Statistical Misspecification and the Reliability of Inference: the simple t-test in the presence of Markov dependence*

Aris Spanos
Department of Economics,
Virginia Tech, USA

Last revision: November 2009

Abstract

The aim of this paper is to consider the problem of *unreliable statistical inference* caused by the presence of statistical misspecification, and discuss the merits of alternative ways to address the problem like invoking generic robustness results or/and using nonparametric inference. For simplicity the discussion focuses on the t-test for hypotheses concerning the mean in the context of the simple Normal model, with the misspecification coming in the form of *Markov Dependence (MD)*. By deriving explicitly the *nominal* and *actual* error probabilities, it is shown that the presence of MD turns the t-test into an unreliable procedure. It is argued that invoking traditional robustness arguments can often be very misleading and, in general, this strategy does not address the unreliability of inference problem, even if one were to use the *actual error probabilities*. A more appropriate strategy is to *respecify* the *original statistical model* to account for the misspecification, and test the hypotheses of interest using an inference procedure that is optimal in the context of the respecified model. It is shown that the presence of MD gives rise to the *Autoregression* (AR(1)) as the respecified model, and one can test the original hypotheses concerning the mean. The optimal t-test in the context of the AR(1) is shown to be related but different from the original and the modified t-tests.

Keywords: misspecification, respecification, statistical adequacy, reliability of inference, nominal error probabilities, actual error probabilities, t-test, robustness, optimal inference, Markov dependence, AR(1), Fieller transformation, nonparametric methods

JEL classification: C12, C13, C51, C52

*Paper published in *The Korean Economic Review*, 2009, 25, 165-213.

1 Introduction

From its humble beginnings of using least-squares to 'fit' a single equation between two variables in the early 20th century (see Morgan, 1990), econometrics has developed into a powerful array of sophisticated statistical tools and procedures for modeling highly complicated dynamic multi-equation systems using time series, cross-section and panel data (see Greene, 2008). Combined with a rapid accumulation of new economic data, together with the widespread use of statistical software on personal computers, the publication of applied econometric papers has been growing exponentially over the last two decades.

Unfortunately for econometrics the trustworthiness of the empirical evidence has not improved since the early 20th century. If anything, the chronic problem of the untrustworthiness of published empirical evidence seems to have deteriorated as computing power has become more readily available. One can make a strong case that as the 21st century unfolds, the applied econometric literature is filled with a compendium of 'study-specific', 'period-specific', and largely *untrustworthy evidence*, which collectively provide a completely inadequate empirical foundation for economics; see Spanos (2006). What are the main sources of this untrustworthiness?

In the same paper, Spanos has argued that the primary sources of the untrustworthiness of empirical evidence are:

1. Inaccurate data: data $\mathbf{x}_0$ are marred by systematic errors imbued by the collection and compilation process.

2. Incongruous measurement: data $\mathbf{x}_0$ do not pertinently measure the concepts $\boldsymbol{\xi}$ envisioned by the particular theory-model (Spanos, 1995).

3. Substantive inadequacy: the circumstances envisaged by the theory-model differ 'systematically' from the actual phenomenon of interest.

4. Statistical misspecification: certain probabilistic assumptions comprising the statistical model (premises of induction) are invalid for data $\mathbf{x}_0$.

The main objective of this paper is to focus on statistical misspecification because it constitutes the single most crucial source of untrustworthy evidence. As shown below, even simple forms of statistical misspecification can easily ruin the reliability of inference. In particular, the discussion will shed light on three interrelated issues:
(i) how the problem of unreliable inferences arises under statistical misspecification,
(ii) how one should address the misspecification problem in practice, and
(iii) how the invocation of generic *robustness* arguments or/and the use of *nonparametric* methods is often misplaced when dealing with statistical misspecification.

The main argument is that the most effective way to secure the trustworthiness of empirical evidence is via thorough *Mis-Specification* (M-S) *testing* in order to assess the statistical adequacy of the prespecified (implicitly or explicitly) statistical model. Moreover, when any departures from the statistical model assumptions are detected, the proper way to address the potential unreliability of inference problems is to *respecify* the original model with a view to secure the adequacy of the new model.

Several widely used error-fixing strategies and invocations of generic robustness have been shown to be highly misleading in practice, and invariably make matters worse, not better. This applies to 'error-fixing' strategies like error-autocorrelation and heteroskedasticity corrections, heteroskedasticity/autocorrelation consistent standard errors (Greene, 2008), etc; see Spanos and McGuirk (2001).

Section 2 places the problem of the reliability of inference — as it pertain to model-based frequentist inference — in the broader context of empirical modeling in economics, distinguishing clearly between statistical and substantive adequacy. This is because researchers in economics often confuse statistical inadequacy with the problem of the unrealisticness of their structural models. Section 3 introduces the simple Normal model around which the discussion revolves and brings out the role of model assumptions in determining the relevant sampling distributions for inference purposes. In section 4 we consider how the presence of Markov dependence affects the reliability of the well-known t-test giving rise to disparities between the nominal and actual type I and II error probabilities. These results are then extended to confidence interval estimation in section 5. Section 6 considers the question of testing the hypotheses reliably by respecifying the simple Normal model with a view to account for the Markov dependence; the respecified model comes in the form of the *Autoregressive* (AR(1)) model. The optimal t-test in the context of the AR(1) model is derived and compared with the original and modified t-tests giving rise to the *relevant*, *nominal* and *actual error probabilities*, respectively. Section 7 utilizes the results of sections 5 and 7 to raise questions concerning the soundness of certain well-known arguments in favor of utilizing robust and/or nonparametric inference methods in order to deal with the unreliability of inference problem.

2 Model-based inference and its reliability

Since the early 20th century econometric modeling has been increasingly relying on frequentist model-based statistical inference pioneered by R. A. Fisher (1922). He initiated a change of paradigms in statistics by recasting the then dominating *Bayesian-oriented induction by enumeration* (Pearson, 1920), relying on large sample approximations, into a *frequentist model-based induction*, relying on finite sampling distributions. He proposed to view the data $\mathbf{x}_0{:=}(x_1, x_2, ..., x_n)$ as a realization of: (a) a 'random sample' from (b) a pre-specified 'hypothetical infinite population' and made the initial choice (specification) of the statistical model a response to the question:

"Of what population is this a random sample?" (Fisher, 1922, p. 313), emphasizing that: 'the adequacy of our choice may be tested posteriori' (ibid., p. 314).

Fisher's notion of a prespecified statistical model can be formalized in terms of the stochastic process $\{X_k,\ k{\in}\mathbb{N}\}$, underlying data $\mathbf{x}_0$. This takes the form of parameterizing the probabilistic structure of $\{X_k,\ k{\in}\mathbb{N}\}$ to specify a *statistical model*:

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

$f(\mathbf{x};\boldsymbol{\theta})$ denotes the joint *distribution of the sample* $\mathbf{X}{:=}(X_1, ..., X_n)$ that encapsulates the whole of the probabilistic information in $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, by giving a general description

of the probabilistic structure of $\{X_k,\ k\in\mathbb{N}\}$ (Doob, 1953). $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is chosen to provide an idealized description of the mechanism that generated data $\mathbf{x}_0$ with a view to appraise and address the substantive questions of interest. The basic idea is to construct statistical models using probabilistic assumptions that 'capture' the chance regularities in the data with a view to adequately account for the underlying data-generating mechanism; see Spanos (1999).

The quintessential *example* of a statistical model is *the simple Normal model*:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\text{: } X_k \backsim \mathsf{NIID}(\mu,\sigma^2),\ \boldsymbol{\theta}\text{:}=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+,\quad k=1,2,...,n,..., \tag{2}$$

where '$\backsim \mathsf{NIID}(\mu,\sigma^2)$' stands for 'distributed as Normal, Independent and Identically Distributed, with mean μ and variance σ^2.

The statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ plays a pivotal role in model-based frequentist inference in so far as it determines what constitutes *a legitimate*:

(a) event — any well-behaved (Borel) functions of the sample $\mathbf{X}$ —

(b) assignment of probabilities to legitimate events via $f(\mathbf{x};\boldsymbol{\theta})$,

(c) data $\mathbf{x}_0$ for inference purposes,

(d) hypothesis and/or inferential claim, and

(e) optimal inference procedure and the associated error probabilities.

Formally an event is legitimate when it belongs to the σ-field generated by $\mathbf{X}$ (Billingsley, 1995). Legitimate data come in the form of data $\mathbf{x}_0$ that can be realistically viewed as a truly typical realization of the process $\{X_k,\ k\in\mathbb{N}\}$, as specified by $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. Legitimate hypotheses and inferential claims are invariably about the data-generating mechanism and framed in terms of the unknown parameters $\boldsymbol{\theta}$. Moreover, the optimality (effectiveness) of the various inference procedures depends on the validity of the probabilistic assumptions constituting $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$; see Spanos (1999).

A major problem with statistical modeling since the 1920s has been articulated succinctly by Rao (2004):

> "The current statistical methodology is mostly model-based, without any specific rules for model selection or validating a specified model." (p. 2)

The methodological framework proposed in Spanos (1986, 1988, 1995, 2010) aspires to remedy this crucial weakness. The idea is to develop a methodology of **Specification, Mis-Specification (M-S) testing** and **Respecification** with a view to secure **statistical adequacy**: how to specify and validate statistical models, how to probe model assumptions, isolate the sources of departures, and account for them in a respecified model to be used as a basis for primary statistical inferences.

Statistical adequacy refers to the validity–vis-à-vis data $\mathbf{x}_0$–of the probabilistic assumptions comprising the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in question, and provides *the* sole criterion for 'when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ accounts for the (recurring) regularities in data $\mathbf{x}_0$.'

Error-reliability. Statistical adequacy renders the relevant error probabilities ascertainable by ensuring that the *nominal* error probabilities for assessing substantive claims are approximately equal to the *actual* ones. The surest way to draw an invalid inference is to apply a .05 significance level test when its actual – due to

misspecification – type I error probability is closer to .99. Despite its obvious importance, securing the reliability of inference has been largely neglected by the modern statistics literature for several reasons.

In addressing the question, 'how can one assess the adequacy of $\mathcal{M}_\theta(\mathbf{x})$ *a posteriori*?' one had to face two difficult hurdles. The first was to specify $\mathcal{M}_\theta(\mathbf{x})$ explicitly using a complete set of testable – vis-a-vis data $\mathbf{x}_0$ – assumptions. The second, and more difficult, had to do with delineating the role of *substantive* subject matter information in specifying $\mathcal{M}_\theta(\mathbf{x})$; see Lehmann (1990). The pivotal difficulty, which has ravaged the trustworthiness of empirical modeling in the social sciences, arises when one imposes the substantive information (theory) on the data at the outset. The end result is often a statistically and substantively *inadequate* estimated model, but one has no way to delineate the two sources of error:

$$\text{is the substantive information erroneous? or the inductive premises mispecified?} \tag{3}$$

This is an example of a classic problem in philosophy of science known as Duhem's problem; see Mayo (1996). To place it in proper context let us briefly discuss how it has stumped econometric modeling since the early 20th century.

2.1 Statistical misspecification vs. the realisticness issue

As argued in Spanos (2009), empirical modeling in economics has been largely dominated by the Pre-Eminence of Theory (PET) perspective since Ricardo (1817). This perspective asserts that modeling takes the form of constructing simple idealized models which capture certain key aspects of the phenomenon of interest, with a view to use such models to shed light or explain such phenomena, as well as gain insight concerning potential alternative policies. From the PET perspective the role of the data is only subordinate in the sense that it can help to instantiate such models (assumed to be true) by quantifying them.

A key point in Spanos (2009) is that the PET perspective proponents conflate:

(a) the unrealisticness – vis-à-vis the phenomenon of interest – of the substantive assumptions comprising the theory-model in question, with

(b) the inappropriateness – vis-à-vis the data in question – of the probabilistic assumptions comprising the statistical model that defines the underlying premises for inductive inferences.

The contrast between the unrealisticness of a theory-model and the adequacy of the statistical model is crucial because the types of errors one should probe for and guard against are very different in the two cases. Crudely put, one pertains to the substantive and the other to the statistical adequacy. Unfortunately, the PET perspective ignores this distinction and often foists the theory-model on the data at the outset giving rise to estimated models which are both *statistically* and *substantively inadequate*, giving rise to the Duhemian ambiguity in (3).

The distinction between statistical and substantive adequacy is important because the presence of statistical misspecification will undermine any prospect of reliably probing potential substantive errors/omissions. Statistical adequacy is necessary for

being able to ascertain the reliability of any inductive inference pertaining to substantive questions of interest. Without it no *learning from data* is possible because one effectively revokes the ascertainment of the reliability of statistical inference rendering it tantamount to a crystal ball procedure! This does not pertain to the realisticness of the theory as such. Having said that, it is important to note that the realisticness of a theory issue should be discussed by juxtaposing the theory-model to the phenomenon of interest as it relates to the particular data $\mathbf{z}_0$, using a statistically adequate model.

The realisticness of the theory is an issue that pertains to the substantive adequacy of the estimated model vis-à-vis the phenomenon of interest, i.e. whether the model in question provides a veritable explanation for that phenomenon. Substantive adequacy concerns the extent to which the estimated model accounts for all systematic aspects of the reality it aims to explain in a statistically and substantively adequate way, shedding light on the phenomenon of interest, i.e. 'learning from data'. Such inadequacy can easily arise from impractical *ceteris paribus* clauses, external invalidity issues, missing confounding factors, false causal claims, etc.; see Guala (2005), Hoover (2006). Securing substantive adequacy calls for additional probing of (potential) errors in bridging the gap between theory and data. However, without securing statistical adequacy first, such probing is likely to be misleading because the statistical procedures employed cannot be trusted to yield reliable inferences; one might as well use a crystal ball for that!

For the proponents of the PET perspective the crystal ball takes the form of assessing whether the estimated model in question "works" or not, where the metric "works" is defined in terms of a variety of criteria that do not include statistical adequacy! These criteria often include certain statistical indicators, such as goodness-of-fit and goodness-of-prediction statistics, as well as several subjective judgements pertaining to the model's capacity to 'shed light' and/or confirm preconceived beliefs by the modeler. Leaving aside the subjective judgements for the moment, what is not appreciated enough in this literature is that, without statistical adequacy, statistical measures of 'goodness' are meaningless artifacts; see Spanos (1989). Indeed, Friedman's (1953), p. 8, highly influential and widely rehearsed catchphrase:

> "Viewed as a body of substantive hypotheses, theory is to be judged by its predictive power for the class of phenomena which is intended to 'explain'."

begs the question:

▶ How can one reliably appraise the predictive power of a theory when the reliability of the very tools (statistical inference) used in that assessment is at best unknowable and at worst highly questionable?

Friedman goes on to make a case for confronting theory with data:

> "Only factual evidence can show whether it is 'right' or 'wrong'(p. 8)

which also begs a related question:

▶ How does one establish 'factual evidence' without statistical adequacy?

Raw data are a far cry from reliable evidence one can use to confront theories with; see Spanos (2010).

2.2 Addressing the Duhemian ambiguity

The key to dealing with this crucial Duhemian ambiguity in (3) is to distinguish, *ab initio*, between *statistical* and *substantive* information; Spanos (1986). The underlying rationale is that statistical adequacy is a precondition for securing the reliability of the inference procedures used in appraising substantive adequacy because error-reliability [the *actual* error probabilities approximate closely the *nominal* ones]. A statistically misspecified model will lead inductive inferences astray.

The big hurdle in getting a handle on the reliability of inference has been to establish a notion of 'statistical information' that can be untangled, at least ab initio, from substantive information.

A. A purely probabilistic construal of a statistical model

Spanos (1986) proposed a notion of *statistical information* that relates directly to the chance regularity patterns (distribution, dependence and heterogeneity) exhibited by data $\mathbf{x}_0$ when the latter is viewed as a realization of a *generic* – free from any substantive information – stochastic process $\{X_k,\ k\in\mathbb{N}:=(1,2,...)\}$. This notion enables one to put forward a purely probabilistic construal of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ by viewing it as a particular *parameterization* of the probabilistic structure of a process $\{X_k,\ k\in\mathbb{N}\}$; Spanos (1995).

Substantive subject matter information usually enters empirical modeling in the form of a *structural model,* say $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$, which constitutes an *estimable form* of a theory, in view of the specific data $\mathbf{x}_0$.

B. Reconciling substantive and statistical information

The statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is built exclusively on *statistical systematic information* in data $\mathbf{x}_0$, and is selected so as to meet two interrelated aims:

(I) to account for the chance regularities in data $\mathbf{x}_0$ by choosing a probabilistic structure for the stochastic process $\{X_k,\ k\in\mathbb{N}\}$ underlying $\mathbf{x}_0$ so as to render it a 'typical realization' thereof, and

(II) to parameterize the probabilistic structure of $\{X_k,\ k\in\mathbb{N}\}$ in the form of an adequate statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ that would *embed* $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ in its context, via *reparametrization/restriction* $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$; formal assessment of the latter provides a way to reconcile the two sources of information; Spanos (1990, 2007).

The Probabilistic Reduction (PR) approach (Spanos, 1989) provides the framework for securing these objectives by:

(i) specifying $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in terms of a *complete* list of (internally consistent) probabilistic assumptions, in a form that is testable vis-à-vis data $\mathbf{x}_0$, and

(ii) supplementing that with a *statistical generating mechanism* (GM) to provide a bridge between the statistical and substantive information.

3 The optimality and reliability of inference

The primary motivation underlying the Probabilistic Reduction perspective is that in statistics the *reliability* of any inference procedure (estimation, testing and predic-

tion) depends crucially on the validity of the *premises*: the probabilistic assumptions comprising the *statistical model* in the context of which the inference takes place. Assuming the validity of such premises, the optimality of inference methods in *frequentist statistics* is defined in terms of their capacity to give rise to valid inferences (*trustworthiness*), which is assessed in terms of the associated error probabilities: how often these procedures lead to erroneous inferences. In the case of Confidence Interval (CI) estimation the assessment is often gauged in terms of minimizing *the coverage error probability*: the probability that the interval does *not* contain the true value of the unknown parameter(s). In the case of hypothesis testing the assessment is ascertained in terms of minimizing *the type II error probability*: the probability of accepting the null hypothesis when false, for a given *type I error probability*; see Cox and Hinkley (1974). Hence, the *reliability* of a frequentist inference procedure depends on two interrelated pre-conditions:

(a) adopting optimal inference procedures, in the context of

(b) a statistically adequate model.

It is also well-known that when any of the probabilistic assumptions comprising the premises of a statistical model are invalid, the *reliability of inference* procedures is called into question; see Pearson (1931), Bartlett (1935) for early discussions.

What is less well-known is how the unreliability of inference manifests itself in empirical modeling. In frequentist statistics, the unreliability of inference is reflected in the *difference* between the *nominal* error probabilities, derived under the assumption of valid premises, and the *actual* error probabilities, derived taking into consideration the particular departure(s) from the premises; see Spanos and McGuirk (2001). Indeed, this difference provides a way to assess the extent of the unreliability of inference. In the terminology of statistical 'robustness', this difference provides a measure of the *sensitivity* of the inference procedure to the particular departure from the model assumptions; see Box (1953), Box and Tiao (1973), Staudte and Sheather (1990), inter alia.

The main argument of this paper is that *reliable* and *precise inferences* are the result of utilizing the *relevant error probabilities* obtained by ensuring (a)-(b). In practice, the unreliability of inference problem often stems from the inability to utilize the *relevant error probabilities* arising from being unaware of the presence of departures from the premises. However, even if one were in a position to utilize the actual error probabilities, that, by itself, does not address the unreliability of inference problem in general. This is because the presence of misspecification calls into question, not only the appropriateness of the nominal error probabilities, but also the optimality of the original inference procedure; without (b), (a) makes little sense. Hence, the unreliability of inference problem is better addressed by *respecifying* the original statistical model and utilizing inference methods that are optimal in the context of the new (adequate) premises; see Spanos (1986).

The distinctions between *nominal*, *actual* and *relevant error probabilities* is important because the traditional discussion of *robustness* compares the actual with the

nominal error probabilities, but downplays the interconnection between (a) and (b) above. Indeed, well-rehearsed mantras like:

"All models are misspecified, to 'a greater or lesser extent', because they are, by definition mere idealizations and approximations. Moreover, slight departures from the assumptions will only lead to minor deviations from the optimal inferences,"

are shown to be highly misleading in practice. It is argued that invoking generic robustness results often amounts to 'glossing over' the unreliability of inference problem instead of addressing it.

3.1 The simple Normal model and the role of inductive premises

The discussion of statistical misspecification and respecification will focus on the very simple statistical model, in order to derive analytical results, but these results can be easily extended to the linear regression and related models that dominate econometric modeling. The model of focus is the simple Normal model as specified in table 1 in terms of a statistical *Generating Mechanism* (GM) and the probabilistic assumptions [1]-[4]; see Spanos (1999).

Table 1 - Simple Normal Model

$$\begin{array}{lll} \textit{Statistical GM:} & X_t=\mu+u_t, & t\in\mathbb{N}. \\ [1] \text{ Normality:} & X_t \backsim \mathsf{N}(.,.), & \left.\begin{array}{l}\\\\\end{array}\right\} \ t\in\mathbb{N}. \\ [2] \text{ constant mean:} & E(X_t):=\mu, & \\ [3] \text{ constant variance:} & Var(X_t):=\sigma^2, & \\ [4] \text{ Independence:} & \{X_t,\ t\in\mathbb{N}\} \text{ is an independent process.} & \end{array} \tag{4}$$

It is well known (see Cox and Hinkley, 1974) that the estimators:

$$\overline{X}=\tfrac{1}{n}\sum_{t=1}^{n}X_t,\ \ s^2=\tfrac{1}{(n-1)}\sum_{t=1}^{n}(X_t-\overline{X})^2,$$

have certain optimal properties like consistency, unbiasedness, efficiency, sufficiency, etc. These properties stem from their sampling distributions:

$$\overline{X} \backsim \mathsf{N}(\mu,\tfrac{\sigma^2}{n}), \qquad \tfrac{(n-1)s^2}{\sigma^2} \backsim \chi^2(n-1), \tag{5}$$

where '$\backsim \mathsf{N}(\mu,\frac{\sigma^2}{n})$' stands for distributed as Normal with mean μ and variance $\frac{\sigma^2}{n}$' and '$\chi^2(n-1)$' denotes the chi-square distribution with $n-1$ degrees of freedom. What is often not appreciated enough is that these distributional results hold only when the model assumptions [1]-[4] are valid. In particular, assumptions [1]-[4] are explicitly invoked in deriving the mean and variance of these estimators as indicated below:

$$\begin{array}{rl} E(\overline{X}) & \overset{[2]}{=}\tfrac{1}{n}\sum_{t=1}^{n}\mu=\mu, \\ Var(\overline{X}) & \overset{[4]}{=}\tfrac{1}{n^2}\sum_{t=1}^{n}Var(X_t)\overset{[3]}{=}\tfrac{1}{n^2}\sum_{t=1}^{n}\sigma^2=\tfrac{\sigma^2}{n}, \\ E(s^2) & \overset{[2]\&[4]}{=}\tfrac{1}{(n-1)}\left[\sum_{t=1}^{n}E\left(X_t-\mu\right)^2-nE\left(\overline{X}-\mu\right)^2\right]\overset{[3]}{=}\sigma^2, \\ Var(s^2) & \overset{[4]}{=}\tfrac{1}{(n-1)^2}\sum_{t=1}^{n}Var(X_t-\overline{X})^2\overset{[1]\&[3]}{=}\tfrac{2\sigma^4}{(n-1)}. \end{array} \tag{6}$$

The validity of [1]-[4] is also needed for Student's (1908) famous result:

$$\frac{\sqrt{n}(\overline{X}-\mu)}{s} \backsim \mathsf{St}(n-1), \tag{7}$$

where $\mathsf{St}(m)$ denotes the Student's t distribution with m degrees of freedom.

Point estimation is often considered *inadequate* for the purposes of scientific inquiry because a 'good' point estimator $\widehat{\theta}_n(\mathbf{X})$, by itself, does not provide any measure of the reliability and precision associated with the estimate $\widehat{\theta}_n(\mathbf{x}_0)$. This is the reason why $\widehat{\theta}_n(\mathbf{x}_0)$ is often accompanied by some significance test result (e.g. p-value) associated with the *generic* hypothesis $\theta=0$.

Interval estimation rectifies this crucial weakness of point estimation by providing the relevant error probabilities associated with inferences pertaining to 'covering' the true value of θ. This comes in the form of the Confidence Interval (CI):

$$\mathbb{P}\left(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})\right) \quad =1-\alpha, \tag{8}$$

where the statistics $L(\mathbf{X})$ and $U(\mathbf{X})$ denote the lower and upper (random) bounds that 'covers' the true value θ^* with probability $(1-\alpha)$, or equivalently, the 'coverage error' probability is α.

Example. In the case of the simple Normal model (table 1):

$$\mathbb{P}\left(\overline{X}-c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}}) \leq \mu \leq \overline{X}+c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}})\right) \quad =1-\alpha, \tag{9}$$

provides a $(1-\alpha)$ CI for μ. The evaluation of the coverage probability $(1-\alpha)$ is based on (7).

What is often not appreciated sufficiently about estimation in general, and CIs in particular, is the underlying reasoning that gives rise to sampling distribution results such as (5) and (7). The reasoning that underlies estimation is *factual,* based on evaluating the relevant sampling distributions 'under the True State of Nature' (TSN), i.e. the *true* data-generating mechanism: $\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}^*)\}$, $\mathbf{x}\in\mathbb{R}^n_X$, where $\boldsymbol{\theta}^*$ denotes the true value of the unknown parameter(s) $\boldsymbol{\theta}$. Hence, the generic CI in (8) is more accurately stated as:

$$\mathbb{P}\left(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X});\ \theta{=}\theta^*\right)=1-\alpha, \tag{10}$$

where $\theta{=}\theta^*$ denotes 'evaluated under the TSN'. The remarkable thing about factual reasoning is that one can make probabilistic statements like (10), with a precise error probability (α), *without* knowing the true θ^*.

Example. In the case of the simple Normal model, the distributional results (5) and (7) are more accurately stated as:

$$\overline{X} \overset{\text{TSN}}{\backsim} \mathsf{N}\left(\mu_*, \tfrac{\sigma^2_*}{n}\right), \quad \tfrac{(n-1)s^2}{\sigma^2_*} \overset{\text{TSN}}{\backsim} \chi^2(n-1), \quad \tfrac{\sqrt{n}(\overline{X}-\mu^*)}{s} \overset{\text{TSN}}{\backsim} \mathsf{St}(n-1), \tag{11}$$

where $\boldsymbol{\theta}^*{:=}(\mu_*,\sigma^2_*)$ denote the 'true' values of the unknown parameters $\boldsymbol{\theta}{:=}(\mu,\sigma^2)$.

Prediction is similar to estimation in terms of its underlying factual reasoning, but it differs from it in so far as it is concerned with finding the most representative

value of X_k beyond the observed data, say X_{n+1}. An optimal predictor of X_{n+1} is given by:

$$\widehat{X}_{n+1}{=}\overline{X}, \tag{12}$$

whose reliability can be calibrated using the sampling distribution of the prediction error:

$$\widehat{u}_{n+1}{=}(X_{n+1}-\overline{X})\overset{\text{TSN}}{\backsim}\mathsf{N}\left(0,\sigma_*^2(1+\tfrac{1}{n})\right), \tag{13}$$

to construct a $(1{-}\alpha)$ prediction interval:

$$\mathbb{P}\left(\overline{X}-c_{\frac{\alpha}{2}}(s\sqrt{(1+\tfrac{1}{n})})\leq X_{n+1}\leq\overline{X}+c_{\frac{\alpha}{2}}(s\sqrt{(1+\tfrac{1}{n})});\boldsymbol{\theta}{=}\boldsymbol{\theta}^*\right)\ {=}1{-}\alpha. \tag{14}$$

Hypothesis testing. In contrast to estimation, the reasoning underlying hypothesis testing is *hypothetical*. The sampling distribution of a test statistic is evaluated under several hypothetical scenarios concerning the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, referred to as 'under the null' and 'under the alternative' hypotheses of interest.

Example. Consider testing the hypotheses):

$$H_0\text{:}\ \mu\leq\mu_0\ \text{vs.}\ H_1\text{:}\ \mu>\mu_0, \tag{15}$$

in the context of the simple Normal model. What renders the hypotheses in (15) legitimate is that: (i) they pose questions concerning the underlying data-generating mechanism, (ii) they are framed in terms of the unknown parameter $\boldsymbol{\theta}$, and (iii) in a way that partitions $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

The N-P test for the hypotheses (15) $T_1(\alpha){:=}\{\tau(\mathbf{X}),C_1(\alpha)\}$, where:

$$\tau(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}-\mu_0)}{s},\ C_1(\alpha){=}\{\mathbf{x}\text{:}\ d(\mathbf{x})>c_\alpha\}, \tag{16}$$

can be shown to be Uniformly Most Powerful (UMP) in the sense that, its type I error probability (significance level) is:

$$\text{[a]}\ \alpha{=}\max_{\mu\leq\mu_0}\mathbb{P}(\mathbf{x}\text{:}\ \tau(\mathbf{X})>c_\alpha;H_0){=}\mathbb{P}(\mathbf{x}\text{:}\ \tau(\mathbf{X})>c_\alpha;\mu{=}\mu_0), \tag{17}$$

and among all the α-level tests $T_1(\alpha)$ has highest *power* (Lehmann, 1986):

$$\text{[b]}\ \mathbb{P}(\mathbf{x}\text{:}\ \tau(\mathbf{X})>c_\alpha;\mu{=}\mu_1),\ \textit{for all}\ \mu_1>\mu_0,\ \mu_1{=}\mu_0{+}\gamma,\ \gamma\geq0); \tag{18}$$

In this sense, a UMP test provides the most effective α-level probing procedure for detecting any discrepancy ($\gamma\geq0$) of interest from the null.

To evaluate the error probabilities in (17) and (18) one needs to derive the sampling distribution of $\tau(\mathbf{X})$ under several *hypothetical* values of μ relating to (6):

$$\text{[a]}\ \tau(\mathbf{X})\overset{\mu=\mu_0}{\backsim}\mathsf{St}(n{-}1),\qquad\text{[b]}\ \tau(\mathbf{X})\overset{\mu=\mu_1}{\backsim}\mathsf{St}(\delta(\mu_1);n{-}1),\ \text{for any}\ \mu_1>\mu_0, \tag{19}$$

where $\delta(\mu_1){=}\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$ is known as the non-centrality parameter. The sampling distribution in (19)[a] is also used to evaluate Fisher's (1935) p-value:

$$p(\mathbf{x}_0){=}\mathbb{P}(\mathbf{x}\text{:}\ \tau(\mathbf{X})>\tau(\mathbf{x}_0);\mu{=}\mu_0), \tag{20}$$

where a small enough $p(\mathbf{x}_0)$ can be interpreted as indicating discordance with H_0.

Comparing the sampling distributions in (19) with those in (11) brings out the key difference between hypothetical and factual reasoning: in the latter case there is

only one unique scenario, but in hypothetical reasoning there is usually an infinity of scenarios. The remarkable thing about hypothetical reasoning is that one can pose sharp questions by comparing $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}_X^n$, for different hypothetical values of $\boldsymbol{\theta}$, with $\mathcal{M}^*(\mathbf{x}_0)$, to learn about $\mathcal{M}^*(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}_X^n$. This often elicits more informative answers from $\mathbf{x}_0$ than factual reasoning. This difference is important in understanding the nature of the error probabilities associated with each type of inference as well as in interpreting the results of these procedures.

In particular, factual reasoning can only be used pre-data to generate the relevant error probabilities, because when data $\mathbf{x}_0$ is observed (i.e. post-data) the unique factual scenario has been realized and the sampling distribution in question becomes degenerate. This is the reason why the p-value in (20) is a well-defined post-data error probability, but one cannot attach error probabilities to an observed CI: $(L(\mathbf{x}_0) \leq \theta \leq U(\mathbf{x}_0))$. In contrast, the scenarios in hypothetical reasoning are equally relevant to both pre-data and post-data assessments. Indeed, one can go a long way towards delineating some of the confusions surrounding frequentist testing, as well as addressing some of the criticisms leveled against it — statistical vs. substantive significance, with a large enough n one can reject any null hypothesis, no evidence against the null is *not* evidence for it — using post-data error probabilities to provide an evidential interpretation of frequentist testing based on the severity rationale; see Mayo and Spanos (2006) for further discussion.

Numerical example. Let us assume that μ_0=0, s=1, n=100, and α=.05 (c_α=1.66). The power of the t-test (16) at different values of μ=μ_1 is given in table 2.

Table 2 - Power of $T_1(.05)$ for different $\mu_1 > \mu_0$

μ_1	.01	.02	.05	.1	.15	.2	.3	.4	.5	.6
δ	0.1	0.2	0.5	1.0	1.5	2.0	3.0	4.0	5.0	6.0
$\pi(\mu_1)$	.061	.074	.121	.258	.437	.637	.911	.991	.999	1.0

As expected, the power of $T_1(\alpha)$ increases as the discrepancy between μ_1 and μ_0 increases, *ceteris paribus*.

4 Hypothesis Testing and misspecification

Consider the case where assumption [4] (see table 1) *is false*, and the sample is **Markov dependent**, in the sense that:

$$[5] \quad \text{Markov dependence:} \quad Corr(X_i, X_j)=\rho^{|i-j|}, \quad -1<\rho<1, \ i \neq j, \ i,j=1,...n. \tag{21}$$

The choice of this form of dependence is made on the basis of accumulated empirical evidence that most economic time series data exhibited such a form of dependence with a positive ρ in the range $.5 \leq \rho \leq .99$.

Misspecification affects the reliability of test $T_1(\alpha)$ by altering the sampling distribution of $\tau(\mathbf{X})$ under H_0 and H_1. This, in turn, distorts its nominal error probabilities (type I and II) rendering them different from the actual ones based on the correct

distributions that allow for the misspecification. Let us consider the consequences of this particular departure for the *error probabilities* of the test $T_1(\alpha)$, beginning with the sampling distribution of $\overline{X}$.

4.1 Misspecification and the sampling distribution of $\overline{X}$

If we return to the derivations in (6), it's easy to see that the Normality of the sampling distribution of $\overline{X}$ will not be affected, the mean of $\overline{X}$ will stay the same as in (6), but its variance will change because assumption [4] no longer holds. In particular, the covariance will no longer be zero but:

$$Cov(X_i, X_j)=\sigma^2\rho^{|i-j|} \neq 0,\ i \neq j,\ i,j=1,...n. \tag{22}$$

Hence, instead of (6) the variance of $\overline{X}$ takes the form (see Anderson, 1971):

$$\begin{array}{ll} Var(\overline{X}) & =\frac{1}{n^2}\left(\sum_{i=1}^{n} Var(X_i)+2\sum_{i=1}^{n}\sum_{j=i+1}^{n} Cov(X_i,X_j)\right)= \\ & =\frac{1}{n^2}\left(\sum_{i=1}^{n}\sigma^2+2n\sigma^2\sum_{k=1}^{n}(1-\frac{k}{n})\rho^k\right),\ \text{for } k=|i-j|, \\ & =\frac{\sigma^2}{n}\left(1+\frac{2\rho[n(1-\rho)-1+\rho^n]}{n(1-\rho)^2}\right)=\frac{\sigma^2}{n}c_n(\rho), \end{array}$$

$$c_n(\rho)=\left(1+\frac{2\rho(n(1-\rho)-1+\rho^n)}{n(1-\rho)^2}\right)=\left(\frac{1+\rho}{1-\rho}-\frac{2\rho(1-\rho^n)}{n(1-\rho)^2}\right). \tag{23}$$

This implies that the *actual* sampling distribution of $\overline{X}$ – assuming (21) instead of [4] – is no longer as in (5), but instead:

$$\overline{X} \backsim \mathsf{N}\left(\mu, \frac{\sigma^2 c_n(\rho)}{n}\right). \tag{24}$$

4.2 Misspecification and the sampling distribution of s^2

The effect of the presence of dependence (21) on the sampling distribution of s^2 is more complicated. It remains chi-square (since Normality is retained), but its mean and variance are very different from those in (6). In particular, the mean of s^2 takes the form (see Anderson, 1971):

$$\begin{array}{ll} E(s^2) & =\sigma^2-\frac{1}{n}\left[\sigma^2+2\sum_{k=1}^{n-1}(1-\frac{k}{n})Cov(X_i,X_j)\right],\ \text{for } k=|i-j|, \\ & =\sigma^2-\left[\frac{\sigma^2}{n}+\frac{2\sigma^2}{n}\sum_{k=1}^{n-1}(1-\frac{k}{n})\rho^k\right]=\sigma^2\left(1-\frac{1}{n}-\frac{2\rho(n(1-\rho)-1+\rho^n)}{n^2(1-\rho)^2}\right)= \\ & =\sigma^2\left(\frac{n(n-1)(1-\rho)^2-2\rho(n(1-\rho)-1+\rho^n)}{n^2(1-\rho)^2}\right)=\sigma^2 d_n(\rho), \end{array}$$

$$d_n(\rho)=\left(\frac{n(n-1)(1-\rho)^2-2\rho(n(1-\rho)-1+\rho^n)}{n^2(1-\rho)^2}\right). \tag{25}$$

Using the relationship between the mean and variance of a chi-square distribution, we can deduce that the actual sampling distribution of s^2, is no longer as in (5), but takes the form:

$$s^2 \backsim \left(\frac{\sigma^2 d_n(\rho)}{n-1}\right)\chi^2(n-1) \quad \text{or} \quad \frac{(n-1)s^2}{\sigma^2 d_n(\rho)} \backsim \chi^2(n-1). \tag{26}$$

4.3 Misspecification and the error probabilities of the t-test

It is important to emphasize at the outset of the discussion that follows that the situation envisaged is a scenario where one applies the t-test (16) assuming that assumptions [1]-[4] are valid, but in fact [4] is invalid and instead (22) holds.

It's clear from the above derivations of the sampling distributions of $\left(\overline{X}, s^2\right)$ under Markov dependence (21), that the sampling distribution of $\tau(\mathbf{X})$ will also be affected. The pivotal quantity (7) is no longer a ratio whose numerator is standard Normal and the denominator, under the square bracket, is chi-square distributed. In view of (24) and (26) that ratio takes the form:

$$h^*(\mathbf{X};\mu)=\frac{\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma\sqrt{c_n(\rho)}}}{\sqrt{\frac{(n-)s^2}{(n-1)\sigma^2 d_n(\rho)}}}=\frac{\sqrt{n}(\overline{X}-\mu)}{s\sqrt{\lambda_n(\rho)}} \backsim \mathsf{St}(n-1), \tag{27}$$

$$\lambda_n(\rho)=\frac{c_n(\rho)}{d_n(\rho)} =\frac{n\left[n(1-\rho)^2+2\rho(n(1-\rho)-1+\rho^n)\right]}{[n(n-1)(1-\rho)^2-2\rho(n(1-\rho)-1+\rho^n)]}. \tag{28}$$

Hence, the *actual* distribution of $\tau(\mathbf{X})$ under H_0 becomes:

$$\tau(\mathbf{X})=\frac{\sqrt{n}(\overline{X}-\mu_0)}{s} \overset{H_0}{\backsim} \sqrt{\lambda_n(\rho)}\mathsf{St}\left(n-1\right), \quad \text{or } \tau^*(\mathbf{X})=\frac{\sqrt{n}(\overline{X}-\mu_0)}{s\sqrt{\lambda_n(\rho)}} \overset{H_0}{\backsim} \mathsf{St}\left(n-1\right). \tag{29}$$

Table 3 - Type I error probability of $T_1(\alpha)$, **Misspecification** $Corr(X_i,X_j)=\rho^{\vert i-j\vert},\ -1<\rho<1$					
$0<\rho_+<1$	$\sqrt{\lambda_n(\rho_+)}$	$\alpha^*(\rho_+)$	$\alpha^*(\rho_-)$	$\sqrt{\lambda_n(\rho_-)}$	$-1<\rho_-<0$
0.0	1.0	.050	.050	1.0	0.0
.05	1.057	.060	.043	0.956	−.05
.1	1.111	.069	.035	0.909	−.1
.2	1. 232	.090	.023	0.821	−.2
.3	1.371	.114	.013	0.738	−.3
.4	1.538	.142	.007	0.659	−.4
.5	1.747	.172	.003	0.582	−.5
.6	2.022	.207	.001	0.505	−.6
.7	2.416	.247	.0001	0.426	−.7
.8	3.069	.295	.00000	0.341	−.8
.9	4.563	.358	.00000	0.240	−.9
.99	16.881	.461	.00000	0.0905	−.99

In view of (29), the *actual* type I error (α^*) is likely to be different from the *nominal* value (α). To find the **actual type I error probability** we need to evaluate the tail area of the distribution of $\tau(\mathbf{X})$ beyond c_α=1.66. In view of (29) one can deduce:

$$\alpha^*=\mathbb{P}(\tau(\mathbf{X})>c_\alpha;H_0)=\mathbb{P}\left(Z>\frac{c_\alpha}{\sqrt{\lambda_n(\rho)}};\mu=\mu_0\right), \quad \text{where } Z\backsim \mathsf{St}\left(n-1\right).$$

The results in table 3 show that the discrepancy between the actual and nominal type I error probability depends crucially on the sign of ρ.

A. For $0<\rho_+<1$ the actual type I error probability $\alpha^*(\rho_+)$ *increases* as $\rho_+ \to 1$.

B. For $-1<\rho_-<0$ the actual type I error probability $\alpha^*(\rho_-)$ *decreases* as $\rho_- \to -1$.

It must be emphasized that in both cases the reliability of the t-test in (16) is undermined in so far as the actual error probability is different from the nominal one. Its assumed trustworthiness relating to the type I error (α) has been compromised. One will apply the t-test in (16) thinking that it will reject a true null hypothesis only 5% of the time when, in fact, its erroneous rejection frequency is either much higher or much lower! In both cases, not knowing α^* will lead to unreliable inferences.

Turning our attention to the type II error probability, in view of (24) and (26), the *actual* distribution of $\tau(\mathbf{X})$ under H_1 is now:

$$\tau(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}-\mu_0)}{s} \overset{\mu=\mu_1}{\backsim} \sqrt{\lambda_n(\rho)}\mathsf{St}\left(\delta;n-1\right), \quad \text{or } \tau^*(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}-\mu_0)}{s\sqrt{\lambda_n(\rho)}} \overset{\mu=\mu_1}{\backsim} \mathsf{St}\left(\delta;n-1\right), \quad (30)$$

where $\delta=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$. This suggests that the *actual power* of the t-test in (16) should be evaluated (for $Z \backsim \mathsf{St}\left(n-1\right)$) using:

$$\pi^*(\mu_1)=\mathbb{P}(\tau(\mathbf{X})>c_\alpha;H_1(\mu_1))=\mathbb{P}\left(Z>\tfrac{1}{\sqrt{\lambda_n(\rho)}}\left[c_\alpha-\tfrac{\sqrt{n}(\mu_1-\mu_0)}{s}\right];\mu=\mu_1\right).$$

In direct analogy to the significance level, the presence of Markov dependence affects the power of the t-test differently depending on the sign of ρ.

4.4 Positive dependence and the power of the t-test

Table 4 - Power $\pi^*(\mu_1)$ of $T_1(\alpha)$ under Misspecification

ρ	$\sqrt{\lambda_n(\rho)}$	$\pi^*(.01)$	$\pi^*(.02)$	$\pi^*(.05)$	$\pi^*(.1)$	$\pi^*(.15)$	$\pi^*(.2)$	$\pi^*(.3)$	$\pi^*(.4)$
0.0	1	.061	.074	.121	.258	.437	.637	.911	.991
.05	1.057	.072	.085	.138	.267	.440	.626	.896	.985
.1	1.111	.082	.096	.149	.277	.443	.620	.885	.981
.2	1.232	.104	.119	.174	.297	.448	.608	.860	.970
.3	1.371	.129	.145	.200	.316	.454	.598	.835	.955
.4	1. 538	.156	.172	.226	.334	.459	.587	.807	.934
.5	1.747	.187	.203	.254	.353	.464	.567	.778	.908
.6	2.022	.221	.236	.284	.372	.469	.556	.745	.875
.7	2.416	.260	.274	.316	.393	.474	.515	.710	.832
.8	3.069	.306	.318	.353	.415	.479	.544	.668	.776
.9	4.563	.367	.375	.400	.443	.486	.530	.615	.695
.99	16.881	.463	.466	.473	.484	.496	.508	.532	.555

In table 4 the power of the t-test is evaluated for different discrepancies $(\mu_1-\mu_0)$ and different positive values of ρ. The main conclusion that emerges from this table is that the UMP property of the t-test is ruined under positive Markov dependence. To be more specific, under this misspecification, the t-test has become a very unreliable procedure because as $\rho \to 1$, its capacity to detect *small discrepancies* $(.01, .02, .05, .1, .15)$ *increases* (hypersensitized), but its capacity to detect *large*

discrepancies $(.2, .3, .4, ...)$ *decreases* (desensitized). The threshold: $\mu^{\dagger}=\mu_0+\frac{c_\alpha s}{\sqrt{n}}$, separating 'small' $(\mu_1 \leq \mu^{\dagger})$ from 'large' $(\mu_1 > \mu^{\dagger})$ discrepancies is incidental, depending on: (i) the prespecified α, (ii) the magnitude of s, and (iii) the sample size n; in the above example $\mu^{\dagger}$=0.166. For 'small' discrepancies $(\mu_1 \leq \mu^{\dagger})$ the power increases toward an *upper bound* of .5, and for 'large' discrepancies $(\mu_1 > \mu^{\dagger})$ the power decreases toward the *lower bound* of .5 as $\rho \to 1$. When H_0 is *rejected* one would not know if it's because H_0 is false or a hypersensitized t-test is picking up on truly 'trivial' discrepancies.

Envisioning the test metaphorically as a smoke alarm and the null as 'no fire', the t-test has been transformed into a *defective smoke alarm* which has the tendency to go off when burning toast, but it will not be triggered off by the smoke generated when a house is fully ablaze; see Mayo (1996), p. 403. If we combine this with its enhanced proclivity to go off when nothing is burning, the t-test has become a (practically) useless smoke alarm!

4.5 Negative dependence and the power of the t-test

In direct analogy to table 4, table 5 shows the power of the t-test evaluated for different discrepancies $(\gamma=\mu_1-\mu_0)$ as they relate to different *negative* values of ρ. The main conclusion that emerges from this table is that, similarly to table 4, the UMP property of the t-test has been ruined by the presence of negative Markov dependence. At first sight it looks as though the probativeness of this test has been enhanced because it its capacity to detect *small discrepancies* $(.01, .02, .05, .1, .15)$ *decreases*, but its capacity to detect *large discrepancies* $(.2, .3, .4, ...)$ *increases*.

Table 5 - Power $\pi^*(\mu_1)$ of $T_1(\alpha)$ under Misspecification

ρ	$\sqrt{\lambda_n(\rho)}$	$\pi^*(.01)$	$\pi^*(.02)$	$\pi^*(.05)$	$\pi^*(.1)$	$\pi^*(.15)$	$\pi^*(.2)$	$\pi^*(.3)$	$\pi^*(.4)$
0.0	1	0.061	.074	.121	.258	.437	.637	.911	.991
-.05	0.956	.053	.065	.114	.246	.434	.639	.918	.992
-.1	0.909	.045	.056	.102	.235	.430	.645	.928	.994
-.2	0.821	.030	.039	.080	.212	.423	.660	.947	.997
-.3	0.738	.019	.025	.060	.187	.414	.677	.964	.999
-.4	0.659	.010	.015	.041	.160	.404	.696	.978	.9999
-.5	0.582	.004	.007	.025	.130	.392	.720	.988	1.00
-.6	0.505	.001	.002	.012	.097	.376	.749	.995	1.00
-.7	0.426	.0002	.0004	.004	.062	.354	.787	.999	1.00
-.8	0.341	.0000	.0000	.0004	.028	.320	.839	.9999	1.00
-.9	0.240	.0000	.0000	.0000	.004	.253	.920	1.00	1.00
-.99	0.091	.0000	.0000	.0000	.0000	.041	.9999	1.00	1.00

A moment's reflection, however, indicates that this apparent improvement is illusory. For *large enough* values of $|\rho|$, the t-test has *very low capacity* to detect a whole range of discrepancies $(\mu_0 < \mu_1 \leq \mu^{\dagger})$ even if present. For 'small' discrepancies $(\mu_1 \leq \mu^{\dagger})$

the power decreases toward a *lower bound* of 0, and for 'large' discrepancies ($\mu_1 > \mu^\dagger$) the power increases toward an *upper bound* of 1.0 as $\rho \to -1$.

The problem is that when H_0 is *not rejected*, one does not know whether it's because H_0 is true or the discrepancy lies within the 'small' range ($\mu_0 < \mu_1 \leq \mu^\dagger$), and thus under the radar of this test.

Doesn't the increase in power for larger discrepancies ($\mu_1 > \mu^\dagger$) compensate for its insensitivity to smaller discrepancies? It does not, because the threshold $\mu^\dagger$ is incidentally determined and depends on unknown parameters; a *substantive* discrepancy of interest could easily be within the 'small' range! This argument calls into question the conventional wisdom as expressed by Staudte and Sheather (1990), that "we can live with negative dependence, but should not use the t-test in the presence of positive dependence" (ibid., p. 168). As argued in Mayo and Spanos (2006), the reliability of inference depends crucially on being able to ascertain correctly the relevant error probabilities associated with the particular inference.

5 Confidence Intervals and misspecification

As mentioned above, the sampling distribution underlying *Confidence Intervals* (CIs) for μ is that of the *pivotal quantity*:

$$h(\mathbf{X};\mu^*)=\tfrac{\sqrt{n}(\overline{X}-\mu^*)}{s} \overset{\text{TSN}}{\backsim} \mathsf{St}(n-1), \tag{31}$$

where the evaluation is under the 'true state of nature (tsn)'; μ^* being the 'true' value of μ. Using this distribution we can derive a $(1-2\alpha)$ two-sided CI of the form:

$$\mathbb{P}\left(\mu : \mu\in CI(\mathbf{X})\right)=\mathbb{P}\left(\overline{X}-c_\alpha(\tfrac{s}{\sqrt{n}}) \leq \mu^* < \overline{X}+c_\alpha(\tfrac{s}{\sqrt{n}})\right)=1-2\alpha, \tag{32}$$

of length $2c_\alpha(\frac{s}{\sqrt{n}})$; 2α denotes the *nominal coverage error probability*.

Example. For $\overline{x}$=0.6, α=.025, c_α=1.984, s=1 and n=100, the observed 95% CI (of length 0.397) is:

$$CI(\mathbf{x}_0)=[0.402,\ 0.798]\,. \tag{33}$$

However, when assumption [4] is false, and instead (21) is the appropriate assumption, any inference based on (32) is likely to be unreliable. This is because the actual sampling distribution of the *pivotal quantity* $h^*(\mathbf{X},\mu)$ is now:

$$h^*(\mathbf{X},\mu)=\tfrac{\sqrt{n}(\overline{X}-\mu^*)}{s\sqrt{\lambda_n(\rho)}} \overset{\text{TSN}}{\backsim} \mathsf{St}(n-1). \tag{34}$$

Hence, the *actual coverage probability* of the CI (of length $2c_\alpha(\frac{s\sqrt{\lambda_n(\rho)}}{\sqrt{n}})$) becomes:

$$\mathbb{P}\left(\mu:\mu\in CI(\mathbf{X})\right)=\mathbb{P}\left(\overline{X}-c_\alpha(\tfrac{s\sqrt{\lambda_n(\rho)}}{\sqrt{n}}) \leq \mu^* < \overline{X}+c_\alpha(\tfrac{s\sqrt{\lambda_n(\rho)}}{\sqrt{n}})\right)=1-2\alpha^*. \tag{35}$$

As in the case of hypothesis testing, the nature of unreliability that the presence of Markov dependence afflicts on the above CI depends on the sign of ρ. Let us consider the two cases separately.

5.1 Positive dependence and the observed CI

Example. In view of (35), when the Markov correlation is ρ=.8, $\sqrt{\lambda_n(\rho)}$=3.069, the *actual* .95 observed CI turns out to be:

$$CI(\mathbf{x}_0)=[-.009,\ 1.209]\,,$$

which is not only different from the *nominal* one in (33), but its *actual length* has increased by a factor of $\sqrt{\lambda_n(\rho)}$ to 1.20. In addition, the *actual coverage error probability* of the original CI $CI(\mathbf{X})$ is not 2α but $2\alpha^*$:

$$(1-2\alpha^*)=\mathbb{P}\left(-1.984\leq\frac{\sqrt{n}(\overline{X}-\mu^*)}{s}<1.984\right)=\mathbb{P}\left(-.647\leq\frac{\sqrt{n}(\overline{X}-\mu^*)}{s\sqrt{\lambda_n(\rho)}}<.647\right)=.481.$$

What we thought was an observed $CI(\mathbf{x}_0)$ ([0.402, 0.798]), arising from .95 coverage probability CI $CI(\mathbf{X})$, turns out to be located somewhere else ([−.009, 1.209]) and arising from a CI with *actual* coverage probability of only .481. The table 6 below shows how misleading the *nominal* CI can be for different values of ρ.

It is interesting to note that if the above observed CIs were to be used as *surrogate tests*, i.e. reject any null hypothesis for μ whose value falls outside the observed CI, the results of table 6 indicate substantial scope for misleading inferences for larger values of ρ; see Mayo and Spanos (2006).

Table 6 - CIs under Misspecification

ρ	$\sqrt{\lambda_n(\rho)}$	*Actual* observed CI	$(1-2\alpha^*)$
0.0	1	[0.402, 0.798]	.950
.05	1.057	[0.390, 0.810]	.936
.1	1.111	[0.380, 0.820]	.923
.2	1.232	[0.356, 0.844]	.889
.3	1.371	[0.328, 0.872]	.849
.4	1.538	[0.295, 0.905]	.800
.5	1.747	[0.253, 0.947]	.741
.6	2.022	[0.199, 1.001]	.671
.7	2.416	[0.121, 1.079]	.586
.8	3.069	[−.009, 1.209]	.481
.9	4.563	[−0.305, 1.505]	.335
.99	16.881	[−2.749, 3.949]	.093

5.2 Negative dependence and the observed CI

Example. $\overline{x}$=0.6, α=.025, c_α=1.984, s=1 and n=100, and the nominal .95 observed CI is:

$$CI(\mathbf{x}_0)=[0.402,\ 0.798]\,, \tag{36}$$

of length 0.397. In view of (35), when the Markov correlation is ρ=−.8, $\sqrt{\lambda_n(\rho)}$=0.341, the *actual* .95 observed CI is:

$$CI(\mathbf{x}_0)=[.532,\ .668]\,.$$

This is not only different from the *nominal* one in (33), but the *actual length* of the observed CI has decreased by a factor of $\sqrt{\lambda_n(\rho)}$ to 0.135. In addition, the actual coverage error probability of the original $CI(\mathbf{X})$ is now $2\alpha^*$=0 since:

$$(1-2\alpha^*)=\mathbb{P}\left(-1.984\leq\frac{\sqrt{n}(\overline{X}-\mu^*)}{s}<1.984\right)=\mathbb{P}\left(-5.818\leq\frac{\sqrt{n}(\overline{X}-\mu^*)}{s\sqrt{\lambda_n(\rho)}}<5.818\right)=1.0.$$

Table 7 reports all the observed CIs together with their actual coverage probabilities for different values of ρ within the range $-1<\rho<0$. As one can see, the actual coverage probability $(1-2\alpha^*)\to 1$, and the actual observed CIs narrow down as $\rho\to -1$.

Table 7 - CIs under Misspecification

ρ	$\sqrt{\lambda_n(\rho)}$	*Actual* observed CI	$(1-2\alpha^*)$
0.0	1	[0.402, 0.798]	.950
−.05	0.956	[0.410, 0.790]	.959
−.1	0.909	[0.420, 0.780]	.969
−.2	0.821	[0.437, 0.763]	.983
−.3	0.738	[0.454, 0.746]	.992
−.4	0.659	[0.469, 0.731]	.997
−.5	0.582	[0.485, 0.715]	.999
−.6	0.505	[0.500, 0.700]	1.00
−.7	0.426	[0.515, 0.685]	1.00
−.8	0.341	[0.532, 0.668]	1.00
−.9	0.240	[0.552, 0.647]	1.00
−.99	0.091	[0.582, 0.618]	1.00

In summary, the presence of Markov dependence in the sample has rendered inferences based on CIs unreliable in two interrelated ways. *First*, the observed CI $CI(\mathbf{x}_0)$ is misplaced because the actual sampling distribution of the pivotal quantity is different from the assumed one. *Second*, the actual coverage probability is different from the nominal coverage. For positive dependence $(0<\rho<1)$ the actual is less than the nominal but for negative dependence $(-1<\rho<0)$ the opposite is true.

The main conclusion emerging from tables 6 and 7 is that when one uses the nominal CI (33) as a basis of inference, it's likely that these inferences will be unreliable because the *actual* $(2\alpha^*)$ and *nominal* (2α) *coverage error probabilities* are different. The fact that in the case of negative dependence $(-1<\rho<0)$ the actual coverage probability is greater than the nominal, although seemingly a good thing, it still contributes to the unreliability of inference which stems from the inability to ascertain the actual error probabilities correctly; see Mayo and Spanos (2006).

6 Re-establishing the reliability of inference

In this section we consider the question of addressing the unreliability of inference problem by respecifying the original model to account for the misspecification in

question. To bring out the potential problems we compare the optimal test in the context of the respecified model with a modification of the original t-test in (16) to allow for the effects of misspecification on its sampling distributions under both H_0 and H_1. In practice the latter procedure is favored because respecification is often considered to raise even more daunting problems than the misspecification itself; see Kennedy (2008). Spanos (1986, 1989, 2000) has proposed a general way to respecify statistical models, known as the *Probabilistic Reduction* (PR) approach, which renders respecification much more manageable by partitioning the set of all possible models. Moreover, the use of implicit statistical parametrizations in conjunction with the PR approach, enables one to test the original hypotheses concerning the mean in the context of the respecified model.

6.1 Respecification: the AR(1) model

Let us return to the question of addressing the misspecification problem in the case where assumption [4] (table 1) is false, and instead the sample is *Markov dependent*:

$$Corr(X_i, X_j)=\rho^{|i-j|},\ \text{for}\ -1<\rho<1,\ \text{for all}\ i\neq j,\ i,j=1,...n. \tag{37}$$

This departure suggests that the simple Normal model is no longer appropriate and the question that naturally arises is whether one can specify a more appropriate statistical model in the context of which (42) can be tested reliably.

As argued in Spanos (1999), ch. 15, under (37) the appropriate statistical model suggested by the Probabilistic Reduction approach comes in the form of the **Autoregressive (AR(1)) model,** as specified in table 8.

Table 8 - Normal AutoRegressive Model

Statistical GM:		$X_t=\alpha_0+\alpha_1X_{t-1}+\varepsilon_t,\ t\in\mathbb{N}.$	
[1]	Normality:	$(X_t \mid X_{t-1})\backsim \mathsf{N}(.,.),$	$t\in\mathbb{N}.$
[2]	Linearity:	$E(X_t \mid X_{t-1})=\alpha_0+\alpha_1X_{t-1},$	
[3]	Homoskedasticity:	$Var(X_t \mid X_{t-1})=\sigma_0^2,$	
[4]	Markov dependence:	$\{X_t,\ t\in\mathbb{N}\}$ is a Markov process,	
[5]	t-invariance:	$(\alpha_0,\alpha_1,\sigma_0^2)$ are *not* changing with t,	

(38)

The details of this respecification can be summarized as follows.

The simple Normal model is the appropriate model when the stochastic process $\{X_t,\ t\in\mathbb{N}\}$ is Normal, Independent and Identically Distributed (NIID), because its joint distribution can be reduced to a product of marginal distributions as follows:

$$D(X_1,X_2,\cdots,X_n;\boldsymbol{\phi})\overset{\mathsf{I}}{=}\prod_{t=1}^{n}D_t(X_t;\boldsymbol{\varphi}_t)\overset{\mathsf{IID}}{=}\prod_{t=1}^{n}D(X_t;\boldsymbol{\varphi}),\ \text{for all}\ \mathbf{x}\in\mathbb{R}^n, \tag{39}$$

The statistical model (see table 1) is specified exclusively in terms of $D(X_t;\boldsymbol{\varphi})$, assumed to be Normal, with a statistical Generating Mechanism (GM):

$$X_t=E(X_t \mid \mathcal{D}_0)+u_t,\ t\in\mathbb{N},$$

where $\mathcal{D}_0=\{S,\varnothing\}$ is the trivial field; S and $\varnothing$ denote the sure and impossible events.

If one were to replace the Independence assumption with that of Markov (M) dependence, the probabilistic reduction in (39) is no longer appropriate. Replacing Independence (I) with Markov (M) dependence and extending Identically Distributed (ID) to Stationarity (S), the appropriate reduction takes the form:

$$\begin{aligned} D(X_1,X_2,\cdots,X_n;\boldsymbol{\phi}) \overset{\mathsf{M}}{=} & D_1(X_1;\boldsymbol{\psi}_1)\textstyle\prod_{t=2}^{n} D_t(X_t|X_{t-1};\boldsymbol{\psi}_t)= \\ & \overset{\mathsf{M\&S}}{=} D_1(X_1;\boldsymbol{\psi}_1)\textstyle\prod_{t=2}^{n} D(X_t|X_{t-1};\boldsymbol{\psi}),\ \text{for all } \mathbf{x}\in\mathbb{R}^n. \end{aligned} \tag{40}$$

This gives rise to the Autoregressive (AR(1)) model (see table 8), specified in terms of $D(X_t|X_{t-1};\boldsymbol{\psi})$, with a statistical Generating Mechanism (GM):

$$X_t=E(X_t \mid \sigma(X_{t-1}))+\varepsilon_t,\ t\in\mathbb{N},$$

where $\sigma(X_{t-1})$ is the sigma-field generated by X_{t-1}. Due to Markovness, the joint distribution can be defined in term of the bivariate distribution $D(X_t,X_{t-1};\boldsymbol{\phi})$:

$$\begin{pmatrix} X_t \\ X_{t-1} \end{pmatrix} \backsim \mathsf{N}\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma(0) & \sigma(1) \\ \sigma(1) & \sigma(0) \end{pmatrix}\right),$$

which can be used to derive the *statistical GM:*

$$X_t=\alpha_0+\alpha_1X_t+\varepsilon_t,\ t\in\mathbb{N},$$

with the underlying *parametrization* $(\alpha_0,\alpha_1,\sigma_0^2)$:

$$\begin{aligned} &\alpha_0=E(X_t)-\alpha_1E(X_{t-1})=\mu(1-\alpha_1)\in\mathbb{R},\quad \alpha_1=\tfrac{Cov(X_t,X_{t-1})}{Var(X_{t-1})}=\left(\tfrac{\sigma(1)}{\sigma(0)}\right)=\rho\in(-1,1), \\ &\sigma_0^2=\sigma(0)-\tfrac{[\sigma(1)]^2}{\sigma(0)}=\sigma^2\left(1-\alpha_1^2\right)\in\mathbb{R}_+. \end{aligned} \tag{41}$$

This brings out the relationship between the AR(1) parameters $(\alpha_0,\alpha_1,\sigma_0^2)$ and (μ,σ^2), the parameters of the simple Normal model (table 1), as well as the Markov dependence parameter ρ; see Spanos (1999). The presence of μ, as part of the implicit parametrization of (α_0,α_1), renders possible the testing of the hypotheses (42) in the context of the AR(1) model (38).

6.2 Testing the mean in the context of the AR(1) model

The first question one needs to answer is how the original hypotheses of interest (15):

$$H_0:\mu\leq\mu_0 \quad \text{vs.} \quad H_1:\mu>\mu_0, \tag{42}$$

can be embedded into the AR(1) model. In light of (41), the idea is to find a reparameterization that isolates the parameter of interest μ. The obvious reparameterization $\mu=\frac{\alpha_0}{(1-\alpha_1)}$ will give rise to a *Fieller type* problem (see Cox, 1967), which can be addressed using a two step *reparametrization*.

Step 1: Reparametrize the original statistical GM into:

$$\Delta X_t=\beta_0+\beta_1X_{t-1}+v_t,\ t\in\mathbb{N}, \tag{43}$$

where the parameters (β_0, β_1) take the form:

$$\beta_0=E(\Delta X_t)-\beta_1 E(X_{t-1})=-\beta_1\mu, \qquad \beta_1=\tfrac{Cov(\Delta X_t, X_{t-1})}{Var(X_{t-1})}=(\alpha_1-1)<0, \tag{44}$$

$$\mu=-\tfrac{\beta_0}{\beta_1} \Rightarrow [\beta_0+\beta_1\mu]=0. \tag{45}$$

How does one relate (45) to (42)?

Step 2: Reparametrize the statistical GM (43) into the *null Autoregression*:

$$\Delta X_t=\gamma_0+\beta_1\left(X_{t-1}-\mu_0\right)+w_t, \quad t\in\mathbb{N}, \tag{46}$$

where, by definition:

$$\gamma_0 \;=E(\Delta X_t)-\beta_1\left(E(X_{t-1})-\mu_0\right)=-\beta_1\left(\mu-\mu_0\right)=\left(\beta_0+\beta_1\mu_0\right). \tag{47}$$

Hence, in view of (44), (42) can be equivalently recast into:

$$H_0: \gamma_0\leq 0 \quad \text{against} \quad H_1: \gamma_0>0. \tag{48}$$

The test statistic for (48) takes the generic form: $\tau_0(\mathbf{X})=\frac{\widehat{\gamma}_0}{\sqrt{\widehat{Var}(\widehat{\gamma}_0)}}$. Let us unpack this.

Re-writing (46) in the form:

$$y_t=\gamma_0+\beta_1 Z_t+w_t, \quad \text{where} \quad y_t=\Delta X_t,\ Z_t=(X_{t-1}-\mu_0), \tag{49}$$

the least-squares (and MLE) estimators of (γ_0, β_1) are:

$$\widehat{\gamma}_0=\overline{y}-\widehat{\beta}_1\overline{Z}, \qquad \widehat{\beta}_1=\frac{\sum_{t=1}^n(y_t-\overline{y})\left(Z_t-\overline{Z}\right)}{\sum_{t=1}^n\left(Z_t-\overline{Z}\right)^2},$$

with variances:

$$Var(\widehat{\gamma}_0)=\sigma_0^2\left(\tfrac{1}{n}+\frac{\overline{Z}^2}{\sum_{t=1}^n\left(Z_t-\overline{Z}\right)^2}\right), \qquad Var(\widehat{\beta}_1)=\sigma_0^2\left(\sum_{t=1}^n\left(Z_t-\overline{Z}\right)^2\right)^{-1}.$$

Circularity assumption. To simplify the algebra we will assume that the Markov process $\{X_t,\ t\in\mathbb{N}\}$ is *circular*, i.e. $x_0=x_n$; see Anderson (1971). The result of this simplification is that the sample moments:

$$\begin{array}{l}\overline{X}_n=\frac{1}{n}\sum_{t=1}^n X_t, \quad \overline{X}_{n-1}=\frac{1}{n}\sum_{t=1}^n X_{t-1} \;\Rightarrow \overline{X}_n=\overline{X}_{n-1}:=\overline{X},\\ s_n^2=\frac{1}{n-1}\sum_{t=1}^n\left(X_t-\overline{X}_n\right)^2, \;\; s_{n-1}^2=\frac{1}{n-1}\sum_{t=1}^n\left(X_{t-1}-\overline{X}_{n-1}\right)^2 \;\Rightarrow_n^2=s_{n-1}^2:=s^2.\end{array}$$

This simplification will render the results that follow *approximate,* in general, but it will be illuminating for our purposes because it enables one to relate directly the test statistic $\tau_0(\mathbf{X})$ with $\tau(\mathbf{X})$ and $\tau^*(\mathbf{X})$ of sections 3-4. In view of the fact that:

$$\begin{array}{rl}\sum_{t=1}^n\left(y_t-\overline{y}\right)\left(Z_t-\overline{Z}\right) & =\sum_{t=1}^n[\left(X_t-\overline{X}\right)-\left(X_{t-1}-\overline{X}\right)]\left(X_{t-1}-\overline{X}\right)= \\ & =\sum_{t=1}^n\left(X_t-\overline{X}\right)\left(X_{t-1}-\overline{X}\right)-\sum_{t=1}^n\left(X_{t-1}-\overline{X}\right)^2, \\ \sum_{t=1}^n\left(Z_t-\overline{Z}\right)^2 & =\sum_{t=1}^n\left(X_{t-1}-\overline{X}\right)^2, \quad \overline{Z}=\left(\overline{X}-\mu_0\right),\end{array}$$

we can deduce that:

$$\widehat{\beta}_1 \;=\frac{\sum_{t=1}^n\left(X_t-\overline{X}\right)\left(X_{t-1}-\overline{X}\right)}{\sum_{t=1}^n\left(X_{t-1}-\overline{X}\right)^2}-1=(\widehat{\alpha}_1-1)<0, \tag{50}$$

$$\widehat{\gamma}_0(\mu_0) \;=\left[1-\frac{\sum_{t=1}^{n}\left(X_t-\overline{X}\right)\left(X_{t-1}-\overline{X}\right)}{\sum_{t=1}^{n}\left(X_{t-1}-\overline{X}\right)^2}\right]\left(\overline{X}-\mu_0\right)=(1-\widehat{\alpha}_1)\left(\overline{X}-\mu_0\right), \tag{51}$$

where $\widehat{\alpha}_1$ is the least-squares estimator of the Markov coefficient α_1 :

$$\widehat{\alpha}_1 \;=\frac{\sum_{t=1}^{n}\left(X_t-\overline{X}\right)\left(X_{t-1}-\overline{X}\right)}{\sum_{t=1}^{n}\left(X_{t-1}-\overline{X}\right)^2}=\widehat{\rho}. \tag{52}$$

The variance of the least-squares estimator $\widehat{\gamma}_0$ takes the form:

$$Var(\widehat{\gamma}_0(\mu_0)) \;=\sigma_0^2\left[\frac{\sum_{t=1}^{n}\left(X_{t-1}-\overline{X}\right)^2+n\left(\overline{X}-\mu_0\right)^2}{n\sum_{t=1}^{n}\left(X_{t-1}-\overline{X}\right)^2}\right]=\frac{\sigma_0^2}{n}\left[\frac{\sum_{t=1}^{n}(X_{t-1}-\mu_0)^2}{\sum_{t=1}^{n}\left(X_{t-1}-\overline{X}\right)^2}\right],$$

using the equality $\sum_{t=1}^{n}\left(X_{t-1}-\overline{X}\right)^2+n\left(\overline{X}-\mu_0\right)^2=\sum_{t=1}^{n}\left(X_{t-1}-\mu_0\right)^2$. This implies that:

$$\frac{\widehat{\gamma}_0(\mu_0)}{\sqrt{Var(\widehat{\gamma}_0)}} \;=\frac{(1-\widehat{\alpha}_1)\left(\overline{X}-\mu_0\right)}{\sqrt{\frac{\sigma_0^2}{n}\left[\frac{\sum_{t=1}^{n}(X_{t-1}-\mu_0)^2}{\sum_{t=1}^{n}(X_{t-1}-\overline{X})^2}\right]}},$$

which can be transformed into a test statistic by replacing σ_0^2 with an unbiased and consistent estimator:

$$s_0^2 \;=\frac{1}{n-2}\sum_{t=1}^{n}(\Delta X_t-\widehat{\gamma}_0-\widehat{\beta}_1\left(X_{t-1}-\mu_0\right))^2=(1-\widehat{\alpha}_1^2)\left[\frac{1}{n-2}\sum_{t=1}^{n}\left(X_{t-1}-\overline{X}\right)^2\right], \tag{53}$$

giving rise to the test statistic:

$$\begin{aligned}\tau_0(\mathbf{X}) \;&=\frac{\left(\sqrt{\sum_{t=1}^{n}\left(X_{t-1}-\overline{X}\right)^2}\right)}{\sqrt{\sum_{t=1}^{n}(X_{t-1}-\mu_0)^2}}\left[\frac{(1-\widehat{\alpha}_1)\sqrt{n}\left(\overline{X}-\mu_0\right)}{\sqrt{\frac{1}{n-2}(1-\widehat{\alpha}_1^2)\sum_{t=1}^{n}\left(X_{t-1}-\overline{X}\right)^2}}\right]= \\ &=\frac{(1-\widehat{\alpha}_1)}{\sqrt{(1-\widehat{\alpha}_1^2)}}\left[\frac{\sqrt{n}\left(\overline{X}-\mu_0\right)}{s(\mu_0)}\right]=\frac{\sqrt{n}\left(\overline{X}-\mu_0\right)}{s(\mu_0)\sqrt{r_n(\widehat{\rho})}},\end{aligned} \tag{54}$$

$$s^2(\mu_0)=\frac{1}{n-2}\sum_{t=1}^{n}\left(X_{t-1}-\mu_0\right)^2, \qquad r_n(\rho) \;=\frac{(1+\rho)}{(1-\rho)},$$

since $\rho=\alpha_1$; see (41). The relevant sampling distributions of $\tau_0(\mathbf{X})$ are:

$$\tau_0(\mathbf{X})\overset{H_0}{\backsim}\mathsf{St}(n-2), \quad \tau_0(\mathbf{X})\overset{H_1}{\backsim}\mathsf{St}(\delta_0;n-2), \quad \delta_0=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma(\mu_0)\sqrt{r_n(\rho)}}, \text{ for } \mu_1>\mu_0,$$

where $\sigma^2(\mu_0)\overset{H_1}{=}E(s^2(\mu_0))$. These distributions can be used to define the *relevant* (in the context of the AR(1) model) *t-test* $T_0(\alpha)$, specified in terms of:

$$\tau_0(\mathbf{X})=\frac{\sqrt{n}\left(\overline{X}-\mu_0\right)}{s(\mu_0)\sqrt{r_n(\widehat{\rho})}}, \qquad C_1^0(\alpha)=\{\mathbf{x}:\tau_0(\mathbf{x})>c_\alpha\}. \tag{55}$$

What renders this test's error probabilities *relevant* is the conjunction of (a) its optimality in the context of (b) a statistically adequate AR(1) model.

6.3 Comparing the relevant, modified and original t-tests

As one can see from (55), this test is directly related to the *original* (in the context of the simple Normal model) *t-test* $T_1(\alpha)$:

$$\tau(\mathbf{X})=\frac{\sqrt{n}(\overline{X}-\mu_0)}{s}, \qquad C_1(\alpha)=\{\mathbf{x}:\tau(\mathbf{x})>c_\alpha\}, \tag{56}$$

as well as the *modified* (to allow for the Markov dependence) *t-test* in the context of the simple Normal model based on $T^*(\alpha)$:

$$\tau^*(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}-\mu_0)}{s\sqrt{\lambda_n(\widehat{\rho})}}, \qquad C_1^*(\alpha)=\{\mathbf{x}:\tau^*(\mathbf{x})>c_\alpha\}, \tag{57}$$

where $\widehat{\rho}=\widehat{\alpha}_1$ (see (52)) and $\lambda_n(\rho)$ is given in (28).

It is important to emphasize, at this stage, that for evaluating the *nominal* error probabilities we used (56), and for evaluating the *actual* error probabilities we used a hybrid of (56) and (57); the sampling distributions of $\tau^*(\mathbf{X})$ under H_0 and H_1, in conjunction with $C_1(\alpha)$ (not $C_1^*(\alpha)$) to demonstrate that the original t-test is, inadvertently, highly defective. The 'modified' (57), as well as the 'optimal' t-test (55), do not share these major defects, but are they equally reliable?

Consider the scenario where one uses the modified t-test (57) as a way to address the reliability of inference problem; its implementation requires only the estimation of ρ using (52). The main differences between the two tests, (55) and (57), are:

(i) the degrees of freedom, $(n-2)$ vs. $(n-1)$,

(ii) the scaling factors, $\sqrt{r_n(\rho)}$ vs. $\sqrt{\lambda_n(\rho)}$,

(iii) the estimators of σ^2, $s^2(\mu_0)$ vs. s^2;

$s^2(\mu_0)$ can be viewed as the *constrained* (under H_0) and s^2 as the *unconstrained* estimator of σ^2. Moreover, one can show that:

$$r_n(\rho)\leq\lambda_n(\rho), \text{ for } -1<\rho<1, \qquad s^2(\mu_0)\geq s^2, \text{ for large enough } n, \tag{58}$$

where the latter follows from identity (**??**). It's clear from the inequalities in (58) that the differences (i)-(iii) will affect, not only the type I error, but also the power of the two tests, rendering them dissimilar.

Hence, using the 'modified' t-test in (57) one *will not,* in general, address the *reliability of inference* problem. The primary reason being that the 'modified' t-test is unlikely to be an 'optimal' test in the context of the new premises; in the next section a more extreme case is discussed. Some of the implications of this argument on the traditional use of robustness are unfolded in the next section.

7 Robustness and the reliability of inference

Robustness, first defined by Box (1953), refers to the sensitivity of inference procedures (estimators, tests, predictors) to departures from the model assumptions. A procedure is said to be robust against certain departure(s) from the model assumptions when the inference is not '*very sensitive*' to the presence of '*modest departures*' from the premises; some assumptions 'do not hold, to a greater or lesser extent'. Since the premises of inference are never exactly 'true', it seems only reasonable that one should evaluate the sensitivity of the inference method to 'modest departures'. At the level of hypothetical reasoning, evaluating the difference between the nominal and actual error probabilities, provides a very natural way to assess the sensitivity of one's inference tools to potential departures from the premises. Establishing the degree of 'insensitivity' that renders the reliability of an inference procedure 'tolerable' in specific circumstances is an extremely difficult task, but that is not the only difficulty facing one invoking robustness arguments in practice.

7.1 Attesting robustness at the practical level

The results reported in tables 3-7 indicate most clearly that being able to evaluate whether the sensitivity of an inference procedure is within tolerable limits (however decided), requires precise and reliable information concerning the *form* and *magnitude of the departure*; vague and imprecise information will not do, unless one remains at the level of hypothetical reasoning. This raises questions concerning the use of robustness arguments in practice when such information is *not* usually readily available. As argued in section 5, securing reliable information concerning the form and magnitude of departures in practice presupposes thorough **misspecification (M-S) testing** and **respecification**.

In the case under discussion this amounts to testing assumptions [1]-[4] of the simple Normal model thoroughly and ensuring that the only departure present is with respect to [4] *independence*. This can be detected using a variety of M-S tests including the von Neumann ratio, the Durbin-Watson and the runs (up and down) tests; see Spanos (1999), ch. 15. The first two M-S tests assume an alternative of the particular Markov form [5] (see (37)), and their test statistics depend crucially on $\widehat{\rho}$ as given in (52). However, rejecting the null hypothesis [4], using any one of the three M-S tests, only provides evidence *against* [4], it does not provide evidence *for* [5]; see Spanos (1999). Indeed, all three tests would reject [4] even if the form of dependence present is different from [5], say:

$$[6]\ Corr(X_i,X_j)=\rho,\ \text{for}\ \ -\tfrac{1}{(n-1)}<\rho<1,\ \ \text{for all}\ i\neq j,\ i,j=1,...n. \tag{59}$$

Note that (59) is the type of dependence assumed by the **random effects model**; see Greene (2008). Moreover, $\widehat{\rho}$ the estimated value of ρ, can be very misleading if one does not have a clear idea as to the form of dependence present. Let us elaborate on this.

Numerical example. One might consider the presence of $\widehat{\rho}$=0.1 as practically 'harmless' if the form of dependence assumed present is [5], because, glancing at tables 3 and 4, the effect on the type I error and power of the t-test (56) does not seem substantial: $\alpha^*(\rho{=}.1){=}.069$ ($\alpha{=}.05$), and the power distortions range from $\pi^*(.05){-}\pi(.05){=}.028$ to $\pi^*(.3){-}\pi(.3){=}{-}.026$. However, if the form of dependence present is actually [6], then the effects on the reliability of the t-test (56) are *significantly greater*. As shown by Arnold (1990), the relevant sampling distributions of $\tau_0^*(\mathbf{X})$ under [6] are:

$$\text{(i)}\ \tau_0^*(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}-\mu_0)}{s\sqrt{\ell_n(\rho)}}\overset{H_0}{\backsim}\mathsf{St}(n{-}1)\quad \text{(ii)}\ \tau_0^*(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}-\mu_0)}{s\sqrt{\ell_n(\rho)}}\overset{H_0}{\backsim}\mathsf{St}(\delta;n{-}1), \tag{60}$$

where $\delta{=}\frac{\sqrt{n}(\mu_2-\mu_0)}{\sigma\sqrt{\ell_n(\rho)}}$ and $\ell_n(\rho){=}(\frac{1+(n-1)\rho}{(1-\rho)})$. From (60) the size distortion is:

$$\alpha^*(\rho{=}.1){=}.317\ (\alpha{=}.05),$$

which is substantial, and the power distortions are sizeable, ranging from:

$$\pi^*(.01){-}\pi(.01){=}.266\ \text{to}\ \pi^*(.3){-}\pi(.3){=}-.291.$$

This should also serve as a warning against misleading arguments based on 'slight departures' from the premises can only have 'minor effects' on the reliability of inference; what is 'slight' and 'minor' depend crucially on the model assumptions contemplated.

In view of these potential differences in the reliability of inference, one needs to go beyond M-S testing and establish the form of dependence present. In the above case this will require thorough M-S testing of assumptions [1]-[5] of the AR(1) model (table 8) in order to establish its statistical adequacy; see Spanos (1999), ch. 15. It is important to note that if [6] is the correct form of dependence, the AR(1) model will be misspecified; its residuals will exhibit 'lingering' dependence. Hence, determining the presence of Markov dependence, and attaining a reliable estimate of the sign and magnitude of ρ, takes a lot of systematic and thorough statistical analysis. Having gone through exhaustive M-S testing and respecification, it makes little sense to return to the original statistical model to consider the question of robustness; assessing the sensitivity of an inference method whose optimality was established on the basis of original statistical model, which ignored the departure. It makes more sense to test the hypothesis of interest in the context of the statistically adequate model, and avoid the pitfalls of using sub-optimal inference procedures, as shown in section 5.

7.2 Invoking generic robustness arguments in practice

The above discussion also raises the broader question of 'how one can utilize a number of established robustness results in practice'. It is well known that the t-test is *not* robust to the presence of dependence, and the above discussion demonstrated that. There are numerous papers in the statistical literature, however, demonstrating the robustness of the t-test to departures from the Normality assumption [1] (see table 1) toward other symmetric distributions; see Geary (1936), Box (1953), Box and Andersen (1955), Scheffe (1959) inter alia. To utilize this robustness result one needs to test and reject the Normality assumption as well as establish the symmetry of the underlying distribution. For the sake of the argument let us assume that one was able to do all that and conclude, after thorough M-S testing and respecification, that the underlying distribution is closer to the Uniform rather than the Normal. *Can one assume under this scenario that inferences based on the t-test are likely to be reliable?* The answer is 'not necessarily' because using the t-test ignores the optimality issue.

Assuming that the *uniform* (not the Normal) is the appropriate distribution, i.e.

$$X_k \backsim \mathsf{U}(a-\mu, a+\mu), \qquad f(x)=\tfrac{1}{2\mu},\ (a-\mu) \leq x \leq (a+\mu)\,,\ \mu > 0,$$

the robustness of the t-test to symmetric departures is *not* very comforting in this case. This is because the optimal test for the hypothesis:

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

is no longer the t-test, but the test based on (see Box and Tiao, 1973):

$$|w(\mathbf{X})| = \left| \frac{(n-1)\left(\left(\frac{1}{2}\right)\left[X_{[1]}+X_{[n]}\right]-\mu_0\right)}{\left(\frac{1}{2}\right)\left[X_{[n]}-X_{[1]}\right]} \right| \overset{H_0}{\backsim} \mathsf{F}(2, 2(n-1)), \qquad C_1:=\{\mathbf{x} : |w(\mathbf{x})| > c_\alpha\}, \quad (61)$$

where $\left(X_{[1]}, X_{[n]}\right)$ denote the smallest and the largest element in the ordered sample $\left(X_{[1]}, X_{[2]}, ..., X_{[n]}\right)$, and $\mathsf{F}(2, 2(n-1))$ the F distribution with 2 and $2(n-1)$ degrees of freedom; see Neyman and Pearson (1928). In view of this, one can argue that the *relevant error probabilities* are no longer the actual ones associated with the t-test 'corrected' to account for the departure, but the ones associated with the test based on (61). Hence, in cases where the *actual* error probabilities differ significantly from the *relevant* error probabilities based on the 'optimal' test, the robustness argument can lead one astray.

A related argument, often used as a 'selling point' for nonparametric methods, when viewed in light of the discussion of the previous section, loses some of its appeal. This argument uses robustness to make the case that the Wilcoxon test:

(a) is more robust than the t-test because it's based on weaker assumptions, and

(b) is almost as good in terms of power as the t-test under Normality, but significantly better under non-Normal distributions. The conventional wisdom in the nonparametric literature is that the *asymptotic efficiency* of the Wilcoxon-Mann-Witney test relative to the t-test is (i) .864 for any continuous distribution which is symmetric around the median, (ii) .955 when the underlying distribution is Normal, (iii) 1.0 when the distribution is Uniform, and (iv) ∞ when the underlying distribution is Cauchy; see Hettmansperger (1984).

This argument is misleading because it misses the point that when the appropriate distribution is either the Uniform or the Cauchy, the t-test is clearly *inappropriate*. In the case of the Uniform distribution the appropriate test is based on (61), and in the case of the Cauchy distribution the t-test is *not* even *definable* because the mean and variance do *not* exist! Hence, using the t-test statistic (defined in terms of $\overline{X}$ and s) makes absolutely no statistical sense, rendering the comparison meaningless. Moreover, it is often insufficiently realized that weaker assumptions (i) are no more immune to misspecification than stronger assumptions, and (ii) often lead to less precise inferences even when they are valid. Not to mention that a simple t-plot will often be sufficient to assess the appropriateness of the Normal, Uniform and Cauchy distribution assumptions vis-a-vis the data; see Spanos (1999), ch. 5.

In summary, robustness arguments and nonparametric methods can lead one astray when no information about the *form* and *structure* of *potential misspecifications* is available, because they can lull one into a false sense of security; weaker assumptions are no more valid than stronger ones. Statements like 'the t-test is robust to symmetric non-Normality' and 'slight departures from the premises can only have minor effects on the reliability of inference', can be misleading in practice because they downplay the problems of how one can (i) affirm the presence of such misspecifications, and subsequently, (ii) address the reliability of inference problem.

7.3 'All models are wrong, but some are useful'

Equally misleading is the widely quoted Box (1979) aphorism: "All models are wrong, but some are useful." (ibid. p. 202), especially when it's taken out of context. Despite

its widespread appeal, this catchphrase is at best a truism and at worst downright misguiding. According to Cox (1995), p. 456:

"it does not seem helpful just to say that all models are wrong. The very word model implies simplification and idealization. The idea that complex physical, biological or sociological systems can be exactly described by a few formulae is patently absurd. The construction of idealized representations that capture important stable aspects of such systems is, however, a vital part of general scientific analysis ..."

Models, by definition, involve abstraction, simplification and idealization of the real-world phenomenon they aim to describe/explain. In the sense that a model is *not* an exact replica of the reality it aims to describe/explain, the above catchphrase constitutes an uninteresting truism because to claim otherwise is deluding oneself.

The question that naturally arises is whether one can use the slogan to justify using a *misspecified* statistical model as a basis for primary inferences. The short answer is *no* because any departures from the statistical model assumptions, such as [1]-[5], will invalidate both the optimality and reliability of any inferences relating to $\boldsymbol{\theta}$. Hence, a statistical model being "wrong" in the sense of being statistically misspecified [in the sense that it *could not have given rise to data* $\mathbf{z}_0$] renders it practically useless as a basis of inference. "Wrong", however, might be used to refer to *this statistical model does not adequately describe the actual data generating mechanism*. In such a case "wrongness" could refer to the potential *substantive inadequacies* of a statistically adequate model. Substantive inadequacies vis-a-vis the phenomenon of interest can be very misleading for explanatory purposes, and they need to be assessed on the basis of a statistically adequate model; see Spanos (2007, 2010).

Box (1979) summarized the main argument of his paper as follows:

"It is argued that the present emphasis by statistical researchers on ad hoc methods of robust estimation is mistaken. Classical methods of estimation should be retained using models that more appropriately represent reality. Attention should not be confined merely to discrepancies arising from outliers and heavy tailed distributions but should be extended to include serial dependence, need for transformations, and other problems." (see ibid. p. xv)

So much for Box encouraging the use of misspecified models! To the contrary, he has been the most steadfast advocate of statistical adequacy and diagnostic checking:

"No statistical model can safely be assumed adequate. Perspicacious criticism employing diagnostic checks must therefore be applied." (Box, 1980, p. 383)

In several publications Box advocates viewing empirical modeling as an iterative process which begins with a tentative model, whose statistical adequacy is assessed using diagnostic checks, and when inadequacies are detected the model is respecified, and iteration begins anew; see Box and Jenkins (1970). Hence, any charge that Box is exhorting the use of misspecified models as a basis of inference is both misplaced and unjustifiable.

Although all statistical models rely heavily on idealizations and approximations, their inductive premises need to be adequate — they need to account for the probabilistic regularities in the data — for reliable inferences and trustworthy evidence.

8 Conclusions

For incisive and reliable inferences in model-based frequentist inference one requires:

(a) optimal inference procedures, based on

(b) statistically adequate models.

Statistical adequacy ensures the reliability of inference and optimality ensures its precision and incisiveness. A *statistically misspecified model*, by definition, does not account for some relevant systematic statistical information in the data, giving rise to discrepancies between nominal and actual error probabilities rendering the reliability of inference unreliable. Given that the ultimate objective of inductive inference is to learn from the data about the underlying data-generating mechanism, it is clear that a statistically misspecified model is not conducive to such learning (Spanos, 2006).

There is disagreement, however, on how realistic such a dual objective is in practice since, in the presence of misspecification, optimality loses its appeal. *Robustness* arguments are often used as a way to steer a middle ground, sacrificing some optimality for the sake of using inference procedures which are less vulnerable to certain forms of misspecification because they rely on weaker model assumptions, e.g. Normality being attenuated to any symmetric distribution. This defensive attitude has misleadingly encouraged practitioners to utilize *nonparametric methods* of inference as a way to pay less attention to ensuring statistical adequacy at the expense of less precise inferences; see Spanos (2001).

Affirming the form and structure of potential misspecifications requires a more aggressive stance, which encourages one to face the statistical adequacy issue head on by using thorough M-S testing and respecification analysis; see Spanos (2000, 2006), Mayo and Spanos (2004). That is, one should test the probabilistic assumptions comprising the statistical model in question thoroughly, and if any of them are found wanting, go the extra mile to respecify in order to secure a statistically adequate model. Having established a statistically adequate model, it makes little sense to return to the original (misspecified) model to address the unreliability of inference problem by utilizing the *actual* error probabilities in conjunction with the original (misspecified) model. This ignores the fact that the original inference method is usually sub-optimal, or even inappropriate, in the context of the respecified model, and the *relevant error probabilities* can be different from the *actual* ones.

The above conclusions apply equally well to a much broader context of statistical modeling, including the use of robust estimators in regression and related models. Indeed, the above comments are particularly relevant for the discussions concerning the use of robust estimators of the *asymptotic covariance matrix* in the presence of heteroskedasticity/autocorrelation when testing hypotheses about regression coefficients. As shown in Spanos and McGuirk (2001), the use of such robust estimators does nothing to ameliorate the unreliability of inference problem arising from the presence of heteroskedasticity and/or autocorrelation in the regression residuals. These conclusions also call into question the strategy to use semiparametric and nonparametric procedures as a way to sidestep the statistical misspecification problem.

References

[1] Anderson, T. W. (1971), *The Statistical Analysis of Time Series*, Wiley, NY.

[2] Arnold, S. F. (1990), *Mathematical Statistics*, Prentice-Hall, NJ.

[3] Bartlett, M. S. (1935), "Some Aspects of the Time-Correlation Problem in Regard to Tests of Significance," *Journal of the Royal Statistical Society*, 98:536-543.

[4] Box, G. E. P. (1953), "Non-Normality and Tests on Variance," *Biometrika*, 318-335.

[5] Box, G. E. P. (1979), "Robustness in the Strategy of Scientific Model Building," in *Robustness in Statistics*, ed. by Launer, R. L. and G. N. Wilkinson, Academic Press, NY.

[6] Box, G. E. P. and S. L. Andersen (1955), "Permutation Theory in the Derivation of Robust Criteria and the Study of Departures from Assumptions," *Journal of the Royal Statistical Society, B*, 17: 1-34.

[7] Box, G. E. P. and G. M. Jenkins (1970), *Time series analysis: forecasting and control*, Holden-Day, San Francisco.

[8] Box, G. E. P. and G. C. Tiao (1973), *Bayesian Inference in Statistical Analysis*, Wiley, NY.

[9] Cox, D. R. (1967), "Fieller's Theorem and a Generalization," *Biometrika,* 54: 567-572.

[10] Cox, D. R. (1995), Comment on "Model Uncertainty, Data Mining and Statistical Inference," by C. Chatfield, *Journal of the Royal Statistical Society*, A, 158: 419-466.

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

[12] Doob, J. L. (1953), *Stochastic Processes*, Wiley, NY.

[13] Friedman, M. (1953), *Essays in Positive Economics*, University of Chicago Press, Chicago.

[14] Geary, R. C. (1936), "The Distribution of Student's t Ratio for Non-Normal Samples," *Journal of the Royal Statistical Society, Supplement*, 3: 178- 184.

[15] Greene, W. H. (2008), *Econometric Analysis*, 6th ed., Prentice Hall, NJ.

[16] Guala, F. (2005), *The Methodology of Experimental Economics*, Cambridge University Press, Cambridge.

[17] Hettmansperger, T. P. (1984), *Statistical Inference Based on Ranks*, Wiley, NY.

[18] Hoover, K. D. (2006), "The Methodology of Econometrics," pp. 61-87, in *New Palgrave Handbook of Econometrics*, vol. 1, ed. T. C. Mills and K. Patterson, Macmillan, London.

[19] Lehmann, E. L. (1986), *Testing statistical hypotheses,* 2nd edition, Wiley, NY.

[20] Kennedy, P. (2008), *A Guide to Econometrics*, 6th edition, MIT Press, MA.

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

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

[23] Mayo, D. G. and Spanos, A. (2006), "Severe testing as a basic concept in a Neyman–Pearson philosophy of induction," *British Journal for the Philosophy of Science*, 57: 323–57.

[24] Morgan, M. S. (1990), *The history of econometric ideas*, Cambridge University Press, Cambridge.

[25] Neyman, J. and E. S. Pearson (1928), "On the Use and Interpretation of Certain Test Criteria for purposes of Statistical Inference," *Biometrika*, 20:175-240.

[26] Pearson, E. S. (1931), "The Analysis of Variance in Cases of Non-Normal Variation," *Biometrika*, 23: 114-133.

[27] Pearson, K. (1920), "The Fundamental Problem of Practical Statistics," *Biometrika*, XIII: 1-16.

[28] Ricardo, D. (1817), Principles of Political Economy and Taxation, vol. 1 of The Collected Works of Davie Ricardo, ed. P. Sraffa and M. Dobb, Cambridge University Press, Cambridge.

[29] Scheffe, H. (1959), *The Analysis of Variance*, Wiley, NY.

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

[31] Spanos, A. (1989), "On re-reading Haavelmo: a retrospective view of econometric modeling", *Econometric Theory*, 5: 405-429.

[32] Spanos, A. (1990), "The Simultaneous Equations Model revisited: statistical adequacy and identification", *Journal of Econometrics*, 44: 87-108.

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

[34] Spanos, A. (1999), *Probability Theory and Statistical Inference: econometric modeling with observational data*, Cambridge University Press, Cambridge.

[35] Spanos, A. (2000), "Revisiting Data Mining: 'hunting' with or without a license," *The Journal of Economic Methodology*, 7: 231-264.

[36] Spanos, A. (2001), "Parametric versus Non-parametric Inference: Statistical Models and Simplicity," ch. 11, pp. 181-206 in *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple*, edited by A. Zellner, H. A. Keuzenkamp and M. McAleer, Cambridge University Press.

[37] Spanos, A. (2006), "Econometrics in Retrospect and Prospect," pp. 3-58 in Mills, T.C. and K. Patterson, *New Palgrave Handbook of Econometrics*, vol. 1, MacMillan, London.

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

[39] Spanos, A. (2009), "The Pre-Eminence of Theory vs. the European CVAR Perspective in Macroeconometric Modeling," in The Open-Access, Open-Assessment E-Journal, 3, 2009-10. http://www.economics-ejournal.org/economics/journalarticles/2009-10.

[40] Spanos, A. (2010), "Theory Testing in Economics and the Error Statistical Perspective," pp. 202-246 in *Error and Inference*, edited by D.G. Mayo and A. Spanos, Cambridge University Press, Cambridge.

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

[42] Staudte, R. G. and S. J. Sheather (1990), *Robust Estimation and Testing*, Wiley, NY.

[43] Student (1908), "The Probable Error of the Mean," *Biometrika*, 6: 1-25.