CENTRE FOR ECONOMETRIC ANALYSIS
CEA@Cass

http://www.cass.city.ac.uk/cea/index.html

Cass Business School
Faculty of Finance
106 Bunhill Row
London EC1Y 8TZ

Philosophy of Econometrics

Aris Spanos

CEA@Cass Working Paper Series
WP–CEA–03-2008

Philosophy of Econometrics

Aris Spanos
Department of Economics,
Virginia Tech, Blacksburg,
VA 24061, USA

December 2007

Abstract

1. Introduction
2. What philosophical/methodological issues?
3. Philsophy of science and empirical modeling:
   Logical positivism/empiricism
   The downfall of logical positivism/empiricism
   The new experimentalism
   Learning from error
4. The Error-Statistical perspective:
   Statistical inference and its philosophical foundations
   Statistical Induction and its underlying reasoning
   Severe testing reasoning
   A statistical model has 'a life of its own'
   Bridging the gap between theory and data
5. Philosophical/methodological issues pertaining to econometrics:
   The Bayesian approach to inductive inference
   Statistical model specification vs. model selection
   Reliability/precision of inference and robustness
   Weak assumptions and the reliability/precision of inference
   'Error-fixing' strategies and data mining
   Substantive 'fixing' strategies and theory fishing
   Pre-test bias and all that!
6. Conclusions

1 Introduction

Philosophy of econometrics is concerned with the systematic (meta-)study of general principles, strategies and philosophical presuppositions that underlie empirical modeling with a view to evaluate their effectiveness in achieving the primary objective of 'learning from data' about economic phenomena of interest. In philosophical jargon it is a core area of the philosophy of economics, which is concerned primarily with *epistemological* and *metaphysical* issues pertaining to the empirical foundations of economics. In particular, it pertains to *methodological* issues having to do with the effectiveness of methods and procedures used in empirical inquiry, as well as *ontological* issues concerned with the worldview of the econometrician. Applied econometricians, grappling with the complexity of bridging the gap between theory and data, face numerous philosophical/methodological issues pertaining to transmuting noisy and incomplete data into reliable evidence for or against a hypothesis or a theory.

The main aim of this paper is to attempt a demarcation of the intended scope of a philosophy of econometrics with a view to integrate its subject matter into the broader philosophy of science discourses. An important objective is to bring out the potential value of a bidirectional relationship between philosophy of science and applied fields in the social sciences. Econometrics can benefit from the broader philosophical discussions on 'learning from data', and philosophy of science can enrich its perspective by paying more attention to the empirical modeling practices in disciplines like econometrics where observational data are the rule, not the exception.

The philosophy of econometrics, as an integral part of economic modeling, is currently at its infancy, with most econometricians being highly sceptical about the value of philosophical/methodological discussions in empirical modeling. The focus in the econometric literature since the early 1960s has been primarily on technical issues concerned with extending estimation and testing propositions associated with the Classical Linear Regression (CLR) and related models in a number of directions. These modifications/extensions are theory-dominated and motivated primarily by the objective of 'quantifying theoretical relationships'. As a result, the focus is on (a) inherent problems such as endogeneity/simultaneity, heterogeneity, heteroskedasticity and non-linearity, and (b) different types of data (time series, cross-section and panel); see Greene (2000), Kennedy (2003).

Discussions of econometric methodology have been primarily 'local' affairs (see Granger, 1990, Hendry et al, 1990, Hendry, 2000, Leamer 1988, Pagan, 1987, Sims (1980), Spanos, 1988, 1989), where no concerted effort was made to integrate the discussions into the broader philosophy of science discussions concerning empirical modeling; some notable recent exceptions are Hoover (2002, 2006), Keuzenkamp (2000) and Stigum (2003). In certain respects, other social sciences, such as psychology, sociology or even political science have been more cognizant of philosophical/methodological issues pertaining to statistical inference and modeling; see Morrison and Henkel (1970), Lieberman (1971) and Harlow et al (1997).

The methodology of economics literature, although extensive, so far has focused primarily on issues such as the status of economic assumptions, the structure of economic theories, falsification vs. verification, Kuhnian paradigms vs. Lakatosian research programs, the sociology of scientific knowledge, realism vs. instrumentalism, 'post-modernist' philosophy, etc.; see Backhouse (1994), Blaug (1992), Davis et all (1998), Maki (2001, 2002) and Redman (1991). Even in methodological discussions concerning the relationship between economic theories and reality, econometrics is invariably neglected (Caldwell, 1994, p. 216) or misrepresented (Lawson, 1997). Indeed, one can make a case that, by ignoring the philosophical issues pertaining to *empirical* modeling, the literature on economic methodology has painted a rather lopsided picture of the relevance of the current philosophy of science in availing philosophical/methodological problems frustrating economics in its efforts to achieve the status of an empirical science. When assessing the current state of philosophy of science and its value for economic methodology Hands (2001) argued that philosophy of science is "currently in disarray on almost every substantive issue" and provides "no reliable tool for discussing the relationship between economics and scientific knowledge." (p. 6). I consider such admonitions unhelpful and believe that parts of current philosophy of science focusing on 'learning from data' (see Chalmers, 1999, Hacking, 1983, Mayo, 1996) have a lot to contribute toward redeeming the credibility of economics as an empirical science.

The current state of the empirical foundations of economics render Popper's (1959) picturesque metaphor of "piles in a swamp" seem charitable, because a closer look at the published empirical evidence over the last century reveals heaps of untrustworthy estimates and test results which (a) provide a tenuous, if any, connection between economic theory and observable economic phenomena, and (b) facilitate no veritable learning from data; see Spanos (2006a). This state of affairs has played an important role in the decision concerning to choice of themes to be discussed in this paper as constituting the core areas of a philosophy of econometrics.

In section 2, a simple empirical example is used to bring out the diversity and complexity of philosophical/methodological issues raised by such modeling attempts in applied econometrics. To set the scene for the discussion that follows several philosophical/methodological issues raised by the current textbook approach to econometrics are highlighted. Section 3 attempts to provide a summary of 20th century philosophy of science, focusing primarily on aspects of that literature that pertain to empirical modeling. Section 4 proposes that the *error-statistical perspective* (see Mayo and Spanos, 2008) provides a most appropriate framework for a philosophy of econometrics. This perspective is presented as a refined modification/extension of the Fisher-Neyman-Pearson approach to statistical induction, which can be used to effectively address some of the inveterate philosophical problems and issues that have bedeviled frequentist statistical inference since the late 1930s. The error-statistical approach is then used in section 5 to reflect briefly, as well as shed new light, on a number of philosophical/methodological problems pertaining to econometrics.

2 What philosophical/methodological issues?

To give some idea as to the kind of philosophical/methodological issues raised by empirical modeling in economics, let us consider the following basic question:

When do data $\mathbf{z}_0$ provide evidence for or against a hypothesis or a theory H?

In econometric modeling it is often insufficiently realized how many different philosophical/methodological issues such a question raises, or how difficult it is to give satisfactory answers. To bring out some of these issues let us revisit Moore's (1914, pp. 62-88) estimated 'statistical demand' curve for corn:

$$y_t = \underset{(2.175)}{7.219} - \underset{(.083)}{0.699}x_t + \widehat{u}_t,\ R^2{=}.622, \quad s{=}14.447, \quad n{=}45, \tag{1}$$

based on annual observations for the period 1866-1911, where $x_t{=}\frac{100(p_t-p_{t-1})}{p_t}$ and $y_t{=}\frac{100(q_t-q_{t-1})}{q_t}$, p_t- average price per bushel, q_t- production in bushels; standard errors in brackets; ibid. . In view of the fact that:

(i) the estimated coefficients appear to be statistically significant:

$$\tau(\widehat{\beta}_0){=}\tfrac{7.219}{2.175}{=}3.319 \Rightarrow \beta_0{\neq}0, \qquad \tau(\widehat{\beta}_1){=}\tfrac{.699}{.083}{=}8.422 \Rightarrow \beta_1{\neq}0, \tag{2}$$

(ii) they have the "correct" signs ($\widehat{\beta}_0 > 0$, $\widehat{\beta}_1 < 0$), and

(iii) and the goodness-of-fit is reasonably high ($R^2{=}.622$),

one might consider the empirical results in (1) as providing confirmatory evidence *for* the 'demand schedule':

$$Q^D = \beta_0 + \beta_1 P, \quad \beta_0 > 0,\ \beta_1 < 0. \tag{3}$$

Such a confirmation claim, however, will be premature and unwarranted before one assesses the reliability of these inferences by probing the different ways they might be in error and ascertaining that such errors are absent. What errors?

(I) Statistical Misspecification. A first serious source of potential error is *statistical misspecification*. The statistical inference results (i)-(iii) are reliable when the estimated model in (3) is *statistically adequate*: the probabilistic assumptions:

$$\{1\}\ u_t \backsim \mathsf{N}(.,.), \quad \{2\}\ E(u_t){=}0, \quad \{3\}\ Var(u_t){=}\sigma^2, \quad \{4\}\ E(u_t u_s){=}0,\ t{\neq}s, \quad t,s{=}1,...,n,$$

underlying the Linear Regression model, are valid for the particular data $\mathbf{z}_0{:=}\{(x_t, y_t),\ t{=}1,...,n\}$. A typical set of Mis-Specification (M-S) tests (see Spanos and McGuirk, 2001) is reported in table 1, with the p-values in square brackets. The tiny p-values indicate serious departures from assumptions {2}–{4}, rendering the inferences concerning the sign and the magnitude of the coefficients (β_0, β_1) *unwarranted*.

Table 1 - Misspecification tests	
Non-Normality:	$D'AP = 3.252[.197]$
Non-linearity:	$F(2,41){=}19.532[.000001]^*$
Heteroskedasticity:	$F(2,41){=}14.902[.000015]^*$
Autocorrelation:	$F(2,41){=}18.375[.000011]^*$

(4)

In view of the M-S testing results in table 1, the estimated model in (1) constitutes an *unreliable basis* for inference and the claims (i)-(iii) are empirically unwarranted. The unreliability arises from the fact that when any of the assumptions {1}-{4} are invalid, the relevant *nominal* and *actual error probabilities* are likely to be very different. Applying a .05 significance level t-test, when the actual type I error is .98, renders the test highly unreliable; see Spanos and McGuirk (2001).

The question that naturally arises at this stage is 'how many published applied econometric papers over the last 50 years are likely to pass the *statistical adequacy test*?' The surprising answer is 'very few', raising serious doubts about the trustworthiness of the mountains of evidence accumulated in econometrics journals during this period; see Spanos (2006a). Indeed, in most cases the modeler is not even aware of all the probabilistic assumptions constituting the statistical model used as a basis of his/her inference. What makes matters worse is that statistical inadequacy is only one of several potential sources of error that could render empirical evidence untrustworthy.

(Ii) Inaccurate data. A second source of potential error is *inaccurate data*: data $\mathbf{z}_0$ are marred by *systematic errors* imbued by the collection/compilation process; see Morgenstern (1963). Such systematic errors are likely to distort the statistical regularities and give rise to misleading inferences. The discussion of the data in Moore (1914) gives enough clues to suspect that inaccurate data is likely to be another serious source of error contributing to the unreliability of any inference based on (1). In particular, the averaging of different prices over time and taking proportional differences is likely to distort their probabilistic structure and introduce systematic errors into the data; see Abadir and Talmain (2002).

(III) Incongruous measurement. A third source of potential error is *incongruous measurement*: data $\mathbf{z}_0$ do not adequately quantify the concepts envisioned by the theory. This, more than the other sources of error, is likely to be the most serious one ruining the trustworthiness of Moore 'statistical demand' in (1). Moore's contention that x_t and y_t provide adequate quantification for the theoretical variables 'quantify demanded' (Q^D) and the corresponding 'price' (P) is altogether unconvincing. The gap between, on one hand, the intentions to buy Q^D_{it}, at some point in time t, and the set of hypothetical prices P_{it}, i=$1, 2, ..., m$, and, on the other, the quantities transacted q_t and the corresponding observed prices p_t, over time t=$1, 2, ..., n$, cannot possibly be bridged by the 'proportional change' transformation; see Spanos (1995) .

(IV) Substantive inadequacy. A fourth source of potential error is *substantive inadequacy (external invalidity)*: the circumstances envisaged by the theory in question differ 'systematically' from the *actual* data generating mechanism. This inadequacy can easily arise from impractical *ceteris paribus* clauses, missing confounding factors, false causal claims, etc.; see Guala (2005), Hoover (2006). Substantive adequacy concerns the extent to which the estimated model 'captures' the aspects of the reality it purports to explain in a statistically and substantively adequate way, shedding light on the phenomenon of interest, i.e. 'learning from data'.

Given the potentially grievous detrimental effects of the other sources of error on the trustworthiness of the inference based on (1), raising questions about its substantive inadequacy seems rather gratuitous.

In view of the seriousness of all these errors, taking the estimated regression in (1) at face value and drawing any inferences seems like a very bad idea. An interesting question to consider is how a textbook econometrician is likely to proceed when faced with the empirical results reported in (1).

2.1 Reflecting on textbook econometrics

In practice, the methodological framework adopted in traditional textbook econometric modeling does *not* include systematic probing for errors as part of the accepted rules and strategies for learning from data. Unfortunately, this methodological framework is implicit and it's usually adopted without examination as part and parcel of learning econometrics.

The emphasis in textbook econometrics is *not* on probing for potential errors at each stage of the modeling, but on 'quantifying a theoretical model' which puts the focus on adopting the weakest possible probabilistic structure that would 'justify' a method yielding 'consistent' estimators of the parameters of interest. In particular, the cornerstone of the textbook approach, the Gauss-Markov (G-M) theorem – as well as analogous theorems concerning the 'optimality' of Instrumental Variables (IV), Generalized Method of Moments (GMM) and non-parametric methods – distance themselves from strong probabilistic assumptions, in particular, Normality, in an attempt to gain greater generality for certain inference propositions. The rationale is that the reliance on weaker probabilistic assumptions will render OLS, IV and GMM-based inferences less prone to statistical misspecifications and thus more reliable; see Greene (2000). This rationale raises very interesting philosophical/methodological questions that need to be discussed and appraised. For instance:

– what does one accomplish, in terms of generality, by not assuming Normality in the Gauss-Markov (G-M) and related theorems?
– can one use the G-M theorem as a basis for reliable inferences?
– how do weaker assumptions give rise to more reliable inferences?
– how does one ensure the reliability of an inference when the premises are not testable, as in the case of non-parametric inference? and
– Does reliance on consistent and asymptotically Normal estimators suffice for reliable inferences?"

In view of this textbook econometric perspective, the question that naturally arises is "what would a traditional econometrician do when faced with the empirical results in (1)?" An ostensible diagnostic checking that relies on a small number of traditional M-S tests, such as the skewness-kurtosis (S-K), the Durbin-Watson and the White heteroskedasticity (W) tests:

$$S\text{-}K=2.186[.335],\ D-W=2.211,\ W(2,42)=15.647[.000], \tag{5}$$

reveals a clear departure from assumption [3]. In textbook econometrics, however, when any of the error assumptions [1]-[4] are found wanting, conventional wisdom recommends a sequence of 'error-fixing' procedures which are designed to remedy the problem; see Greene (2000). A textbook econometrician faced with the results in (5) is likely to count his/her blessings because they do not seem to show devastating departures from assumptions [1]-[4]. The presence of heteroskedasticity, according to the conventional wisdom, will only affect the efficiency of $(\widehat{\beta}_0, \widehat{\beta}_1)$; unbiasedness and consistency still hold. The departure is supposed to be 'accounted for' by employing the so-called Heteroskedasticity Consistent Standard Errors (HCSE). In view of the fact that $\mathsf{HCSE}(\widehat{\beta}_0)$=2.363, $\mathsf{HCSE}(\widehat{\beta}_1)$=.108, these inferences are usually declared 'robust' to the departure from [3].

These conventional wisdom recommendations raise many interesting philosophical/methodological problems with a long history in philosophy of science, such as ad-hoc modifications, double-use of data, curve-fitting, pre-designation vs. post-designation, etc. Interesting questions raised by the above textbook strategies are:

– are the 'error-fixing' procedures justified on statistical grounds?
– is 'error-fixing' the best way to respecify a statistically inadequate model?
– what kind of robustness/reliability does the use of HCSE bring about?
– are the various specification searches justified statistically?
– how thorough should M-S testing be to avert any data mining charges?
– how does one decide what M-S tests are the most appropriate to apply in a particular case?
– how does one distinguish between legitimate and illegitimate double-use of data?

Another set of issues likely to be raised by practitioners of textbook econometrics relate to the *simultaneity* problem between y_t and x_t. The contention is that the endogeneity of x_t (arising from the demand/supply theory) calls into question the substantive validity of (1), and the only way to render the empirical results meaningful is to account for that. This amounts to bringing into the modeling additional variables $\mathbf{W}_t$, such as rainfall and the prices of complementary and substitute commodities, which could potentially influence the behavior of both x_t and y_t. This reasoning gives rise to an implicit reduced form (Spanos, 1986):

$$y_t = \pi_{10} + \pi_{11}^{\top}\mathbf{w}_t + \varepsilon_{1t}, \qquad x_t = \pi_{20} + \pi_{21}^{\top}\mathbf{w}_t + \varepsilon_{2t},\ t\in\mathbb{N}. \tag{6}$$

Again, this modeling strategy raises interesting methodological issues which are often neglected. For example:

– how does a mixture of statistical significance and theoretical meaningfulness renders a model "best"?
– in what sense does the IV amplification of the model in (6) alleviate the statistical inadequacy problem for (1)?
– how does the substantive information in (6) relate to the statistical information unaccounted for by (1)?,

– how does one chooses the 'optimal' instruments $\mathbf{W}_t$ in (6)?
– what conditions would render the IV-based inference for (β_0, β_1) any more reliable than OLS-based inference in (1)?

The above textbook arguments stem from adopting an implicit methodological framework that defines the fundamental ideas and practices which demarcate econometric modeling, and determine the kind of questions that are supposed to be asked and probed, how these questions are to be structured and answered, and how the results of scientific investigations should be reported and interpreted; it establishes the 'norms' of scientific research – what meets the 'standards' of publication in learned journals and what does not. An important task of philosophy of econometrics is to make all these implicit methodological presuppositions *explicit*, as well as evaluate their effectiveness.

3 Philosophy of science and empirical modeling

From the perspective of the philosophy of econometrics, a central question in 20th century philosophy of science has been (see Mayo, 1996):

How do we learn about phenomena of interest in the face of uncertainty and error?

In particular, this raises several interrelated questions:

(a) Is there such a thing as a scientific method?
(b) What makes an inquiry scientific or rational?
(c) How do we appraise a theory vis-a-vis empirical data?
(d) How do we make reliable inferences from empirical data?
(e) How do we obtain good evidence for a hypothesis or a theory?

These are some of the most crucial questions that philosophy of science has grabbled with during the 20th century. For the discussion that follows, it will be convenient to divide 20th century philosophy of science into several periods: 1918-1950s: logical positivism/empiricism (Hempel, Nagel), 1960s-1980s: the downfall of logical empiricism (Quine, Kuhn, Popper, Lakatos), 1980s-1990s: miscellaneous turns (historical, naturalistic, sociological, pragmatic, feminist etc.), 1990s- : new experimentalism and learning from error (Mayo, 1996).

The following discussion is ineluctably sketchy and highly selective with the emphasis placed on philosophical/methodological issues and problems pertaining to empirical modeling. For a more balanced textbook discussion of current philosophy of science see Chalmers (1999), Godfrey-Smith (2003), Machamer and Silberstein (2002), Newton-Smith (2000); for a more economics-oriented perspective see Hands (2001), Redman (1991).

3.1 Logical positivism/empiricism

The tradition that established philosophy of science as a separate sub-field within philosophy during the first half of the 20th century was that of logical positivism/empiricism.

Its roots can be traced back to the 19th century traditions of positivism and empiricism, but what contributed significantly in shaping logical positivism into a dominating school of thought were certain important developments in physics and mathematics in the early 20th century.

In physics the overthrow of Newtonian mechanics by Einstein's theory of relativity (special and general), as well as the predictive success of quantum mechanics, raised numerous philosophical problems and issues that were crying out for new insights and explanations concerning scientific methods and the nature of knowledge; how do we acquire attested knowledge about the world? The re-introduction of the axiomatic approach to mathematics by Hilbert and the inception and development of propositional and predicate logic by Frege, Russel, Whitehead and Wittgenstein, provided a formal logico-mathematical language that promised to bring unprecedented clarity and precision to mathematical thinking in general, and to foundational inquiry in particular. The new formal language of first order predicate logic, when combined with the exhaustive specification of the premises offered by the axiomatic approach, appeared to provide a model for precise and systematic reasoning, and thus an ideal tool for elucidating the many aspects of scientific reasoning and knowledge.

These developments called into question two of the most sanctified pillars of knowledge at the time, Newtonian mechanics and Euclidean geometry. The combination of general relativity and Hilbert's axiomatization of Euclidean geometry left no doubts that our knowledge of geometry cannot be synthetic a priori in Kant's sense.

It's no coincidence that the founding group of logical positivism (Schlick, Hahn, Waismann, Carnap, Neurath, Frank, Reichebach) were primarily mathematicians and physicists who aspired to use physics as their paradigmatic example of a real scientific field. Their aspiration was that this formal logico-mathematical language will help to formalize the structure of scientific theories as well as their relationship to experiential data in precise ways which would avoid the ambiguities and confusions of the natural language. The idea being that a philosophy of science modeled on physics could then be extended and adapted to less developed disciplines, including the social sciences. Not surprisingly, the early primary focus of logical positivism/empiricism was on the *form and structure* of scientific theories as well as *epistemology,* which is concerned with issues and problems about knowledge (meaning, nature, scope, sources, justification, limits and reliability), evidence and rationality. The strong empiricist stance adopted by this tradition marginalized *metaphysics,* which is concerned with issues and problems about the nature and structure of reality. At the same time it elevated empirical meaningfulness to a demarcation criterion between scientific and non-scientific statements and put forward a Hypothetic-Deductive (H-D) form of reasoning as the way science is grounded in observation and experiment, as well as how we acquire knowledge about the world from experience. Viewing a theory h as empirically interpretable (via correspondence rules) deductive axiomatic system, H-D reasoning,in its simplest form, boils down to assessing the empirical validity of certain observational implications $\mathbf{e}$ of h. If $\mathbf{e}$ turns out to be true, it provides

confirmatory evidence for the (probable) validity of h :

$$\begin{array}{c} \text{If } h \text{ then } \quad \mathbf{e} \\ \mathbf{e}, \\ \hline \therefore \text{ (probably) } h \text{ is true} \end{array} \tag{7}$$

The above argument is deductively invalid (known as affirming the consequent fallacy), but it provided the basis of (inductive) confirmation for logical empiricists; see Nagel (1961), Hempel (1965).

From the perspective of empirical modeling, a major weakness of the logical empiricist tradition was its failure to put forward a satisfactory explanation of how we learn from experience (induction). The tradition's simplistic confirmation reasoning in (7) as a means to assess the truth of a hypothesis h, in conjunction with the inadequacy of the inductive logics devised to evaluate the relative support of completing hypotheses, contributed significantly to the tradition's demise by the 1970s. Their attempts to formalize induction as primarily a logical relationship $C(\mathbf{e}, h)$ between evidence $\mathbf{e}$ – taken as objectively given – and a hypothesis h, failed primarily because they did not adequately capture the complexity of the relationship between h and $\mathbf{e}$ in scientific practice. Indeed, an enormous amount of hard work and ingenuity go into fashioning a testable form h of a hypothesis of interest, and establishing experiential facts $\mathbf{e}$ from noisy, finite and incomplete data $\mathbf{x}_0$, as well as relating the two. Their view of theory confirmation as a simple logical argument which involves two readily given statements, h - the hypothesis of interest and $\mathbf{e}$ – the experiential facts, was not just overly simplistic, but misleading in so far as neither h or $\mathbf{e}$ are straight forward nor readily available in actual scientific practice. Moreover, hypotheses or theories expressed as a set of sentences in an axiomatic system of first order logic are not easily amenable to empirical analysis. Not surprisingly, the inductive logics of logical empiricists were plagued by several paradoxes (ravens, grue), and they had little affinity to the ways practicing scientists learn from data. This was particularly true of learning from data in statistical induction as developed by Fisher in the early 1920s and extended by Neyman and Pearson in the early 1930s.

3.2 The downfall of logical empiricism

Part of the appeal of logical positivism/empiricism stemmed from the fact that there was something right-headed about their presumption that the distinguishing features of science, as opposed to other forms of human activity, can be found in observation and experiment; that knowledge about the world is secure only when it can be tested against observation and experiment. However, their answers to the above crucial questions (a)-(e) in the first half of the 20th century turned out to be inadequate and unconvincing. The tradition's undue reliance on formal logics, axiomatization, the analytic-synthetic and theoretical-observational distinctions, were instrumental in undermining its credibility and its leadership role in philosophy of science. The view that scientific theories and research activity can be codified in terms of these

idealized tools turned out to be overly optimistic. By the early 1970s there was general consensus that logical empiricism was not only inadequate but also untenable. The downfall of logical empiricism was hastened by critics such as Quine, Popper and Kuhn who pinpointed and accentuated these weaknesses.

Quine (1953, 1960) contributed to the downfall of logical empiricism in a number of ways, but the most influential were: (i) his undermining of the analytic-synthetic distinction, (ii) his reviving and popularizing of Duhem's (1906) theses that (a) 'no hypothesis can be tested separately from an indefinite set of auxiliary hypotheses' and (b) 'crucial experiments that could decide unequivocally between competing theories do not exist', and (iii) his initiating the naturalistic turn.

His revisiting of Duhem's theses became known as the Quine-Duhem problem which gave rise to an inveterate conundrum:

(I) The underdetermination of theory by data − the view that there will always be more than one theory *consistent* with any body of empirical data.

Naturalism constitutes an epistemological perspective that emphasizes the 'continuity' between philosophy and science in the sense that the methods and strategies of the natural sciences are the best guides to inquiry in philosophy of science; there is no higher tribunal for truth and knowledge than scientific practice itself. Philosophy should study the methods and findings of scientists in their own pursuit of knowledge, while heightening its evaluative role.

Popper (1959, 1963) replaced the confirmation argument in (7) with a falsification argument, based on *modus tollens* (a deductively valid argument):

$$\frac{\begin{array}{c}\text{If } h \text{ then } \quad \mathbf{e} \\ \text{not-}\mathbf{e},\end{array}}{\therefore \text{ not-}h \text{ is true}} \tag{8}$$

His falsificationism was an attempt to circumvent the problem of induction as posed by Hume, as well as replace confirmation as a demarcation criterion with falsifiability: a hypothesis h is scientific if and only it's falsifiable by some potential evidence $\mathbf{e}$, otherwise it's non-scientific.

Popper's falsificationism was no more successful in explaining how we learn from experience than the inductive logics it was designed to replace for a variety of reasons. The most crucial was Duhem's problem: the premises h entailing $\mathbf{e}$ is usually a combination of a primary hypothesis H of interest and certain auxiliary hypotheses, say $A_1,\ A_2, ..., A_m$. Hence, not-h does not provide a way to distinguish between not-H and not-A_k, k=1, ..., m. As a result, one cannot apportion blame for the failure to observe $\mathbf{e}$ to any particular sub-set of the premises $(H, A_1,\ A_2, ..., A_m)$. *Second*, Popper's falsification does not allow one to learn anything positive about h using the data. When several 'genuine' attempts to refute h fail to do so, one cannot claim that h is true, or justified, or probable or even reliable. A Popperian can only claim that hypothesis h is the "best tested so far" and that it is *rational to accept* it (tentatively) because it has survived 'genuine' attempts to falsify it. *Third*, any attempt to measure the degree of 'corroboration' – credibility bestowed on h

for surviving more and more 'genuine' attempts to refute it – brings back the vary problem of induction falsificationism was devised to circumvent.

Despite the failure of falsificationism to circumvent induction as capturing the way we learn from experience, there is something right-minded about Popper's intuition underlying some of his eye-catching slogans such as "Mere supporting instances are as a rule too cheap to be worth having", "tests are severe when they constitute genuine attempts to refute a hypothesis" and "we learn from our mistakes". This intuition was garnered and formalized by Mayo (1996) in the form of severe testing, but placed in the context of frequentist statistical induction.

Kuhn (1962, 1977) undermined the logical empiricist tradition by questioning the wisdom of abstracting scientific theories and the relevant experiential data from their historical and a social context, arguing that the idealized formal models did not capture the real nature and structure of science in its ever-changing complexity. Partly motivated by Duhem's problem he proposed the notion of a *scientific paradigm* to denote the set of ideas and practices that define a scientific discipline during a particular period of time, and determine what is to be observed and scrutinized, the kind of questions that are supposed to be asked and probed, how these questions are to be structured, and how the results of scientific investigations should be interpreted. Using the notion of *normal science* within a paradigm, Kuhn questioned the positivist account of cumulative growth of knowledge, arguing that old paradigms are overrun by new ones which are usually 'incommensurable' with the old.

As a result of the extended controversy that ensued, Kuhn's ideas had an important influence on the development of philosophy of science to this day, and his legacy includes a number of crucial problems such as:

(II) Theory-dependence of observation. An observation is theory-laden, if, either the statement expressing the observation employs or presupposes certain theoretical concepts or knowing the truth of the observation statement requires the truth of some theory.

The theory-ladeness of data problem has to do with whether data can be considered an unbiased or neutral source of information when assessing the validity of theories, or whether data are usually 'contaminated' by theoretical information in a way which prevents them from fulfilling that role.

(III) Relativism refers to the view that what is true or a fact of nature is so only relative to some overarching conceptual framework of which the truth of fact of the matter is expressible or discoverable. The idea that the truth of justification of a claim, or the applicability of a standard or principle, depends on one's perspective.

(IV) The social dimension of science. What makes science different form other kinds of inquiry, and renders it especially successful, is its unique social structure. This unique social structure has an important role to play in establishing scientific knowledge.

Kuhn's move to 'go large' from a scientific theory to an all-encompassing scientific

paradigm was followed by **Lakatos** (1970) and **Laudan** (1977) who proposed the notions of a scientific research programme and a research tradition, respectively, in their attempts to avoid the ambiguities and unclarities, as well as address some of the failings of Kuhn's original notion of a scientific paradigm.

Despite the general understanding that logical empiricism was no longer a viable philosophical tradition, by the 1980s there was no accord as to which aspects of logical empiricism were the most problematic, or how to modify/replace the basic tenets of this tradition; there was no consensus view on most of the crucial themes in philosophy of science including the form and structure of theories, the nature of explanation, confirmation , theory testing, growth of knowledge, or even if there is such a thing as a scientific method; see Suppe (1977). This disagreement led to a proliferation of philosophical dictums like "anything goes", "evidence and confirmation are grounded on rhetoric or power", which began to gain appeal in certain disciplines, but especially in the social sciences where rock-solid scientific knowledge is more difficult to establish. This was part of the broader movement of miscellaneous turns (historical, sociological, pragmatic, feminist, social constructivist, discursivist, etc.) aspiring to influence the tradition that will eventually emerge to replace logical empiricism; see Hands (2001).

3.3 The new experimentalism

By the 1980s, the combination of Duhem's problem, the underdetermination conundrum and the theory-dependence of observation problem, made theory appraisal using empirical data seem like a hopeless task.

As mentioned above, establishing $\mathbf{e}$ (or not-$\mathbf{e}$) as *observational facts* constitutes one of the most difficult tasks in scientific research because the raw data $\mathbf{x}_0$ (experimental and observational) contain uncertainties, noise and are never in plenitude necessitated. Indeed, the raw data $\mathbf{x}_0$ usually need to be discerningly modeled to separate the systematic (signal) from the non-systematic (noise) information, as well as provide a measure of the reliability of inference based on $\mathbf{x}_0$. Such modeling is often vulnerable to numerous errors that would render $\mathbf{e}$ far from being 'objectively given facts'.

The first concerted effort in philosophy of science to study the process generating the raw data $\mathbf{x}_0$ and establish observational facts $\mathbf{e}$ (or not- $\mathbf{e}$) was made by the "new experimentalism" tradition; Hacking (1983), Ackermann (1985), Mayo (1996, 1997) – see Chalmers (1999) for an excellent summary. Using the piece-meal activities involved and the strategies used in successful experiments, Hacking (1983) argued persuasively against the theory-dominated view of experiment. He made a strong case that in scientific research an experiment can have a 'life of its own' that is independent of 'large-scale theory', and thus alleviating the theory-dependence of observation problem. In addition, scientists employ a panoply of practical step-by-step strategies for eliminating error and establishing the 'factual basis of experimental effects' without 'tainting' from large-scale theory.

3.4 Learning from error

Mayo (1996) proposed a formalization of these research activities and strategies for detecting and eliminating errors using the Neyman-Pearson testing as the quintessential inductive framework, supplemented with a post-data evaluation of inference based on severe testing reasoning. Contrary to the Popperian and growth of knowledge traditions' call for 'going bigger' (from theories to paradigms, to scientific research programs and research traditions), in order to deal with such problems as theory-laden observation, underdetermination and Duhem-Quine, Mayo argues that theory testing should be piece-meal and thus we should 'go smaller':

> "The fact that theory testing depends on intermediate theories of data, instruments, and experiment, and that the data are theory laden, inexact and "noisy", only underscores the necessity for numerous local experiments, shrewdly interconnected." (Mayo, 1996, p. 58)

Mayo's attempt to put forward an epistemology of experiment includes, not only how observational facts $\mathbf{e}$ are established using experimental controls and learning from error, but also how the hypothesis of interest h is fashioned into an estimable form appropriate to face the tribunal of $\mathbf{e}$. This comes in the form of a hierarchy of interconnected models: 'primary, experimental and data models' (p. 128).

In her proposed framework an integral component of the modeling procedure includes questions about 'what data are relevant', 'how the data were generated', 'how can the relevant data be adequately summarized in the form of data models' etc. The reliability of evidence is assessed at all three levels of models by using error-statistical procedures based on learning from error reasoning. The primary tool for these assessments is the notion of severity, which assesses, not the degree of support for a hypothesis, but rather the ability of the testing procedure to detect discrepancies from that hypothesis. Probability attaches not to hypotheses but to testing procedures, to inform us of their probativeness and capacity to detect errors.

Mayo (1996) made a strong case that there is a domain of 'experimental knowledge' that can be reliably established independent of high-level theory and the continuity of scientific progress consists in part of the steady build up of claims that pass severe tests. The answers she provided to the questions (a)-(e) posed above are distinctly different from those of logical empiricism as well as the other post-received view 'large-scale theory' traditions.

What makes the error-statistical approach appropriate as a methodological framework for empirical modeling is primarily because it provides a framework which adequately captures the complexity of the gap between theory and observation in scientific practice and focuses on the 'learning from error' procedures that underlie the fashioning of a testable form of a hypothesis of interest H, as well as establishing experiental facts (reliable inferences) $\mathbf{e}$ from noisy, finite and incomplete data $\mathbf{x}_0$. In addition, it proposes a general way to bridge the gap between theory and data using a chain of complecting models (primary, experimental, data), and harnesses the power of modern statistical inference and modeling to bear upon the problems and issues

raised by our attempt to come to grips with learning from experience, including the question 'When do data $\mathbf{x}_0$ provide evidence for or against H?

The fundamental intuition is encapsulated by the **Severity Principle (SP)**:

Data $\mathbf{x}_0$ do NOT provide good evidence for hypothesis H if $\mathbf{x}_0$ is used in conjunction with a test procedure T which has a very low capacity to uncover discrepancies from H when present; see Mayo (1996).

To appreciate the formalization of this intuition the discussion next focuses on statistical induction where hypothesis H, test T and data $\mathbf{x}_0$ are clearly defined in the context of a statistical model $\mathcal{M}$ describing the process that generated $\mathbf{x}_0$.

4 The Error-Statistical perspective

The **Error Statistical perspective** was proposed by Mayo (1996) as a coherent framework for inductive inference spearheaded by 'learning from error', which can accommodate a refined elaboration of the *Fisher-Neyman-Pearson approach* to statistical inference. The crucial features of the perspective are:

(i) Emphasizing the **learning from data** (about the phenomenon of interest) objective of empirical modeling.

(ii) Paying due attention to the **validity of the premises** of induction via *statistical adequacy*, using thorough *misspecification testing* and *respecification*.

(iii) Emphasizing the central role of **error probabilities** in assessing the reliability (capacity) of inference, both *pre-data* as well as *post-data*.

(iv) Supplementing the original framework with a *post-data* assessment of inference in the form of **severity evaluations** in order the provide an inferential construal of tests.

(v) Bridging the gap between theory and data using a *sequence of interconnected models,* **theory** (primary), **structural** (experimental), **statistical** (data) built on two different, but related, sources of information: *substantive subject matter* and *statistical information* (chance regularity patterns).

(vi) Actively encouraging the **thorough probing** of the different ways an inductive inference might be **in error**, by localizing the error probing in the context of the different models (theory, structural, statistical and empirical).

The next few sub-sections elaborate on these key features as a prelude to using the error-statistical perspective to elucidate a number of philosophical/methodological issues pertaining to econometrics in section 5.

4.1 Statistical inference and its philosophical foundations

One of the primary reasons for proposing the error-statistical perspective as providing the appropriate framework for a sound philosophy of econometrics is that it gives convincing answers to numerous philosophical/methodological problems and issues that have bedeviled frequentist statistical inference since the 1930s. To motive this, we

consider a birds eye view of the discussions pertaining to the philosophical foundations of frequentist inference.

The modern approach to frequentist (classical) statistics was pioneered by Fisher (1921, 1922) as model-based statistical induction, anchored on the notion of a statistical model. Fisher (1925, 1934), almost single-handedly, created the current theory of 'optimal' point estimation and formalized significance testing based on the p-value reasoning. Neyman and Pearson (1933) proposed an 'optimal' theory for hypothesis testing, by modifying/extending Fisher's significance testing; see Pearson (1966). Neyman (1937) proposed an 'optimal' theory for interval estimation analogous to N-P testing. Broadly speaking, the probabilistic foundations of frequentist statistics, as well as the technical apparatus associated with statistical inference methods, were largely in place by the late 1930s, but its philosophical foundations concerned with the proper form of the underlying inductive reasoning were in a confused state. Fisher was arguing for 'inductive inference', spearheaded by his significance testing in conjunction with p-values and his fiducial probability for interval estimation. Neyman was arguing for 'inductive behavior' based on N-P testing and confidence interval estimation in conjunction with pre-data error probabilities; see Mayo (2005).

The last exchange between these pioneers of frequentist statistics took place in the mid 1950s (see Fisher, 1955, Neyman, 1956, Pearson, 1955) and left the philosophical foundations of the field in a state of confusion with many more questions than answers.

Philosophical/methodological questions.

What are the differences between a Fisher significance test and a N-P test?
Does a proper test require the specification of an alternative hypothesis?
What about goodness-of-fit tests like Pearson's?
Do the notions of type II error probability and power apply to Fisher-type tests?
What about the use of error probabilities post-data? Is the p-value a legitimate error probability?
Is there a relationship between p-values and posterior probabilities?
Does Fisher's fiducial distribution give rise to legitimate error probabilities?
Can one distinguish between different values of the unknown parameter within an observed Confidence Interval (CI)?
Can one infer substantive significance from an observed CI?
In what sense does conditioning on an ancillary statistic enhance the precision and data-specificity of inference?

In addition to these questions which were primarily concerned with:
(a) the proper form of inductive reasoning underlying frequentist inference,
(b) it was not at all obvious under what circumstances one could safeguard the coarse accept/reject decisions against:
(i) the fallacy of acceptance: interpreting accept H_0 [no evidence against H_0] as evidence for H_0,
(ii) the fallacy of rejection: interpreting reject H_0 [evidence against H_0] as evidence for H_1; the best known example of this is the conflating of

statistical with substantive significance.

Fisher's use of the p-value to reflect the 'strength of evidence' against the null, was equally susceptible to the fallacy of rejection since the p-value often goes to zero as the sample size $n \to \infty$. Moreover, interpreting a p-value which is not 'small enough' as evidence for H_0 would render it susceptible to the fallacy of acceptance.

The subsequent literature on frequentist statistics shed very little (if any) additional light on these philosophical/foundational issues. Not surprisingly, due to the absence of any guidance from statistics or philosophy of science, the practitioners in several disciplines came up with their own 'pragmatic' ways to deal with the philosophical puzzles bedeviling the frequentist approach. Indeed, the above questions gave rise to a numerous debates (see Harper and Hooker, 1976), which were especially heated in the social sciences like psychology, sociology and education (Morrison and Henkel, 1971, Lieberman, 1971), and more recently re-discovered in economics (McCloskey, 1985). This resulted in a hybrid of the Fisher and N-P inference accounts which is "inconsistent from both perspectives and burdened with conceptual confusion." (Gigerenzer 1993, p. 323). This inconsistent hybrid eventually acquired a life of its own in these separate fields and led to widespread abuses of these methods that continue unabated to this day; see Harlow et al (1997), Thompson (1999).

The literature in philosophy of science missed the obvious connection between the Popperian and Fisherian versions of falsification, and largely ignored the important developments in frequentist statistics. In direct contrast to the practitioners' extensive use of statistics in almost all scientific fields, by the early 1950s logical empiricism had adopted a largely Bayesian perspective on inductive inference with Carnap's confirmatory logics (logical relations between statements and evidence – going back to Keynes, 1921) dominating the evidential accounts in philosophy of science; see Neyman (1957). Hacking (1965) criticized the Neyman-Pearson approach to testing for its incompleteness. The *pre-data* (before-trial) error probabilistic account of inference, although adequate for assessing optimality, is inadequate for a *post-data* (after-trial) evaluation of the inference reached; see ibid., pp. 99-101.

By the early 1960s the confused state of the philosophical underpinnings of frequentist inference, especially as it relates to its underlying inductive reasoning, began to be used as evidence for its philosophical destitution and the superiority of Bayesian inference. The latter does away with error probabilities altogether by focusing exclusively on the likelihood function $L(\theta; \mathbf{x}_0) \propto f(\mathbf{x}_0; \theta)$, $\theta \in \Theta$, which depends only on the data $\mathbf{x}_0$, and ignores the distribution of the sample $f(\mathbf{x}; \theta)$ for $\mathbf{x} \neq \mathbf{x}_0$, for $\mathbf{x} \in \mathbb{R}^n_X$ -the sample space; this is often expressed with a hint of sarcasm as "taking no account of other possible realizations that might have been observed but were not"; see Jeffreys (1961), Savage (1962), Pratt (1961).

Despite the weaknesses in its philosophical foundations and the criticisms from the Bayesian perspective, the frequentist approach to statistical inference continued to dominate applied research in most scientific fields during the 1970s and 1980s. Indeed, its extensive application raised additional philosophical/methodological problems, the

most important of which are:

(c) the role of substantive subject matter information,

(d) statistical model specification vs. model selection,

(e) data mining, pre-test bias, double-use of data,

(f) multiple hypotheses in testing and the relevant error probabilities,

(g) model validation.

The main thesis of this paper is that most of the above issues can be addressed in the context of an enhanced Fisher-Neyman-Pearson (F-N-P) frequentist framework named *error-statistical* by Mayo (1996). The error-statistical perspective addressing the problems (a)-(g) depends crucially on blending the Fisher and Neyman-Pearson (N-P) testing perspectives to weave a coherent frequentist inductive reasoning anchored firmly on *error probabilities*. The key to this coalescing is provided by recognizing that Fisher's p-value reasoning is based on a *post-data error probability*, and Neyman-Pearson's type I and II errors reasoning is based on *pre-data error probabilities*. In the coalescing, both pre-data and post-data error probabilities fulfill crucial complementary roles. The post-data component of this coalescing was proposed by Mayo (1991) in the form of *severe testing reasoning*. It is important to emphasize that the error-statistical perspective provides a broader *methodology of error inquiry* that encourages detecting and identifying the different ways an inductive inference could be in error by applying effective methods and procedures which would detect such errors when present with very high probability. Indeed, it is argued that this perspective offers a philosophy of econometrics that can address numerous philosophical/methodological issues currently bedevilling econometric modeling.

In an attempt to shed light on the inductive reasoning underlying the error-statistical perspective the next sub-section elaborates on the form of induction underlying the Fisher-Neyman-Pearson approach to statistical inference and contrasts that with induction-by-enumeration as well as Bayesian induction.

4.2 Statistical Induction and its underlying reasoning

Fisher (1922) pioneered a recasting of statistical induction from Karl Pearson's *induction-by-enumeration* in the context of an inverse probability (Bayesian) setup, to a model-based induction in the context of a purely frequentist frame-up. His recasting included two interrelated innovations.

The first was to replace the inverse probability approach, giving rise to a posterior distribution, with a frequentist approach based on the sampling distributions of relevant statistics. This changeover is well-known and widely discussed; see Stigler (1986), Hald (1998). The only aspect of the recasting that it's still somewhat controversial is the extent to which the frequency definition of probability is circular or not (see Keuzenkamp, 2000), an issue that will be touched upon below.

The second, was to transform the primitive form of induction-by-enumeration, whose reliability was based on a priori stipulations, into a refined model-based induction with 'ascertainable error probabilities' valuating its reliability. Indeed, Fisher

initiated a general way to quantify the uncertainty associated with inference by:

(a) *embedding* the *material experiment* into a *statistical model*, and

(b) use the latter to ascertain the (*frequentist) error probabilities* associated with particular inferences in its context. The form of induction envisaged by Fisher (1922, 1935a-b) is one where the reliability of the inference stems from the 'trustworthiness' of the procedure used to arrive at the inference. A very similar form of model-based induction was proposed much earlier by Peirce (1878), but his ideas were way ahead of his time and did not have any direct influence on either statistics or philosophy of science; see Mayo (1996).

4.2.1 Induction by enumeration vs. model-based induction

Induction by enumeration seeks to generalize observed *events*, such as '80% of A's are B's', beyond the data in hand. In particular, the form of inference based on it takes the form:

"*Straight-rule:* if the proportion of red marbles from a sample of size n is (m/n), infer that approximately a proportion (m/n) of all marbles in the urn are red"
(see Salmon, 1967, p. 50)

The reliability of this inference is thought to depend on the *a priori* stipulations of (i) the 'uniformity' of *nature* and (ii) the 'representativeness' of the sample (Mills, 1924, pp. 550-2). In addition, there was an emphasis on 'large enough samples' stemming from the fact that under (i)-(ii) one can show that, as $n \to \infty$, the observed proportion (m/n) converges in probability to the true proportion θ; see Pearson (1920).

Fisher's model-based statistical induction extends the intended scope of induction-by-enumeration by replacing its focus on *events* and associated probabilities with modeling the *mechanism* that underlies the generation of the observed data – a prespecified statistical model – capturing all possible events and associated probabilities. For example, the statistical model underlying the above straight-rule is the *simple Bernoulli model* where the outcome $X{=}1$ denotes the event 'the marble is red', with $\mathbb{P}(X{=}1){=}\theta$, and $X{=}0$ the event 'the marble is not red', with $\mathbb{P}(X{=}0){=}1{-}\theta$, i.e.

$$X_k \backsim \mathsf{BerIID}\left(\theta,\ \theta(1{-}\theta)\right),\ k{\in}\mathbb{N} := \{1,2,...\}, \tag{9}$$

where 'BerIID' reads 'Bernoulli, Independent (I) and Identically Distributed (ID)'. The data $\mathbf{x}_0{:=}(1,0,0,...,1)$ are interpreted as a 'typical' realization of the sample $\mathbf{X} := (X_1, X_2, ..., X_n)$ generated by the process in (9). The inference concerning the proportion θ of red marbles in the urn amounts to choosing the point estimator $\widehat{\theta}_n(\mathbf{X}){=}\frac{1}{n}\sum_{k=1}^{n} X_k$, as yielding a representative value for θ; note that the estimate is $\widehat{\theta}_n(\mathbf{x}_0){=}\left(\frac{m}{n}\right)$. The claim that (m/n) converges in probability to θ is more formally stated in terms of $\widehat{\theta}_n(\mathbf{X})$ being a *consistent* estimator of θ :

$$\lim_{n\to\infty} \mathbb{P}\left(|\widehat{\theta}_n(\mathbf{X}) - \theta| < \epsilon\right) = 1, \qquad \text{for any } \epsilon > 0. \tag{10}$$

Viewed from **Fisher's model-based perspective** the straight-rule inference is fraught with potential unreliability problems. *First*, the inference in the form of

a point estimate is rather weak without some measure of reliability; one needs to calibrate the qualifier 'approximately'. *Second*, reliance on consistency alone provides no assurance for reliable inference for a given sample size n. *Third,* the soundness of the premises of inference, upon which the reliability of inference depends, relies on the validity of the priori stipulations (i)-(ii).

Fisher's recasting of statistical induction addresses these issues in an most effective manner. To begin with he replaces the a priori stipulations (i)-(ii) with the premises of inference which are *testable* vis-a-vis data $\mathbf{x}_0$, the probabilistic assumptions constituting the underlying statistical model given in (9):

[1] $X_k \backsim \mathsf{Ber}(.,.)$,

[2] $\{X_k,\ k\in\mathbb{N}\}$ is Identically Distributed (ID), and

[3] $\{X_k,\ k\in\mathbb{N}\}$ is Independent (I),

Hence, the *soundness* of the premises is no longer a matter of faith in stipulations (i)-(ii), but it can, and should, be empirically established *a posteriori*; see Fisher (1922), p. 314. Having specified the *statistical model* (assumptions [1]-[3]) explicitly, Fisher would then proceed to derive the *sampling distribution* $f(\widehat{\theta}_n(\mathbf{x}))$ under assumptions [1]-[3]:

$$\widehat{\theta}_n(\mathbf{X}) \backsim \mathsf{Bin}\left(\theta,\ \tfrac{\theta(1-\theta)}{n}\right), \quad \text{for any } n>1, \tag{11}$$

where 'Bin' stands for a 'Binomial' distribution. $f(\widehat{\theta}_n(\mathbf{x}))$, $\mathbf{x}\in\{0,1\}^n$ gives a complete description of the probabilistic structure of $\widehat{\theta}_n(\mathbf{X})$, and provides the error probabilities needed to assess the reliability of any inference concerning θ.

In Fisher's model-based induction, *learning from data* takes the form of applying effective (optimal) inference methods whose error probabilities – how often these methods give rise to erroneous inferences – are ascertainable. Indeed, one can use the same error probabilities to shed light on the weaknesses of induction-by-enumeration.

Consistency can now be seen as a *minimal* (optimal) property of an estimator, which, by itself, does not ensure the reliability of inference for a given n. Although (10) invokes the sampling distribution of $\widehat{\theta}_n(\mathbf{X})$, it only concerns its behavior as $n\to\infty$, providing at best crude upper bounds for the uncertainty associated with any inference concerning θ. In contrast, using (11) one can invoke more pertinent *finite sample* properties (valid for any $n>1$) to assess the optimality of $\widehat{\theta}_n(\mathbf{X})$ as an estimator of θ, such as *unbiasedness*, *sufficiency* and *full efficiency*; see Cox and Hinkley (1974). More importantly, the uncertainty associated with any inference concerning θ can now be 'quantified' in terms of the ascertainable *error probabilities*. For example, instead of a point estimate $\widehat{\theta}_n(\mathbf{x}_0)=\left(\frac{m}{n}\right)$ one can define the two-sided Confidence Interval (CI) for θ :

$$\mathbb{P}\left(\widehat{\theta}_n(\mathbf{X})-c_{\frac{\alpha}{2}}s_n\leq\theta\leq\widehat{\theta}_n(\mathbf{X})+c_{\frac{\alpha}{2}}s_n\right)\ =1-\alpha,$$

where $s_n^2=\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}$ which quantifies the relevant 'approximation' invoked by the straight rule via the coverage probability $1-\alpha$. For n=10, and m=2, i.e. $\widehat{\theta}_n(\mathbf{x}_0)$=.2,

the observed 95% CI for θ is $[-.048 \leq \theta \leq .448]$; for simplicity we use the Normal approximation $c_{\frac{\alpha}{2}}$=1.96. Using this result, one can show that the straight-rule inference is likely to be unreliable (imprecise) because the width of this observed CI (0.496) indicates that the point estimate is not very precise and the fact that it includes 0 suggests that θ is statistically indistinguishable from zero; see Lehmann (1986).

Bayesian induction. An alternative interpretation to the straight-rule can be given in the context of Bayesian approach to inference. This approach postulates the same premises (9) as the frequentist inference, but in addition it assumes a uniform prior for θ, i.e. $\pi(\theta) \backsim U(0,1)$. The combination of this prior and the likelihood function gives rise to a Beta *posterior* of the form:

$$\pi(\theta \mid \mathbf{x}_0) \propto \pi(\theta) \cdot L(\theta; \mathbf{x}_0) = \left[\theta^{n\overline{x}_n}(1-\theta)^{n(1-\overline{x}_n)}\right],$$

and $\overline{x}_n = \left(\frac{m}{n}\right)$, being the mode of $\pi(\theta \mid \mathbf{x}_0)$, can be interpreted as the *Bayesian estimator* of θ. For n=10, and m=2, $\overline{x}_n$=.2, and a .95 *credible interval* for θ yields $(.060 \leq \theta \leq .517)$; see Schervish (1995). In this context, *learning from data* takes the form of revising somebody's *prior beliefs,* represented by $\pi(\theta)$, in light of the sample information in $L(\theta; \mathbf{x}_0)$.

4.2.2 The frequency interpretation of probability

The frequentist interpretation of probability pioneered in statistics by Fisher (1921) has been questioned by Bayesians ever since as circular; see Keuzenkamp (2000). The objective of this sub-section is to argue that the circularity charge is misplaced.

The basic formal result invoked for the frequency interpretation is the *Strong Law of Large Numbers (SLLN)*; a stronger version of (10). This theorem states that *under certain restrictions on the probabilistic structure of the process* $\{X_k,\ k \in \mathbb{N}\}$, it follows that:

$$\mathbb{P}\left(\lim_{n\to\infty}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = p\right) = 1. \tag{12}$$

The first SLLN was proved by Borel in 1909 in the case of a Bernoulli, IID process, but since then the result in (12) has been extended to hold with much less restrictive probabilistic structure, including $\{X_k,\ k \in \mathbb{N}\}$ being a *martingale difference* process; see Spanos (1999), pp. 476-481. The result in (12) is often invoked to define the *frequentist probability* of an event A:=$\{X$=1$\}$ via:

$$P(A) := \lim_{n\to\infty}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = p. \tag{13}$$

To what extent does (12) provide a justification of the frequentist interpretation of probability as given in (13)? The issue often raised is that this justification is *circular*: it uses *probability* to define *probability*; see Lindley (1965), p. 5. This is denied by some notable mathematicians including Renyi (1970, p. 159) who draws a clear distinction between the intuitive description in (13), and the purely a mathematical result in (12), dismissing the circularity charge as based on conflating the two; see Renyi (1970), p. 159. Indeed, a closer look at (12) reveals that the mathematical theory underlying the result is that of a Lebesgue measure.

4.2.3 Statistical induction: factual vs. hypothetical reasoning

The difference in the nature of reasoning between estimation and testing has caused numerous confusions in the literature, especially as it relates to the relevant error probabilities of different procedures (estimation, testing prediction), as well as the interpretation of the inference results. The optimality of inference methods in frequentist statistics is defined in terms of their capacity to give rise to valid inferences (trustworthiness), evaluated in terms of the associated *error probabilities*: how often these procedures lead to erroneous inferences.

To simplify the discussion the simple Normal model where:

$$X_k \backsim \mathsf{NIID}\left(\mu,\ \sigma^2\right),\ k\in\mathbb{N}, \tag{14}$$

is used for illustration purposes. All forms of inductive inference assume that the prespecified statistical model, generically specified by:

$$\mathcal{M}_\theta(\mathbf{x})=\{f(\mathbf{x};\theta),\ \theta\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X,$$

is *valid*, giving rise to the distribution of the sample $f(\mathbf{x};\theta)$, and then proceed to draw inferences concerning the underlying data generating mechanism via the unknown parameters $\theta:=(\mu,\sigma^2)$. In the case of the simple Normal model $f(\mathbf{x};\theta)$ is:

$$f(x_1,x_2,...,x_n;\theta)=\prod_{k=1}^{n}\frac{e^{-\frac{(x_k-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}=(\tfrac{1}{\sigma\sqrt{2\pi}})^n\exp\left\{-\tfrac{\sum_{k=1}^n(x_k-\mu)^2}{2\sigma^2}\right\},\ \mathbf{x}\in\mathbb{R}^n.$$

Inferences concerning θ are based on the sampling distribution of a particular statistic, say $Y_n=h(X_1,X_2,...,X_n)$ (estimator, test statistic, predictor), which theoretically can be derived from $f(\mathbf{x};\theta)$ via:

$$\mathbb{P}(Y_n\leq y)=F(y)=\underbrace{\int\int\cdots\int}_{\{h(x_1,x_2,...,x_n)\leq y\}}f(x_1,x_2,...,x_n;\theta)dx_1dx_2\cdots dx_n. \tag{15}$$

Statistical inference. It is well known (Cox and Hinkley, 1974) that in this case the statistics:

$$\overline{X}_n=\tfrac{1}{n}\sum_{k=1}^{n}X_k,\qquad s^2=\tfrac{1}{(n-1)}\sum_{k=1}^{n}(X_k-\overline{X})^2, \tag{16}$$

constitute 'optimal' estimators of (μ,σ^2), with sampling distributions:

$$\overline{X}_n\backsim\mathsf{N}\left(\mu,\tfrac{\sigma^2}{n}\right),\qquad s^2\backsim\tfrac{\sigma^2}{(n-1)}\chi^2(n-1), \tag{17}$$

where $\chi^2(n-1)$ denotes the chi-square distribution with $(n-1)$ degrees of freedom. To these sampling distributions, one should add Gosset's (1908) famous result:

$$\tfrac{\sqrt{n}(\overline{X}_n-\mu)}{s}\backsim\mathsf{St}(n-1), \tag{18}$$

which inspired Fisher (1921, 1922) to pioneer the model-based frequentist approach to statistical inference. What is often not appreciate enough is that these sampling distributions are interpreted very differently in inference, depending on the nature of the underlying form of reasoning employed in each case.

The reasoning used in estimation and prediction is known as *factual* because it concerns evaluation of $\mathcal{M}_\theta(\mathbf{x})$ under the True State of Nature (TSN), but the reasoning underlying hypothesis testing is known as *hypothetical* because it is based on conjectural scenarios concerning $\mathcal{M}_\theta(\mathbf{x})$. To illustrate this, let us focus on (18); analogous results hold for (17). As (18) stands it constitutes a **pivot** – a function which depends on both the sample and parameter spaces – whose interpretation demands to be spelled out under the different forms of reasoning.

(i) **Factual reasoning** relies on evaluating the sampling distribution of a statistic under the TSN and interpreting what actually happened as an instantiation of that.

(a) For μ_* true value of μ :

$$\tau(\mathbf{X};\mu)=\tfrac{\sqrt{n}(\overline{X}_n-\mu_*)}{s} \overset{\mathsf{TSN}}{\backsim} \mathsf{St}(n-1). \tag{19}$$

(b) For a given (arbitrary) value of μ, say μ_1 :

$$\tau_1(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_1)}{s} \overset{\mathsf{TSN}}{\backsim} \mathsf{St}(\delta_1;n-1),\ \delta_1=\tfrac{\sqrt{n}(\mu_*-\mu_1)}{\sigma_*},$$

where $\mathsf{St}(\delta_1;n-1)$ denotes a non-central Student's t distribution with non-centrality parameter $\delta_1=\frac{\sqrt{n}(\mu_*-\mu_1)}{\sigma_*}$; σ^2_* denotes the true value of σ^2.

(ii) **Hypothetical reasoning** relies on comparing the sampling distribution of a statistic under different hypothetical scenarios with what actually happened.

(c) For a given value of μ, say μ_0 :

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

(d) For a different (given) value of μ, say $\mu_1\neq\mu_0$:

$$\tau(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s} \overset{\mu=\mu_1}{\backsim} \mathsf{St}(\delta;n-1),\quad \tau_1(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_1)}{s} \overset{\mu=\mu_1}{\backsim} \mathsf{St}(n-1), \tag{20}$$

where $\delta=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$. The results in (a)-(d) frame statistical inductive reasoning in the sense that they provide the basis for evaluating the relevant error probabilities.

Inductive reasoning in estimation. In point estimation an 'optimal' estimator $\widehat{\theta}$ selects the most 'representative' value of the unknown parameter θ, in the sense that it gives rise to values as close to close to θ_* (the true θ), as possible. The form of reasoning that underlies estimation is that of *factual reasoning,* where the sampling distribution in (17) are understood as evaluated under the TSN, with (μ_*,σ^2_*) denoting the 'true' values of (μ,σ^2) , whatever those happen to be. The difficulty with this form of inductive inference is that rendering the error probabilities ascertainable requires one to know (μ_*,σ^2_*) . To circumvent this problem properties of point estimators, such as *Unbiasedness* and minimum *Mean Square Error (MSE)*, are usually defined in terms of a quantifier that involves *all possible values* of θ :

$$E(\widehat{\theta})=\theta^2,\ MSE(\widehat{\theta})=E(\widehat{\theta}-\theta)^2,\quad \text{for all } \theta\in\Theta.$$

However, a moment's reflection suggests that there is something wrong-headed about the use of this quantifier because it gives rise to dubious results like $\overline{X}_n$ is *not* better, on MSE grounds, than $\widetilde{\mu}$=7405926 (which ignores the data completely), as an estimator of μ. This is because $\overline{X}_n$ is *not* better for all $\theta\in\Theta$ than $\widetilde{\mu}$ since:

$$MSE(\overline{X}_n)=\tfrac{\sigma^2}{n}>MSE(\widetilde{\mu}) \text{ for } \mu\in\left(7405926-\tfrac{\sigma^2}{n},\ 7405926+\tfrac{\sigma^2}{n}\right)$$

i.e. for values of μ close enough to $\widetilde{\mu}$. In this sense, point estimators and their optimal properties do *not* provide enough information to evaluate the *reliability* of a particular point estimate for a given n.

Interval estimation (Confidence Interval (CI)) attempts to rectify this deficiency by providing a way to evaluate the probability of covering the true value θ^* of θ, without knowing θ^*. In the case of the simple Normal model:

$$\mathbb{P}\left(\overline{X}_n - c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}}) \leq \mu \leq \overline{X}_n + c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}}); \mu{=}\mu^*\right) {=}1{-}\alpha, \tag{21}$$

where '$\mu{=}\mu^*$' indicates evaluation under the TSN. This result stems from (19), and evaluation of (21) under $\mu \neq \mu^*$ yields α - the *coverage error* probability: the probability that the interval does *not* contain μ^*.

Inductive reasoning in prediction. Prediction takes the estimated model $\mathcal{M}_{\widehat{\theta}}(\mathbf{x})$ as given and seeks a best guesstimate for observable *events* beyond the observation period, say $X_{n+1}{=}x_{n+1}$, in the form of a predictor $\widehat{X}_{n+1} = h(\mathbf{X})$. The prediction error is defined by $e_{n+1}{=}(X_{n+1}{-}\widehat{X}_{n+1})$, and its sampling distribution is evaluated under the *true state of nature* (TSN), as in the case of estimation, relying on (i)(a).

Example. In the case of the simple Normal model (table 1), the prediction error takes the form:

$$e_{n+1}{=}(X_{n+1}{-}\overline{X}_n) \overset{\text{TSN}}{\backsim} \mathsf{N}\left(0,\ \sigma_*^2(1{+}\tfrac{1}{n})\right) \ \Rightarrow\ \tfrac{e_{n+1}}{s\sqrt{(1+\frac{1}{n})}} \overset{\text{TSN}}{\backsim} \mathsf{St}\,(n{-}1)\,.$$

This can be used to construct a prediction CI of the form:

$$\mathbb{P}\left(\overline{X} - c_{\frac{\alpha}{2}} s\sqrt{(1+\tfrac{1}{n})} \leq x_{n+1} \leq c_{\frac{\alpha}{2}} s\sqrt{(1+\tfrac{1}{n})}; \mu{=}\mu^*, \sigma^2{=}\sigma_*^2\right) {=}1{-}\alpha, \tag{22}$$

where $(1{-}\alpha)$ denotes the coverage probability for the realized value of X_{n+1}.

Inductive reasoning in hypothesis testing. In contrast to estimation and prediction, the reasoning underlying hypothesis testing is *hypothetical*, relying on (ii)(c)-(d). For instance, in testing the *hypotheses*:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0, \tag{23}$$

the test statistic $\tau(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n{-}\mu_0)}{s}$ is evaluated under numerous (often infinite) hypothetical scenarios:

$$\tau(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n{-}\mu_0)}{s} \overset{H_0}{\backsim} \mathsf{St}(n{-}1), \qquad \tfrac{\sqrt{n}(\overline{X}_n{-}\mu_0)}{s} \overset{H_1(\mu_1)}{\backsim} \mathsf{St}(\delta; n{-}1), \text{ for any } \mu_1{\neq}\mu_0, \tag{24}$$

where $\delta{=}\frac{\sqrt{n}(\mu_1{-}\mu_0)}{\sigma}$; these follow from (20) These sampling distributions are then used to define the *type I and II error probabilities* via:

$$\mathbb{P}\left(|\tau(\mathbf{X})| \ > c_{\frac{\alpha}{2}};\ H_0\right) {=}\alpha, \qquad \mathbb{P}\left(|\tau(\mathbf{X})| \ \leq c_{\frac{\alpha}{2}};\ H_1(\mu_1)\right) {=}\beta(\mu_1), \text{ for } \mu_1{>}\mu_0, \tag{25}$$

as well as the *power* of the test:

$$\pi(\mu_1) = \mathbb{P}(\tau(\mathbf{X}) > c_\alpha;\ H_1(\mu_1)){=}1{-}\beta(\mu_1), \text{ for all } \mu_1 > \mu_0.$$

It can be shown that the above test is Uniformly, Most Powerful Unbiased; see Lehmann (1986), Cox and Hinkley (1974).

Notwithstanding the well-known (mathematical) duality between hypothesis testing and *interval estimation:*

$$C_0(\alpha)=\{\mathbf{x}:|\tau(\mathbf{x})|\leq c_{\frac{\alpha}{2}}\}\Leftrightarrow CI(\mathbf{X};\alpha)=\{\mu:|\tau(\mathbf{X};\mu)|\leq c_{\frac{\alpha}{2}}\},$$

(see Lehmann, 1986), there is a crucial difference in the interpretation of the two types of inference, stemming from their underlying reasoning. In factual reasoning that there is only *one* scenario, but usually in hypothetical reasoning there is an *infinite* number of possible scenarios. This has two important implications. *First*, due to the legion of hypothetical scenarios, testing poses sharper questions and often elicits more precise answers. *Second*, the error probabilities associated with hypothetical reasoning are properly defined *post-data* as well, but those associated with factual reasoning become *degenerate*. This is because factual reasoning inevitably involves the TSN (μ^*), and thus post-data the inference is either true or false; the relevant probabilities are either *one* (1) or *zero* (0). Which situation is instantiated in a particular case can only be assessed when the true value μ^* is known!

This crucial difference between post-data error probabilities has many important implications for statistical inference in the sense that it can be used to shed light on several philosophical/methodological issues mentioned above.

4.2.4 Post-data error probabilities

The post-data degeneracy of the factual error probabilities is the reason why one cannot distinguish between different values of μ within the observed CI:

$$[\overline{x}_n-c_{\frac{\alpha}{2}}(s/\sqrt{n}),\ \overline{x}_n+c_{\frac{\alpha}{2}}(s/\sqrt{n})], \tag{26}$$

using probabilistic arguments; $\overline{x}_n$ denotes the observed value of $\overline{X}_n$. This is because, post-data the observed CI covers the true μ with probabilities 0 or 1. This brings out the fallacy in often made claims like:

> "... the [parameter] is much more likely to be near the middle of the confidence interval than towards the extremes." (Altman et al, 2000, p. 22).

Indeed, one cannot provide proper post-data *coverage probabilities* associated with:

$$\mu\geq\overline{x}_n-c_{\frac{\alpha}{2}}(s/\sqrt{n})\text{ or }\mu\leq\overline{x}_n+c_{\frac{\alpha}{2}}(s/\sqrt{n}), \tag{27}$$

beyond the uninformative degenerate ones. Similarly fallacious is the often invoked argument that one can evaluate proper post-data coverage error probabilities for CIs using a sequence of CIs by changing α. The truth of the matter is that there is no valid way one can transfer the pre-data error probability to the observed CI, since post-data all coverage probabilities become degenerate! This can be seen in diagram below where 3 typical observed CIs are shown for different coverage probabilities, and it's clear that there is no probabilistic statement of the type (27) one can make that will be consistent with all three cases!

μ_*

1. $\vdash - - -\overline{x}_n - -- \dashv$-99%
   $\vdash - - \overline{x}_n - - \dashv$-80%
   $\vdash -\overline{x}_n - \dashv$-50%
   $\vdash \overline{x}_n \dashv$-30%
2. $\vdash - - - - \overline{x}_n - - - - \dashv$-99%
   $\vdash - - -\overline{x}_n - -- \dashv$-80%
   $\vdash - - \overline{x}_n - - \dashv$-50%
   $\vdash \overline{x}_n \dashv$-30%
3. $\vdash - - - - -\overline{x}_n - - - -- \dashv$-99%
   $\vdash - - - - \overline{x}_n - -- \dashv$-80%
   $\vdash - - \overline{x}_n - - \dashv$-50%
   $\vdash \overline{x}_n \dashv$-30%

On the other hand, hypothetical (testing) reasoning relies on comparing what actually happened to what it is expected under different hypothetical values of μ; it does not involve TSN (μ_*). Hence, there is nothing to prevent one from evaluating, post-data, claims of the form:

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

using well-defined post data error probabilities associated with the testing reasoning underlying the above t-test. In view of that, it is evident that one can evaluate the probability of claims of the form given in (27) by relating μ_1 to whatever values one is interested in, including $\overline{x}_n \pm c_{\frac{\alpha}{2}}(s/\sqrt{n})$ for different α, using hypothetical (not factual) reasoning. This is exactly how the severity assessments (see next), based on post-data testing error probabilities, circumvent the problem facing observed CIs and provide an effective way to evaluate inferential claims for different values of μ belonging to such an interval. Indeed, one might go as far as to conjecture that Fisher's (1935c, 1956) fiducial reasoning could not succeed precisely because the factual reasoning cannot deliver what he had in mind, but the severe testing reasoning can because it is hypothetical.

4.3 Severe testing reasoning

Practitioners in a variety of disciplines have long felt that the smaller the p-value the better the accord of $\mathbf{x}_0$ *with* H_1, but the dependence of $p(\mathbf{x}_0)$ on the sample size made that intuition very difficult to flesh out correctly. A way to formalize this intuition and bridge the gap between the coarse accept/reject rule and the evidence for or against a hypothesis warranted by the data was proposed by Mayo (1991) in the form of a post-data evaluation of inference using the notion of severity.

4.3.1 Post-data severity evaluations

A hypothesis H passes a *severe test* T with data $\mathbf{x}_0$ if,

(S-1) $\mathbf{x}_0$ agrees with H, and

(S-2) with very high probability, test T would have produced a result that accords less well with H than $\mathbf{x}_0$ does, if H were false.

The inferential interpretation stems from the fact that H passing test T provides good evidence for inferring H (is correct) to the extent that T severely passes H with data $\mathbf{x}_0$. By evaluating the severity of a test T, as it relates to claim H and data $\mathbf{x}_0$, we learn about the kind and extent of errors that T was (and was not) highly capable of detecting, thus informing one of errors ruled out and errors still in need of further probing. Thus, from the thesis of *learning from error*, it follows that a severity assessment allows one to determine whether there is evidence for (or against) claims; see Mayo (1996).

In order to see how the above notion of severity can be formalized let us return to the simple Normal model (14) and, to simplify the notation, consider the one-sided *hypotheses*:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0. \tag{28}$$

It is well-known that the test defined by:

$$T_\alpha{:=}\left\{\tau(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s},\ C_1(\alpha){=}\{\mathbf{x}:\tau(\mathbf{x})>c_\alpha\}\right\},$$

is Uniformly Most Powerful (UMP); see Lehmann (1986). Depending on whether this test has given rise to accept or reject H_0 with data $\mathbf{x}_0$, the post-data evaluation of that inference takes the form of:

$$\begin{aligned}\mathsf{Sev}(T_\alpha;\mathbf{x}_0;\mu\leq\mu_1) &= \mathbb{P}(\tau(\mathbf{X})>\tau(\mathbf{x}_0);\mu>\mu_1),\\ \mathsf{Sev}(T_\alpha;\mathbf{x}_0;\mu>\mu_1) &= \mathbb{P}(\tau(\mathbf{X})\leq\tau(\mathbf{x}_0);\mu\leq\mu_1),\end{aligned} \tag{29}$$

respectively, where $\mu_1{=}\mu_0{+}\gamma$, for $\gamma\geq 0$. The severity evaluation introduces a *discrepancy parameter* in order to define the relevant *inferential claims* associated when accepting $(\mu\leq\mu_1)$ or rejecting H_0 $(\mu>\mu_1)$. In the case of accept, the idea is to establish the *smallest discrepancy* γ from H_0, and in the case of reject establish the *largest discrepancy* γ from H_0, that is licensed by data $\mathbf{x}_0$. The discrepancy parameter γ plays a crucial role in the severity assessment because it reflects what Fisher called the 'strength of evidence' for or against H_0 warranted by data $\mathbf{x}_0$.

The severity assessment allows for a post-data objective interpretation of any N-P test result that bridges the gap between the coarse accept/reject decision and the evidence for or against the null warranted by the data; it can be applied to any (properly defined) N-P test. When the severity evaluation of a particular inferential claim, say $\mu\leq\mu_0{+}\gamma$, is very high (close to one), it can be interpreted as indicating that this claim is warranted to the extent that the test has ruled out discrepancies larger than γ; the underlying test would have detected a departure from the null as large as γ almost surely, and the fact that it didn't suggests that no such departures were present. Viewing N-P tests from the severe testing perspective, suggests that the value of confining *error probabilities* at small values is not only the desire to have a good track record in the *long run*, but also because of how this lets us severely probe,

and thereby learn about, the process that gave rise to data $\mathbf{x}_0$. This emphasizes the *learning from error* by applying highly probative procedures. Severity takes the pre-data error probabilities as calibrating the generic capacity of the test procedure and custom-tailors that to the particular case of data $\mathbf{x}_0$ and the relevant inferential claim H, rendering the post-data evaluation test-specific, data-specific and claim-specific, hence the notation in (29).

The chronic fallacies of acceptance and rejection associated with N-P testing, alluded to above, can also be addresses using severe testing reasoning; see Mayo (1996), Mayo and Spanos (2006).

4.3.2 Severe testing and the p-value

Viewed from the severity perspective the p-value can be interpreted as a post-data error probability that lacks the discrepancy parameter refinement. To see this let us consider a severe-testing interpretation of using a small p-value, say $p = .01$, to infer that data $\mathbf{x}_0$ provide evidence against H_0. Such a small p-value indicates that $\mathbf{x}_0$ *accords with* H_1, and the question is whether it provides evidence for a particular $\mu_1 = \mu_0 + \gamma$, $\gamma \geq 0$, in H_1. The severe-testing interpretation suggests that $H_1 : \mu > \mu_0$ has passed a severe test because the probability that test T_α would have produced a result that accords less well with H_1 than $\mathbf{x}_0$ does is:

$$\mathsf{Sev}(T_\alpha; \mathbf{x}_0; \mu > \mu_0) = \mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu \leq \mu_0) \quad = 1 - \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu = \mu_0) = .99,$$

and is very high, for *some* discrepancy $\gamma \geq 0$, but provides no information concerning the *magnitude* of the discrepancy γ licensed by this data $\mathbf{x}_0$.. The severity construal of the p-value also addresses the problem of its dependence on the sample size; see Mayo and Spanos (2006).

4.3.3 Statistical vs. substantive significance

Of particular interest in econometrics is special case of the fallacy of rejection where statistical significance is misinterpreted as substantive significance. It is interesting to note that this problem was first raised by Hodges and Lehmann (1954), but their attempt to address it was not successful. In the case of the hypotheses in (28), rejecting H_0 only establishes the presence of some discrepancy from μ_0, say $\delta > 0$, but it does not provide any information concerning the magnitude of δ. The severity evaluation $\mathsf{Sev}(T_\alpha; \mathbf{x}_0; \mu > \mu_1)$ associated with the claim that $\mu > \mu_1 = \mu_0 + \gamma$, for some $\gamma \geq 0$, can be used to establish the warranted discrepancy γ^*, and then proceed to assess whether γ^* is also substantively significant or not; see Mayo and Spanos (2006).

The error-statistical perspective also elucidates the comparisons between p-values and CIs and can be used to explain why the various attempts to relate p-value and observed confidence interval curves (see Birnbaum, 1961, Kempthorne and Folks (1971), Poole, 1987) were unsuccessful. In addition, it can be used to shed light on the problem of evaluating 'effect sizes' (see Rosenthal et al, 1999) sought after in some applied fields like psychology and epidemiology; see Spanos (2004).

4.4 A statistical model has 'a life of its own'

Perhaps the most crucial feature of error-statistics is its reliance on error probabilities, pre-data, to evaluate the capacity of an inference procedure, and post-data the evidential warrant of a claim. For such evaluations to be reliable, however, one needs to ensure the validity of the underlying statistical model $\mathcal{M}_\theta(\mathbf{x})$ demarcating the premises of inference when viewed in conjunction with data $\mathbf{x}_0$:=$(x_1, x_2, ..., x_n)$.A crucial precondition for ensuring statistical adequacy is a complete specification of a statistical model in terms of probabilistic assumptions which are testable vis-a-vis data $\mathbf{x}_0$; e.g. table 2. *Statistical adequacy* is tantamount to affirming the assumption that data $\mathbf{x}_0$ constitute a 'truly typical realization' of the stochastic process represented by $\mathcal{M}_\theta(\mathbf{x})$. In the context of the error-statistical approach statistical adequacy is assessed using thorough *Mis-Specification (M-S) testing*: probing for departures from the probabilistic assumptions comprising $\mathcal{M}_\theta(\mathbf{x})$ vis-a-vis data $\mathbf{x}_0$.

Table 2 - The simple Normal model

Statistical GM:		$X_k = \mu + u_k,\ k\in\mathbb{N}$,
[1]	Normal:	$X_k \backsim \mathsf{N}(.,.)$, for all $k\in\mathbb{N}$,
[2]	Constant mean:	$E(X_k) = \mu$, for all $k\in\mathbb{N}$,
[3]	Constant variance:	$Var(X_k) = \sigma^2$, for all $k\in\mathbb{N}$,
[4]	Independence:	$\{X_k,\ k\in\mathbb{N}\}$ - independent process.

(30)

Denoting the set of all possible models that could have given rise to data $\mathbf{x}_0$ by $\mathcal{P}(\mathbf{x})$, the generic form of M-S testing is:

$$H_0 : f_*(\mathbf{x})\in\mathcal{M}_\theta(\mathbf{x}), \quad \text{vs.} \quad H_1 : f_*(\mathbf{x})\in [\mathcal{P}(\mathbf{x})-\mathcal{M}_\theta(\mathbf{x})], \tag{31}$$

where $f_*(\mathbf{x})$ denotes the 'true' joint distribution of the stochastic process $\{X_t,\ t\in\mathbb{N}\}$. The specification of the null and alternatives in (31) indicates most clearly that M-S testing is probing outside the boundaries of $\mathcal{M}_\theta(\mathbf{x})$, in contrast to N-P testing which is probing within this boundary; see Spanos (1999).

The problem that needs to be addressed for (31) to be implementable is to particularize $\overline{\mathcal{M}_\theta(\mathbf{x})}$:= $[\mathcal{P}(\mathbf{x})-\mathcal{M}_\theta(\mathbf{x})]$ representing the set of all possible alternative models. This can be as specific as a broader statistical model $\mathcal{M}_\psi(\mathbf{x})$ that parametrically encompasses $\mathcal{M}_\theta(\mathbf{x}) \subset \mathcal{M}_\psi(\mathbf{x})$, or as vague as a direction of departure from $\mathcal{M}_\theta(\mathbf{x})$, which might only be implicitly determined, such as a goodness-of-fit test.

The hypothetical reasoning underlying M-S tests is similar to Fisher's *significance test reasoning*: data $\mathbf{x}_0$ provide evidence for a departure from a null hypothesis H_0 in so far as the value of a distance statistic $d(\mathbf{x}_0)$ is 'improbably far' from what would have been expected if H_0 were true, i.e. $\mathcal{M}_\theta(\mathbf{x})$ is true. In the case where the alternative is specified in terms of an encompassing model $\mathcal{M}_\psi(\mathbf{x})$, $d(\mathbf{X})$ can be chosen using power, but in the case where $\overline{\mathcal{M}_\theta(\mathbf{x})}$ is not explicitly specified, the chosen form of $d(\mathbf{X})$ defining 'improbably far', defines the implicit alternative to be the direction of departure from $\mathcal{M}_\theta(\mathbf{x})$ with maximum power; see Davidson and MacKinnon (1987). In a M-S test the primary role for the particularized alternative

is to determine the form of the distance function, and hence the power of the test. In that sense, rejection of the null in an M-S test cannot (should not) be interpreted as evidence for the particularized alternative, implicit or explicit. The validity of a particularized alternative such as $\mathcal{M}_\psi(\mathbf{x})$ needs to be established on its own merit; $\mathcal{M}_\psi(\mathbf{x})$ shown to be statistically adequate vis-a-vis data $\mathbf{x}_0$. Therefore, accepting the particularized (explicit) alternative in a M-S test constitutes a classic example of the fallacy of rejection.

How to choose (or create) a battery of M-S tests to probe for possible departures from $\mathcal{M}_\theta(\mathbf{x})$ as thoroughly as possible, and at the same time avoid circularity or infinite regress, raises both philosophical/methodological and technical issues and problems beyond the scope of this paper; see Spanos (1999), Mayo and Spanos (2004, 2006). Having said that, it is important to comment briefly on whether M-S testing involves illegitimate double-use of data (see Kennedy, 2003) in the sense that the same data is used to draw inferences concerning θ as well as test the validity of the model assumptions (e.g. [1]-[4], table 2). Although nobody can deny that M-S testing involves double-use of data, the charge of illegitimacy can be challenged on several grounds; see Spanos (2000), Mayo and Spanos (2004). One such ground is purely statistical in nature and relies on separating the sample information into two orthogonal components, one used for inference and the other for model validation purposes. Spanos (2007c) showed that for a variety of statistical models, including the simple Normal and the Linear Regression models, the distribution of the sample $f(\mathbf{x};\theta)$ can be reduced to:

$$f(\mathbf{x};\theta) \propto f(\mathbf{s};\theta)\cdot f(\mathbf{r}), \qquad \forall\,(\mathbf{s},\mathbf{r})\in\mathbb{R}^m_s\times\mathbb{R}^{n-m}_r. \tag{32}$$

where $\mathbf{S}(\mathbf{X})$ is a *minimal sufficient* and $\mathbf{R}(\mathbf{X})$ a *maximal ancillary statistic* for θ, with the two statistics being independent. The crucial argument for relying on $f(\mathbf{r})$ for model validation purposes is that the probing for departures from $\mathcal{M}_\theta(\mathbf{x})$ is based on error probabilities that do not depend on the true θ. This argument holds 'approximately' in the case of statistical models whose inference is based on asymptotic Normality, which comprises the overwhelming majority of statistical models of interest in econometrics.

From a methodological perspective the separation in (32) reflects the drastically different questions posed for inference and model adequacy purposes. The question posed for model adequacy purposes is: "do data $\mathbf{x}_0$ represent a truly typical realization of the stochastic mechanism specified by $\mathcal{M}_\theta(\mathbf{x})$?" or equivalently "does the model $\mathcal{M}_\theta(\mathbf{x})$ account for the systematic (statistical) information in data $\mathbf{x}_0$?" In contrast, the questions posed by inferences concerning θ takes the model $\mathcal{M}_\theta(\mathbf{x})$ as given (assumed to be adequate for data $\mathbf{x}_0$), and probe their validity *within* its boundaries; see Spanos (1999).

When any departures from the statistical model assumptions are detected, the next step is to *respecify* $\mathcal{M}_\theta(\mathbf{x})$, by choosing a different model $\mathcal{M}_\varphi(\mathbf{x})$ which accounts for the systematic information left unaccounted for by the original model. For all three facets of statistical modeling, *specification*, *M-S testing* and *respecification,* data plots

(t-plots, scatter plots, P-P and Q-Q plots), as well as non-parametric methods, play a crucial role in guiding one through the type of statistical regularities exhibited by the data; see Spanos (1999, 2000, 2006).

Statistical knowledge. What is important for theory testing purposes is that a *statistically adequate model*, built exclusively on statistical information is independent (separate) from the substantive information, and therefore, it can be used to provide the broader inductive premises for evaluating its adequacy. The independency stems from the fact that the statistical model is erected on purely probabilistic information by capturing the 'chance regularities' exhibited by data $\mathbf{x}_0$, when the latter is viewed as a realization of a generic stochastic process $\{X_k,\ k \in \mathbb{N}\}$, disregarding what X_k measures. In this sense, a statistically adequate model provides a form of *statistical knowledge*, analogous to what Mayo (1996) calls *experimental knowledge*, against which the substantive information should be appraised. The notion of a statistically adequate model formalizes the sense in which data $\mathbf{x}_0$ have 'a life of its own', separate from the one ideated by some theory.

The notion of *statistical knowledge* separate from substantive knowledge has been denied vehemently by mainstream economics for centuries (see Spanos, 2008a), but it constitutes a crucial step in enhancing the reliability of inference in empirical modeling. Indeed, in disciplines which rely primarily on observational data, a statistically adequate model provides a crucial necessary step in assessing the validity of any substantive subject matter information and offers an opportunity for what Schumpeter (1954) called: "in order to know precisely what there is to explain." He went on to extol the importance of a better understanding of data and statistical methods:

> "It is impossible to understand statistical figures without understanding how they have been compiled. It is equally impossible to extract information from them or to understand the information specialists extract for the rest of us without understanding the methods by which this is done – and the epistemological backgrounds of these methods. Thus, an adequate command of modern statistical methods is a necessary (but not sufficient) condition from preventing the modern economist from producing nonsense ... our stake in these methods is too great for us to leave judgment on the virtues and shortcomings ... to specialists." (Schumpeter, 1954, p. 14)

Indeed, one can go as far as to suggest that the one thing that unites critics of textbook econometrics like Hendry (1995, 2000) and Sims (1980), is the call for allowing the data 'to have a life of its own' beyond the one envisaged by some theory or other. In contrast, textbook econometricians dispute the very notion of statistical information. To dispel this myth consider the data exhibited in figures 1 and 2. Viewing the t-plot in figure 1 as a realization of a generic stochastic process $\{X_k,\ k\in\mathbb{N}\}$ (free from any substantive information), it is not unreasonable to conjecture that $\mathbf{x}_0$ constitutes a typical realization of a NIID process, for which the simple Normal model is a particular parameterization; it can be verified that, indeed, assumptions [1]-[4] (table 2) are valid for the data in question. On the other hand if the data

are the ones shown in figure 2, it is reasonable to conjecture that the simple Normal model will be *misspecified* because the t-plot of data $\mathbf{y}_0$ exhibit cycles which indicate departures from the ID assumption; see Spanos (1999), ch. 5. A more appropriate probabilistic structure for the underlying stochastic process $\{Y_k,\ k \in \mathbb{N}\}$ might be that it's Normal, Markov and Stationary. A statistical model that can parameterize such a process is the *AutoRegressive* (AR(1)) model:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + u_t,\ t \in \mathbb{N}; \tag{33}$$

see Spanos (1999, 2001b, 2005) for further details.

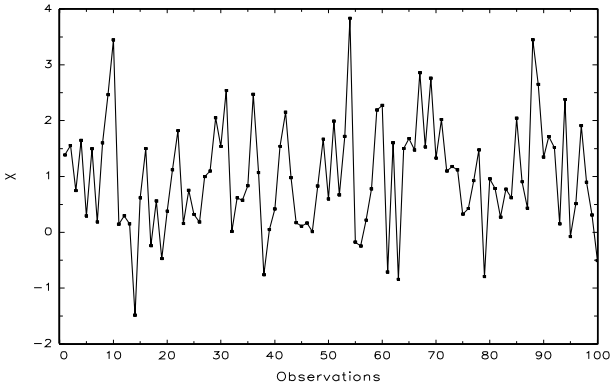

Fig. 1: t-plot of x_t

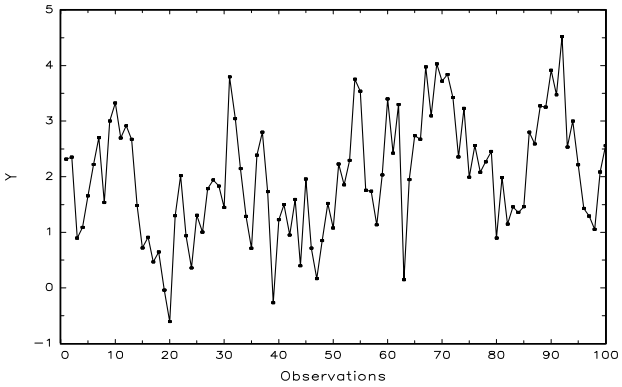

Fig. 2: t-plot of y_t

4.5 Bridging the gap between theory and data

Another aspect of modeling that the error-statistical approach differs appreciably from other perspectives is in terms of how the *statistical* and *substantive information* are integrated without compromising the credibility of either source of information. The problem is viewed more broadly as concerned with bridging the gap between theory and data using a *chain of complecting models,* theory (primary), structural (experimental), statistical (data) built on two different, but related, sources of information: *substantive subject matter* and *statistical information* (chance regularity patterns); see Spanos (2006a). Disentangling the role played by the two sources of information has been a major problem in modern statistics (see Lehmann, 1990, Cox, 1990, Spanos, 2006c). The error-statistical perspective provides a framework in the context of which these sources of information are treated as complementary, and the chain of interconnected models can be used to disentangle their respective roles. *Ab initio,* the statistical information is captured by a statistical model and the substantive information by a *structural model*. The connection between the two models is that a structural model acquires statistical operational meaning when embedded into an adequate statistical model. Let us flesh out some of the details.

4.5.1 How theory relates to a statistical model

The term theory is used generically as any claim hypothesized to elucidate a phenomenon of interest. When one proposes a *theory* to explain the behavior of an observable

variable, say y_k, one demarcates the segment of reality to be modeled by selecting the primary influencing factors $\mathbf{x}_k$, aware that there might be numerous other potentially relevant factors ξ_k (observable and unobservable) influencing the behavior of y_k. A *theory model* is used to denote an idealized mathematical representation of a theory, say:

$$y_k = h^*(\mathbf{x}_k, \xi_k), \quad k\in\mathbb{N}. \tag{34}$$

A model, in general, denotes any idealized mathematical representation of a phenomenon of interest, that facilitates 'learning' about that phenomenon. The guiding principle in selecting the variables in $\mathbf{x}_k$ is to ensure that they collectively account for the *systematic* behavior of y_k, and the omitted factors ξ_k represent non-essential disturbing influences which have only a non-systematic effect on y_k. The potential presence of a large number of contributing factors $(\mathbf{x}_k, \xi_k)$ explains the conjuring of *ceteris paribus* clauses. This line of reasoning transforms the theory model (34) into a *structural (estimable) model* of the form:

$$y_k = h(\mathbf{x}_k; \phi) + \epsilon(\mathbf{x}_k,\xi_k), \quad k\in\mathbb{N}, \tag{35}$$

where $h(.)$ denotes the postulated functional form, ϕ stands for the structural parameters of interest. The *structural error term*, defined to represent all unmodeled influences:

$$\{\epsilon(\mathbf{x}_k,\xi_k) = y_k - h(\mathbf{x}_k;\phi),\ k\in\mathbb{N}\}, \tag{36}$$

is viewed as a function of both $\mathbf{x}_k$ and ξ_k. For (36) to provide a meaningful model for y_k the error term needs to be non-systematic: an *IID* (non-systematic) stochastic process $\{\epsilon(\mathbf{x}_k,\xi_k),\ k\in\mathbb{N}\}$ satisfying the properties:

$$\text{[i]} \quad \epsilon(\mathbf{x}_k,\xi_k) \ \backsim \mathsf{IID}(0,\sigma^2),\ \forall(\mathbf{x}_k,\xi_k)\in\mathbb{R}_{\mathbf{x}}\times\mathbb{R}_{\xi}.\ \big\} \tag{37}$$

In addition, one needs to ensure (see Spanos, 1995) that the generating mechanism (35) is 'nearly isolated' in the sense that the unmodeled component $(\epsilon(\mathbf{x}_k,\xi_k))$ is *uncorrelated* with the modeled influences $(h(\mathbf{x}_k;\phi))$:

$$\text{[ii]} \quad E(\epsilon(\mathbf{x}_k,\xi_k)\cdot h(\mathbf{x}_k;\phi))=0,\ \forall(\mathbf{x}_k,\xi_k)\in\mathbb{R}_{\mathbf{x}}\times\mathbb{R}_{\xi}.$$

Looking at assumptions [i]-[ii] it is clear that they are empirically non-testable because their confirmation would involve *all possible values* of both $\mathbf{x}_k$ and ξ_k. To render them testable one needs to embed this structural into a statistical model; a crucial move that often goes unnoticed. Whether a structural model can be embedded into a statistical model or not depends crucially on the nature of the available statistical data and their relation to the theory in question; sometimes the gap between them might be unbridgeable.

The nature of the embedding itself depends crucially on whether the data $\mathbf{Z}_0{:=}(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n)$ are the result of an experiment or they are non-experimental (observational) in nature, but the aim in both cases is to find a way to transform the structural error $\epsilon(\mathbf{x}_k,\xi_k)$, for all $(\mathbf{x}_k,\xi_k)\in\mathbb{R}_{\mathbf{x}}\times\mathbb{R}_{\xi}$ into a *generic white noise error process* without the quantifier.

Experimental data. In the case where one can perform experiments, controls and 'experimental design' techniques such as *replication*, *randomization* and *blocking,* can often be used to 'neutralize' and 'isolate' the phenomenon from the potential effects of ξ_k by ensuring that the uncontrolled factors cancel each other out; see Fisher (1935). The objective is to transform the structural error into a *generic* IID process:

$$\left(\epsilon(\mathbf{x}_k,\xi_k)\Big\|\begin{smallmatrix}\text{controls,}\\ \text{experimental}\\ \text{design}\end{smallmatrix}\right)=\varepsilon_k\backsim \mathsf{IID}(0,\sigma^2),\qquad k=1,...,n. \tag{38}$$

This in effect embeds the structural model (35) into a *statistical model* of the form:

$$y_k=h(\mathbf{x}_k;\theta)+\varepsilon_k,\quad \varepsilon_k\backsim \mathsf{IID}(0,\sigma^2),\ k=1,2,...,n, \tag{39}$$

where the statistical error term ε_k in (39) is qualitatively very different from the structural error term $\epsilon(\mathbf{x}_k,\xi_k)$ in (35), because ε_k is no longer a function of $(\mathbf{x}_k,\xi_k)$, and its assumptions are rendered empirically testable; see Spanos (2006a). A widely used special case of (39) is the *Gauss Linear model*; see Spanos (1986), ch. 18.

In contrast to a *structural model*, which relies on substantive subject matter information, a statistical model relies on the statistical information in $D(\mathbf{Z}_1,\mathbf{Z}_2,...,\mathbf{Z}_n;\phi)$ – the statistical universe of discourse – that 'reflects' the chance regularity patterns exhibited by the data. Hence, once $\mathbf{Z}_k{:=}(y_k,\mathbf{X}_k)$ is chosen by some theory or theories, a statistical model takes on 'a life of its own' in the sense that it constitutes an 'idealized' probabilistic description of a (vector) stochastic process $\{\mathbf{Z}_k,\ k{\in}\mathbb{N}\}$, giving rise to data $\mathbf{Z}_0$, chosen to ensure that this data represent a 'truly typical realization' of $\{\mathbf{Z}_k,\ k{\in}\mathbb{N}\}$. This statistical information, coming in the form of chance regularity (recurring) patterns, has an objective ontology which can be independently verified. Whether the data exhibit temporal dependence and/or heterogeneity is not something one can fake or falsify, and exists independently of one's beliefs; see Spanos (1999). This purely probabilistic construal of a statistical model takes the sting out of the theory-ladeness of observation charge since theory information is deliberately ignored when data $\mathbf{Z}_0$ are viewed as a realization of a generic stochastic process $\{\mathbf{Z}_k,\ k{\in}\mathbb{N}\}$; see Spanos (2006a-c) for further details

Observational data. In this case, the observed data on $\mathbf{z}_t:=(y_t,\mathbf{x}_t)$ are the result of an ongoing actual data generating process. The embedding in this case is different in the sense that the experimental control and intervention are replaced by judicious *conditioning* on an appropriate information set $\mathfrak{D}_t$ chosen so as to transform the structural error into a generic white-noise statistical error:

$$(u_t|\mathfrak{D}_t)\backsim \mathsf{IID}(0,\sigma^2),\ t=1,2,...,n. \tag{40}$$

Spanos (1999) demonstrates how sequential conditioning provides a general way to decompose orthogonally a stochastic process $\{\mathbf{Z}_t,\ t{\in}\mathbb{N}\}$ into a systematic component μ_t and a *martingale difference process* u_t relative to a conditioning information set $\mathfrak{D}_t$; a modern form of a white-noise process.

A widely used special case of (40) is the *Normal/Linear Regression model* given in table 3, where the testable assumptions [**1**]-[**5**] pertain to the probabilistic structure

of the observable process $\{(y_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t),\ t{\in}\mathbb{N}\}$ and $\mathfrak{D}_t := (\mathbf{X}_t{=}\ \mathbf{x}_t)$. This model can be formally shown to arise from a probabilistic reduction of the joint distribution $D(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_n; \phi) \rightsquigarrow \prod_{t=1}^{n} D(y_t|\ \mathbf{X}_t; \psi_1)$; see Spanos (1986).

Table 3 - The Normal/Linear Regression Model

Statistical GM:	$y_t = \beta_0 + \beta_1^\top \mathbf{x}_t + u_t,\ t{\in}\mathbb{N},$
[1] Normality:	$(y_t \mid \mathbf{X}_t = \mathbf{x}_t) \backsim \mathsf{N}(.,.),$
[2] Linearity:	$E(y_t \mid \mathbf{X}_t = \mathbf{x}_t) = \beta_0 + \beta_1^\top \mathbf{x}_t$, linear in $\mathbf{x}_t$,
[3] Homoskedasticity:	$Var(y_t \mid \mathbf{X}_t = \mathbf{x}_t) = \sigma^2$, free of $\mathbf{x}_t$,
[4] Independence:	$\{(y_t \mid \mathbf{X}_t = \mathbf{x}_t),\ t{\in}\mathbb{N}\}$ is an independent process,
[5] t-invariance:	$\theta := (\beta_0, \beta_1, \sigma^2)$ do not change with t.

5 Philosophical/Methodological issues pertaining to econometrics

The error-statistical perspective has been used to shed light on a number of methodological issues relating to specification, misspecification testing, and respecification, including the role of graphical techniques, structural vs. statistical models, model specification vs. model selection, and statistical vs. substantive adequacy; see Spanos (2006a-c, 2008). In addition, this perspective has been used to illuminate a number of crucial problems in statistics, such as the likelihood princple and the role of conditioning (see Mayo and Cox, 2006, Cox and Mayo, 2008), as well as philosophy of science including the problems of curve-fitting, underdetermination and Duhemian ambiguities; see Mayo (1997, 1997), Mayo and Spanos (2008), Spanos (2007a-b).

In this section the error-statistical perspective is used to shed some new light on a number of different philosophical/methodological issues pertaining to econometrics.

5.1 The Bayesian approach to inductive inference

The Bayesian approach to inference supplements a statistical model with a prior distribution over the parameters $\pi(\theta),\ \theta \in \Theta$, and inference is based on the posterior distribution (see Poirier, 1995):

$$\pi(\theta \mid \mathbf{x}_0) \propto \pi(\theta) \cdot L(\theta; \mathbf{x}_0),\ \theta \in \Theta. \tag{41}$$

The lack of interest in the philosophical foundations of frequentist inference during the 1960s and 1970s, in both statistics and philosophy of science, created the general impression that Bayesian inference occupied the philosophical high ground because of its grounding in the axiomatic approach and its upholding of the likelihood (LP) and coherency principles, while frequentist inference violates both; see Berger and Wolpert (1988). This impression continues to be reiterated, largely unchallenged, to

this day in both statistics (see Berger, 1985, Schervish, 1995, Ghosh et al, 2006), and philosophy of science (see Howson and Urbach, 1993).

A crucial argument for the Bayesian case was based on Birnbaum's (1962) result that two principles developed in the context of frequentist statistics, the *conditionality principle* (CP) and the *sufficiency principle* (SP), when combined give rise to the Likelihood principle (LP). This result was broadly interpreted by Bayesians to imply that recognizing the need for conditional inference leads inevitably to adopting the Bayesian perspective; see Poirer (1995), Ghosh et al (2006). The Bayesian case was built primarily on circumstantial evidence using several examples developed primarily by the frequentist literature for Fisher's (1934) CP. As shown in Spanos (2007d) all the examples, including, including the Welch (1939) uniform, and the mixture of Normals, exhorted by Bayesians to make their case involve some kind of 'rigging' of the statistical model so that it appears as though the CP provides the only way out, when in fact other frequentist principles allow extrication in every case. For example, in the Welch uniform case where $X_k \backsim \mathsf{U}(\theta-.5, \theta+.5)$ the 'rigging' stems from the fact that this distribution is irregular in the sense that its support depends on the unknown parameter θ; see Cox and Hinkley (1974), p. 112. This irregularity creates a constraint between θ and the data $\mathbf{x}_0$, in the sense that post-data $\theta \in A(\mathbf{x}_0)=[x_{[n]}-.5, x_{[1]}+.5]$, where $x_{[n]}=\max(\mathbf{x}_0)$ and $x_{[1]}=\min(\mathbf{x}_0)$, that can be accounted for by post-data error probabilities using truncation, instead of applying the CP.

More recent methodological discussions in some Bayesian circles, calling themselves 'objective' (Bernardo, 2005), shifted their focus away from the earlier foundational principles, and instead they call for 'objectivity' in inference stemming from (i) using the statistical model itself to determine a 'reference' (objective) prior – "One thinks of the objective priors as consensus priors with low information" (Ghosh et al, 2006, p. 147) – (ii) aligning their perspective toward a reconciliation with Fisherian conditionalism, as well as (iii) promoting Bayesian procedures with 'good' frequentist properties; see Berger (2004). The problem is that the moves (i)-(iii) flout the earlier foundational principles subjective Bayesian built their case.

Viewing the Bayesian approach from the error-statistical perspective, raises several philosophical/methodological issues. *First*, by focusing exclusively on $\mathbf{x}_0$ the Bayesian approach leaves no room for assessing the validity of the statistical model defining the likelihood function. This is because Mis-Specification (M-S) testing requires Fisher type significance test reasoning which involves entertaining counterfactual scenarios beyond the observed data $\mathbf{x}_0$ and/or the pre-specified model. Hence, any departures from the statistical model assumptions will invalidate the likelihood function and result in misleading inferences based on $\pi(\theta \mid \mathbf{x}_0)$, irrespective of the choice of the prior $\pi(\theta)$; see Spanos and McGuirk (2001). *Second*, Cox and Mayo (2008) call into question the apparent LP dilemma facing a frequentist to either renounce sufficiency or renounce error probabilities altogether "an illusion." (p. ##). Indeed, Mayo (2008) goes much further than simply raise questions about the cogency of the LP for frequentist inference. She subjects Birnbaum's (1962) "proof"

to a careful logical scrutiny and shows that the underlying argument is fallacious. Intuitively, the source of the fallacy is the misuse of the notion of frequentist sufficiency in the proofs of the LP. *Third*, the choice of the 'reference' priors by 'Objective' (O) Bayesians (see Berger, 2004, Bernardo, 2005), require evaluations which involve the whole of the sample space, leading to violations of the likelihood principle and the stopping rule principle – long embraced as fundamental for, and as logically entailed by, the Bayesian paradigm (see Berger and Wolpert, 1988). *Fourth*, claims by O-Bayesians that their inference methods often enjoy good frequentist properties is somewhat misleading because the notion of error probability is quite different in the two perspectives. Hence, echoing Cox and Mayo (2008) it's not at all obvious in what sense posterior 'error probabilities' are comparable to proper error probabilities.

The question that naturally arises is: "if one renounces the principles cherished by subjective Bayesians (the likelihood, the stopping rule and the coherence), marginalizes the use of prior information as largely untrustworthy, and seeks procedures with 'good' error probabilistic properties (whatever that means), what is there left to render the inference Bayesian, apart from a misconceived provision that the only way to provide an evidential account of inference is to attach probabilities to hypotheses?"

5.2 Statistical model specification vs. model selection

As argued in section 4.2, from the error-statistical pespective the *problem of specification*, as originally envisaged by Fisher (1922), is one of choosing a statistical model $\mathcal{M}_\theta(\mathbf{x})$ so as to render the particular data $\mathbf{x}_0$ a *truly typical realization* of the stochastic process $\{X_k,\ k\in\mathbb{N}\}$ parameterized by $\mathcal{M}_\theta(\mathbf{x})$. This problem is addressed by evaluating $\mathcal{M}_\theta(\mathbf{x})$ in terms of whether it is statistically adequate – it accounts for the regularities in the data; its probabilistic assumptions are validated vis-a-vis data $\mathbf{x}_0$. In cases where the original model is found wanting one should respecify and assess model adequacy until a validated model is found; see Spanos (2006c).

The model validation problem is generally acknowledged in statistics:

"The current statistical methodology is mostly model-based, without any specific rules for model selection or validating a specified model." (Rao, 2004, p. 2)

Over that last 25 years or so, Fisher's specification problem has been recast in the form of *model selection* which breaks up the problem into two stages where, a broad family of models is selected first, and then the particular model within that family is chosen using certain normed-based (goodness-of-fit) criteria; see Rao and Wu (2001). The quintessential example of such a model selection procedure is the Akaike Information Criterion (AIC), and variations/extensions thereof; see Burnham and Anderson (2002). Such norm-based model selection encompasses several procedures which are motivated by mathematical approximation such as curve-fitting by least-squares, structural estimation using GMM as well as nonparametric procedures.

Spanos (2007d) argues that model selection procedures, as currently understood, do *not* address the original problem, as envisaged by Fisher (1922), and invariably gives rise to unreliable inferences because:

(a) model selection begins with a prespecified family of models assumed to nest the 'true' model, without implementing any form of model validation,
(b) even when the 'true' model is nested within the prespecified family of models, satisfying these norm-based criteria does not ensure the reliability of inferences because the associated error probabilities are unknown, and
(c) the model chosen does not necessarily account for the regularities in the data, since the evaluation is based primarily on fit.

This is illustrated in Spanos (2007b) where the Kepler and Ptolemy models for the motion of the planets are compared in terms of goodness-of-fit vs. statistical adequacy. It is shown that, despite the excellent fit of the Ptolemaic model, it does not 'account for the regularities in the data', contrary to conventional wisdom; see Glymour (1980), Laudan (1977). In contast, the statistical adequacy of the Kepler model renders it a statistical model with a life of its own, regardless of its substantive adequacy which stems from Newton's law of universal graviation. Substantive subject matter information is crucially important in learning from data about phenomena of interest, but no learning can take place in the context of statistically misspecified models, irrespective of their theoretical meaningfulness or excellent fit. Substantive information can potentially increase the precision of inference in cases where it is data-validated in the context of a statistically adequate model.

The basic argument is that high goodness-of-fit is neither necessary nor sufficient for reliable inference. Indeed, selection procedures using norm-based criteria do not yield statistical models that 'account for the regularities in the data'. This is because: (i) model selection begins with a prespecified family of models assumed to nest the 'true' model, without implementing any form of model validation, (ii) even when the 'true' model is nested within the prespecified family, satisfying their norm-based criteria does not ensure the reliability of inferences because their associated error probabilities are unknown, and (iii) the model chosen does not necessarily account for the regularities in the data, since the evaluation is based primarily on fit; see Spanos (2007d). One can argue that securing *statistical adequacy* addresses both objectives associated with the model selection procedures: selecting a prespecified family of models, and determining the 'best' model within this family. As a result, norm-based model selection procedures are rendered superfluous and potentially misleading. Indeed, norm-based model selection makes statistical sense only when one is comparing between different statistical models whose (statistical) adequacy has already been established and the task is to compare them on substantive adequacy grounds; see Spanos (2006b).

5.3 Robustness and the reliability/precision of inference

It is well known in statistics that the *reliability* of any inference procedure (estimation, testing and prediction) depends crucially on the validity of the *premises*: the model probabilistic assumptions. The *trustworthiness* 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.

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. Indeed, this difference provides a measure of *robustness*: the sensitivity of the inference procedure to the particular departure from the model assumptions; see Box (1979).

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). Condition (a) ensures the approximate equality of the nominal and actual error probabilies, hence the reliability of inference, and (b) secures the high capacity of the inference procedure. What is often not appreciated enough in practice is that without (b), (a) makes little sense. An example of this is given by the traditional textbook econometrics way of dealing with departures from the homoskedasticity assumption ([3], table 3), by adopting the HCSE for the least squares estimators of the coefficients; see section 2.1. In contrast, in the context of the error-statistical approach the unreliability of inference problem is addressed, not by using actual error probabilities in the case of misspecification, but by *respecifying* the original statistical model and utilizing inference methods that are optimal in the context of the new (adequate) premises; see Spanos (1986, 2005).

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. When the problem of statistical misspecification is raised, the response is often a variant of the following argument invoking robustness:

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

This seemingly reasonable argument is shown to be highly misleading when one attempts to *quantify* 'slight' departures and 'minor' deviations. It is argued that invoking robustness often amounts to 'glossing over' the unreliability of inference problem instead of addressing it; see Spanos (2005).

Example. Assume that data $\mathbf{x}_0$ constitute a 'truly typical realization' of the stochastic process represented by the simple Normal model (table 2), but it turns out that assumption [4] is actually invalid, in the sense that:

$$Corr(X_i, X_j) = \rho, \quad 0<\rho<1, \ i \neq j, \quad i, j = 1, ...n. \tag{42}$$

As argued above, this is likely to render inference, such as the t-test, based on this model *unreliable.* Let $\mu_0 = 0$, $n = 100$, $\alpha = .05$, $c_\alpha = 1.66$. Table 4 shows that the presence of even some tiny correlation ($\rho = .05$) will induce a sizeable discrepancy between the *nominal* ($\alpha = .05$) and *actual type I error probability* ($\alpha^* = .25$); this discrepancy increases with ρ.

Table 4 - Type I error of t-test							
ρ	**0.0**	.05	.10	.30	.50	.75	.90
α^*-actual	**.05**	.249	.309	.383	.408	.425	.431

Similarly, the presence of dependence will also distort the power of the t-test. As shown in table 5, as $\rho \to 1$ the power of the t-test increases for small discrepancies from the null, but it decreases for larger discrepancies. That is, the presence of correlation would render a powerful smoke alarm into a *faulty one,* being triggered by burning toast but not sounding until the house is fully ablaze; see Mayo (1996).

Table 5 - Power $\pi^*(\mu_1)$ of the t-test					
ρ	$\pi^*(.02)$	$\pi^*(.05)$	$\pi^*(.1)$	$\pi^*(.2)$	$\pi^*(.4)$
0.0	**.074**	**.121**	**.258**	**.637**	**.991**
.05	.276	.318	.395	.557	.832
.1	.330	.364	.422	.542	.762
.3	.397	.418	.453	.525	.664
.5	.419	.436	.464	.520	.630
.75	.434	.447	.470	.516	.607
.9	.439	.452	.473	.514	.598

The above example illustrates how misleading the invocation of robustness can be when one has no way of quantifying 'slight' departures and 'minor' deviations. Spanos (2005) discusses a more realistic example where the temporal dependence is Markov $Corr(X_i, X_j) = \rho^{|i-j|}$, $|\rho| < 1$, where respecification gives rise to the AR(1) model in (33).

5.4 Weak assumptions and the reliability/precision of inference

The current approbation in textbook econometrics for using the GMM (Hall, 2005) and non-parametric methods (Pagan and Ullah, 1999), is often justified in terms of the rationale that the broad premises assumed by these methods are less vulnerable to misspecification and thus often lead to more reliable inferences. Indeed, these methods are often motivated by claims that weak probabilistic assumptions provide a way to overcome unreliability. Matyas (1999, p. 1) went as far as to argue that, "the crises of econometric modeling in the seventies" ... was "precipitated by reliance on highly unrealistic strong probabilistic assumptions", and the way forward is to abandon such assumptions in favor of weaker ones. As argued in Spanos (2006a), this rationale is highly misleading in so far as broader premises give rise to less precise inferences without any guarantee of reliability, because they invariably invoke non-tested and non-testable (differentiability of unknown density functions and boundedness conditions) assumptions, or/and asymptotic results of unknowable pertinence. Moreover, contrary to commonly used claims data plots (t-plots, scatter plots, etc.) convey a

good deal of information pertaining to the underlying distributions and associated functional forms; see Spanos (1999), ch. 5.

The quintessential example of this perspective is the *Gauss-Markov (G-M) theorem* in the context of the Classical Linear model:

$$\begin{gathered}\mathbf{y}=\mathbf{X}\beta+\mathbf{u},\\ (1)\ E(\mathbf{u})=\mathbf{0},\ (2)\ E(\mathbf{uu}^{\top})=\sigma^2\mathbf{I}_n,\ (3)\ rank(\mathbf{X})=k.\end{gathered} \tag{43}$$

This theorem establishes that the OLS $\widehat{\beta}=(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{y}$ is Best Linear Unbiased Estimator (BLUE) of β under assumptions (1)-(3), *without* invoking Normality: (4) $\mathbf{u}\backsim \mathsf{N}(.,.)$. In addition to BLUE being of very limited value since it secures the relative efficiency of $\widehat{\beta}$, the G-M theorem yields an unknown sampling distribution for $\widehat{\beta}$, i.e. $\widehat{\beta}\backsim \overset{?}{\mathsf{D}}(\beta,\sigma^2(\mathbf{X}^{\top}\mathbf{X})^{-1})$, which provides a poor basis for hypothesis testing and other forms of inference that involve error probabilities. Finite sample inference can only be based on inequalities like Chebyshev's which often turn out to be very crude and imprecise; Spanos (1999), ch. 10. As a result, practitioners usually invoke the central limit theorem in order to use the approximation $\widehat{\beta}\backsimeq \mathsf{N}(\beta,\sigma^2(\mathbf{X}^{\top}\mathbf{X})^{-1})$, but one has no way of knowing how good this approximation is for the particular sample size n; unless one is prepared to do a thorough job with probing for departures from the premises of the Linear Regression model as given in table 3; see Spanos (2006a).

As argued in Spanos (1999), ch. 10, there is a lot of scope for non-parametric inference in empirical modeling, such as in exploratory data analysis and M-S testing, but not for providing the premises of inference when reliability and precision are the primary objectives.

5.5 Statistical 'Error-fixing' strategies and data mining

A number of different activities in empirical modeling are often described as unwarranted 'data mining' when the procedures followed undermine the trustworthiness of the evidence they give rise to.

Typically a textbook econometrician begins with a theory model, more or less precisely specified, and proceeds to specify a statistical model in the context of which the quantification will take place, by viewing the theory model as its systematic component and attaching a *white noise error* as its non-systematic component. This implicitly assumes that the chosen data provide apposite observations for the concepts envisaged by the theory. Usually, the estimated model does not give rise to the "expected" results in the sense that it often yields 'wrong' signs, insignificant coefficients for crucial variables, as well as indications that some of the model assumptions, (see (43)) are invalid. What does one do next? According to Wooldridge (2006):

"When that happens, the natural inclination to try different models,
different estimation techniques, or perhaps different subsets of data
until the results correspond more closely to what was expected." (ibid., p. 688)

This describes the well-known textbook 'error-fixing' strategy which takes the form of estimating several variants of the original model by modifying the underlying

assumptions (using OLS, GLS, GMM, IV), guided by a combination of diagnostic checking and significance testing of the coefficients, in the hope that one of these variants will emerge as the "best" model, and then used as a basis of inference. What is "best" is conventionally left vague, but it's understood to comprise a combination of statistical significance and theoretical meaningfulness.

The statistical 'error-fixing' strategies are based on a textbook repertoire of recommendations which arise from relaxing the G-M assumptions (1)-(3) (see (43)) one at a time, and seeking 'optimal' estimators under a particular departure. For example, when the *no-autocorrelation* assumption in (2) is invalid and instead $E(\mathbf{uu}^\top) = \Omega \neq \sigma^2\mathbf{I}_n$, the recommendation is twofold. Either to retain the OLS estimator $\widehat{\beta}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ and utilize the HCSE for inference purposes, or to use a Feasible Generalized Least Squares (FGLS) estimator based on the autocorrelation-corrected model where: $u_t = \rho u_{t-1} + \varepsilon_t$. When the *homoskedasticity* assumption in (2) is invalid a similar twofold recommendation is prescribed where one 'fixes' the problem by either retaining the OLS estimator $\widehat{\beta}$ and uses the HCSE for inference, or estimates the heteroskedastic variances using an auxiliary regression, $\widehat{u}_t^2 = c_0 + \mathbf{c}_1^\top\mathbf{z}_t + v_t$, and applies weighted least squares. As argued by Greene (2000), p. 521:

> "It is rarely possible to be certain about the nature of the heteroskedasticity in regression model. In one respect, the problem is only minor. The weighted least squares estimator is consistent regardless of the weights $[\mathbf{z}_t]$ used, as long as the weights are uncorrelated with the disturbances."

This claim is clearly misleading when one realizes that the regression and skedastic functions are the first two moments of the same conditional distribution $f(y_t|\mathbf{x}_t;\psi)$, whose structure is determined by the underlying joint distribution $f(y_t,\mathbf{x}_t;\varphi)$; see Spanos and McGuirk (2001).

In practice one is encouraged to try out different forms for the weights $\mathbf{z}_t$ and pick the one with the "best" results. When such statistical 'error-fixing' recommendations are tried out, one is supposed to keep one eye on the 'theoretical meaningfulness' of the estimated variants and choose between them on the basis of what can be rationalized both statistically and substantively. It is widely acknowledged that these 'error-fixing' strategies constitute problematic forms of data mining:

> "Virtually all applied researchers search over various models before finding the "best" model. Unfortunately, this practice of data mining violates the assumptions we have made in our econometric analysis." (Wooldridge, 2006, p. 688)

The end result is that such 'error-fixing' misuses data in ways that 'appear' to provide empirical (inductive) *support* for the theory in question, when in fact the inferences are usually unwarranted. These 'error-fixing' procedures illustrate the kind of problematic use of the data to construct (ad hoc) a model to account for an apparent 'anomaly' (departures from error assumptions) that naturally gives rise to skepticism; this is known as pejorative 'double-use' of data.

These strategies, driven by the search for an 'optimal' estimator for each different set of error assumptions (OLS, GLS, FGLS, IV, GMM, etc.), ignore the fact that

model assumptions, such as [1]-[5] (table 3), are interrelated and thus the various 'anomalies' are often misdiagnosed, and the ad hoc 'fixes' of specific error assumptions lead to exacerbating (not ameliorating) the reliability of inference (see Spanos, 1986, 2000, Spanos and McGuirk, 2001). For instance, when autocorrelated *residuals* are interpreted as autocorrelated *errors,* any inference based on the 'autocorrelation-corrected' model' is likely to be unreliable because the latter model is often as misspecified as the original; see Spanos (1986), McGuirk and Spanos (2004). As shown by Spanos and McGuirk (2001), the HCSE do very little, if anything, to ameliorate the reliability of inference is practice. The general reasoning flaw in this *respecification* strategy is that by adopting the alternative hypothesis in a misspecification test commits the fallacy of rejection. More often than not, after such 'error-fixing' takes place - by choosing the 'optimal' estimator that goes with the new set of error assumptions - one often ends up (unwittingly) with another misspecified model (see Mayo and Spanos, 2004). This misspecified model, however, is then used as a basis for deciding the sign and significance of key coefficients in order to secure theoretical meaningfulness, giving rise to unreliable inferences.

Viewed from the error-statistical perspective, each step in the above 'error-fixing' strategies fosters further errors, and ignores existing one (see section 2), with the modeler unwittingly worsening the overall trustworthiness of the evidence these strategies give rise to. Moreover, the modeler focuses on 'saving the theory' by retaining the systematic component and ignoring alternative theories which might fit the same data equally well or even better. By focusing the 'error-fixing' strategies the textbook perspective overlooks the ways the systematic component may be misspecified. In addition, incomplete specifications of statistical models (assumption [5] in table 3) are not conducive to securing statistical adequacy. This should be contrasted with warranted 'data mining', such as the use of graphical techniques and M-S testing, in context of the error-statistical where they enhance the reliability of the inferences reached; see Spanos (2000), Mayo and Spanos (2004).

The error-statistical perspective suggests that once certain departures from the original model assumptions are established, the way to proceed is not to use the actual error probabilities, but to respecify the original model and construct a new optimal inference procedure based on the respecified model; see Spanos (2005, 2006a).

5.6 Substantive 'fixing' strategies and theory 'fishing'

Since the 1970s the question most often posed in seminars to any presenter of an applied econometrics paper, when discussing the estimation of any linear regression:

$$y_t = \beta_0 + \beta_1 x_t + u_t, \quad t \in \mathbb{N}, \tag{44}$$

is 'did you account for simultaneity in your model?' The estimated model in (1) provides a perfect target for the cognoscenti of textbook econometrics. The right answer is supposed to be 'yes I did and here are my Instrumental Variables (IV) estimates'. The discussion would invariably move to whether the particular set of

chosen instruments, say $\mathbf{W}_t$, are 'optimal' or not, and the correct answer to that is expected to be a good 'explanation' of why it is reasonable to assume that:

$$\text{(i) } E(X_t u_t) \neq 0 \text{ in (44), (ii) } E(\mathbf{W}_t \varepsilon_{2t}) = \mathbf{0}, \text{ (iii) } Cov(\mathbf{W}_t, X_t) \neq \mathbf{0} \text{ in (6)},$$

conditions which ensure that the IV estimator of β_1 is at least consistent. A comparison between the OLS and IV estimates if often used as an indication of how serious the simultaneity problem is, and the choice between the two estimators (models) is often made on the basis of a combination of statistical significance of key coefficients like β_1 and theoretical meaningfulness. With these criteria in mind, the cognoscenti of textbook econometrics search through several sets of instruments $\mathbf{W}_t$, and choose as 'optimal' the set that meets their expectations, and then they forge an 'explanation' for this choice. This is a textbook substantive 'error-fixing' strategy which is nothing short of theory fishing that usually gives rise to unreliable inferences with probability one. This is because such a procedure is rife with potential errors and one has no way of detecting or avoiding them.

Viewed in the context of the error-statistical approach, the problem begins with conditions (i) and (ii) which are clearly unverifiable, giving the impression that the choice of 'optimal' instruments is a matter of rhetoric; it is not! The choice of instruments is not just a matter of giving a persuasive 'story' why the set of instruments $\mathbf{W}_t$ one happens to choose satisfies (i)-(iii). As argued in Spanos (1986, 2007a) the choice of optimal instruments also depends on the *statistical adequacy* of the system of equations in (6) in conjunction with the confirmation of (iii) and (iv) $Cov(\mathbf{W}_t, y_t) \neq \mathbf{0}$ in its context.

To illustrate these textbook arguments let us return to the estimated model in (1) and consider the following set of instruments $\mathbf{W}_t$:=$(W_{1t}, W_{2t}, W_{3t}, W_{4t}, W_{5t})$ where W_{1t} - price of oats, W_{2t} - output of oats, W_{3t} - price of potatoes, W_{4t} - out of potatoes, W_{5t} - rainfall; all prices and output series denote per cent proportional changes. Re-estimating (1) using the IV method yields:

$$y_t = \underset{(2.179)}{7.180} - \underset{(.090)}{0.689} x_t + \widetilde{u}_t, \; R^2 = .622, \quad s = 14.450, \quad n = 45, \tag{45}$$

showing only minor differences between the OLS and IV estimates. In the textbook econometrics tradition this is interpreted as an excellent indication that the original estimates are robust to simultaneity. However, looking at the overidentifying restrictions test for (45), $F(4, 39)$=13.253[.0000007], indicates that such an inference might be premature; the restrictions are strongly rejected. The truth of the matter is that none of the t-ratios, and F-statistics invoked in the above arguments is statistically meaningful unless they are based on statistically adequate models. Not surprisingly, both estimated equations (1) and (45) are seriously statistically misspecified. More importantly, the statistical meaningfulness of (45) depends crucially on the statistical adequacy of the implicit reduced form in (6). Using several M-S tests for this system of equations (see Spanos, 1986, ch. 24, Spanos, 1990) one can easily verify that it is misspecified – assumptions [1]-[2], [4]-[5] (table 3) are invalid – calling into question the reliability of all inferences, including that of the overidentifying restrictions test.

Hence, the substantive 'error-fixing' strategy of invoking simultaneity and using IV estimators does is not usually remedy the initial statistical misspecification of (1) problem, but instead it enhances the unreliability of inference by bringing into the statistical analysis additional equations which are also statistically misspecified.

5.7 Pre-test bias and all that!

In the context of the error-statistical approach described in section 4, a number of modeling procedures, such as Mis-Specification (M-S) testing (diagnostic checking) and respecification with a view to find a statistically adequate model, are often criticized by the textbook econometrics perspective as illegitimate data mining which induces biases in the resulting inferences. The most widely used argument against the error-statistical approach is the charge that the process of M-S testing and respecification suffers from pre-test bias; see Kennedy (2003).

5.7.1 Pre-test bias or befuddlement?

To bring out the gist of the pre-test bias argument consider the following two alternative models:

$$\begin{array}{ll} \mathcal{M}_0: & y_t=\beta_0+\beta_1 x_t+u_t, \\ \mathcal{M}_1: & y_t=\beta_0+\beta_1 x_t+u_t,\ u_t{=}\rho u_{t-1}+\varepsilon_t, \end{array} \tag{46}$$

where the choice between them will be decided on the basis of the Durbin-Watson (D-W) test for the hypotheses:

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

This comparison is then transformed into a choice between two estimators of β_1 which is then formalized in decision-theoretic terms using the *pre-test estimator* $\ddot{\beta}_1$:

$$\ddot{\beta}_1=\lambda\widehat{\beta}_1+(1{-}\lambda)\widetilde{\beta}_1,\ \text{where}\ \lambda=\begin{cases} 1, & \text{if } H_0 \text{ is accepted,} \\ 0, & \text{if } H_0 \text{ is rejected,} \end{cases} \tag{47}$$

This constitutes a convex combination of the two alternative estimators $(\widehat{\beta}_1,\widetilde{\beta}_1)$; $\widehat{\beta}_1$ is the OLS estimator under H_0 $(\mathcal{M}_0)$, and $\widetilde{\beta}_1$ is the GLS estimator under H_1 $(\mathcal{M}_1)$. It turns out that the sampling distribution of $\ddot{\beta}_1$ is often non-Normal, suffering from bias and has a highly complicated variance; see Mittelhammer et al (2000). The pre-test argument warns that when these effects are ignored by using either $\widehat{\beta}_1$ or $\widetilde{\beta}_1$, one uses the 'wrong' error probabilities, giving rise to unreliable inferences.

When the pre-test bias argument, based on (46), is viewed in the context of the error-statistical approach, it becomes clear that its methodological grounding is questionable for two reasons. *First*, adopting the alternative in a M-S test is an example of the classic *fallacy of rejection*: evidence against the null is misinterpreted as evidence for the alternative; one should *never* accept the alternative in a M-S test without further testing. The validity of the alternative model $\mathcal{M}_1$ needs to be established separately by testing its own assumptions; see Spanos (2000, 2001a). That is, in this case the pre-test bias argument is the result of a misguided attempt to formalize a fallacy! *Second*, it misconstrues a M-S testing problem (testing the

validity of assumption [4] table 3) by viewing it as an estimation problem whose relevant error probabilities are evaluated using a loss function. As argued in section 4.2, the error probabilities for estimation and testing are very different in nature and conflating the two can lead to major befuddlements.

5.7.2 Omitted Variables bias

A question that might arise from the discussion of the above example is the extent to which the criticisms of the pre-test bias argument depend on the fact that (46) was essentially an M-S testing problem. The short answer is that it does not. The real problem with the pre-test bias argument is that it conflates two very different error probabilities, by replacing testing with estimation. To see this consider the classic omitted variables problem where the following two alternative models are compared:

$$\mathcal{M}_0 : y_t = \beta_0 + \beta_1 x_{1t} + u_t, \qquad \mathcal{M}_1 : y_t = \alpha_0 + \alpha_1 x_{1t} + \alpha_2 x_{2t} + \varepsilon_t, \tag{48}$$

and the decision will be made on the basis of the t-test for the hypotheses:

$$H_0 : \alpha_2 = 0, \text{ vs. } H_1 : \alpha_2 \neq 0; \tag{49}$$

see Leeb and Potscher (2005). This example is different from (46) in so far as the latter poses a question concerning *statistical adequacy*, but (48) poses a question concerning *substantive adequacy:* does model $\mathcal{M}_0$ provide an adequate explanation for the behavior of y_t? The hypotheses in (49) raise the crucial problem of *confounding:* whether the estimated model M_0 omitted a certain potentially important factor X_{2t} misidentifying the influence of X_{1t} on y_t, and thus giving rise to misleading inferences.

Formulating this problem as one of pre-test estimation based on $\ddot{\beta}_1 = \lambda\widehat{\beta}_1 + (1-\lambda)\widehat{\alpha}_1$, where λ is given in (47), $(\widehat{\beta}_1, \widehat{\alpha}_1)$ denote the OLS estimators of (β_1, α_1) is problematic for several reasons. *First*, the parameterizations of (β_1, α_1) are very different; one is *not* estimating the same parameter in the two cases; see Spanos (1986). *Second*, the framing of the problem in terms of a choice between two point estimators is inadequate for the task, because it automatically (mis-)interprets accept and reject the null as evidence for $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively; committing both classic fallacies of acceptance and rejection; see section 4.1. *Third*, the pre-test bias is evaluated in terms of estimation error probabilities, like the Mean Square Error, and the associated sensitivity analysis can be shown to be too crude for reliable answers to the question of confounding. When the confounding issue is posed a N-P testing problem, one can show that there are eight alternative scenarios (different answers) in (48), depending on the non-zero values of $Cov(y_t, X_{1t})$, $Cov(y_t, X_{2t})$, $Cov(X_{1t}, X_{2t})$, which cannot be be distinguished by the traditional estimation and the associated sensivity analysis. *Fourth*, the comparison in (48) gives rise to reliable inferences only to the extent that $\mathcal{M}_1$ in (48) is statistically adequate, ensuring that the N-P tests employed to distinguish between the different scenarios are reliable; their actual error probabilities approximate well the nominal ones. Note that no such presupposition is invoked in the case of (46). *Lastly*, posing the confounding question as a testing issue in the context of the error-statistical approach enables one to guard against the fallacies of

acceptance/rejection by supplementing the accept/reject decisions with a post-data evaluation of inference based on severe testing; see Spanos (2006b).

6 Conclusions

The current state of applied econometrics, viewed as the empirical understructure of economics, calls for much greater attention to be paid to the philosophical foundations of empirical modeling. Like political arithmetic towards the end of the 18th century (see Spanos, 2008a), current econometrics runs a great risk of losing credibility as a way to provide empirical grounding for economic theorizing and policy analysis. The accumulation of mountains of untrustworthy empirical evidence over the last century is a symptom of major weaknesses in the current methodological framework for empirical modeling in economics. The current textbook approach to econometric modeling pays little, if any, attention to ensuring the reliability of inference by probing for and eliminating all potential errors that could lead the inference astray, including the *data inaccuracy*, *incongruous measurement* and *substantive inadequacy*. The emphasis on 'quantifying theoretical relationships' and the 'error-fixing' strategies endanger the trustworthiness of the empirical evidence they give rise to.

An attempt has been made in this paper to bring out some of these weaknesses and make constructive suggestions on how the reliability of inductive inference in econometrics can be improved by viewing empirical modeling in a richer and more refined methodological framework known as the error-statistical approach; see Mayo and Spanos (2008). This approach provides a coherent inductive reasoning for frequentist statistics and focuses on 'learning from data' about phenomena of interest by employing reliable procedures based on ascertainable error probabilities; both pre-data and post-data. The error-statistical inductive reasoning strongly encourages the probing of the different ways an inference might be in error, and has been used in this paper to shed light on several important methodological issues which concern the nature, interpretation, and justification of methods and models that are relied upon to learn from data.

In particular, the error-statistical perspective: (a) views empirical modeling as bridging the *gap between theory and data*, using a chain of complecting models, (b) views statistical models in terms of testable probabilistic assumptions concerning the observable processes, (c) affords the data 'a life of its own' in the context of a *statistically adequate* model, (d) allows for both statistical and substantive information to play important complementary roles without compromising the credibility of either, and (e) encourages *error probing* at all levels (models). Securing statistical as well as substantive adequacy can contribute significantly to 'learning from data' and establish economics as an empirical science; see Spanos (2006a-c).

References

[1] Abadir, K. and G. Talmain, (2002), "Aggregation, Persistence and Volatility in a Macro Model," *Review of Economic Studies*, **69**, 749-779.

[2] Ackermann, R., J. (1985), *Data, Instruments and Theory: a Dialectical Approach to Understanding Science*, Princeton University Press, Princeton.

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

[4] Backhouse, R. E. (1994), *New Directions in Economic Methodology*, Routledge, London.

[5] Bernardo, J. M. (2005), "Reference Analysis," pp. 17–90 in *Handbook of Statistics*, vol. 25: Bayesian Thinking, Modeling and Computation, D. K. Dey and C. R. Rao, (eds.), Elsevier, North-Holland.

[6] Berger, J. (2004) "The Case for Objective Bayesian Analysis," *Bayesian Analysis*, 1: 1–17.

[7] Berger, J. and Wolpert, R. (1988), *The Likelihood Principle*, 2d ed., Institute of Mathematical Statistics, Hayward, CA.

[8] Birnbaum, A. (1961), "Confidence Curves: An Omnibus Technique for Estimation and Testing," *Journal of the American Statistical Association*, **294**: 246-249.

[9] Birnbaum, A. (1962), "On the Foundations of Statistical Inference" (with discussion), *Journal of the American Statistical Association*, **57**: 269-306.

[10] 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.

[11] Blaug, M. (1992), The Methodology of Economics, Cambridge University Press, Cambridge.

[12] Caldwell, B. (1994), *Beyond Positivism: Economic Methodology in the Twentieth Century*, 2nd ed., George Allen & Unwin, London.

[13] Carnap, R. (1950/1962), Logical Foundations of Probability, 2nd ed., The University of Chicago Press, Chicago.

[14] Chalmers, A. F. (1999), *What is this thing called Science?*, 3rd ed., Hackett, Indianapolis.

[15] Cox, D. R. (1990), "Role of Models in Statistical Analysis," *Statistical Science*, **5**: 169-174.

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

[17] Cox, D. R. and D. G. Mayo (2008), "Objectivity and Conditionality in Frequentist Inference," forthcoming in *Error and Inference*, D. G. Mayo and A. Spanos (eds.), Cambridge University Press, Cambridge.

[18] Davidson, R. and J. G. MacKinnon (1987), "Implicity alternatives and the local power of test statistics," *Econometrica*, 55: 1305-1329.

[19] Davis, J. B. , D. W. Hands, U. Maki (1998), *The Handbook of Economic Methodology*, (eds.), Edward Elgar, Cheltenham.

[20] Duhem, P. (1914), *The Aim and Structure of Physical Theory*, English translation published by Princeton University Press, Princeton.

[21] Fisher, R. A. (1921), "On the 'Probable Error' of a Coefficient of Correlation Deduced from a small sample," *Metron*, 3-32.

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

[23] Fisher, R. A. (1925), "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society*, 22: 700-725.

[24] Fisher, R. A. (1934), "Two New Properties of Maximum Likelihood," *Proceedings of the Royal Statistical Society*, A, 144: 285-307.

[25] Fisher, R. A. (1935a) "The logic of inductive inference", *Journal of the Royal Statistical Society*, **98**, 39-54, with discussion pp. 55-82.

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

[27] Fisher, R. A. (1935c), "The Fiducial Argument in Statistical Inference," *Annals of Eugenics*, 6: 391-398.

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

[29] Fisher, R. A. (1956), *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh.

[30] Giere, R. N. (1999), *Science Without Laws*, The University of Chicago Press, Chicago.

[31] Gigerenzer, G. (1993) "The superego, the ego, and the id in statistical reasoning," pp. 311-39 in Keren, G. and C. Lewis (eds.), *A Handbook of Data Analysis in the Behavioral Sciences: Methodological Issues*, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

[32] Ghosh, J. K., M. Delampady and T. Samanta (2006), *An Introduction to Bayesian Analysis: Theory and Methods*, Springer, NY.

[33] Glymour, C. (1980), *Theory and Evidence*, Princeton University Press, NJ.

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

[35] Godambe, V. P. and D. A. Sprott (1971), *Foundations of Statistical Inference: a Symposium*, Holt, Rinehart and Winston, Toronto.

[36] Godfrey-Smith, P. (2003), *Theory and Reality: An Introduction to the Philosophy of Science*, The University of Chicago Press, Chicago.

[37] Granger, C. W. J. (1990), (ed.) *Modelling Economic Series*, Clarendon Press, Oxford.

[38] Greene, W. H. (2000), *Econometric Analysis*, 4th ed., Prentice Hall, NJ.

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

[40] Hacking, I. (1983), *Representing and Intervening*, Cambridge University Press, Cambridge.

[41] Hald, A. (1998), *A History of Mathematical Statistics from 1750 to 1930,* Wiley, NY.

[42] Hall, A. R. (2005), *Generalized Method of Moments*, Oxford University Press, Oxford.

[43] Hands, W. D. (2001), *Reflection without Rules: Economic Methodology and Contemporary Science Theory*, Cambridge University Press, Cambridge.

[44] Harlow, L. L., S. A. Mulaik and J. H. Steiger (1997), *What if there were no Significance Tests?* Mahwah, Erlbaum, NJ.

[45] Harper, W. L. and C. A. Hooker (1976), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II: Foundations and Philosophy of Statistical Inference*, Reidel, Dordrecht.

[46] Hempel, C. G. (1965), *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, Mcmillan, New York.

[47] Hendry, D. F. (1995), *Dynamic Econometrics*, Oxford University Press, Oxford.

[48] Hendry, D. F. (2000), *Econometrics: Alchemy or Science?*, 2nd ed., Blackwell, Oxford.

[49] Hendry, D. F., E. E. Leamer and D. J. Poirier (1990), "The ET dialogue: a conversation on econometric methodology," *Econometric Theory*, **6**, 171-261.

[50] Hodges, J. L. and E. L. Lehmann (1954), "Testing the Approximate Validity of Statistical Hypotheses," *Journal of the Royal Statistical Society*, B, **16**: 261-268.

[51] Hoover, K. D. (2001), *Causality in Macroeconomics*, Cambridge University Press, Cambridge.

[52] Hoover, K. D. (2002), "Econometrics and Reality," in Maki, U. (2002), pp. 152-177.

[53] Hoover, K. D. (2006), "The Methodology of Econometrics," in Maki, U. (2002), pp. 152-177.

[54] Howson, C. and P. Urbach (1993), *Scientific Reasoning: The Bayesian Approach*, 2nd ed., Open Court, Chicago, IL.

[55] Keynes, J. M. (1921), *A Treatise on Probability*, MacMillan, London.

[56] Jeffreys, H. (1939/1961), *Theory of Probability*, 3rd ed., Oxford University Press, Oxford.

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

[58] Kennedy, P. (2003), *A Guide to Econometrics*, 5th edition, MIT Press, MA.

[59] Keuzenkamp, H. A. (2000), *Probability, Econometrics and Truth*, Cambridge University Press, Cambridge.

[60] Kuhn, T. (1962), *The Structure of Scientific Revolutions*, The University of Chicago Press, Chicago.

[61] Kuhn, T. (1977), *The Essential Tension: Selected Studies in Scientific Tradition and Change*, The University of Chicago Press, Chicago.

[62] Lakatos, I. (1970), "Falsification and the Methodology of Scientific Research Programms," in Lakatos and Musgrave (1970), pp. 91-196.

[63] Lakatos, I. and A. Musgrave (eds.) (1970), *Criticism and Growth of Knowledge*, Cambridge University Press, Cambridge.

[64] Laudan, L. (1977), *Progress and Its Problems: Towards a Theory of Scientific Growth*, Berkeley: University of California Press.

[65] Lawson, T. (1997), *Economics and Reality*, Routledge, London.

[66] Leamer, E. E. (1978), *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Wiley, New York.

[67] Leeb, H. and B. M. Pötscher (2005), "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, **21**, 21-59.

[68] Lehmann, E. L. (1986), *Testing Statistical Hypotheses*, 2nd ed., Wiley, NY.

[69] Lehmann, E. L. (1990), "Model specification: the views of Fisher and Neyman, and later developments", *Statistical Science*, **5**, 160-168.

[70] Lieberman, B. (1971), *Contemporary Problems in Statistics: a Book of Readings for the Behavioral Sciences*, Oxford University Press, Oxford.

[71] Lindley, D. V. (1965), *Introduction to Probability and Statistics from the Bayesian Viewpoint*, Cambridge University Press, Cambridge.

[72] Machamer, P. and M. Silberstein (2002), *The Blackwell Guide to the Philosophy of Science*, Blackwell, Oxford.

[73] Maki, U. (2001), *The Economic World View: Studies in the Ontology of Economics*, Cambridge University Press, Cambridge.

[74] Maki, U. (2002), *Fact and Fiction in Economics*, Cambridge University Press, Cambridge.

[75] Matyas, L. (1999), (editor), *Generalized Method of Moments Estimation*, Cambridge University Press, Cambridge.

[76] Mayo, D. G. (1991), "Novel Evidence and Severe Tests", *Philosophy of Science,* **58**, 523-552.

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

[78] Mayo, D. G. (1997), "Duhem's Problem, the Bayesian Way, and Error Statistics, or "What's Belief Got to Do with It?", *Philosophy of Science*, **64**, 222-244.

[79] Mayo, D. G. (2005), "Philosophy of Statistics," in S. Sarkar and J. Pfeifer (eds.), *Philosophy of Science: An Encyclopedia*, London: Routledge, pp. 802–15.

[80] Mayo, D. G. (2008), 'An Error in the Argument From WCP and S to the SLP: Discussion," forthcoming in *Error and Inference*, D. G. Mayo and A. Spanos (eds.), Cambridge University Press, Cambridge.

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

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

[83] Mayo, D. G. and D. R. Cox (2006), "Frequentist statistics as a theory of inductive inference," pp. 96-123 in *The Second Erich L. Lehmann Symposium – Optimality*, Lecture Notes-Monograph Series, Volume 49, Institute of Mathematical Statistics.

[84] Mayo, D. G. and A. Spanos (2008), "Error Statistics," forthcoming in *Philosophy of Statistics*, the *Handbook of Philosophy of Science*, Elsevier (editors) D. Gabbay, P. Thagard, and J. Woods.

[85] McCloskey, D. N. (1985), *The Rhetoric of Economics*, University of Wisconsin, Madison.

[86] McGuirk, A. and A. Spanos (2004), "Revisiting Error Autocorrelation Correction: Common Factor Restrictions and Granger Non-Causality," Virginia Tech working paper.

[87] Mills, F. C. (1924), *Statistical Methods*, Henry Holt and Co., NY.

[88] Mills, T.C. and K. Patterson, (2006), *New Palgrave Handbook of Econometrics*, vol. 1, MacMillan, London.

[89] Mittelhammer, R. C., G. C. Judge and D. J. Miller (2000), *Econometric Foundations*, Cambridge University Press, Cambridge.

[90] Moore, H. L. (1914), *Economic Cycles - Their Laws and Cause*, McMillan, NY.

[91] Morgenstern, O. (1963), *On the accuracy of economic observations*, 2nd edition, Princeton University Press, New Jersey.

[92] Morrison, D. E. and R. E. Henkel (1970), *The Significance Test Controversy: A Reader*, Aldine, Chicago.

[93] Nagel, E. (1961), *The Structure of Science*, Hackett, Indianapolis.

[94] Newton-Smith, W. H. (ed.) (2000), *A Companion to the Philosophy of Science*, Blackwell, Oxford.

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

[96] Neyman, J. (1956), "Note on an Article by Sir Ronald Fisher," *Journal of the Royal Statistical Society*, B, **18**, 288-294.

[97] Neyman, J. (1957), "Inductive Behavior as a Basic Concept of Philosophy of Science," *Revue Inst. Int. De Stat.*, **25**: 7-22.

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

[99] Pagan, A.R. (1987), "Three econometric methodologies: a critical appraisal", *Journal of Economic Surveys,* **1**, 3-24. Reprinted in C. W. J. Granger (1990).

[100] Pagan, A.R. and A. Ullah (1999), *Nonparametric Econometrics*, Cambridge University Press, Cambridge.

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

[102] Pearson, E. S. (1955), "Statistical Concepts in the Relation to Reality," J*ournal of the Royal Statistical Society, Series B*, **17**, 204-207.

[103] Pearson, E. S. (1966), "The Neyman-Pearson Story: 1926-34," in *Research Papers in Statistics: Festschrift for J. Neyman*, ed. by F. N. David, Wiley, NY, pp. 1-23.

[104] Peirce, C. S. (1878), "The Probability of Induction," *Popular Science Monthly*, **12**, 705-718.

[105] Poirier, D. J. (1995), *Intermediate Statistics and Econometrics*, MIT Press, Cambridge.

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

[107] Popper, K. R. (1959), *The Logic of Scientific Discovery*, Hutchinson, London.

[108] Popper, K. R. (1963), *Conjectures and Refutations,* Routledge and Kegan Paul, London.

[109] Pratt, J. W. (1961), "Review of 'Testing Statistical Hypotheses' by E. L. Lehmann," *Journal of the American Statistical Association*, **56**: 163-166.

[110] Quine, W. V. (1953), *From the Logical Point of View*, Harvard University Press, Cambridge.

[111] Quine, W. V. (1960), *World and Object*, The MIT Press, Cambridge.

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

[113] Rao, C. R. and Y. Wu (2001) "On Model Selection," pp. 1-64 in *Model Selection*, ed. by P. Lahiri, Institute of Mathematical Statistics, Lecture Notes-Monograph series, vol. 38, Beachwood, OH.

[114] Redman, D. A. (1991), *Economics and the Philosophy of Science*, Oxford University Press, Oxford.

[115] Renyi, A. (1970), *Probability Theory*, North-Holland, Amsterdam.

[116] Rosenthal, R., R. L. Rosnow, D. B. Rubin (1999), *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach*, Cambridge University Press, Cambridge.

[117] Salmon, W. (1967), *The Foundations of Scientific Inference*, University of Pittsburgh Press.

[118] Savage, L. (ed.) (1962), *The Foundations of Statistical Inference: A Discussion*. London, Methuen.

[119] Schervish, M. J. (1995), *Theory of Statistics*, Springer-Verlag, NY.

[120] Schumpeter, J. A. (1954), *History of Economic Analysis*, Oxford University Press, Oxford.

[121] Sims, C. A. (1980), "Macroeconomics and Reality," *Econometrica*, **48**, 1-48.

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

[123] Spanos, A. (1988), "Towards a Unifying Methodological Framework for Econometric Modelling", *Economic Notes*, 107-34; reprinted in Granger (1990).

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

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

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

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

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

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

[130] Spanos, A. (2001b), "Time series and dynamic models," ch. 28, pp. 585-609, *A Companion to Theoretical Econometrics*, edited by B. Baltagi, Blackwell Publishers, Oxford.

[131] Spanos, A. (2004), "Confidence Curves, Consonance Intervals, P-value Functions vs. Severity Evaluations," Working Paper, Virginia Tech.

[132] Spanos, A. (2005), "Misspecification, Robustness and the Reliability of Inference: the simple t-test in the presence of Markov dependence," Working Paper, Virginia Tech.

[133] Spanos, A. (2006a) "Econometrics in Retrospect and Prospect," in Mills, and Patterson (2006), pp. 3-58.

[134] Spanos, A. (2006b) "Revisiting the Omitted Variables Argument: Substantive vs. Statistical Adequacy," *Journal of Economic Methodology*, **13**: 179-218.

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

[136] Spanos, A. (2007a), "The Instrumental Variables Method revisited: On the Nature and Choice of Optimal Instruments," pp. 34-59 in *Refinement of Econometric Estimation and Test Procedures*, ed. by G. D. A. Phillips and E. Tzavalis, Cambridge University Press, Cambridge.

[137] Spanos, A. (2007b), "Curve-Fitting, the Reliability of Inductive Inference and the Error-Statistical Approach," forthcoming *Philosophy of Science*.

[138] Spanos, A. (2007c), "Sufficiency and Ancillarity Revisited: Testing the Validity of a Statistical Model" Working Paper, Virginia Tech.

[139] Spanos, A. (2007d), "Revisiting the Welch Uniform Model: A case for Conditional Inference?" Working Paper, Virginia Tech.

[140] Spanos, A. (2008a), "Statistics and Economics," forthcoming in the *New Palgrave Dictionary of Economics*, 2nd edition, edited by Steven N. Durlauf and Roger E. Backhouse, MacMillan, London.

[141] Spanos, A. (2008b), "Statistical Model Specification vs. Model Selection: Akaike-type Criteria and the Reliability of Inference," Working Paper, Virginia Tech.

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

[143] Spanos, A. and A. McGuirk (2004), "Revisiting Error Autocorrelation Correction: Common Factor Restrictions and Granger Non-Causality," Virginia Tech working paper.

[144] Stigler, S. M. (1986), *The History of Statistics: the Measurement of Uncertainty before 1900,* Harvard University Press, Cambridge, Massachusetts.

[145] Stigum, B. P. (2003), *Econometrics and the Philosophy of Economics*, Princeton University Press, Princeton.

[146] Suppe, F. (1977), *The Structure of Scientific Theories*, 2nd ed., University of Illinois Press, Urbana.

[147] Thompson, B. (1999), "If Statistical Significance Tests are Broken/Misused, what Practices should Supplement or Replace them?" *Theory and Psychology*, **9**:167-183.

[148] Welch, B. L. (1939), "On Confidence Limits and Sufficiency, and Particular Reference to Parameters of Location," *Annals of Mathematical Statistics*, 10, 58-69.

[149] Wooldridge, J. M. (2006), *Introductory Econometrics: a modern approach*, Thomson, South-Western.