{16}$$

The *ideal scenario* for replicating frequentist inference results is to select a typical sample size, say n=100, and simulate a large number, say N=50000, of such data (sample realizations) $\mathbf{z}^{(k)}\mathbf{:=}(\mathbf{z}_1,...,\mathbf{z}_n)^\top$, $\mathbf{z}_t{:=}(x_t,y_t)$ $k=1,2,...,N$, using the true statistical GM in (14). For each replication one can evaluate all the statistics in (16) and create empirical distributions that approximate their sampling distributions. The empirical distributions of $(\widehat{\beta}_0,\widehat{\beta}_1,s^2)$ will approximate the distributions in (16) but the particular values of these estimates will vary widely, as their standard errors attest. For approximating the true values of the parameters (β_0=1.5, β_1=.5, σ^2=.755) one needs to use the arithmetic mean of the empirical distributions; see Spanos and McGuirk (2001). Similarly, the empirical distributions of $(\tau_0(\mathbf{z}),\tau_1(\mathbf{z}))$ can be used to approximate both the empirical type I error as well as the power of the test for different discrepancies from any null values; using the discrepancies $(\widehat{\beta}_0-1.5)$ and $(\widehat{\beta}_1-.5)$ one can evaluate the empirical type I errors, which are close to α.

A key feature of replicability stemming from the such ideal scenario simulations is that for any particular sample realization $\mathbf{z}^{(k)}$, the point estimates and the observed test statistics take different values over a certain range, and the accuracy noted above is exhibited by the *sample mean* of the realized values over the N (large enough) different realizations. Hence, even in the ideal case where: (i) the statistical model is adequate (its assumptions are valid), (ii) the same sample size n, and (iii) the same generating mechanism, the inferential results, including the parameter estimates as well as the observed test statistics of the same hypothesis are *not* going to be *identical*, because they represent different realizations from the same sampling distributions. On the other hand, the empirical relevant pre-data error probabilities are likely to approximate closely the nominal ones because they are based on relative frequencies that involve averaging.

The p-value. In light of the fact that the p-value is a post-data error probability, implies that the value $p(\mathbf{z}_0)$ does not lend itself to the same replicability set up because $d(\mathbf{x}_0)$ depends crucially on the particular set of estimates, which change for each n data set $\mathbf{Z}^{(k)}\mathbf{:=}(\mathbf{z}_1,...,\mathbf{z}_n)^\top$, $k=1,2,...,N$. The closest to replicability for the p-value $p(\mathbf{z}_0)$ arises when viewed as a statistic $p(\mathbf{Z})$ for different realizations of $\mathbf{Z}$. It is well known that $p(\mathbf{Z})\overset{H_0}{\backsim}\mathsf{U}(0,1)$, i.e. it is Uniformly distributed over the range [0,1] (Cox and Hinkley, 1974). Assuming that $p(\mathbf{z}^{(k)})$ is evaluated for each data set of size n, then it is expected that the empirical distribution of $p(\mathbf{Z})$ will be close to uniform. This suggests that $p(\mathbf{z}_0)$ is not replicable in practice, and at best only the threshold is replicable (significance level) α, i.e. $\mathbb{P}(p(\mathbf{Z})\leq\alpha;H_0)\leq\alpha$. Hence, even under the

ideal conditions (i)-(iii) above, one should not expect the statistical significant results to coincide in different studies. When one allows for much smaller replications N, different sample sizes n and unknown statistical adequacy of the underlying statistical model, replicability in the sense described above become a much more elusive target.

4 Frequentist testing: averting confusions

4.1 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 abuse of significance testing as the main culprit. His argument revolves around a posterior probability measure, the Positive Predictive Value (PPV), adapted from medical screening that aims to evaluate the reliability of medical diagnostic tests for detecting a disease in patients that revolves around the notions of 'false positive' and 'false negative' for the screening devices; see Fletcher and Fletcher (2005). The PPV is defined in terms of H_0: no disease, F=H_0 is false, R=test rejects H_0, and takes the Bayesian probability formulation:

$$\text{PPV}=\Pr(F|R)=\frac{\text{number of true positive detections}}{\text{number of positive detections}}=\frac{\Pr(R|F)\Pr(F)}{\Pr(R|F)P(F)+\Pr(R|\overline{F})P(\overline{F})}, \quad (17)$$

where $\Pr(R|F)$ and $\Pr(\overline{R}|\overline{F})$ are referred to as 'sensitivity' and 'specificity', respectively. Sensitivity aims to measure the proportion of correct rejections of H_0 (when false), specificity aims to measure the proportion correct acceptances of H_0 when true. Both depend crucially on 'prevalence' $\Pr(F)$ that aims to measure the proportion of false H_0 in a certain population. To make sense of the PPV in hypothesis testing, as opposed to medical screening devices, one needs to imagine that there is a population of null hypotheses for a particular discipline, a proportion of which are false, say 20%! .Unfortunately, the 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 highly misleading and fundamentally 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. In the context of frequentist testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, '$H_0$ is true or false' represent *hypothetical scenarios* under which the sampling distribution of the test statistic $(d(\mathbf{X}))$ is evaluated. Hence, the claim: "... 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. Moreover, pretending that it is a matter of notational ignores the fact that assigning probabilities to θ via 'H_0 is true or false' is illegitimate in the context of frequentist testing.

Second, in frequentist testing there is no such thing as discipline wide false positive/negative proportions that revolve around generic tests and generic null hypothe-

ses analogous to medical screening devices. One can assert that the false positive of this screening device is 7% and the prevalence of this illness for this population is 20%, but the analogical reasoning used to transfer such notions to frequentist testing is completely misplaced for several reasons. Frequentist testing is *local* in the sense that it depend crucially on the particular $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, the relevant test $(d(\mathbf{X}),\ C_1)$ and the particular data $(\mathbf{x}_0)$, 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 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 a value in Θ_1=Θ−Θ_0, see (11). It calibrates the capacity of the test to detect different discrepancies in Θ_1.

Third, the bias inducing abuses of the p-value, finger pointed as the main culprit, take place at the level of an individual study, based on a particular statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. In contrast, the PPV postulates implicitly an imaginary meta-model of discipline-wide null hypotheses and decisions, and assigns a posterior measure of *untrustworthiness by association* to the overall performance of diagnostic screening in that field that revolves around $\Pr(F)$, the proportion of false null hypotheses; a meaningless notion in the context of frequentist testing. Unfortunately, this meta-model is not just imaginary, it defines the inductive premises of inference in a way that renders its appropriateness a matter of speculation.

Finally, ensuring that every practitioner in a particular discipline refrains from any form abuse (p-hacking, multiple testing, cherry-picking, etc.) or misinterpretation of p-values, will be a good starting point, but nowhere near enough to guarantee that the end result will be trustworthy empirical evidence. This becomes obvious when the main sources of untrustrworthiness are recalled: (i) statistical misspecification, (ii) poor implementation of inference procedures, and (iii) unwarranted evidential interpretations of their inferential results.

4.2 The large n problem

The large n problem is that as n increases $p(\mathbf{x}_0)$ decreases, and thus there is always a large enough n to reject H_0 however small the adopted threshold c>0, i.e., when $(\theta^*-\theta_0)\neq 0$, $p(\mathbf{x}_0)\underset{n\to\infty}{\to}0$. Hence, 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 practice gives rise to 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 (e.g. large n). 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 (e.g. small n).

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 (10) 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. 3a-b. 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.

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.

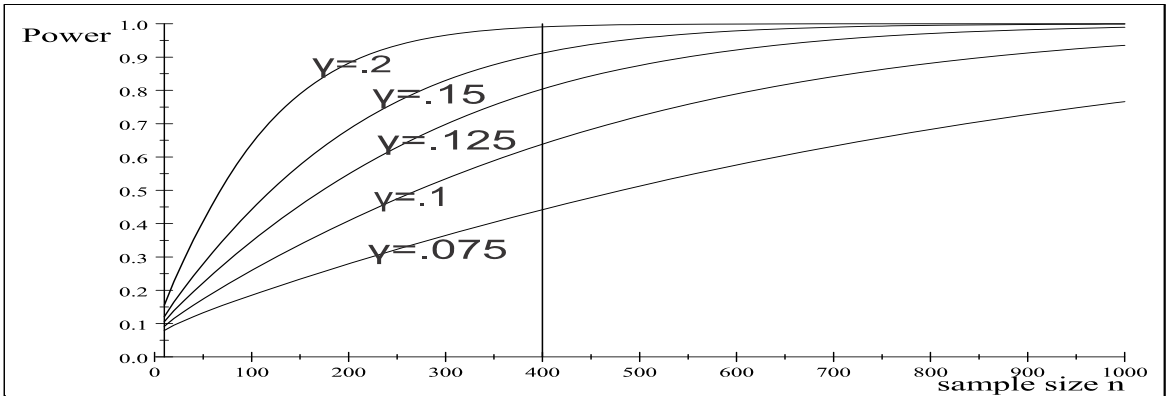

Fig. 3a: Power for different n and discrepancies $\gamma\sigma$ from the null

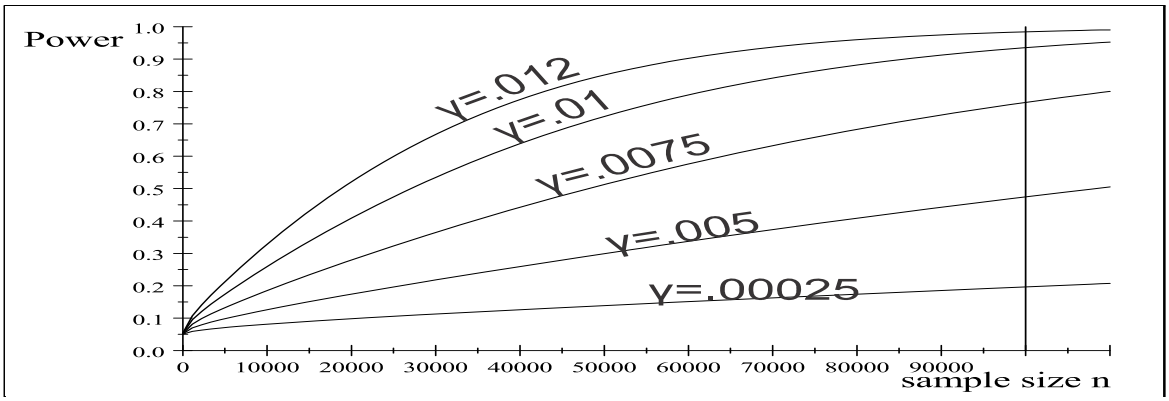

Fig. 3b: Power for different n and discrepancies $\gamma\sigma$ from the null

To counter the decrease in the p-value as n increases, some textbooks advise practitioners to use rules of thumb based on decreasing α as n increases; see Lehmann and Romano (2005). Good (1988) suggests standardizing the p-value $p(\mathbf{x}_0)$ to n=100

using the formula: $p_{100}(\mathbf{x}_0)=\min\left(.5,\left[p(\mathbf{x}_0)\cdot\sqrt{n/100}\right]\right),\ n>40$; see table 2.

Table 2: 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

4.3 Redefining statistical significance

More recently, there has been a call by a group of statisticians in several fields to replace the traditional .05 threshold with a smaller one, .005, as a way to improve the reproducibility of empirical results; see Benjamin et al. (2018). The justification for this proposal is based on three arguments. The first stems from a comparison between the two-sided p-value of .05 and .005 and the corresponding Bayes factor using a particular H_1. This comparison suggests that a two-sided p-value of .05 (.005) is evidentially weak (strong) when H_0 and H_1 are evaluated using the Bayes factor. The second justification stems from selecting particular values for the conditional probabilities associated with the PPV in (17), to show that a lower $\Pr(R|\overline{F})$ would increase the PPV. The third justification is that empirical evidence from large scale replications indicate that studies with $p(\mathbf{x}_0)<.005$ are more likely to replicate than those based on $p(\mathbf{x}_0)<.05$.

When viewing the call by reformers to replacing the .05 threshold with .005 in the context of the above discussion, it becomes clear that it has no substantive merit for several reasons. First, the question that naturally arises is 'what does the Bayesian perspective on evidence have to do with p-values?' Bayesian evidence stems from comparing the ration of posterior probabilities $\pi(\theta|\mathbf{x}_0)$ associated with H_0: $\theta\in\Theta_0$ and H_1: $\theta\in\Theta_1$, assigned to all $\theta\in\Theta_i,\ i=0,1$. That has no bearing on the p-value as a measure of discordance – evaluated under H_0 – of all outcomes $\mathbf{x}\in\mathbb{R}^n_X$ that are more discordant with H_0 than $d(\mathbf{x}_0)$ is. Indeed, the most agregious part of this argument is the misinterpretation of the Bayes factors or the p-values as measures of the 'strength of evidence' for or against H_0 and H_1; see Spanos (2013). Leaving that aside, any legitimate frequentist evidence for or against a hypothesis or an inferential claim should stem from the probativeness (capacity) of the testing procedure, as calibrated by α and the power function, with the relevant probabilities defined via $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$; see (32). Hence, the comparison of the tail area of the sampling distribution of $d(\mathbf{X})$ for $\mathbf{x}\in\mathbb{R}^n_X$ beyond $d(\mathbf{x}_0)$ (p-value), with the ratio of the ratio of posterior probabilities $\pi(\theta|\mathbf{x}_0)$ assigned to $\theta\in\Theta_i,\ i=0,1$, seems an ill-conceived!

Second, the threshold $p(\mathbf{x}_0)<\alpha$ was never meant to be either arbitrary or fixed for all significance testing; see Fisher (1925; 1935). Indeed, the N-P framing of frequentist testing showed that a proper interpretation of the results needs to take into account the tradeoff between the type I and II error probabilities. Hence, the call to replace one arbitrary threshold (.05) with a much smaller (.005) misses the quintessence of this tradeoff and its dependence on the sample size n; see figures 3a-b. In an attempt to compensate for the loss of power when replacing .05 with .005, Benjamin et al.

(2018) call for increasing n to ensure high power of .8 for a particular, but often arbitrary point alternative H_1. This runs afoul an essential element of N-P testing of viewing the framing of H_0 and H_1 as a partition of the parameter space, and not as two cherry-picked values. Also, a two-sided p-value ignores its post-data nature.

The argument based on the PPV has already been questioned on the basis that the latter is framed in terms of conditional probabilities whose semblance to the proper error probabilities alluded to by this argument is more apparent than real.

Third, the claim that the threshold $p(\mathbf{x}_0) < .005$ will result in better replicability of empirical evidence, is seriously flawed because any decrease (not just .005) in the threshold is likely to improve replicability, but only when all the other sources of untrustrworthiness are ignored, including (i) statistical misspecification, (ii) poor implementation of inference procedures, and (iii) unwarranted evidential interpretations of their inferential results.

4.4 Repacing the p-value with observed CIs

It is often claimed by the reformers that the p-value should be replaced by the analogous observed CI because the latter is: (i) less vulnerable to the large n problem and (ii) more informative than the p-value since it provides a measure of the 'effect size'; Cohen (1994) recommendation to practitioners: "routinely report effect sizes in the form of confidence intervals" (p. 1002). A closer look at these claims reveals that they are misleading.

Claim (i) is false because a CI is equally vulnerable to the large n problem as the p-value since the expected length of a consistent CI shrinks to zero as $n \to \infty$. In the case of (13): $E\left([\overline{X}_n+c_{\frac{\alpha}{2}}(\frac{s}{\sqrt{n}})]-[\overline{X}_n-c_{\frac{\alpha}{2}}(\frac{s}{\sqrt{n}})]\right)=2c_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}}) \underset{n\to\infty}{\to} 0.$

The fact that n plays a crucial role in defining the length of an observed CI calls into question claim (ii) that it provides a reliable measure of the 'effect size'.

Further light can be shed on both claims, using the mapping between the p-value and the corresponding observed CI stemming from placing the null value μ_0 on its boundary. For the CI (13), using the relationship $c_\alpha=F^{-1}(1-\alpha)$, F-cumulative distribution function (cdf) of the Student's t:

$$\mu_0=\overline{x}_n\pm c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}}) \to c_\alpha=|\tfrac{\sqrt{n}(\overline{x}_n-\mu_0)}{s}| \to \alpha(\mu_0)=\mathbb{P}(|\tau(\mathbf{X})|>|\tau(\mathbf{x}_0)|;\ \mu=\mu_0), \tag{18}$$

where $\alpha(\mu_0)$ denotes the smallest significance level at which H_0: $\mu=\mu_0$ is rejected. Indeed, the mapping in (18) relates to another issue with observed CIs that pertain to several moves to assign probabilities that render different points within such an interval more or less likely. This move is fallacious because *post-data* the coverage probability $(1-\alpha)$, evaluated under $\mu=\mu^*$, makes no statistical sense; the observed CI either contains μ^* or it doesn't. This stems from the fact that there is no post-data coverage probability since the factual event $\mu=\mu^*$ has played out. The apparent assignment of probabilities to different values of μ within the observed CI is attempted by extending (18) to all values $\mu_1\neq\mu_0$ and holding $\overline{x}_n$ constant:

$$\mu_1=\overline{x}_n\pm c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}}) \to \alpha(\mu_1)=\mathbb{P}(|\tau(\mathbf{X})|>|\tau(\mathbf{x}_0)|;\ \mu=\mu_1), \tag{19}$$

This family of curves has been rediscovered by several statisticians. Initially by Birnbaum (1961) who called it *an omnibus confidence curve*, later by Kempthorne and Folks (1971) naming it *a consonance interval curve*, and more recently by Poole (1987), calling it a *p-value curve*. What is often insufficiently appreciated is that (19) has nothing to do with coverage probability since its evaluation stems from a post-data hypothetical (testing) reasoning, under μ=μ_1. Moreover, placing such curves over observed CIs gives the misleading impression of assigning probabilities to the different values of μ, with $\overline{x}_n$ assigned the highest value. For instance, Altman, et. al (2000) argue: "... the difference between population means is much more likely to be near the middle of the confidence interval than towards the extremes." (p. 22).

Equally misplaced is the move to place the likelihood function $L(\mu,\sigma^2;\mathbf{x}_0)$ over the observed CI (Cumming, 2012) in a misguided attempt to justify assigning likelihoods to different value of μ, since $L(\mu,\sigma^2;\mathbf{x}_0)$ has nothing to do with coverage probability; the relevant error probability associated with CIs. One might as well ignore the observed CI altogether and just use $L(\mu,\sigma^2;\mathbf{x}_0)$ to infer the relative likelihood of different values of μ. This is a bad idea since any inference based on $L(\mu,\sigma^2;\mathbf{x}_0)$ is plagued with a serious flaw: the Maximum Likelihood estimate, $\widehat{\mu}_{MLE}(\mathbf{x}_0)$, is always the *maximally likely value,* irrespective of values designated by H_0 and H_1 or any substantive information. This distorts any likelihood-based comparisons of relative likelihood, including Bayes factors; see Spanos (2013).

5 Evidential interpretation of accept/reject results

The severity evaluation provides a more formal way to account for the sample size in framing an evidential interpretation for the accept/reject results. 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 (power) of the particular test in detecting discrepancies from H_0 in the direction of H_1. 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).

5.1 Post-data Severity evaluation

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 and the accept/reject results can be adequately addressed using the post-data severity assessment to provide an evidential account for frequentist testing. The severity evaluation 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.

Example 3. Consider testing the hypotheses:

$$H_0\text{: } \theta \leq \theta_0 \text{ vs. } H_1\text{: } \theta > \theta_0, \text{ where } \theta_0\text{=}.5, \tag{20}$$

in the context of the simple Bernoulli model with θ=$\mathbb{P}(X$=1):

$$X_t \backsim \mathsf{BerIID}(\theta, \theta(1-\theta)),\ \theta\in[0,1],\ t=1,2,...,n,..., \tag{21a}$$

using data on n=10514 newborns during 1993 in Cyprus, 5442 boys (X=1) and 5072 girls (X=0). The relevant test statistic is $d(\mathbf{X})=\sqrt{n}(\widehat{\theta}_n-\theta_0)/\sqrt{\theta_0(1-\theta_0)}$ with $\widehat{\theta}_n=\frac{1}{n}\sum_{i=1}^n x_i=\overline{x}_n$=.518. Using significance level α=.01 $\Rightarrow c_\alpha$=2.326:

$$d(\mathbf{x}_0)=\left(\sqrt{10514}(\tfrac{5442}{10514}-\tfrac{1}{2})/\sqrt{.5(.5)}\right)=3.608[.00015], \tag{22}$$

with the p-value, in square brackets, indicating *reject* H_0. Since $d(\mathbf{x}_0)$=3.608>0, the relevant inferential claim is (Spanos, 2013):

$$\theta > \theta_1=\theta_0+\gamma, \quad \text{for some } \gamma \geq 0. \tag{23}$$

To establish the particular discrepancy γ warranted by data $\mathbf{x}_0$, the evaluation is:

$$SEV(T_\alpha;\theta>\theta_1)=\mathbb{P}(d(\mathbf{X}) \leq d(\mathbf{x}_0);\ \theta=\theta_1),\ \text{for } \theta_1=\theta_0+\gamma,\ \gamma\geq 0, \tag{24}$$

$$\begin{aligned}&\left([d(\mathbf{X})-\delta(\theta_1)]\,/\sqrt{V(\theta_1)}\right)\overset{\theta=\theta_1}{\backsim}\simeq \mathsf{N}(0,1),\ \text{for } \theta_1>\theta_0,\\ &\delta(\theta_1)=\left[\sqrt{n}(\theta_1-\theta_0)/\sqrt{\theta_1(1-\theta_1)}\right]\geq 0,\ V(\theta_1)=\tfrac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)},\ 0<V(\theta_1)\leq 1.\end{aligned} \tag{25}$$

The objective of SEV($T_\alpha;\theta>\theta_1$) is to determine the *largest discrepancy* $\gamma\geq$0 warranted by data $\mathbf{x}_0$ for a prespecified level of severity, say SEV($T_\alpha;\theta>\theta_1$)$\geq$.85.

Table 3: Severity Evaluation of 'Reject H_0' with $(T_\alpha;\mathbf{x}_0)$

γ	.01	.012	.0125	.013	.014	.015	.0176	.02	.025	.03
θ_1=.5+γ	.51	.512	.5125	.513	.514	.515	.5176	.52	.525	.53
$\mathsf{Sev}(\theta>\theta_1)$	.940	.874	.852	.827	.782	.703	.500	.311	.064	.005

Table 3 evaluates (24) for different discrepancies, indicating that the largest discrepancy from the null warranted by $\mathbf{x}_0$ is:

$$\gamma^* \leq .01254,\ \text{since } SEV(T_\alpha;\theta>.51254)=.85. \tag{26}$$

That is, the testing-based *effect size* with data $\mathbf{x}_0$ and severity .85 is $\gamma^* \leq .01254$.

5.2 Warranted discrepancy vs. estimation-based effect sizes

The warranted discrepancy γ^* should be contrasted with *estimation-based* effect size estimates whose primary aim is to get a more appropriate measure of the 'magnitude of the scientific effect'. Although there is no agreement in the literature about the most appropriate effect size estimate, Cohen's (1988) g is: g=(.5176−.5)=.0176, which is much larger than γ^*=.01254 and $SEV(T_\alpha;\theta > .5176)$=.50.
This reveals the arbitrariness of grading such effects sizes as small, medium and large. Using Cohen's benchmarks, .0176 is tiny, since the benchmark for small is g_s=.05, despite the fact that $\gamma^*\leq$.01254 implies substantive significance.

As a general rule, it is never a good idea to view the point estimate, say $\widehat{\theta}(\mathbf{x}_0)=\overline{x}_n$, as (approximately) coinciding with θ^*, since it represents just a single value from the sampling distribution of $\widehat{\theta}(\mathbf{X})$. That is why point estimation does *not* output an inferential claim that $\overline{x}_n$ approximates θ^* sufficiently close. The optimality of an estimator $\widehat{\theta}(\mathbf{X})$ is calibrated in terms of its sampling distribution; unbiasedness, consistency, efficiency, etc. To get a reliable value for θ^* one needs to use a large number of replicated data and evaluate the empirical mean of the sampling distribution of $\widehat{\theta}(\mathbf{X})$; see example 4 section 6.2. This calls into question the appropriateness of any estimation-based effect sizes, more generally, since they are based on a single realization.

Statistical vs. substantive significance. The warranted discrepancy γ^* from the null H_0 is as far as the statistical information in data $\mathbf{x}_0$ can take an inferential claim; γ^* cannot be proclaimed substantively significant. Whether γ^* is *substantively significant* can only be established using substantive information in conjunction with the inferential claim: $\theta > \theta_1$=θ_0+γ, $\gamma^* \leq$ .01254. In human biology the substantively determined sex ratio at birth is: $\theta^\star\simeq$.5122; Hardy (2002).
In light of that, the above warranted discrepancy $\gamma^* \leq$ .01254, is also *substantively significant* since $\theta^\star\simeq.5122 < \theta_0+\gamma^* \leq .51254$. This raises the broader problem of relating the statistical and substantive information in coherent modeling framework; see section 6.

5.3 Addressing questionable practices in N-P testing

Example 3 (continued). [i] Consider switching the 1-sided hypotheses in (20) with:

$$H_0\text{: } \theta=\theta_0 \text{ vs. } H_1\text{: } \theta\neq\theta_0, \text{ where } \theta_0=.5, \tag{27}$$

which changes the threshold to $c_{\frac{\alpha}{2}}$=2.575. This does not effect the p-value or the severity evaluation, since $d(\mathbf{x}_0)$=3.608[.00015]>0 renders the left tail irrelevant.

[ii] What happens if one were to replace the hypotheses in (20) with the simple-vs-simple hypotheses:

$$H_0\text{: } \theta=\tfrac{1}{2} \text{ vs. } H_1\text{: } \theta=\tfrac{18}{35}. \tag{28}$$

Given that H_0 and H_1 do not constitute a partition of Θ:=$[0,1]$, the framing in (28) is *improper*; see Spanos (2013).

Having said that, if one were interested in the discrepancy $(\frac{18}{35}-\frac{1}{2})=\frac{1}{70}$=.014286, then its post-data severity could be evaluated using table 3. The inferential claim $\gamma\leq\frac{1}{70}$ has severity: $SEV(T_\alpha;\theta>(.5+\frac{1}{70})$=.5143)=.75.

[iii] What if one were to replace the hypotheses in (20) with:

$$H_0\text{: } \theta \leq \theta_0 \text{ vs. } H_1\text{: } \theta > \theta_0, \text{ where } \theta_0{=}\tfrac{18}{35}, \tag{29}$$

with α=.01 $\Rightarrow$ c_α=2.326? Since the test statistic yields:

$$d_B(\mathbf{x}_0){=}\left[\sqrt{10514}(\tfrac{5442}{10514}{-}\tfrac{18}{35})/\sqrt{.5(.5)}\right]{=}.679[.249], \tag{30}$$

the null will now be accepted, but the post-data severity evaluation will not change because $d_B(\mathbf{x}_0)$=.679>0 indicates that θ_1 will remain the same; γ will change in table 3 since the relevant inferential claim is now $\theta{>}\theta_1{=}(\tfrac{18}{35}{+}\gamma)$.

Table 3A: Severity Evaluation of 'Accept H_0' with $(T_\alpha;\mathbf{x}_0)$

γ	-.0043	-.002	-.0018	-.013	-.0003	.0007	.003	.006	.011	.016
$\theta_1{=}\frac{18}{35}{+}\gamma$	.51	.512	.5125	.513	.514	.515	.5176	.52	.525	.53
$\mathsf{Sev}(\theta{>}\theta_1)$	.940	.874	.852	.827	.782	.703	.500	.311	.064	.005

[iv] What if one were to change α=.01 in case [iii] to α=.25 $\Rightarrow$ c_α=.674? The null $\theta_0{=}\frac{18}{35}$ will now be rejected, but the severity evaluation remains the same since $d_B(\mathbf{x}_0)$>0.

In summary, the severity evaluation remains invariant to the reframing of H_0 and H_1, and any changes in α, as long as the framing constitutes a partition of Θ.

5.4 Severity vs. observed Confidence Intervals (CIs)

The post-data severity evaluation can be used to address the issue of degenerate post-data coverage error probability, resulting the impossibility to distinguish between different values of μ within an observed CI; Mayo and Spanos (2006).

This is achieved by replacing:

(i) the *factual* reasoning underlying CIs with the *hypothetical* reasoning, and

(ii) the inferential claims of overlaying the true θ^* with *post-data* severity-based inferential claims.

In the case of example 3, this takes the form of placing θ_1 on the relevant boundary (in light of $d(\mathbf{x}_0)$>0) of the observed CI as in (19):

$$\theta > \theta_1{=}\overline{x}_n{-}c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}}) \text{ with } SEV(T_\alpha;\theta{>}\theta_1){=}\mathbb{P}(d(\mathbf{X}) \leq d(\mathbf{x}_0);\ \theta{=}\theta_1). \tag{31}$$

A moment's reflection, however, suggests that the changes (i)-(ii) have eliminated any connection between on observed CI and (31). The severity assessment of $\theta > \theta_1$ does not assign probabilities to the observed CI or pertain to θ_1, but to the inferential claim in (31) to establish the warranted discrepancy $\gamma^*{=}\theta_1{-}\theta_0$ in light of $\mathbf{x}_0$. The severity evaluation probability has nothing to do with the coverage probability. Indeed, the equality of the tail areas stems from the mathematical duality, but that does not imply inferential duality. Many points in any observed CI have very low severity; see Mayo and Spanos (2006).

6 Statistical vs. substantive adequacy

6.1 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:

$$F(Y_n \leq y)=\underbrace{\int\int\cdots\int}_{\{\mathbf{x}:\ g(\mathbf{x})\leq y\}} f(\mathbf{x};\boldsymbol{\theta})d\mathbf{x},\ \forall y\in\mathbb{R}. \tag{32}$$

This can derail the optimality of these procedures (e.g. inconsistency) and induce *inconsistency* in estimators and/or 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 all forms of statistical inference, including Bayesian, nonparametric as well as Akaike-type model selection procedures, are vulnerable to statistical misspecifiation; see Spanos (2010c).

6.2 Statistical misspecification and unreliable inferences

Example 4. Spanos and McGuirk (2001) illustrate these effects using a simulation based on a Linear Regression (LR) model (N=10000 replications) under two different scenarios.

Scenario 1: **statistically adequate model**. The true and the estimated model coincide:

$$Y_t=1.5+.5x_t+u_t,\ u_t\backsim\mathsf{N}(0,.75),\ t=1,2,...,n. \tag{33}$$

The key conclusions under scenario 1 are as follows:

(i) The empirical sampling distributions of the estimators are *highly accurate,* with their mean closely approximating the true parameter values:

$$\widehat{E}(\widehat{\beta}_0)=1.502,\ \widehat{E}(\widehat{\beta}_1)=.499,\ \widehat{E}(\widehat{\sigma}^2)=.751,\ \widehat{E}(R^2)=.253,$$

where $\widehat{E}(.)$ denotes the empirical mean of the sampling distribution.

(ii) The actual empirical type I error probabilities associated of the t-tests are very close to the nominal (α=.05), .049 and .047 for a sample size n=50.

(iii) The accuracy of both sets of results in (i)-(ii), improves as n increases from n=50 to n=100.

Scenario 2: **statistically misspecified model**: the estimated model is (33), but the true model is: $Y_t=1.5+\underline{.13t}+.5x_t+u_t,\ u_t\backsim\mathsf{N}(0,.75),\ t=1,2,...,n.$

The key conclusions under scenario 2 are as follows:

(iv) The empirical sampling distributions of the estimators are *highly inaccurate* (symptomatic of *inconsistent* estimators) with their mean very different from the true parameter values: $\widehat{E}(\widehat{\beta}_0)=.462,\ \widehat{E}(\widehat{\beta}_1)=1.959,\ \widehat{E}(\widehat{\sigma}^2)=2.945,\ \widehat{E}(R^2)=.979.$

(v) There is a huge discrepancy between the nominal (α=.05) and actual empirical type I error probabilities, .774 and 1.0, for a sample size n=50.

(vi) The accuracy of both sets of results in (iv)-(v), worsens significantly as n increases from n=50 to n=100.

One might argue that the above simulation results are based on a departure from the LR model that seems rather artificial because it is very easy to detect any trends in data by just *glancing* at their t-plots. As demonstrated in section 6.3, this misspecification can easily arise in more subtle ways that are not at all obvious by just looking at the t-plots of the data.

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).

6.3 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 pertaining to the phenomenon of interest, and could range from simple conjecture to intricate *substantive* (structural) model. Such information has an important and multifaceted role to play, including demarcating the crucial aspects of the phenomenon of interest to be studied, e.g. suggesting the relevant variables and data, etc. In contrast, statistical information stems from the *chance regularities* exhibited by the selected data $\mathbf{x}_0$. 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**'. This slogan was coined by George Box (1979), but it is often taken out context and misinterpreted as rendering statistical misspecification inevitable. The 'wrongness' Box referred to, however, was not statistical but substantive: "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 asserts that substantive models are not exact pictures of reality. 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 *[respecification]* in other needed directions." It is worth noting that diagnostic checking was initially proposed in Box and Jenkins (1970).

In summary, it is one thing to claim that 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:
[a] statistical adequacy: does $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ account for the chance regularities in $\mathbf{x}_0$?
[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 flaws in capturing adequately the phenomenon of interest, such as missing confounding factors, systematic approximation errors, etc. In this sense, probing for substantive adequacy is a considerably more complicated problem, which 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 $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0},\ \boldsymbol{\theta}\in\boldsymbol{\Theta},\ \boldsymbol{\varphi}\in\boldsymbol{\Phi}$.

6.4 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 4).

Table 4: Empirical modeling and Inference

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

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. Imagine trying to construct a sailboat from a pile of wooden planks while sailing it!

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. Hence, a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is selected to meet two interrelated aims. First, to account for the chance regularities in data $\mathbf{x}_0$ by choosing appropriate probabilistic assumptions for $\{X_t,\ t\in\mathbb{N}\}$ with a view to render $\mathbf{x}_0$ a 'typical realization' thereof. Second, to parameterize $(\boldsymbol{\theta}\in\Theta)$ with a view to embed$\pi\pi$ the substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ 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.

7 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.

1. Untestable and implicit probabilistic assumptions. Practitioners rarely have a complete list of testable probabilistic assumptions defining statistical models. For instance, the incompleteness of the traditional specification of the LR model (table 6) becomes apparent when compared with a complete and testable set of probabilistic assumptions in table 6. Assumption [5] and the parameterization in table 6 are only implicit in table 6; see McGuirk and Spanos (2009).

Table 6: 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:	$(u_t\|X_t=x_t)\backsim \mathsf{N}(.,.)$,	$t\in\mathbb{N}$.
{2}	Zero mean:	$E\left(u_t\|X_t=x_t\right)=0$,	
{3}	Homoskedasticity:	$Var\left(u_t\|X_t=x_t\right)=\sigma^2$,	
{4}	Zero correlation:	$\{(u_t\|X_t=x_t),\ t\in\mathbb{N}\}$ is uncorrelated,	

Table 7: Normal, Linear Regression model

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

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

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

2. 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 give the practitioner a chance to appraise the appropriateness of different methods and do a much better job in producing trustworthy evidence by testing the validity of the invoked assumptions; see Spanos (2018a).

Unfortunately, empirical modeling is currently dominated by a serious disconnect between these two since the theoretician is practicing *mathematical deduction* and the practitioner uses recipe-like *statistical induction* by transforming formulae into numbers using the data. The theoretician has no real motivation to render the invoked deductive premises 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; see Fisher (1922), p. 314.

3. Parametric vs. nonparametric inference. A particularly pernicious example of weak inductive premises is the use of nonparametric methods as a way to ensure the reliability of inference. A sizeable percentage of practitioners have the erroneous impression, stemming from the misleading claims by advocates, that nonparametric inference imposes no probabilistic assumptions on one's data. That could not be further from the truth! All statistical inference methods rely on three types of probabilistic assumptions: distribution, dependence and heterogeneity. Parametric inference invokes explicit assumptions from all three categories. Nonparametric inference invokes explicit assumptions from the last two categories, and only indirect distribution assumptions in the form of (i) the existence of certain moments up to order $p \geq 1$, and (ii) smoothness restrictions on the unknown density function $f(z)$, $z \in \mathbb{R}_Z$ (continuity, symmetry, differentiability, unimodality, bounded and continuous derivatives of $f(z)$ up to order $m > 1$); see Wasserman (2006). Hence, the only real difference between parametric and nonparametric inference is that the latter replaces easily testable distribution assumptions with untestable indirect distributional assumptions. Worse, nonparametric inference usually relies on *asymptotic sampling distributions*, but, 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). Hence, the trustworthiness of nonparametric inference results depends only on

the validity of the model assumptions, such as [1]-[5] (table 7) for the particular $\mathbf{z}_0$ and nothing else; see Spanos (2018a).

4. 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. The most pernicious 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 LR model (table 6), there are no robustness results for generic departures of the form:

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

The claimed robustness results in the econometric literature are often asymptotic (as $n\to\infty$) and rely on *particularized* forms of departures (section 6.5), without establishing their appropriateness for the particular data; see Greene (2012).

8 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 $\mathbf{x}_0$. Formally, the key differences between N-P and M-S testing are: (i) their objective and (ii) their domain of probing:

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

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

$$\underline{\textbf{Testing outside } \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\text{: probing the validity of its assumptions}} \tag{34}$$

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

8.1 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 (2018a).

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:

$$H_0\text{: all assumptions of } \mathcal{M}_\theta(\mathbf{x}) \text{ are valid for } \mathbf{x}_0, \tag{35}$$

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:

$$\begin{array}{l} H_1\text{: the stated departures from specific assumptions being tested,} \\ \textit{assuming} \text{ the rest of the assumption(s) of } \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}) \text{ hold for data } \mathbf{x}_0. \end{array} \tag{36}$$

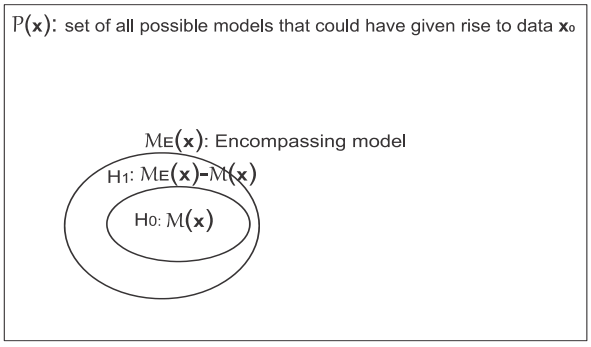

Fig. 4: M-S testing by encompassing

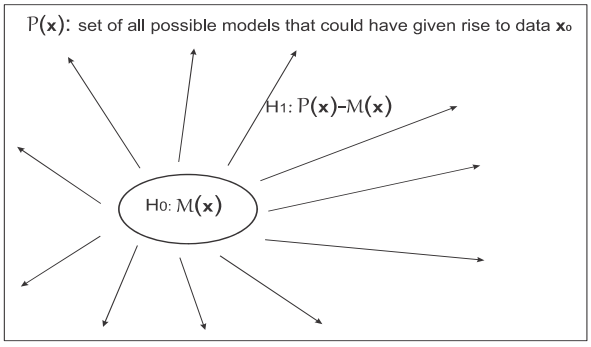

Fig. 5: M-S testing: directions of departures

The particularization of $[\mathcal{P}(\mathbf{x})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ 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. 5); 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. 4) and testing the nesting restrictions; see Spanos (2018a).

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 it provides better evidence for its existence than parametric tests with very high power; see Spanos (2018a).

8.2 Joint M-S testing vs. individual assumption tests

The form of the null and alternative hypotheses for M-S testing in (35)-(36) call for assuming the validity of as fewer assumptions as possible under the alternative (36) to avoid misleading diagnoses. 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*. Indeed, joint M-S testing based on auxiliary regressions has several distinct advantages over other procedures based on individual test statistics; 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 conceal the presence of first order autocorrelation in the residuals by using a Box-Pierce test with p=8 lags. Reporting the estimated auxiliary regressions associated with the joint M-S testing leaves no room for that.

To simplify the discussion, the focus will be on the LR regression model in table 7, 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 7), as they relate to:

$$H_0\text{: } E\left(Y_t \mid X_t{=}x_t\right){=}\beta_0{+}\beta_1 x_t,\ Var\left(Y_t \mid X_t{=}x_t\right){=}\sigma^2. \tag{37}$$

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

$$\widehat{u}_t{=}\delta_0+\delta_1 x_t+\overbrace{\delta_2 t}^{\overline{[5]}}+\overbrace{\delta_3 x_t^2}^{\overline{[2]}}+\overbrace{\delta_4 x_{t-1}+\delta_5 Y_{t-1}}^{\overline{[4]}}+v_{1t}, \tag{38}$$

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 $\delta_5{\neq}0$.

$$\widehat{u}_t^2{=}\gamma_0+\overbrace{\gamma_2 t}^{\overline{[5]}}+\overbrace{\gamma_1 x_t+\gamma_3 x_t^2}^{\overline{[3]}}+\overbrace{\gamma_4 x_{t-1}^2+\gamma_5 Y_{t-1}^2}^{\overline{[4]}}+v_{2t}, \tag{39}$$

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 auxiliary regressions (38)-(39) are probing the residuals $\{(\widehat{u}_t, \widehat{u}_t^2)\ \ t=1,2,...,n\}$ for any lingering systematic information overlooked by the regression and skedastic functions in (37) relating to assumptions [2]-[5].

A more formal justification/derivation of auxiliary regressions such as (38)-(39) can be based on the conditional expectation orthogonality theorem (Williams, 1991):

$$E\left([y{-}E(y|\sigma(\mathbf{X}))]{\cdot}h(\mathbf{X})\right){=}0, \text{ for any Borel-function } h(\mathbf{X}); \tag{40}$$

see Spanos (2018a) for further details and extensions to other statistical models.

It is important to emphasize that the above F-type M-S tests based on (38)-(39) are viewed as testing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ significance tests aiming to detect 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. 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 not a false acceptance. 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 for the task. The rule of thumb for the minimal value of n is: if n is not large enough to test the model assumptions thoroughly, it is not sufficient large for inference purposes.

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 $E(y_t \mid \mathbf{X}_t{=}\mathbf{x}_t){=}h(\mathbf{x}_t),\ \mathbf{x}_t{\in}\mathbb{R}^k_X$, is misspecified, the M-S tests relating to the skedastic function $Var(y_t \mid \mathbf{X}_t{=}\mathbf{x}_t){=}g(\mathbf{x}_t),\ \mathbf{x}_t{\in}\mathbb{R}^k_X$, are likely to give rise to misleading results, since the latter is defined relative to a misspecified $h(\mathbf{x}_t),\ \mathbf{x}_t{\in}\mathbb{R}^k_X$. Moreover, for a reliable test of Normality one should secure the validity of [2]-[5] beforehand using the above auxiliary regressions.

8.3 Example 6. Modeling industrial production

The t-plots in fig. 6-7 depict the level (y_t) and changes ($\Delta y_t{=}y_t{-}y_{t-1}$) of annual data for the US industrial production for the period 1962-2007.

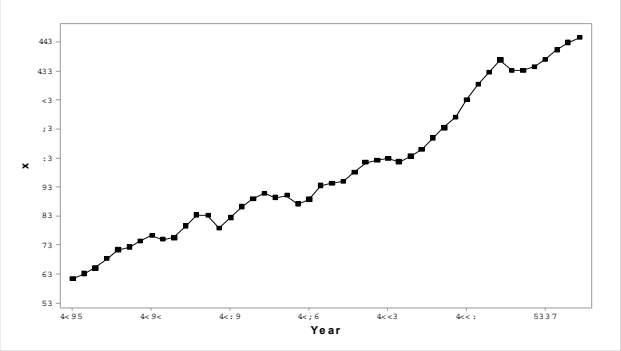

Fig. 6: Industrial production 1962-2007

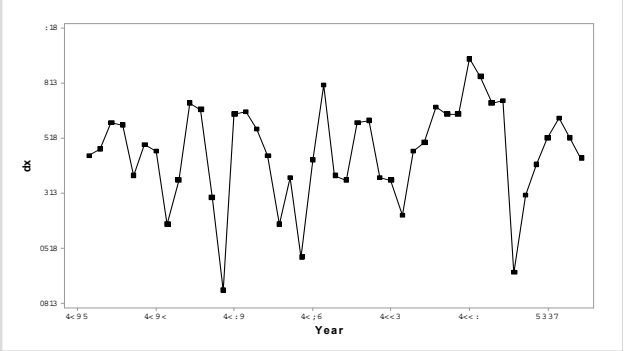

Fig. 7: Changes in industrial production

Looking at fig. 7 one might surmise that the data from the process $\{\Delta y_t,\ t{\in}\mathbb{N}\}$ seem stationary; see Choi (2015). It turns out that such a presumption will be false because the invoked AR(1) model (table 1) is mispecified; assumption [5] turns out to be invalid. To unpack that claim, consider testing for a unit root using the augmented

AR(1) model (Choi, 2015):

$$\Delta y_t = \underset{(1.038)}{1.104} + \underset{(.015)}{.004} y_{t-1} + \underset{(.153)}{.274} \Delta y_{t-1} + \widehat{u}_t, \;\; R^2=.08, \;\; s=2.252, \;\; n=43, \tag{41}$$

The insignificance of the coefficient of y_{t-1}, $\tau(\mathbf{y}_0)=\frac{.004}{.015}=.267[.81]$, when taken at face value, indicates that the stochastic process $\{y_t, \; t\in\mathbb{N}\}$ is integrated of order one. This inference, however, is not unwarranted because (41) is statistically misspecified. The significance of the coefficients of the trend terms in the following auxiliary regression:

$$\widehat{u}_t = \underset{(6.057)}{20.957} + \underset{(3.556)}{12.337} t_s + \underset{(.780)}{2.137} t_0^2 - \underset{(.092)}{.318} y_{t-1} + \underset{(.142)}{.120} \Delta y_{t-1} + \widehat{v}_{1t}, \tag{42}$$

(section 6.2) confirms that a unit root does not fully account for the mean heterogeneity in this data; note that $t_s=(2t-n-1)/(n-1)$ is a scaled version of t and $t_0^2=2t_s^2-1$, are used to avoid near-collinearity problems; see Spanos (2018b).

Respecifying (41) in light of (42) yields a statistically adequate model:

$$\Delta y_t = \underset{(6.067)}{22.057} + \underset{(3.557)}{12.337} t_s + \underset{(.780)}{2.137} t_0^2 - \underset{(.092)}{.315} y_{t-1} + \underset{(.142)}{.393} \Delta y_{t-1} + \widehat{v}_t, \;\; R^2=.3, \;\; s=2.015, \;\; n=43.$$

The significance of the coefficient of y_{t-1} indicates that the mean heterogeneity cannot be accounted for solely by a unit root, reversing the inference based on (41), i.e. $\{y_t, \; t\in\mathbb{N}\}$ is *not* integrated of order one.

8.4 Example 1 (continued). Revisiting the ER hypothesis

Returning to example 1, let us test the adequacy of the estimated AR(2) model in (3) using the auxiliary regressions to probe for the validity of assumptions [1]-[5] of the AR(2) model (table 1):

$$\begin{aligned} \widehat{u}_t = \underset{(.020)}{.003} - \underset{(.035)}{.007} y_{t-1} + \underset{(.033)}{.003} y_{t-2} + \underset{(.033)}{.003} y_{t-3} - \underset{(.033)}{.063} y_{t-4} - \\ - \underset{(.031)}{.002} t - \underset{(.025)}{.008} y_{t-1}^2 - \underset{(.025)}{.004} y_{t-2}^2 + \underset{(.04)}{.03} y_{t-1} y_{t-2} + \widehat{v}_{1t}, \end{aligned} \tag{43}$$

$$\widehat{u}_t^2 = \underset{(.028)}{.203} + \underset{(.041)}{.057} t + \underset{(.032)}{.14} \, y_{t-1}^2 + \underset{(.033)}{.04} \, y_{t-2}^2 + \underset{(.032)}{.105} y_{t-3}^2 + \underset{(.06)}{.35} y_{t-1} y_{t-2} + \widehat{v}_{2t}. \tag{44}$$

The auxiliary regression (43) indicates no departures from assumptions [2]-[5], but auxiliary regression (44) suggests serious departures from assumptions [3] since the coefficients of most quadratic terms are very significant. In light of the validity of the autoregressive function indicated by (43), one can proceed to test the Normality assumption. The Anderson-Darling (A-D) test yields: $A\text{-}D(\mathbf{y}_0)=8.138[.005]$.

8.5 Unwarranted respecification strategies

The aim in respecifying $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is to select more appropriate probabilistic assumptions for the stochastic process $\{\mathbf{Z}_t:=(y_t, \mathbf{X}_t), \; t\in\mathbb{N}\}$ that could account for the chance regularities in $\mathbf{Z}_0$ not accounted for by $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$.

Unfortunately, the traditional 'error-fixing' moves constitute another source of untrustworthy evidence because adopting the particularized alternative H_1 of the M-S test applied is fallacious.

Example: from OLS to GLS. The Durbin and Watson (1950) test for autocorrelation probes the validity of assumption {4} of the LR model (table 6) using the hypotheses:

$$H_0\text{: } \rho{=}0 \text{ vs. } H_1\text{: } \rho{\neq}0, \tag{45}$$

in the context of the encompassing model:

$$\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})\text{: } Y_t{=}\beta_0+\beta_1 x_t+u_t,\ u_t{=}\rho u_{t-1}{+}\varepsilon_t,\ |\rho|{<}1,\ (\varepsilon_t|X_t{=}x_t)\backsim\mathsf{NIID}(0,\sigma_\varepsilon^2). \tag{46}$$

When the D-W test rejects H_0, the traditional respecification is to adopt the alternative (encompassing) LR model (46), by replacing the OLS estimator of the parameters in $\boldsymbol{\theta}$ with 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 (35) is rejected, the only inference that can be drawn is that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is misspecified since the data indicate a generic departure from assumption [4] (table 7). That is, the test indicates that $E(u_t u_s|X_t{=}x_t){\neq}0$, but does *not* provide evidence *for* the particular alternative:

$$H_1\text{: } E(u_t u_s|X_t{=}x_t){=}\left[(\rho^{|t-s|})/(1{-}\rho^2)\right]\sigma_\varepsilon^2,\ t{>}s,\ t,s{=}1,2,...,n, \tag{47}$$

i.e. (46). This is because the D-W test would have rejected {4} for numerous particularized forms of departure within $[\mathcal{P}(\mathbf{z}){-}\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})]$, not just (47). For instance, the alternative encompassing model:

$$\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{z})\text{: } Y_t{=}\alpha_0+\alpha_1 x_t+\alpha_2 Y_{t-1}+\alpha_3 x_{t-1}+v_t,\ (v_t|X_t{=}x_t)\backsim\mathsf{NIID}(0,\sigma_v^2), \tag{48}$$

would have elicited a similar rejection by the D-W test; see Spanos and McGuirk (2001). To establish the validity of (47) one needs to tests the probabilistic assumptions of $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ in (46) and secure their validity; evidence *against* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is *not* evidence *for* $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ or $\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{z})$; see Mayo and Spanos (2004). Indeed, error-fixing gives rise to respecified models with unnecessary and often implausible restrictions; see McGuirk and Spanos (2009).

Robust standard errors. The conventional wisdom on how to deal with departures from [3] Homoskedasticity, is another questionable strategy that replaces the ordinary standard errors for the estimated coefficients with Heteroskedasticity Consistent Standard Errors (HCSE) proposed by White (1980).

Example 1 (continued). The auxiliary regressions (43)-(44) showed that the model in (3) is misspecified; assumptions [1] and [3] are invalid. Replacing the original SEs with the slightly larger HCSEs gives rise to:

$$y_t=\underset{[.019]}{.012}+\underset{[.037]}{.086}y_{t-1}+\underset{[.038]}{.002}y_{t-2}+\widehat{u}_t,\ s^2{=}.549,\ n{=}952, \tag{49}$$

that is as misspecified as (3) and yields equally unreliable inferences.

Recall that in M-S testing one needs to select a particularized alternative H_1 that is invariably a small subset of $[\mathcal{P}(\mathbf{z}){-}\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})]$. This implies that when H_0 is rejected, H_1 is never an option for respecification purposes, without further testing. Hence, M-S testing based on auxiliary regressions, such as (38) and (39), could only provide information about directions of departure from the model assumptions and not what the respecified model might look like. The latter should be decided on statistical adequacy grounds of the respecified models.

8.6 A coherent model respecification strategy

The framework for frequentist modeling and inference articulated in the previous sections offers a broader and more coherent vantage point from the narrow one stemming from the error process $\{(u_t|\mathbf{X}_t{=}\mathbf{x}_t),\ t{\in}\mathbb{N}\}$. This views the LR model as specified in terms of the regression and skedastic functions associated with the conditional distribution $D(y_t \mid \mathbf{X}_t;\boldsymbol{\theta})$: $E(y_t \mid \mathbf{X}_t{=}\mathbf{x}_t){=}h(\mathbf{x}_t),\ Var(y_t \mid \mathbf{X}_t{=}\mathbf{x}_t){=}g(\mathbf{x}_t),\ \mathbf{x}_t{\in}\mathbb{R}^k_X$, 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).

Example 1 (continued). As shown in section 6.4, the M-S testing results suggest that assumptions [1] and [3] are invalid for the estimated AR(2) model in (3). How would one go about respecify the original model with a view to account for these misspecifications? The AR(2) model assumes that the underlying stochastic process $\{y_t,\ t{\in}\mathbb{N}\}$ is Normal, Markov(2) and stationary. Since assumptions [1] and [3] for the conditional process $\{(y_t|\sigma(\mathbf{y}_{t-1})),\ t{\in}\mathbb{N}\}$, where for $\mathbf{y}_{t-1}{:=}(y_{t-1},y_{t-2},...,y_1)$, stem from the Normality assumption for $\{y_t,\ t{\in}\mathbb{N}\}$, a good starting point will be to replace it with a more appropriate distribution.

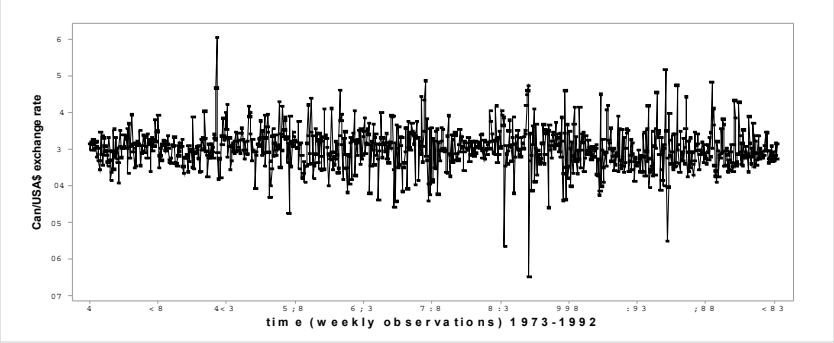

Fig. 8: Exchange rate log-returns

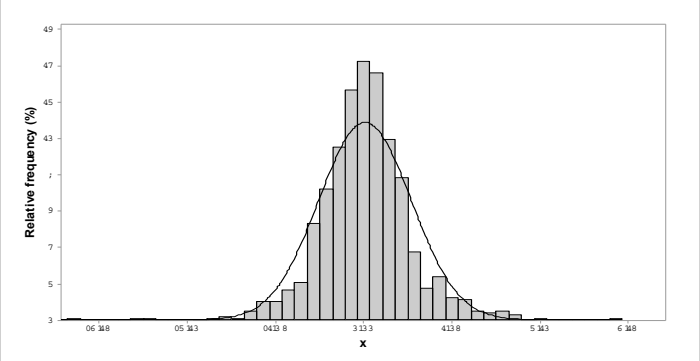

Fig. 9: Histogram of exchange rate log-returns

A closer look at the t-plot and the histogram of the data in figures 8-9 suggests that the underlying distribution is symmetric and leptokurtic; this is confirmed by the estimated skewness and kurtosis coefficients, $\widehat{\alpha}_3=-.06$ and $\widehat{\alpha}_4=7.17$, respectively. In light of the M-S testing results in (43), one should select a symmetric but leptokurtic distribution whose autoregressive function is the same as that of the Normal. An obvious choice is the Student's t distribution that yields the Student's t AR(2) model (table 8); see Spanos (1999). Estimating this model for $\nu=5$ yields:

$$y_t = \underset{(.012)}{.0001} + \underset{(.031)}{.104}y_{t-1} + \underset{(.035)}{.093}y_{t-2} + \widehat{v}_t,\ \hat{\sigma}_0^2=.375,\ n=952, \tag{50}$$

$$\widehat{\omega}_t^2=\widehat{\sigma}_0^2(1+\textstyle\sum_{i=1}^{2}\{\underset{(.011)}{.255}\widetilde{y}_{t-i}-\underset{(.007)}{.021}[\widetilde{y}_{t-i}\widetilde{y}_{t-i-1}+\widetilde{y}_{t-i}\widetilde{y}_{t-i+1}]\}), \tag{51}$$

where $\widetilde{y}_{t-i}:=(\widetilde{y}_{t-i}-\overline{y})$, that turns out to be statistically adequate; see Spanos (1995a).

Table 8: Student's t, AR (p) model

Statistical GM:	$y_t=\beta_0+\sum_{i=1}^{p}\beta_i y_{t-i}+v_t,\ \ t\in\mathbb{N},$
[1] Student's t:	$D(y_t\|\mathbf{y}_{t-1};\theta)$, is Student's t, $\nu+2$ d.f.
[2] Linearity:	$E(y_t\|\sigma(\mathbf{y}_{t-1}^0))=\beta_0+\sum_{i=1}^{p}\beta_i y_{t-i},$
[3] Heteroskedasticity:	$Var(y_t\|\sigma(\mathbf{y}_{t-1}^0))=\omega_p^2(\mathbf{y}_{t-1}),$
where $\omega_p^2(\mathbf{y}_{t-1})=(\frac{\nu\sigma_0^2}{\nu+p-2})$	$\left(1+\sum_{j=1}^{p}\sum_{i=1}^{p}\{\delta_{ij}[y_{t-i}-\mu][y_{t-j}-\mu]\}\right).$
[4] Markov:	$\{y_t,\ t\in\mathbb{N}\}$ is a Markov(2) process,
[5] t-invariance:	$\boldsymbol{\theta}:=(\beta_0,\beta_i,\sigma_0^2,\delta_{ij},\ \mu,\ i,j=1,...,p),$
	$\delta_{ij}=\delta_{kl}$ for$\|i-j\|=\|k-l\|,\ i,j,k,l=1,...,p.$

Statistical replicability. It is important to note that a statistically adequate model such as (50)-(51) is statistically replicable in the sense that it can be used to simulate replicas of the original data $\mathbf{y}_0$ that exhibit the same statistical systematic information (chance regularities), which in turn can be used to reproduce the original findings. In contrast, a statistically misspecified model would not reproduce the original findings when simulated; see Spanos and Mayo (2015).

9 Summary and conclusions

The discussion in this paper questioned the basic presumption 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 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. The advantages of the alternative proposals to replace significance testing, such as using observed CIs and estimation-based effects sizes, are often misconceived. Moroever, these alternatives 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 draws a sharp distinction between the latter and M-S testing.

(b) Pre-data vs. post-data error probabilities. This uncovers the key difference between the significance level and the p-value, neither of which has any information relating to particular discrepancies from H_0. This, in turn, motivates the post-data severity evaluation that provides the discrepancy from H_0 warranted by data $\mathbf{x}_0$. This can be used to (i) address several foundational problems, including the fallacies of acceptance and rejection, and (ii) a principled appraisal of the proposed alternatives to replace p-values.

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

(d) Statistical vs. substantive information/model. This brings out the fact that the bridge between the data and a scientific theory is sound when the adequacy of the invoked statistical model has been secured. Only then the data could be brought to bear upon the veracity of the substantive information by probing substantive adequacy.

The proposed framework offers suggestions to journal editors and referees for a number of ways to ameliorate the trustworthiness of published empirical evidence. First, decline papers that ignore establishing the adequacy of the invoked statistical model(s) for their inferences. Second, call out authors for uninformed implementation of frequentist inference procedures and unwarranted evidential interpretations of their results. Third, demand that authors probe adequately for any potential substantive misspecifications, after they secure the adequacy of the underlying statistical model.

References

[1] Altman, D. G., D. Machin, T. N. Bryant and M. J. Gardner (2000), *Statistics with Confidence*, (eds), British Medical Journal Books, Bristol, UK.

[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] Benjamin, D.J., et. al (2017) "Redefine statistical significance," *Nature Human Behaviour,* 33(1): 6–10. https://doi.org/10.1038/s41562-017-0189-z.

[4] Birnbaum, A. (1961), "A Unified Theory of Estimation, I," *The Annals of Mathematical Statistics*, 32: 112-135.

[5] 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.

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

[7] Choi, I. (2015) *Almost All About Unit Roots*, Cambridge University Press, Cambridge.

[8] Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), Lawrence Erlbaum, NJ.

[9] Cohen, J. (1994) "The Earth is round ($p < .05$)," *American Psychologist*, **49**: 997-1003.

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

[11] Cumming, G. (2012) *Understanding the New Statistics*, Routledge, NY.

[12] Durbin, J. and G.S. Watson, (1950), "Testing for Serial Correlation in Least Squares Regression, I[[, *Biometrika*, **37**: 409-428.

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

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

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

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

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

[18] Gigerenzer, G. (1993) "The superego, the ego, and the id in statistical reasoning", *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, 311-339.

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

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

[21] Haig, B.D. (2016) "Tests of statistical significance made sound," *Educational and Psychological Measurement,* 77: 489–506. https://doi.org/10.1177/0013164416667981.

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

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

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

[25] Kempthorne, O. and L. Folks (1971), *Probability, Statistics, and Data Analysis*, The Iowa State University Press, Ames, IA.

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

[27] 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.

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

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

[30] 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.

[31] 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.

[32] Nosek, B.A. and D.E. Lakens (2014) "A Method to Increase the Credibility of Published Results", *Social Psychology*, 45: 137-141.

[33] Poole, C. (1987), "Beyond the Confidence Interval," *The American Journal of Public Health*, **77**: 195-199.

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

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

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

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

[38] 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.

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

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

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

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

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

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

[45] Spanos, A. (2018a) "Near-collinearity in linear regression revisited: The numerical vs. the statistical perspective," *Communications in Statistics - Theory and Methods*, DOI:10.1080/03610926.1513147.

[46] Spanos, A. and D.G. Mayo (2015) "Error statistical modeling and inference: Where methodology meets ontology," *Synthese*, **192**: 3533-3555.

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

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

[49] Sugden, R. (2005) "Experiments as exhibits and experiments as tests," *Journal of Economic Methodology*, **12**: 291-302.

[50] Wasserman, L. (2006), *All of Nonparametric Statistics*, Springer, NY.

[51] Wasserstein, R.L., and N.A. Lazar (2016) "The ASA's statement on p-values: Context, process, and purpose," *The American Statistician,* 70: 129–133. https://doi.org/10.1080/00031305.2016.1154108.

[52] White, H. (1980) "A Heteroskedasticity-Consistent Covariance Matrix Estimator and Direct Test for Heteroskedasticity," *Econometrica*, **48**: 817-838.

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