

# Philosophy of Econometrics

Aris Spanos

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

May 2007

## Abstract

1. Introduction
2. What philosophical/methodological issues?
3. Philosophy of science and empirical modeling:
  - Logical positivism/empiricism
  - The downfall of logical positivism/empiricism
  - The new experimentalism
4. Statistical inference: philosophical foundations:
  - Statistical induction
  - The nature of different forms of statistical inference
  - Statistical inference and its problems circa 1955
5. The Error-Statistical approach:
  - Severe testing reasoning
  - Statistical adequacy
  - A chain of complementing models: theory  $\longleftrightarrow$  data
6. Problems and issues in philosophy of science:
  - Reflecting on curve-fitting
  - Reflecting on Duhem's problem
7. Philosophical/methodological issues in econometrics:
  - Reliability/precision of inference and robustness
  - Weak assumptions and the reliability/precision of inference
  - 'Error-fixing' strategies and data mining
  - Substantive 'error-fixing' strategies and theory mining
  - Bias-inducing procedures revisited
8. 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 statistical/empirical inference/modeling in their endeavors to transform 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.

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 (asymptotically ‘optimal’) associated with the Classical Linear Regression (CLR) model in a number of directions. These modifications/extensions are motivated primarily by (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).

Discussions of econometric methodology have been primarily ‘local’ affairs (see Gilbert, 1988, Granger, 1990, Hendry et al, 1990, Hendry, 1993, Leamer 1988, Pagan, 1987, 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 (2001) 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 al (1998), Maki (2001, 2002) and Redman (1991). Even in methodological discussions concerning the relationship between economic theories and economic 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 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.

In section 2, a simple empirical example is used to bring out the diversity and multiplicity 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 discusses the philosophical foundations of statistical inference paying particular attention to the form and structure of statistical induction as envisioned by Fisher, Neyman and Pearson during the first half of the 20th century, and discusses some of the philosophical/methodological issues that have confounded statistical testing since the late 1930s. A modification/extension of the Fisher-Neyman-Pearson approach to statistical induction, known as error-statistics, is briefly discussed in section 5 in an attempt to show how it can be used to address some of the inveterate philosophical/methodological mentioned in section 4. The error-statistical approach is then used in section 6 to reflect briefly on certain impor-

tant problems in philosophy of science, such as curve-fitting and Duhem's problem. The same error-statistical perspective is used in section 7 to shed new light on a number of philosophical/methodological problems in econometrics pertaining to certain 'error-fixing' strategies practiced by textbook econometrics.

## 2 What philosophical/methodological issues?

To get 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 a satisfactory answer. To bring out some of these issues let us revisit Moore's (1914) estimated 'statistical demand' curve for corn:

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

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.* pp. 62-88.

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

$$Q^D = \beta_0 + \beta_1 P, \quad \beta_0 > 0, \quad \beta_1 < 0, \quad (2)$$

in its simplest form? Taking the empirical results in (1) at face value, the estimates:

(i) indicate statistically significant coefficients:

$$\tau(\hat{\beta}_0) = \frac{7.219}{2.175} = 3.319 \Rightarrow \beta_0 \neq 0, \quad \tau(\hat{\beta}_1) = \frac{699}{.083} = 8.422 \Rightarrow \beta_1 \neq 0, \quad (3)$$

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

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

Taken together, (i)-(iii) are often construed as providing confirmation for the demand schedule (2). Such a confirmation claim, however, is 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.

**Error source 1.** A first serious source of potential error is *statistical inadequacy*. The statistical inference results (i)-(iii) are reliable only when the estimated model in (2) is *statistically adequate*: the probabilistic assumptions:

$$[1] u_t \sim \mathbf{N}(\cdot, \cdot), \quad [2] E(u_t) = 0, \quad [3] \text{Var}(u_t) = \sigma^2, \quad [4] E(u_t u_s) = 0, \quad t \neq s, \quad t, s = 1, \dots, n,$$

comprising the underlying Linear Regression model, are valid for the particular data  $\mathbf{z}_0 := \{(x_t, y_t), t=1, \dots, n\}$ . 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.

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

(4)

The tiny p-values indicate serious departures from assumptions [2]-[4], rendering the inferences concerning the sign and the magnitude of the coefficients  $(\beta_0, \beta_1)$  *unwarranted*. In view of the M-S testing results, the estimated model in (1) constitutes an *unreliable basis* for inference and the claims (i)-(iii) are unwarranted. The unreliability of the inference 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, if any’, 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.

**Error source 2.** 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 introduce systematic errors into the data; see Abadir and Talmain (2002).

**Error source 3.** 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  $x_t = \frac{100(p_t - p_{t-1})}{p_t}$  and  $y_t = \frac{100(q_t - q_{t-1})}{q_t}$  can 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  $Q_{it}^D$ , at some point in time  $t$ , and the set of hypothetical prices  $P_{it}$ ,  $i=1, 2, \dots, m$ , and, on the other, the quantities transacted  $q_t$  and the corresponding observed prices  $p_t$ , over time

$t=1, 2, \dots, n$ , cannot possibly be bridged by the ‘proportional change’ transformation; Spanos (1995) for further discussion.

**Error source 4.** 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, 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 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 issues relating to probing for potential errors that could potentially render the inference unreliable raised above are usually ignored because the methodological framework adopted in traditional econometric modeling does *not* include such probing as part of the accepted rules and strategies for learning from data. Unfortunately, this methodological framework, referred to as *textbook econometrics*, is implicit and it’s usually adopted without examination as part and parcel of learning econometrics.

The emphasis in textbook econometrics is *not* in probing for potential errors at each stage of the modeling, but in 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. 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 (see Greene, 2000):

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

reveals a clear departure from assumption [3]. In textbook econometrics, 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 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  $(\widehat{\beta}_0, \widehat{\beta}_1)$ . This is supposed to be 'accounted for' by employing the so-called heteroskedasticity consistent (HC) standard errors. In view of the fact that  $HCSE(\widehat{\beta}_0)=2.363$ ,  $HCSE(\widehat{\beta}_1)=.108$ , these inferences appear to be 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 double-use of data, curve-fitting, pre-designation vs. post-designation. Interesting questions raised by the above textbook strategies are:

- 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?
- why does a mixture of statistical significance and theoretical meaningfulness renders a model "best"?
- are the various specification searches justified statistically?
- are the 'error-fixing' procedures justified on statistical grounds?
- is 'error-fixing' the best way to respecify a statistically inadequate model?
- how do we distinguish between legitimate and illegitimate double-use of data?
- what kind of robustness/reliability does the use of HC standard errors bring about?

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,

theorized to potentially influence the behavior of both  $x_t$  and  $y_t$ . This reasoning gives rise to an implicit reduced form:

$$y_t = \pi_{10} + \boldsymbol{\pi}_{11}^\top \mathbf{w}_t + \varepsilon_{1t}, \quad x_t = \pi_{20} + \boldsymbol{\pi}_{21}^\top \mathbf{w}_t + \varepsilon_{2t}, \quad t \in \mathbb{N}. \quad (6)$$

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

- in what sense does (6) alleviate the statistical inadequacy problem?
- 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  $(\beta_0, \beta_1)$  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?

More specifically:

- (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 grappled 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.

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), Dewitt (2004), Giere (1999), Godfrey-Smith (2003), Ladyman (2002), Machamer and Silberstein (2002), Newton-Smith (2000); for a more economics-oriented perspective see Caldwell (1994), 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, Russell, 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 con-

cerned 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  $e$  of  $h$ . If  $e$  turns out to be true, it provides confirmatory evidence for the (probable) validity of  $h$  :

$$\frac{\text{If } h \text{ then } e}{e,} \quad (7)$$

$$\therefore \text{(probably) } h \text{ is true}$$

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 competing hypotheses, contributed significantly to the tradition's demise by the 1970s. Their attempts to formalize induction as primarily a logical relationship  $C(e, h)$  between evidence  $e$  – taken as objectively given – and a hypothesis  $h$ , failed primarily because they did not adequately capture the complexity of the relationship between  $h$  and  $e$  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  $e$  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  $e$  – the experiential facts, was not just overly simplistic, but misleading in so far as neither  $h$  or  $e$  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, to which we will return below.

### 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, 1972) replaced the confirmation argument in (7) with a falsification argument, based on *modus tollens* (a deductively valid argument):

$$\frac{\text{If } h \text{ then } \mathbf{e} \\ \text{not-}\mathbf{e},}{\therefore \text{not-}h \text{ is true}} \quad (8)$$

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  $A_1, A_2, \dots, A_m$ . Hence, not- $h$  does not provide a way to distinguish between not- $H$  and not- $A_k$ ,  $k=1, \dots, m$ . As a result, one cannot apportion blame for the failure to observe  $e$  to any particular sub-set of the premises ( $H, A_1, A_2, \dots, A_m$ ). *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 very 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-ladenness 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 from 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  $e$  (or not- $e$ ) 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  $e$  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  $e$  (or not-  $e$ ) was made by

the "new experimentalism" tradition; Hacking (1983), Franklin (1986), 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.

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':

"... in contrast to the thrust of holistic models, I take these very problems to show that we need to look to the force of low-level methods of experiment and inference. 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  $e$  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 observational facts. This comes in the form of a hierarchy of models aiming to bridge the gap between theory and data in a piecemeal way that enables the modeler to 'learn from error' at each stage of the modeling:

"For each experimental inquiry we can delineate three types of models: models of primary scientific hypotheses, models of data, and models of experiment that link the others by means of test procedures." (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 experiential 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 completing 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 underlying the error-statistical account is that:

if a hypothesis  $H$  ‘passes’ a test  $T$  with data  $\mathbf{x}_0$ , but  $T$  had very low capacity to detect departures from  $H$  when present, then  $\mathbf{x}_0$  does not provide good evidence for the verity of  $H$ ; its passing  $T$  with  $\mathbf{x}_0$  is *not* a good indication that  $H$  is true.

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 Statistical inference: philosophical foundations

Modern frequentist statistics was founded by Fisher (1921, 1922) and extended by Neyman and Pearson (1933). Although the technical aspects of frequentist inference methods like estimation (point and interval), hypothesis testing and prediction were more or less in place by the late 1930s, their philosophical foundations were a source of tension, unclarity and heated discussions among the protagonists well into the 1950s; see Fisher (1955, 1956), Pearson (1955, 1966), Neyman (1956) and Lehmann (1993). The dispute was primarily focused on the form of inference in hypothesis testing with Fisher arguing in favor of his significance testing reasoning using the p-value (inductive inference), and Neyman favoring a more behavioristic interpretation founded upon the type I and II error probabilities and defined in terms of the accept/reject rule (inductive behavior). Fisher disdained these pre-data error probabilities viewing them as inextricably bound up with the ‘long-run’ metaphor of repeating the experiment under identical conditions, and thus only relevant in the case of acceptance sampling. Even though each side of this argument had its merits, neither account of

inductive reasoning provided an adequate account for addressing the question ‘when do data  $\mathbf{x}_0$  provide *evidence* for (or against) a hypothesis  $H$ ?’

Despite these foundational tensions, unclarities and problems, frequentist statistics was widely adopted as a primary tool for empirical modeling in numerous applied fields, including all the social sciences; for a historical/methodological perspective on the common roots and parallel developments between statistics and economics see Spanos (2007c). Indeed, statistics textbooks in the social sciences, especially psychology and sociology glossed over these tensions and presented hypothesis testing as a hybrid of both approaches; see Gigerenzer (1993). This hybrid approach interprets the p-value as a post-data type I error probability which provides more information than the accept/reject rule, but does not resolve the underlying tension. In order to shed light on this tension let us discuss the form and structure of frequentist induction.

## 4.1 Statistical Induction

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; the comparison with Bayesian inference will not be discussed in this paper (see Mayo, 1996, Cox and Mayo, 2007). This changeover is well-known and widely discussed; see Stigler (1986, 2005), Hald (1998, 2007). 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, 2001), 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’ evaluating its reliability. In particular, Fisher initiated a general way to quantify the uncertainty associated with inference by (a) *embedding* the *material experiment* into a *statistical model*, and then (b) use the latter to ascertain the (*frequentist*) *error probabilities* associated with particular inferences in its context. The form of induction envisaged by Fisher 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).

### 4.1.1 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  $n \rightarrow \infty$ , the observed proportion  $(m/n)$  converges in probability to the true proportion  $\theta$ ; see Pearson (1920).

Fisher’s model-based statistical induction extends the intended scope of induction-by-enumeration by replacing its focus on *events* and their probabilities with modeling the *mechanism* that underlies the generation of the observed data, and thus capturing all possible events and their probabilities. For example, the statistical model underlying the above example 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.

$$X_k \sim \text{BerIID}(\theta, \theta(1-\theta)), \quad k \in \mathbb{N}, \quad (9)$$

where ‘BerIID’ reads ‘Bernoulli, Independent and Identically Distributed’. The data  $\mathbf{x}_0 := (1, 0, 0, \dots, 1)$  are interpreted as a ‘typical’ realization of the sample  $\mathbf{X} := (X_1, X_2, \dots, X_n)$  generated by the process in (9).

The inference concerning the proportion  $\theta$  of red marbles in the urn amounts to choosing the point estimator:

$$\hat{\theta}_n(\mathbf{X}) = \frac{1}{n} \sum_{k=1}^n X_k, \quad (10)$$

as yielding a representative value for  $\theta$ ; note that the estimate  $\hat{\theta}_n(\mathbf{x}_0) = (\frac{m}{n})$ . The claim that  $(m/n)$  converges in probability to  $\theta$  is more formally stated in terms of  $\hat{\theta}_n(\mathbf{X})$  being a *consistent* estimator of  $\theta$ :

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( |\hat{\theta}_n(\mathbf{X}) - \theta| < \epsilon \right) = 1, \quad \text{for any } \epsilon > 0. \quad (11)$$

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. Starting with the last weakness, the explicit specification of the underlying statistical model in (9) replaces the stipulations (i)-(ii) with probabilistic assumptions, [1]  $X_k \sim \text{Ber}(\cdot, \cdot)$ , [2]  $\{X_k, k \in \mathbb{N}\}$  is Identically Distributed, and [3]  $\{X_k, k \in \mathbb{N}\}$  is Independent, which are *testable* vis-a-vis data  $\mathbf{x}_0$ . 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 explicitly, Fisher would proceed to derive the *sampling distribution*  $f(\hat{\theta}_n(\mathbf{x}))$  of the estimator  $\hat{\theta}_n(\mathbf{X})$  under assumptions [1]-[3], yielding:

$$\hat{\theta}_n(\mathbf{X}) \sim \text{Bin}\left(\theta, \frac{\theta(1-\theta)}{n}\right), \quad (12)$$

for any  $n > 1$ , where ‘Bin’ stands for a ‘Binomial’ distribution.  $f(\hat{\theta}_n(\mathbf{x}))$ , for all  $\mathbf{x} \in \mathbb{R}_X^n$  gives a complete description of the probabilistic structure of  $\hat{\theta}_n(\mathbf{X})$ , and provides the basis for addressing both of the other weaknesses of induction-by-enumeration.

Consistency can now be seen as a *minimal* property of an estimator, which, by itself, does not ensure the reliability of inference. Although (11) invokes the sampling distribution of  $\hat{\theta}_n(\mathbf{X})$ , it only concerns its behavior as  $n \rightarrow \infty$ , providing at best crude upper bounds for the uncertainty associated with any inference concerning  $\theta$ . In contrast, using the finite sampling distribution in (12) one can invoke more pertinent *finite sample* properties (valid for any  $n > 1$ ) to assess the optimality of  $\hat{\theta}_n(\mathbf{X})$  as an estimator of  $\theta$ , such as *unbiasedness*, *sufficiency* and *full efficiency*; see Cox and Hinkley (1974). More importantly, the uncertainty associated with any inference concerning  $\theta$  can now be ‘quantified’ in terms of the ascertainable *error probabilities*. For example, instead of a point estimate  $\hat{\theta}_n(\mathbf{x}_0) = \left(\frac{m}{n}\right)$  one can define the two-sided Confidence Interval (CI) for  $\theta$  :

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

which quantifies the relevant ‘approximation’ invoked by the straight rule via the coverage probability  $1-\alpha$ . For instance, if  $n = 10$ , and  $m = 2$ , i.e.  $\hat{\theta}_n(\mathbf{x}_0) = .2$ , the straight-rule inference can be shown to be unreliable (imprecise), despite being consistent, because the observed 95% Confidence Interval (CI) for  $\theta$  is  $[-.048 \leq \theta \leq .448]$ ; for simplicity we use  $c_{\frac{\alpha}{2}} = 1.96$  (Normal approximation) and ignore some of the problems associated with this interval (see Agresti, 2002). The width of this observed CI (0.496) indicates that the point estimate is not very precise and the fact it includes 0 suggests that on the basis of this data  $\theta$  is statistically indistinguishable from zero; see Lehmann (1986).

Before we consider some other aspects of Fisher’s recasting of statistical induction it is important to digress briefly in order to consider the question of the frequentist interpretation of probability.

#### 4.1.2 The frequency interpretation of probability

There has been a continuing discussion concerning the credibleness of the frequentist interpretation of probability going back to Fisher (1921) and von Mises (1928). The basic formal result invoked for this interpretation is the *Strong Law of Large Numbers (SLLN)*; a stronger version of (11). This theorem states that *under certain restrictions on the probabilistic structure of the process*  $\{X_k, k \in \mathbb{N}\}$  it follows that:

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{k=1}^n X_k\right) = p\right) = 1. \quad (13)$$

The first SLLN was proved by Borel in 1909 in the case of a Bernoulli, IID process, but since then the result in (13) 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.

This result in (13) can be used to define the *frequentist probability* of an event  $A := \{X=1\}$  via:

$$P(A) := \lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{k=1}^n X_k \right) = p. \quad (14)$$

To what extent does this provide a justification of the frequentist interpretation of probability? The issue often raised is that this justification is *circular*: it uses *probability* to define *probability*! For example, Lindley (1965) argues:

“... there is nothing impossible in  $\frac{m}{n}$  differing from  $p$  by as much as  $2\epsilon$ , it is merely rather unlikely. And the word unlikely involves probability ideas so that the attempt at a definition of ‘limit’ using mathematical limit becomes circular.” (p. 5)

This is denied by some notable mathematicians including Renyi (1970, p. 159) who draws a clear distinction between the intuitive description in (14), and the purely a mathematical result in (13), dismissing the circularity charge as based on conflating the two. Hence, the justification of the above frequentist interpretation of  $\mathbb{P}(A) = p$ , is not derived from any a priori stipulations such as (i)-(ii) above, but stems from the appropriateness of the probabilistic assumptions needed to prove (13); assumptions which are testable vis-a-vis a realization of the process  $\{X_k, k \in \mathbb{N}\}$ .

## 4.2 The nature of different forms of statistical inference

To understand the problem concerning the proper form of *inductive reasoning* in frequentist statistics and the issues raised, it is important to bring out the differences in the nature of induction between alternative forms of inference. With that in mind, consider  $\mathbf{x}_0 = (x_1, x_2, \dots, x_n)$  being a ‘typical realization’ of the simple Normal model as specified in table 1.

**Table 1 - The simple Normal model**

Statistical GM:	$X_k = \mu + u_k, k \in \mathbb{N},$	
[1] Normal:	$X_k \sim \mathbf{N}(\cdot, \cdot),$ for all $k \in \mathbb{N},$	(15)
[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.	

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

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k, \quad s^2 = \frac{1}{(n-1)} \sum_{k=1}^n (X_k - \bar{X})^2, \quad (16)$$

constitute ‘optimal’ estimators of  $(\mu, \sigma^2)$ , with sampling distributions:

$$\bar{X} \sim \mathbf{N} \left( \mu, \frac{\sigma^2}{n} \right), \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1), \quad (17)$$

where  $\chi^2(n-1)$  denotes the chi-square distribution with  $(n-1)$  degrees of freedom. Moreover, for testing the *hypotheses*:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0, \quad (18)$$

the test statistic  $\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s}$ , with a sampling distribution:

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \stackrel{H_0}{\rightsquigarrow} \text{St}(n-1), \quad (19)$$

where  $\text{St}(n-1)$  denotes the Student's t distribution with  $(n-1)$  degrees of freedom. When  $\tau(\mathbf{X})$  is combined with a rejection region  $C_1(\alpha)$  where:

$$C_1(\alpha) = \{\mathbf{x} : |\tau(\mathbf{x})| > c_\alpha\}, \quad C_0(\alpha) = \{\mathbf{x} : |\tau(\mathbf{x})| \leq c_\alpha\}, \quad (20)$$

one can define the well-known t-test; see Lehmann (1986).

Using the well-known duality between hypothesis testing and *interval estimation* based on the acceptance region  $C_0(\alpha)$  (Lehmann, 1986), one can specify the two-sided Confidence Interval (CI) for  $\mu$ :

$$\mathbb{P} \left( \bar{X} - c_{\frac{\alpha}{2}} \left( \frac{s}{\sqrt{n}} \right) \leq \mu \leq \bar{X} + c_{\frac{\alpha}{2}} \left( \frac{s}{\sqrt{n}} \right) \right) = 1 - \alpha. \quad (21)$$

*Prediction* differs from the above inferences in so far as it is concerned with finding the most representative value of  $X_k$  beyond the observed data, say  $X_{n+1}$ . A good predictor for  $X_{n+1}$  in the case of (15) is given by  $\hat{X}_{n+1} = \bar{X}$ .

How are these inference procedures (estimation, testing and prediction) different?

It turns out that their differences arise primarily from the nature of induction involved; the questions posed and the answers being elicited from the data. The traditional statistical literature distinguishes between estimation (point and interval) and hypothesis testing. Chatterjee (2003) describes to the former as *open induction* and to the latter as *hypothetic induction* paraphrasing approvingly the philosopher Day (1961): “Some philosophers regard hypothetic induction more important than open induction for the progress in science ..., since one can give free play to one’s imagination in framing the hypothesis.” (ibid. p. 65). This is a very interesting point that it is not widely appreciated in statistics.

#### 4.2.1 Inductive reasoning in estimation

In point estimation one selects the most ‘representative’ value of the unknown parameter  $\theta$ ; representativeness being formalized in terms of optimal properties such as unbiasedness, efficiency, sufficiency, consistency etc. The form of reasoning that underlies estimation is that of *factual reasoning*, where these properties are defined in terms of the sampling distribution of such estimators evaluated under the ‘true state of nature’ (TSN):

$$\bar{X} \stackrel{\text{TSN}}{\rightsquigarrow} \text{N} \left( \mu_*, \frac{\sigma_*^2}{n} \right), \quad \frac{(n-1)s^2}{\sigma_*^2} \stackrel{\text{TSN}}{\rightsquigarrow} \chi^2(n-1), \quad (22)$$

where  $(\mu_*, \sigma_*^2)$  denote the ‘true’ values of  $(\mu, \sigma^2)$ , whatever those happen to be; (22) is more accurate than (17). The main problem with this form of inductive inference

is that defining error requires one to know  $(\mu_*, \sigma_*^2)$ . To circumvent this problem error probabilities in point estimation, such as *Mean Square Error (MSE)*, are usually defined in terms of a quantifier that involves all possible values of  $\theta$  :

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

There is something wrong-headed about such a definition because the quantifier gives rise to absurd results like  $\bar{X}$  is *not* uniformly better than  $\tilde{\mu} = 7405926$  as an estimator of  $\mu$ ;  $MSE(\bar{X} - \mu) \stackrel{\leq}{\geq} MSE(\tilde{\mu} - \mu)$  for  $\mu \in \mathbb{R}$ , even though  $\tilde{\mu}$  ignores the data completely. As a result, ensuring that an estimator is consistent, unbiased, sufficient or even fully efficient, does *not* provide one with enough information to evaluate the reliability of a point estimate inference.

Interval estimation attempts to rectify this deficiency by providing a way to evaluate the relevant error probabilities associated with an interval estimator covering the true value  $\theta^*$  of the unknown parameter  $\theta$  :

$$\mathbb{P}\left(\bar{X} - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right); \mu = \mu^*\right) = 1 - \alpha, \quad (23)$$

where ‘ $\mu = \mu^*$ ’ indicates that the evaluation of the relevant probabilities are under the TSN; note that  $\mathbb{P}\left(\bar{X} - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right); \mu \neq \mu^*\right) = \alpha$  is the *coverage error* probability, the relevant pivot being:

$$\frac{\sqrt{n}(\bar{X} - \mu_*)}{s} \stackrel{\text{TSN}}{\underset{\sim}{\rightsquigarrow}} \text{St}(n-1). \quad (24)$$

The important difference with point estimation is that (23) holds, whatever the true value  $\mu^*$  is.

#### 4.2.2 Inductive reasoning in hypothesis testing

The questions posed in estimation and testing, as well as the answers elicited, differ because the form of reasoning underlying these procedures are dissimilar. Hypothesis testing poses more probative questions, i.e. whether particular hypothetical values of  $\theta$  are warranted in view of the data. In contrast to estimation, the reasoning underlying hypothesis testing is *counterfactual*. The sampling distribution of a test statistic is evaluated under several hypothetical scenarios. In particular, what would the sampling distribution of the test statistic be if the null or the alternative hypotheses are true? In the above case these scenarios give rise to the counterfactual sampling distribution results:

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \stackrel{H_0}{\rightsquigarrow} \text{St}(n-1), \quad \frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \stackrel{H_1}{\rightsquigarrow} \text{St}(\delta; n-1), \quad \text{for any } \mu_1 > \mu_0, \quad (25)$$

where  $\text{St}(\delta; n-1)$  denotes the non-central Student’s t distribution with  $(n-1)$  degrees of freedom and non-centrality parameter  $\delta = \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}$ . Using (25) one can define the *type I and II error probabilities* via:

$$\mathbb{P}\left(|\tau(\mathbf{X})| > c_{\frac{\alpha}{2}}; H_0\right) = \alpha, \quad \mathbb{P}\left(|\tau(\mathbf{X})| \leq c_{\frac{\alpha}{2}}; H_1(\mu_1)\right) = \beta(\mu_1), \quad \text{for } \mu_1 > \mu_0. \quad (26)$$

Related to the type II error probability is the *power* of the test:

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

in terms of which the optimality of a test is defined. In this case the above test can be shown to be Uniformly, Most Powerful Unbiased; see Lehmann (1986), Cox and Hinkley (1974). A N-P test is defined in terms of a test statistic (distance function)  $\tau(\mathbf{X})$  and a rejection region (see (20)) giving rise the accept/reject rule:

$$\text{accept } H_0 \text{ if } \mathbf{x}_0 \in C_0(\alpha), \quad \text{reject } H_0 \text{ if } \mathbf{x}_0 \in C_1(\alpha). \quad (27)$$

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. The trustworthiness of a Confidence Interval (CI) is ascertained in terms of a single error probability known as the *coverage error probability*: the probability that the interval does *not* contain the true value of the unknown parameter(s). In the case of hypothesis testing the is ascertained in terms of two error probabilities, *type I (II)*: the probability of rejecting (accepting) the null hypothesis when true (false); see Cox and Hinkley (1974).

The factual nature of reasoning in estimation gives one no flexibility in posing questions to the data, but the counterfactual reasoning in testing allows one avid flexibility to pose sharper questions using particular (hypothetical) values for  $\mu$ ; "[it] can give free play to one's imagination in framing the hypothesis". That often elicits more informative answers from the data. To see this, let the statistical model be denoted by  $\mathcal{M}_\theta(\mathbf{x}) = \{f(\mathbf{x}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ , where  $f(\mathbf{x}; \boldsymbol{\theta})$ ,  $\mathbf{x} \in \mathcal{X} := \mathbb{R}_X^n$ , is the distribution of the sample;  $\mathcal{X}$  and  $\Theta$  being the sample and parameter space, respectively. In general,  $\mathcal{M}_\theta(\mathbf{x})$  constitutes a subset of the set of all possible models, say  $\mathcal{P}(\mathbf{x})$ , that could have given rise to the data  $\mathbf{x}_0$ . In point estimation, an optimal estimator  $\hat{\boldsymbol{\theta}}$  of  $\boldsymbol{\theta}$ , reduces the subset  $\mathcal{M}_\theta(\mathbf{x})$  to a point  $\mathcal{M}_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) = \{f(\mathbf{x}; \hat{\boldsymbol{\theta}})\}$ , but leaves the question of the uncertainty associated with this choice open. A *confidence interval* quantifies that uncertainty by yielding a subset  $\mathcal{M}_{\hat{\boldsymbol{\theta}} \pm \epsilon}(\mathbf{x}) = \{f(\mathbf{x}; \boldsymbol{\theta}), \hat{\boldsymbol{\theta}} - \epsilon \leq \boldsymbol{\theta} \leq \hat{\boldsymbol{\theta}} + \epsilon\} \subset \mathcal{M}_\theta(\mathbf{x})$ , but treats all  $\boldsymbol{\theta}$  inside the interval on par and provides no measure of uncertainty associated with the models within the observed CI; it is well-known that one cannot attach the pre-data coverage probability  $(1 - \alpha)$  to the observed CI.

*Hypothesis testing*, partitions the original statistical model  $\mathcal{M}_\theta(\mathbf{x})$  into two subsets  $\mathcal{M}_{\theta_0}(\mathbf{x}) = \{f(\mathbf{x}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_0\}$  and  $\mathcal{M}_{\theta_1}(\mathbf{x}) = \{f(\mathbf{x}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_1\}$ , and poses the question whether the data  $\mathbf{x}_0$  were generated by  $\mathcal{M}_{\theta_0}(\mathbf{x})$  or  $\mathcal{M}_{\theta_1}(\mathbf{x})$ . By narrowing down  $\Theta_0$  hypothesis testing can pose sharp questions to the data and potentially elicit more incisive answers. However, the Neyman-Pearson accept/reject decision does not give the sought after incisive answer to the question of which subset generated the data. Accepting  $H_0 : \boldsymbol{\theta} \in \Theta_0$  does not necessarily mean that that data  $\mathbf{x}_0$  provide

evidence for  $\mathcal{M}_{\theta_0}(\mathbf{x})$  because the test might have had very low power to detect any substantive departures of interest that were present. Similarly, rejecting  $H_0 : \theta \in \Theta_0$  does not necessarily mean that that data  $\mathbf{x}_0$  provide evidence for  $\mathcal{M}_{\theta_1}(\mathbf{x})$  because the test might have been oversensitive to minor departures from the null. These two scenarios give rise to the well known acceptance (rejection) fallacy: interpreting the acceptance (rejection) of the null as evidence *for* the null (alternative). A variant of the rejection fallacy is the well-known confusion between statistical and substantive significance.

### 4.2.3 Inductive reasoning in prediction

Prediction differs from estimation (point and interval) and hypothesis testing in so far as it is *not* posing any questions concerning the data generating mechanism vis-a-vis  $\mathcal{M}_\theta(\mathbf{x})$ . Instead, it takes the estimated model  $\mathcal{M}_{\hat{\theta}}(\mathbf{x})$  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  $\hat{X}_{n+1} = h(\mathbf{X})$ . The prediction error is defined by  $e_{n+1} = (X_{n+1} - \hat{X}_{n+1})$ , and its sampling distribution is evaluated under the *true state of nature* (TSN), as in the case of estimation.

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

$$e_{n+1} = (X_{n+1} - \bar{X}) \stackrel{\text{TSN}}{\underset{\sim}{\sim}} \mathbf{N} \left( 0, \sigma_*^2 \left( 1 + \frac{1}{n} \right) \right), \quad \frac{e_{n+1}}{s\sqrt{\left(1+\frac{1}{n}\right)}} \stackrel{\text{TSN}}{\underset{\sim}{\sim}} \text{St}(n-1),$$

with  $s\sqrt{\left(1+\frac{1}{n}\right)}$  providing a measure of the uncertainty associated with predicting the event  $\{X_{n+1}=x_{n+1}\}$  using  $\hat{X}_{n+1}=\bar{X}$ . This can be used to construct a prediction CI of the form:

$$\mathbb{P} \left( \bar{X} - c_{\frac{\alpha}{2}} s \sqrt{\left(1 + \frac{1}{n}\right)} \leq x_{n+1} \leq c_{\frac{\alpha}{2}} s \sqrt{\left(1 + \frac{1}{n}\right)}; \mu = \mu^* \right) = 1 - \alpha. \quad (28)$$

In summary, in hypothesis testing the probativeness of a test is assessed in terms of its capacity (power) to reject (accept) the null when false (true), with the relevant type I and II errors appraising its imperfectibility to do that, respectively. The relevant error in the case of interval estimation concerns the imperfectibility of the CI interval to cover the true value of the parameter in question. Similarly, in prediction the error concerns the imperfectibility of the CI to cover the realized value of  $X_{n+1}$ .

## 4.3 Statistics and its foundational problems circa 1955

Fisher (1922, 1925, 1935) used the notion of a sampling distribution to construct an almost complete optimal theory of point estimation which included the maximum likelihood method, the finite sample properties of unbiasedness, full efficiency, sufficiency and the asymptotic properties of consistency and asymptotic Normality. However, his significance testing based on the p-value was incomplete. A more complete optimal hypothesis testing was proposed by Neyman and Pearson (1933) as

a modification/extension of Fisher’s theory. Neyman (1937) used a duality result between hypothesis testing confidence interval estimation to propose an optimality theory for the latter.

Neyman (1950) superseded Fisher’s metaphor of an ‘infinite population’ as a basis of a statistical model, with the notion of a ‘stochastic (chance) mechanism’, which extended the intended scope of statistical modeling beyond IID samples. He described the stages of statistical modeling as follows:

“There are three distinct steps in this [statistical modeling] process:

- (i) Empirical establishment of apparently stable long-run relative frequencies ... of events judged interesting, as they develop in nature.
- (ii) Guessing and then verifying the ‘chance mechanism’, the repeated operations of which produces the observed frequencies. This is a problem of ‘frequentist probability theory’. Occasionally, this step is labelled ‘model building’. Naturally, the guessed chance mechanism is hypothetical.
- (iii) Using the hypothetical chance mechanism of the phenomenon studied to deduce rules of adjusting our actions (or decisions) to the observations so as to ensure the highest ‘measure of success’. (Neyman, 1977, p. 99)

In (i) Neyman demarcates the domain of statistical modeling to *stochastic phenomena* which exhibit *chance regularity*, in the form of the long-run stability of relative frequencies from observational data. In (ii) he provides a clear statement concerning the nature of *specification* and model validation, and in (iii) he brings out the role of ascertainable error probabilities in assessing the optimality of inference procedures.

In summary, the crucial features the Fisher-Neyman-Pearson model-based statistical induction can be summarized as follows:

- (i) it models the data generating mechanism itself, ensuring the ampliative nature of the inference,
- (ii) it renders the premises testable vis-a-vis the data in hand,
- (iii) it provides a way to assess the reliability of inference via the error probabilities derived from the relevant sampling distributions,
- (iv) it extends the intended scope of statistical modeling by broadening the notion of a statistical model to include non-IID samples (stochastic processes)
- (v) it formalizes various alternative forms of inference: estimation, confidence intervals, hypothesis testing, prediction.

#### 4.3.1 Philosophical/methodological tensions in statistics

Gigerenzer (1993) described the textbook discussion on hypothesis testing forged in the 1960s as an ‘infelicitous hybrid’ of two fundamentally different approaches to testing, the Fisher and Neyman-Pearson approaches. This hybrid testing adopted the basic N-P framework but supplemented it with the p-value notion. Returning to

the simple Normal model and the hypotheses in (18), the p-value:

$$\mathbb{P}(|\tau(\mathbf{X})| > |\tau(\mathbf{x}_0)|; H_0) = p(\mathbf{x}_0), \quad (29)$$

was used to modify the accept/reject rule (27) into:

$$\text{accept } H_0 \text{ if } p(\mathbf{x}_0) > \alpha, \quad \text{reject } H_0 \text{ if } p(\mathbf{x}_0) \leq \alpha. \quad (30)$$

This forged hybrid would have incensed both Fisher and Neyman, but practitioners felt that the difference  $(p(\mathbf{x}_0) - \alpha) \stackrel{\geq}{\leq} 0$  could be intuitively interpreted as shedding additional light on the ‘strength of evidence’ for or against  $H_0$ . For example, a p-value  $p(\mathbf{x}_0) = .00001$ , indicates much stronger evidence against  $H_0$  than  $p(\mathbf{x}_0) = .048$ . Although there is an element of truth in this intuition, the p-value does not provide an adequate resolution to bridging the gap between the coarse accept/reject rule and an evidential interpretation of the inference result; see Barnett (1999).

The underlying problem is that accepting  $H_0$  in a N-P test does *not* warrant the claim that data  $\mathbf{x}_0$  provide *evidence for*  $H_0$ . Also, rejecting  $H_0$  does *not* warrant the claim that data  $\mathbf{x}_0$  provide *evidence for*  $H_1$ . Doing so gives rise to the well-known *fallacies of acceptance and rejection*, respectively, which have bedeviled hypothesis testing since the mid 1930s.

Similarly, a p-value of  $p(\mathbf{x}_0) = .125$  could not be interpreted as data  $\mathbf{x}_0$  providing evidence for  $H_0$  because its falsifying-orientation renders such an affirmative evidential interpretation unwarranted. Moreover, although a small p-value, say  $p(\mathbf{x}_0) = .01$  provides some evidence for a discrepancy from  $H_0$ , its dependence on the sample size  $n$  renders such an interpretation susceptible to a particular instantiation of the fallacy of rejection: *statistical* significance might not indicate the presence of *substantive* significance; see Mayo (2005).

These problems gave rise to a numerous debates (see Harper and Hooker, 1976), which were especially heated in the social sciences like psychology, sociology and education; see Morrison and Henkel (1970), Lieberman (1971); more recently re-discovered in economics (McCloskey, 1985). As a result of this debates social scientists are currently making a concerted effort aided by journal editors to persuade practitioners to reduce their reliance on *p-value* significance testing and use *confidence intervals* (CIs) instead; see Harlow et al (1997), Altman et al (2000). The justification underlying these efforts is that CIs are more informative than p-values and are less susceptible to some of the *fallacies* and confusions that have befuddled hypothesis testing. The only perceived drawback is that observed CIs cannot discriminate among the values of the unknown parameter within that interval.

Hacking (1965) went further in criticizing the Neyman-Pearson testing arguing that this theory is clearly incomplete. 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. He questioned the role of error probabilities post-data because they are inextricably bound up with the frequentist ‘long-run’ metaphor of repeating the experiment under

identical conditions, and put forward an after-trial likelihood-based support theory. He changed his mind about the appropriateness of the latter in Hacking (1980).

In addition to the problem of:

- (a) the proper form of *inductive reasoning* underlying frequentist inference, and
- (b) the various fallacies the N-P approach is vulnerable to,

statistical modeling has been bedeviled by several additional outstanding philosophical/methodological issues including:

- (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.

According to Rao (2004):

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

It turns out that most of the above issues depend crucially on being able to specify the *relevant* error probabilities unequivocally under a variety of different circumstances and assess the reliability of the *inference in question*. Addressing these problems depends crucially on finding an adequate post-data interpretation of Neyman-Pearson testing that will enable one to give unambiguous answers to the question ‘when do data  $\mathbf{x}_0$  provide *evidence* for (or against)  $H$ ?’ Such a post-data interpretation was proposed by Mayo (1991, 1996) in the form of severe testing reasoning. Using this perspective Spanos (2000, 2001, 2006a-c, 2007b) discusses (c)-(e); Mayo and Cox (2006) discuss (e)-(f).

## 5 The Error-Statistical approach

The term error-statistics was coined by Mayo (1996) in order to describe a refined modification/extension of the Fisher-Neyman-Pearson frequentist approach to inference which can be used addresses some of the inveterate problems mentioned above. In particular, the error-statistical approach emphasizes the *learning from data* objective and supplements the Neyman-Pearson testing framework with a post-data evaluation of inference component based on *severe testing reasoning*. The term is chosen to reflect the central role *error probabilities* play in assessing the reliability of inference, both pre-data as well as post-data.

The post-data supplement enables one to address Hacking’s (1965) pre-data optimality vs. post-data evaluation of inference issue, as well as deals effectively with the various testing fallacies (acceptance, rejection) alluded to above. In addition, it deals with Gigerenzer’s (1993) ‘infelicitous hybrid’ problem by accommodating Fisher’s significance testing reasoning in the same statistical framework, interpreting the p-value as a post-data error probability. The presence/absence of an alternative renders the

Fisher and the N-P approaches to testing complementary and gives rise to no contradictions. As argued in Spanos (1999), testing the validity of the premises of inferences (Mis-Specification (M-S) testing) is more akin to the Fisher testing than to N-P testing. M-S testing is particularly important for the error-statistical approach because it enables one to secure the *statistical adequacy* of the model: the assumptions constituting the statistical model are valid for data  $\mathbf{x}_0$ . Statistical adequacy constitutes a necessary condition for the reliability of inference because it ensures that the actual error probabilities are approximately equal to the nominal ones; any discrepancy between them is likely to give rise to misleading inferences. When a 5% nominal significance level t-test ( $\tau(\mathbf{X}), C_1(.05)$ ) turns out to have a 95% actual significance, it is highly likely that it will give rise to unreliable inferences; see Spanos and McGuirk (2001), Spanos (2005).

Using statistical adequacy as the exclusive criterion for the choice of a statistical model vis-a-vis data  $\mathbf{x}_0$ , goes a long way toward addressing another inveterate problem in empirical modeling, that of model selection based on goodness-of-fit criteria, including Akaike type information criteria. In traditional accounts statistical model specification, as envisaged by Fisher (1922), is often conflated with model selection. The latter, however, is always viewed as based on comparing a number of different prespecified models on some criterion or other, which is a *relativist* comparison. In contrast, statistical adequacy is a non-relativist criterion because it concerns how well a particular statistical model  $\mathcal{M}_\theta(\mathbf{x})$  captures the systematic information in data  $\mathbf{x}_0$ . Model selection makes statistical sense only when one is comparing between different statistical models whose adequacy has already been established; see Spanos (2007b).

The error-statistical approach also elucidates the comparisons between p-values and CIs and can be used to address the problem of ‘effect sizes’ (see Rosenthal et al, 1999) sought after in some applied fields like psychology and epidemiology; see Spanos (2004).

## 5.1 Severe testing reasoning

### 5.1.1 Post-data error probabilities

It is well-known that one cannot attach the coverage probability  $(1-\alpha)$  to the observed CI:

$$\left[\bar{x} - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right), \bar{x} + c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)\right]; \quad (31)$$

see Arnold (1990). The reason is that what constitutes *error* in this case cannot even be defined post-data since the underlying factual reasoning would require one to know the *true value*  $\mu_*$ . Moreover, even if that were possible the post-data *coverage error probability* will be either zero or one, depending on whether the *true value*  $\mu_*$  of the parameter lies or does not lie within the observed CI (31). Hence, factual reasoning does *not* lend itself to defining post-data error probabilities.

In contrast, counterfactual reasoning can be easily adapted to give rise to post-data error probabilities. Indeed, the p-value in (29) can be legitimately viewed as

such. What makes it a post-data error probability is the fact that it can only be defined post-data and its evaluation is based on the same sampling distribution as that of type I error. Reflecting on these two error probabilities, it is clear that, post-data, the *de facto* relevant threshold is no longer the pre-designated  $c_{\frac{\alpha}{2}}$ , but the observed value of the test statistic  $|\tau(\mathbf{x}_0)|$ . The difference between factual and counterfactual reasoning and the associated error probabilities 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; see Spanos (2004).

Users of significance testing 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.

### 5.1.2 Severe testing

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 (table 1) and, to simplify the notation, consider the one-sided *hypotheses*:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0. \quad (32)$$

It is well-known that  $T_\alpha := \{\tau(\mathbf{X}), C_1(\alpha)\}$  (see (25), (20)) defines a Uniformly Most Powerful (UMP) test; 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:

$$\begin{aligned} \text{Sev}(T_\alpha; \mathbf{x}_0; \mu \leq \mu_1) &= \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu \leq \mu_1 \text{ is false}) = \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu > \mu_1), \\ \text{Sev}(T_\alpha; \mathbf{x}_0; \mu > \mu_1) &= \mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu > \mu_1 \text{ is false}) = \mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu \leq \mu_1), \end{aligned} \quad (33)$$

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*  $\gamma$  from  $H_0$ , and in the case of reject establish the *largest discrepancy*  $\gamma$  from  $H_0$ , that is licensed by data  $\mathbf{x}_0$ . The discrepancy parameter  $\gamma$  plays a crucial role in the severity assessment because it reflects what Fisher called the ‘strength of evidence’ for or against the null warranted by data  $\mathbf{x}_0$ .

Viewed from the severity perspective the p-value can be interpreted as a crude 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  $H_0$ .

### 5.1.3 Severe testing and the p-value

Such a small p-value indicates that  $\mathbf{x}_0$  *accords with*  $H_1$ , and the question is whether it provides evidence for  $H_1$ . The severe-testing interpretation suggests that  $H_1$  has passed a severe test because the probability that test  $T_\alpha$  would have produced a result that accords less well with  $H_1$  than  $\mathbf{x}_0$  does (values of  $\tau(\mathbf{x})$  less than  $\tau(\mathbf{x}_0)$ ), if  $H_1$  were false ( $H_0$  true) is:

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

and is very high. The severity construal of the p-value brings out its most crucial weakness: it establishes the existence of *some* discrepancy  $\gamma \geq 0$ , but provides no information concerning the magnitude of the discrepancy  $\gamma$  licensed by data  $\mathbf{x}_0$ . Moreover, the dependence of the p-value on the sample size can belie the warranted discrepancy. The severity evaluation addresses both of these problems (Mayo and Spanos, 2006).

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  $\gamma$ ; the underlying test would have detected a departure from the null as large as  $\gamma$  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 errors* 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 (33). This can be used to provide unambiguous answers to the question posed earlier:

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

The chronic fallacies associated with N-P testing, alluded to above, can also be addressed using Mayo’s post-data severe testing reasoning; see Mayo (1996), Mayo and Spanos (2006).

#### 5.1.4 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 (32), rejecting  $H_0$  only establishes the presence of some discrepancy from  $\mu_0$ , say  $\delta > 0$ , but it does not provide any information concerning the magnitude of  $\delta$ . The severity evaluation  $\text{Sev}(T_\alpha; \mathbf{x}_0; \mu > \mu_1)$  associated with the claim that  $\mu > \mu_1 = \mu_0 + \gamma$ , for some  $\gamma \geq 0$ , can be used to establish the warranted discrepancy  $\gamma^*$ , and then proceed to assess whether  $\gamma^*$  is also substantively significant or not. More generally, the severity evaluation provides a sharper answer to the original question whether the data  $\mathbf{x}_0$  were generated by  $\mathcal{M}_{\theta_0}(\mathbf{x})$  or  $\mathcal{M}_{\theta_1}(\mathbf{x})$ , and addresses the lacuna left by the observed CI, by replacing its factual reasoning with the counterfactual reasoning associated with severe testing; see Mayo and Spanos (2006) for the details.

## 5.2 Statistical adequacy

Perhaps the most crucial feature of error-statistics is its reliance on error probabilities, pre-data, to evaluate the trustworthiness 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  $\mathcal{M}_\theta(\mathbf{x})$  (e.g. table 1) demarcating the premises of inference when viewed in conjunction with data  $\mathbf{x}_0 := (x_1, x_2, \dots, x_n)$ . A crucial precondition for ensuring statistical adequacy is a complete specification of a statistical model in terms of testable assumptions. *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  $\mathbf{x}_0$ .

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 takes the form:

$$H_0 : f_*(\mathbf{x}) \in \mathcal{M}_\theta(\mathbf{x}), \quad \text{vs.} \quad H_1 : f_*(\mathbf{x}) \in [\mathcal{P}(\mathbf{x}) - \mathcal{M}_\theta(\mathbf{x})], \quad (34)$$

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 (34) 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 searching within this boundary.

The problem that needs to be addressed for (34) 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  $\mathcal{M}_\theta(\mathbf{x})$ , or as vague as a direction of departure from  $\mathcal{M}_\theta(\mathbf{x})$ , which might only implicitly be determined; the latter is similar to a goodness-of-fit test.

The counterfactual 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  $\mathbf{x}_0$  is ‘improbably far’ from what would have been expected if  $H_0$  were true. The difference is that in a M-S test  $H_0$  is always the statistical model in question  $\mathcal{M}_\theta(\mathbf{x})$ . ‘Improbably far’ is evaluated in terms of a distance function  $d(\mathbf{X})$  whose form is related to the particularized version of  $\overline{\mathcal{M}_\theta(\mathbf{x})}$ . In the case where this is 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. Even in the case where  $\mathcal{M}_\theta(\mathbf{x})$  is parametrically nested within the alternative model  $\mathcal{M}_\psi(\mathbf{x})$ , a M-S test is unlike a N-P test in so far as the validity of the latter model is not a stipulation; for a N-P test to be reliable – the nominal and actual error probabilities are approximately equal –  $\mathcal{M}_\theta(\mathbf{x})$  needs to be statistically adequate. 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. Hence, 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).

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).

### 5.3 A chain of complecting models: theory $\longleftrightarrow$ data

Another aspect of modeling that the error-statistical approach differs appreciably from the traditional approach 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) for further discussion. Disentangling the role played by the two sources of information has been a major problem in modern statistics (see Lehmann, 1990, Cox, 1990). 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 see some of the details in a generic set up.

#### 5.3.1 How theory links up 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  $\boldsymbol{\xi}_k$  (observable and unobservable) influencing the behavior of  $y_k$ . A *theory model* is used to denote an idealized mathematical representation of a theory, say:

$$y_k = h^*(\mathbf{x}_k, \boldsymbol{\xi}_k), \quad k \in \mathbb{N}. \quad (35)$$

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  $\boldsymbol{\xi}_k$  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  $(\mathbf{x}_k, \boldsymbol{\xi}_k)$  explains the conjuring of *ceteris paribus* clauses. This line of reasoning transforms the theory model (35) into a *structural (estimable) model* of the form:

$$y_k = h(\mathbf{x}_k; \boldsymbol{\phi}) + \epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k), \quad k \in \mathbb{N}, \quad (36)$$

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

$$\{\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k) = y_k - h(\mathbf{x}_k; \boldsymbol{\phi}), \quad k \in \mathbb{N}\}, \quad (37)$$

is viewed as a function of both  $\mathbf{x}_k$  and  $\boldsymbol{\xi}_k$ . For (37) to provide a meaningful model for  $y_k$  the error term needs to be non-systematic: a *white-noise* (non-systematic) stochastic process  $\{\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k), k \in \mathbb{N}\}$  satisfying the properties:

$$\left. \begin{array}{l} \text{[i]} \quad E(\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k))=0, \\ \text{[ii]} \quad E(\epsilon^2(\mathbf{x}_k, \boldsymbol{\xi}_k))=\sigma^2, \\ \text{[iii]} \quad E(\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k) \cdot \epsilon(\mathbf{x}_\ell, \boldsymbol{\xi}_\ell)) = 0, \quad k \neq \ell, \quad k, \ell \in \mathbb{N}, \end{array} \right\} \forall (\mathbf{x}_k, \boldsymbol{\xi}_k) \in \mathbb{R}_{\mathbf{x}} \times \mathbb{R}_{\boldsymbol{\xi}}. \quad (38)$$

In addition to [i]-[iii], one needs to ensure (see Spanos, 1995) that the generating mechanism (36) is ‘nearly isolated’ in the sense that the unmodeled component ( $\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k)$ ) is *uncorrelated* with the modeled influences ( $h(\mathbf{x}_k; \boldsymbol{\phi})$ ):

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

Looking at assumptions [i]-[iv] it is clear that they are empirically non-testable because their confirmation would involve *all possible values* of both  $\mathbf{x}_k$  and  $\boldsymbol{\xi}_k$ . 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  $\mathbf{Z}_0 := (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n)$  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  $\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k)$ , for all  $(\mathbf{x}_k, \boldsymbol{\xi}_k) \in \mathbb{R}_{\mathbf{x}} \times \mathbb{R}_{\boldsymbol{\xi}}$  into a *generic white noise error process* without the qualifier.

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  $\boldsymbol{\xi}_k$  by ensuring that the uncontrolled factors cancel each other out; see Fisher (1935). The objective is to transform the structural error into a generic white noise process:

$$\left( \epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k) \Big| \Big|_{\substack{\text{controls,} \\ \text{experimental} \\ \text{design}}} \right) = \varepsilon_k \sim \text{IID}(0, \sigma^2), \quad k = 1, \dots, n. \quad (39)$$

This in effect embeds the structural model (36) into a *statistical model* of the form:

$$y_k = h(\mathbf{x}_k; \boldsymbol{\theta}) + \varepsilon_k, \quad \varepsilon_k \sim \text{IID}(0, \sigma^2), \quad k = 1, 2, \dots, n, \quad (40)$$

where the statistical error term  $\varepsilon_k$  in (40) is qualitatively very different from the structural error term  $\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k)$  in (36), because  $\varepsilon_k$  is no longer a function of  $(\mathbf{x}_k, \boldsymbol{\xi}_k)$ , and its assumptions are rendered empirically testable; see Spanos (2006a). A widely used special case of (40) 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  $D(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n; \boldsymbol{\phi})$

– 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-ladenness 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

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:

$$(u_t | \mathfrak{D}_t) \sim \text{IID}(0, \sigma^2), \quad t = 1, 2, \dots, n. \quad (41)$$

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  $\mu_t$  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 (41) 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  $D(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n; \phi) \rightsquigarrow \prod_{t=1}^n D(y_t | \mathbf{X}_t; \psi_1)$ ; see Spanos (1986).

<b>Table 3 - The Normal/Linear Regression Model</b>	
<b>Statistical GM:</b>	$y_t = \beta_0 + \beta_1^\top \mathbf{x}_t + u_t, \quad t \in \mathbb{N},$
<b>[1] Normality:</b>	$(y_t   \mathbf{X}_t = \mathbf{x}_t) \sim \text{N}(\cdot, \cdot),$
<b>[2] Linearity:</b>	$E(y_t   \mathbf{X}_t = \mathbf{x}_t) = \beta_0 + \beta_1^\top \mathbf{x}_t,$ linear in $\mathbf{x}_t,$
<b>[3] Homoskedasticity:</b>	$\text{Var}(y_t   \mathbf{X}_t = \mathbf{x}_t) = \sigma^2,$ free of $\mathbf{x}_t,$
<b>[4] Independence:</b>	$\{(y_t   \mathbf{X}_t = \mathbf{x}_t), t \in \mathbb{N}\}$ is an independent process,
<b>[5] t-invariance:</b>	$\boldsymbol{\theta} := (\beta_0, \beta_1, \sigma^2)$ do not change with $t.$

It is important to re-iterate that the error-statistical approach promotes *thorough probing* of the different ways an inductive inference might be *in error*, by localizing the

error probe in the context of the different models, theory, structural and statistical, mentioned above; see Spanos (2006a-b).

The above error-statistical perspective can be used to shed light on a number of methodological issues relating to specification, misspecification testing, and respecification, including the role of preliminary data analysis, structural vs. statistical models, model specification vs. model selection, and statistical vs. substantive adequacy; see Spanos (2006a-c).

## 6 Problems in philosophy of science

In this section we present brief reflections on two problems in philosophy of science when viewed from the error-statistical perspective.

### 6.1 Reflecting on curve-fitting

In its simplest form the curve-fitting problem is how to approximate an unknown functional form between  $y$  and  $x$ , say  $y = h(x)$ , by choosing among a family of curves, say  $y = \sum_{i=0}^m \alpha_i \varphi_i(x)$ ,  $m=1, 2, \dots$ , which can be fitted through the scatter-plot of data points  $\mathbf{z}_0 := \{(x_k, y_k), k=1, \dots, n\}$  in a way that would capture the ‘regularities’ in the data adequately; see Glymour (1980). The conventional wisdom is that there is an infinity of possible curves (models) that can be considered to be ‘consistent with any data’. The crucial problem is how to determine the ‘fittest’ curve if reliability of inference is an important objective.

Viewed from the error-statistical perspective, the current framework for addressing the curve-fitting problem is, on the one hand, the undue influence of the mathematical approximation perspective, and on the other, the insufficient attention paid to the statistical modeling aspects of the problem. Using goodness-of-fit as the primary guiding criterion, the mathematical approximation perspective undermines the reliability of inference objective by giving rise to selection rules which pay insufficient attention to ‘capturing the regularities in the data’. It is argued that high goodness-of-fit, is neither necessary nor sufficient for reliable inference. The contention that one can always fit an  $(n-1)$  degree polynomial, say  $y_k = \sum_{i=0}^{n-1} \alpha_i x_k^i$ , to data  $\mathbf{z}_0$  (see Skyrms, 2000) is statistically fallacious because it ignores the fact the least-squares estimated coefficients  $(\hat{\alpha}_0, \hat{\alpha}_1, \dots, \hat{\alpha}_{n-1})$  are *inconsistent estimators of*  $(\alpha_0, \alpha_1, \dots, \alpha_{n-1})$ . There is one observation for each parameter, rendering inference based on such a model totally unreliable; see Spanos (2007b)

In the context of the error-statistical approach the *fittest curve* is the one which, when embedded in a statistical model, turns out to be *statistical adequate*: its probabilistic assumptions (say, table 3) are valid for data  $\mathbf{z}_0$ . The validity of these assumptions ascertains when a curve ‘captures the ‘regularities’ in the data’ and formalizes the intuitive notion: the residuals contain *no systematic information*.

The *reliability of inference* in the context of the error-statistical approach is achievable because: (i) the premises of inductive inference are rendered *empirically testable*, and (ii) statistically adequate premises ensure that the *nominal* error probabilities approximate closely the *actual* error probabilities. Hence, the relevant *error probabilities* can be used to assess the reliability of inductive inference. In Spanos (2007b) the Kepler and Ptolemy models for the motion of the planets are compared in terms of statistical adequacy and shown that, contrary to conventional wisdom, the two models do not ‘account for the regularities in the data’ equally well; the former is statistically adequate but the latter is not.

## 6.2 Reflecting on Duhem’s problem

Duhem’s problem has been discussed in some detail from the error-statistical perspective by Mayo (1997). The purpose of the brief note below is to relate it to the notion of statistical adequacy.

According to the conventional wisdom in philosophy of science, Duhem’s problem in testing a hypothesis  $h$  using data  $\mathbf{x}_0$  arises from the fact that the observations must be interpreted, and that interpretation always invokes auxiliary hypotheses about phenomena or data, say  $\mathcal{A} := (A_1, A_2, \dots, A_m)$ . Since these are usually part of the broader premises ( $h \& \mathcal{A}$ ), such results cannot provide independent evidence for or against that hypothesis.

In terms of the simple modus tollens argument, Duhem’s problem takes the form:

$$\frac{\text{If } h \& \mathcal{A}, \quad \text{then } \mathbf{e} \\ \text{not-}\mathbf{e},}{\therefore \text{not-}h, \text{ or not-}\mathcal{A}}$$

In the context of the error-statistics data  $\mathbf{x}_0$  are interpreted (as a truly typical realization) in the context of a statistical model  $\mathcal{A}$ , comprising probabilistic assumptions  $(A_1, A_2, \dots, A_m)$ , and  $(h \& \mathcal{A})$  constitutes the embedding of the primary hypothesis of interest into the statistical model.  $(h, \bar{h})$  denotes the parameterization of the primary hypothesis of interest  $h$  in the context of  $\mathcal{A}$ .

**Example.** Let  $\mathcal{A}$  represent assumptions [1]-[4] of the simple Normal model (table 1), and the primary hypothesis is parameterized by  $h : \mu = \mu_0$  vs.  $\bar{h} : \mu \neq \mu_0$ . The argument ‘if  $h \& \mathcal{A}$ , then  $\mathbf{e}$ ’ can be viewed as an instance of N-P testing reasoning where when a test statistic is evaluated under  $h$  it is expected to yield  $\mathbf{e}$ , but if the p-value indicates that data  $\mathbf{x}_0$  is ‘improbably far’ from what would have been expected if  $h$  were true, that gives rise to not- $\mathbf{e}$ . But it could be that one of the assumptions in  $\mathcal{A}$  is false and that might be the source of inferring not- $\mathbf{e}$ . However, in the error-statistical context the purely probabilistic construal of  $\mathcal{A}$  enables one to test the validity of assumptions  $(A_1, A_2, \dots, A_m)$  separately from  $h$ , and establish its statistical adequacy without invoking  $h$ . M-S testing enables us, not only to probe the validity of the assumptions [1]-[4], but also establish which, if any, of them are false. Assuming that [1]-[4] turn out to be valid, the original modus tollens is now

modified into:

$$\frac{\begin{array}{l} \text{If } h \& \mathcal{A}, \text{ then } e \\ \mathcal{A} \& \text{ not-}e, \end{array}}{\therefore \text{not-}h}$$

Severity then enables one to proceed from the inference not- $h$  to establish the warranted evidence for  $\bar{h}$ .

## 7 Methodological issues in econometrics

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

### 7.1 Reliability/precision of inference and robustness

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 probabilistic assumptions comprising the statistical model in the context of which the inference takes place. Based on such premises, the optimality of inference methods in *frequentist statistics* is defined in terms of their capacity to give rise to valid inferences (trustworthiness), which is appraised in terms of the associated error probabilities: how often these procedures lead to erroneous inferences. 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 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). In practice, the unreliability of inference problem often stems from the inability to utilize the *relevant error probabilities* arising from being unaware of the presence of departures from the premises. However, even if one were in a position to utilize the actual error probabilities, that, by itself, does not address the unreliability of inference problem in general. This is because the presence of misspecification calls into question, not only the appropriateness of the nominal error probabilities, but also the optimality of the original inference procedure; without (b), (a) makes little sense. Hence, the unreliability of inference problem is better addressed by *respecifying* the original statistical model and utilizing inference methods that are optimal in the context of the new (adequate) premises; see Spanos (1986).

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

the nominal error probabilities, but downplays the interconnection between (a) and (b) above. 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.

**Example.** Assume that data  $\mathbf{x}_0$  constitute a ‘truly typical realization’ of the stochastic process represented by the simple Normal model (table 1), but it turns out that assumption [4] is actually invalid. Instead, the following form dependence is present:

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

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  $\rho$ .

Table 4 - Type I error of t-test							
$\rho$	<b>0.0</b>	.05	.10	.30	.50	.75	.90
$\alpha^*$ -actual	<b>.05</b>	.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  $\rho \rightarrow 1$  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					
$\rho$	$\pi^*(.02)$	$\pi^*(.05)$	$\pi^*(.1)$	$\pi^*(.2)$	$\pi^*(.4)$
<b>0.0</b>	<b>.074</b>	<b>.121</b>	<b>.258</b>	<b>.637</b>	<b>.991</b>
.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.

Moreover, 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).

## 7.2 Weak assumptions and the reliability 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 of weak probabilistic assumptions as 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 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:

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

The G-M theorem establishes that the OLS  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$  is Best Linear Unbiased Estimator (BLUE) of  $\boldsymbol{\beta}$  under assumptions (1)-(3), *without* invoking Normality ((4)  $\mathbf{u} \sim \mathbf{N}(\cdot, \cdot)$ ). In addition to BLUE being of very limited value The problem, however, is that the G-M theorem yields an unknown sampling distribution ( $\hat{\mathbf{D}}$ ) for  $\hat{\boldsymbol{\beta}}$ , i.e.  $\hat{\boldsymbol{\beta}} \sim \hat{\mathbf{D}}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$ , which provides a poor basis for hypothesis testing and other forms of inference. Finite sample inference can only be based on inequalities like Chebyshev’s and it will be very crude and imprecise; Spanos (1999), ch. 10. As a result, practitioners usually invoke the central limit theorem to use the approximation  $\hat{\boldsymbol{\beta}} \simeq \mathbf{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$ , but one has no way of knowing how good this approximation is for the particular sample size  $n$ ; unless they 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.

### 7.3 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 (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  $E(\mathbf{u}\mathbf{u}^\top) = \Omega \neq \sigma^2\mathbf{I}_n$ , the recommendation is twofold. Either to retain the OLS estimator  $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$  but utilize the autocorrelation consistent standard errors for inference purposes, or to use a Feasible Generalized Least Squares (FGLS) estimator based on the autocorrelation-corrected model where the error terms is assumed to be an AR(1) process  $u_t = \rho u_{t-1} + \varepsilon_t$ . 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  $\hat{\beta}$  but uses the heteroskedasticity consistent standard errors for inference, or estimates the heteroskedastic variances using an auxiliary regression:

$$\hat{u}_t^2 = c_0 + \mathbf{c}_1^\top \mathbf{z}_t + v_t, \quad (44)$$

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  $[z_t]$  used, as long as the weights are uncorrelated with the disturbances.”

In practice one is encouraged to try out different forms for the weights  $z_t$  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 appreciated 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 model 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 autocorrelation/heteroskedasticity consistent standard errors do very little, if anything, to ameliorate the reliability of inference in 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 being oblivious to them, enhancing the overall untrustworthiness of the evi-

dence it gives rise to. Instead 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 and. Moreover, incomplete specifications of statistical models are not conducive to statistical adequacy; assumption [5] (see table 3) is not explicitly stated or tested in practice.

This should be contrasted with warranted "data mining" which arises in cases where, despite appearances to the contrary, the procedure followed enhances the reliability of the inference reached. In the context of the error-statistical framework, the procedures of specification (choosing the original statistical model), M-S testing and respecification, as well as the use of graphical techniques, can be shown to constitute warranted data mining; see Spanos (2000), Mayo and Spanos (2004).

## 7.4 Substantive ‘error-fixing’ strategies and theory mining

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

$$y_t = \beta_0 + \beta_1 x_t + u_t, \quad t \in \mathbb{N}, \quad (45)$$

is ‘did you account for simultaneity in your model?’ For instance, the estimated model in (1) provides a perfect target for the cognoscentis 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 instrumental variables, 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:

(i)  $E(X_t u_t) \neq 0$  in (45), (ii)  $E(\mathbf{W}_t \varepsilon_{2t}) = \mathbf{0}$ , (iii)  $Cov(\mathbf{W}_t, X_t) \neq \mathbf{0}$  in (6);

conditions (ii)-(iii) ensure that the IV estimator of  $\beta_1$  is consistent. A comparison between the OLS and IV estimates is 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  $\beta_1$  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 constitutes theory mining 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.

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 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 proportional changes like  $(y_t, x_t)$ . Re-estimating (1) using the IV method yields:

$$y_t = \underset{(2.179)}{7.180} - \underset{(.090)}{0.689}x_t + \tilde{u}_t, \quad R^2=.622, \quad s=14.450, \quad n=45, \quad (46)$$

showing only minor differences between the OLS and IV estimates. One could take this as an excellent indication that the original estimates are robust and simultaneity is not a problem in this case. However, looking at the overidentifying restrictions test for (46),  $F(4, 39)=13.253[.0000007]$ , 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 (46) are seriously statistically misspecified. More importantly, the statistical meaningfulness of (46) 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] (see 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.

## 7.5 Bias-inducing procedures revisited

In this sub-section, it is argued that certain bias-inducing arguments constitute inappropriate formalizations of well-known testing fallacies; the result of conflating estimation with testing error probabilities.

### 7.5.1 Pre-test bias

The pre-test bias argument is often used in econometrics to raise questions about the appropriateness of various model selection methods based on some form of testing; see Mittelhammer et al (2000). To bring out the gist of the argument consider the following two alternative models:

$$\begin{aligned} \mathcal{M}_0 : \quad & y_t = \beta_0 + \beta_1 x_t + u_t, \\ \mathcal{M}_1 : \quad & y_t = \beta_0 + \beta_1 x_t + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \end{aligned} \quad (47)$$

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

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

The choice between  $\mathcal{M}_0$  and  $\mathcal{M}_1$  is viewed as one of choosing between two estimators of  $\beta_1$  and then formalized in decision-theoretic terms using the *pre-test estimator*  $\ddot{\beta}_1$  :

$$\ddot{\beta}_1 = \lambda \widehat{\beta}_1 + (1-\lambda) \widetilde{\beta}_1, \text{ where } \lambda = \begin{cases} 1, & \text{if } H_0 \text{ is accepted,} \\ 0, & \text{if } H_0 \text{ is rejected,} \end{cases} \quad (48)$$

viewed as 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 complicated variance; see Mittelhammer et al (2000). The pre-test argument suggests that when these effects are ignored one uses the ‘wrong’ error probabilities, giving rise to unreliable inferences.

When the pre-test bias argument, based on (47), is viewed in the context of the error-statistical approach, it becomes clear that the conceptual foundations of this argument are questionable. *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. The validity of the alternative model  $\mathcal{M}_1$  needs to be established separately by testing its own assumptions; see Spanos (2000, 2001). *Second*, it misconstrues an M-S testing problem (testing the validity of assumption [4] table 3) as an estimation problem whose relevant error probabilities are evaluated using a loss function. As argued in section 4, the error probabilities for estimation and testing are very different in nature and conflating the two can lead to major confusions.

### 7.5.2 Omitted Variables bias

A question that arises 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 it 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:

$$\mathcal{M}_0 : y_t = \beta_0 + \beta_1 x_{1t} + u_t, \quad \mathcal{M}_1 : y_t = \alpha_0 + \alpha_1 x_{1t} + \alpha_2 x_{2t} + \varepsilon_t, \quad (49)$$

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

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

see Leeb and Pötscher (2005). This poses a crucial problem of *confounding* which is motivated by concerns that the estimated model  $M_0$  could have 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  $\check{\beta}_1 = \lambda\hat{\beta}_1 + (1-\lambda)\hat{\alpha}_1$ , where  $\lambda$  is given in (48),  $(\hat{\beta}_1, \hat{\alpha}_1)$  denote the OLS estimators of  $(\beta_1, \alpha_1)$  is problematic for several reasons. First, the parameterizations of  $(\beta_1, \alpha_1)$  are very different; one is not estimating the same parameter in the two cases. Second, the framing of the problem in terms of a choice between two point estimators is inadequate for the task, because it mechanically interprets accept and reject as evidence for  $\mathcal{M}_0$  and  $\mathcal{M}_1$ , respectively; committing both classic fallacies of acceptance and rejection. Third, the pre-test bias is evaluated in terms of estimation error probabilities, like the Mean Square Error, but the relevant error probabilities are the ones associated with testing. Indeed, the most effective way to address the confounding problem is to treat it as a N-P testing problem and supplement accept/reject with post-data error probabilities associated with severe testing; see Spanos (2006b).

The comparison of the two models in (47) and (49) differ in so far as the former concerns *statistical adequacy*, but the latter concerns *substantive adequacy*: does model  $\mathcal{M}_0$  provide an adequate explanation for the behavior of  $y_t$ ? One can show that there are eight alternative scenarios (different answers) to the confounding question in (49), depending on the non-zero values of  $Cov(y_t, X_{1t})$ ,  $Cov(y_t, X_{2t})$ ,  $Cov(X_{1t}, X_{2t})$ , which can only be probed adequately using hypothesis testing; estimation is too crude a method. Moreover, the comparison in (49) gives rise to reliable inferences only to the extent that  $\mathcal{M}_1$  in (49) 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. No such presupposition is invoked in the case of (47); see Spanos (2006b).

## 8 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, 2007b), current econometrics runs a great risk of losing credibility as a way to provide empirical foundations to economic theorizing and policy analysis. The accumulation of mountains of untrustworthy empirical evidence in applied econometrics over the last century is a symptom of major weaknesses in the 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 all potential errors that could lead the inference astray. The ‘error-fixing’ strategies endanger the trustworthiness of the empirical evidence they give rise to and often brush aside other forms of potential errors, including the *data inaccuracy*, *incongruous measurement* and *substantive inadequacy*.

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. This approach derives from both modern frequentist statistics and the philosophy of science tradition of new experimentalism, and considers the primary objective of empirical modeling to be ‘learning from data’ about phenomena of interest. Such learning is achieved by employing reliable procedures with ascertainable error probabilities. The form of inductive reasoning adopted is based on the notion of *severe testing*, which strongly encourages the probing of the different ways an inference might be in error. The severe testing reasoning can also 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 observational data.

In particular, the error-statistical framework: (a) views empirical modeling as bridging the *gap between theory and data*, using a chain of complecting models, (b) affords the data ‘a life of their own’, (c) specifies statistical models in terms of testable probabilistic assumptions concerning the observable processes, (d) allows for both statistical and substantive information to play important roles without compromising the integrity of either, and (e) encourages *error probing* at all levels (models); see Spanos (2006a-c). Reliable theory testing can only take place when a substantive claim is assessed in the context of a statistically adequate model of the data. Only then, can ‘learning from data’ contribute significantly towards establishing economics as an empirical science.

## 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] Agresti, A. (2002), *Categorical Data Analysis*, 2nd ed., Wiley, NY.
- [4] Altman, D. G., D. Machin, T. N. Bryant and M. J. Gardner (2000), *Statistics with Confidence*, (eds), British Medical Journal Books, Bristol.
- [5] Backhouse, R. E. (1994), *New Directions in Economic Methodology*, Routledge, London.
- [6] Bennett, J. H. (1990), ed., *Statistical Inference and Analysis: Selected correspondence of R. A. Fisher*, Clarendon Press, Oxford.
- [7] Birnbaum, A. (1961), “Confidence Curves: An Omnibus Technique for Estimation and Testing,” *Journal of the American Statistical Association*, **294**, 246-249.
- [8] 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.

- [9] Blaug, M. (1992), *The Methodology of Economics*, Cambridge University Press, Cambridge.
- [10] Mellor, D. H. (1980), *Science, Belief & Behaviour: Essays in Honor of R. B. Braithwaite*, Cambridge University Press, Cambridge.
- [11] Caldwell, B. (1994), *Beyond Positivism: Economic Methodology in the Twentieth Century*, 2nd ed., George Allen & Unwin, London.
- [12] Cartwright, N. (1983), *How the Laws of Physics Lie*, Clarendon Press, Oxford.
- [13] Chalmers, A. F. (1999), *What is this thing called Science?*, 3rd ed., Hackett, Indianapolis.
- [14] Chatterjee, S. K. (2003), *Statistical Thought : A Perspective and History*, Oxford University Press, Oxford.
- [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 (2007), "Some remarks on the nature of statistical inference," forthcoming in *Error and Inference*, D. G. Mayo and A. Spanos (eds.), Cambridge University Press, Cambridge.
- [18] Davis, J. B. , D. W. Hands, U. Maki (1998), *The Handbook of Economic Methodology*, (eds.), Edward Elgar, Cheltenham.
- [19] Day, J. P. (1961), *Inductive Probability*, Routledge & Kegan Paul, London.
- [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), *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh.
- [24] Fisher, R. A. (1935), *The Design of Experiments*, Oliver and Boyd, Edinburgh.
- [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. (1955), "Statistical methods and scientific induction," *Journal of the Royal Statistical Society*, **B**, **17**, 69-78.
- [27] Giere, R. N. (1999), *Science Without Laws*, The University of Chicago Press, Chicago.
- [28] 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.

- [29] Gilbert, C. L. (1998), "Econometric Methodology," in Davis et al (1998), pp. 111-116.
- [30] Glymour, C. (1981), *Theory and Evidence*, Princeton University Press, NJ.
- [31] Guala, F. (2005), *The Methodology of Experimental Economics*, Cambridge University Press, Cambridge.
- [32] Fisher, R. A. (1956), *Statistical methods and scientific inference*, Oliver and Boyd, Edinburgh.
- [33] Franklin, A. (1986), *The Neglect of Experiment*, Cambridge University Press, Cambridge.
- [34] Godambe, V. P. and D. A. Sprott (1971), *Foundations of Statistical Inference: a Symposium*, Holt, Rinehart and Winston, Toronto.
- [35] Godfrey-Smith, P. (2003), *Theory and Reality: An Introduction to the Philosophy of Science*, The University of Chicago Press, Chicago.
- [36] Granger, C. W. J. (1990), (ed.) *Modelling Economic Series*, Clarendon Press, Oxford.
- [37] Greene, W. H. (2000), *Econometric Analysis*, 4th ed., Prentice Hall, NJ.
- [38] Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge University Press, Cambridge.
- [39] Hacking, I. (1980), "The Theory of Probable Inference: Neyman, Peirce and Braithwaite," in Mellor (1980), pp. 141-160.
- [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] Hald, A. (2007), *A History of Parametric Statistical Inference from Bernoulli to Fisher*, Springer, NY.
- [43] Hall, A. R. (2005), *Generalized Method of Moments*, Oxford University Press, Oxford.
- [44] Hands, W. D. (2001), *Reflection without Rules: Economic Methodology and Contemporary Science Theory*, Cambridge University Press, Cambridge.
- [45] Harlow, L. L., S. A. Mulaik and J. H. Steiger (1997), *What if there were no Significance Tests?* Mahwah, Erlbaum, NJ.
- [46] 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.
- [47] Hempel, C. G. (1965), *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, Mcmillan, New York.
- [48] Hendry, D. F. (1993), *Econometrics: Alchemy or Science?*, 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] Kempthorne, O. and L. Folks (1971), *Probability, Statistics, and Data Analysis*, The Iowa State University Press, Ames, IA.
- [55] Keuzenkamp, H. A. (2000), *Probability, Econometrics and Truth*, Cambridge University Press, Cambridge.
- [56] Kuhn, T. (1962), *The Structure of Scientific Revolutions*, The University of Chicago Press, Chicago.
- [57] Kuhn, T. (1977), *The Essential Tension: Selected Studies in Scientific Tradition and Change*, The University of Chicago Press, Chicago.
- [58] Ladyman, J. (2002), *Understanding Philosophy of Science*, Routledge, London.
- [59] Lakatos, I. (1970), "Falsification and the Methodology of Scientific Research Programms," in Lakatos and Musgrave (1970), pp. 91-196.
- [60] Lakatos, I. and A. Musgrave (eds.) (1970), *Criticism and Growth of Knowledge*, Cambridge University Press, Cambridge.
- [61] Laudan, L. (1977), *Progress and Its Problems: Towards a Theory of Scientific Growth*, Berkeley: University of California Press.
- [62] Lawson, T. (1997), *Economics and Reality*, Routledge, London.
- [63] Leamer, E. E. (1978), *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Wiley, New York.
- [64] Leeb, H. and B. M. Pötscher (2005), "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, **21**, 21-59.
- [65] Lehmann, E. L. (1986), *Testing statistical hypotheses*, 2nd edition, Wiley, New York.
- [66] Lehmann, E. L. (1990), "Model specification: the views of Fisher and Neyman, and later developments", *Statistical Science*, **5**, 160-168.
- [67] Lehmann, E. L. (1993), "The Fisher and Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?" *Journal of the American Statistical Association*, **88**, 1242-9.

- [68] Lieberman, B. (1971), *Contemporary Problems in Statistics: a Book of Readings for the Behavioral Sciences*, Oxford University Press, Oxford.
- [69] Lindley, D. V. (1965), *Introduction to Probability and Statistics from the Bayesian Viewpoint*, Cambridge University Press, Cambridge.
- [70] Machamer, P. and M. Silberstein (2002), *The Blackwell Guide to the Philosophy of Science*, Blackwell, Oxford.
- [71] Maki, U. (2001), *The Economic World View: Studies in the Ontology of Economics*, Cambridge University Press, Cambridge.
- [72] Maki, U. (2002), *Fact and Fiction in Economics*, Cambridge University Press, Cambridge.
- [73] Matyas, L. (1999), (editor), *Generalized Method of Moments Estimation*, Cambridge University Press, Cambridge.
- [74] Mayo, D. G. (1991), "Novel Evidence and Severe Tests", *Philosophy of Science*, **58**, 523-552.
- [75] Mayo, D. G. (1996), *Error and the Growth of Experimental Knowledge*, The University of Chicago Press, Chicago.
- [76] 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.
- [77] Mayo, D. G. (2005), 'Philosophy of Statistics,' in S. Sarkar and J. Pfeifer (eds.), *Philosophy of Science: An Encyclopedia*, London: Routledge, pp. 802–15.
- [78] Mayo, D. G. and A. Spanos (2004), "Methodology in Practice: Statistical Misspecification Testing", *Philosophy of Science*, **71**, 1007-1025.
- [79] 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.
- [80] 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.
- [81] McCloskey, D. N. (1985/1998), *The Rhetoric of Economics*, 2nd ed., University of Wisconsin, Madison.
- [82] McGuirk, A. and A. Spanos (2004), "Revisiting Error Autocorrelation Correction: Common Factor Restrictions and Granger Non-Causality," Virginia Tech working paper.
- [83] Mills, F. C. (1924), *Statistical Methods*, Henry Holt and Co., New York.
- [84] Mills, T.C. and K. Patterson, (2006), *New Palgrave Handbook of Econometrics*, vol. 1, MacMillan, London.
- [85] Mises, R. von (1928/1981), *Probability, Statistics and Truth*, Dover, New York.

- [86] Mittelhammer, R. C., G. C. Judge and D. J. Miller (2000), *Econometric Foundations*, Cambridge University Press, Cambridge.
- [87] Moore, H. L. (1914), *Economic Cycles - Their Laws and Cause*, McMillan, New York.
- [88] Morgenstern, O. (1963), *On the accuracy of economic observations*, 2nd edition, Princeton University Press, New Jersey.
- [89] Morrison, D. E. and R. E. Henkel (1970), *The Significance Test Controversy: A Reader*, Aldine, Chicago.
- [90] Nagel, E. (1961), *The Structure of Science*, Hackett, Indianapolis.
- [91] Newton-Smith, W. H. (ed.) (2000), *A Companion to the Philosophy of Science*, Blackwell, Oxford.
- [92] 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.
- [93] Neyman, J. (1950), *First Course in Probability and Statistics*, Henry Holt, New York.
- [94] Neyman, J. (1952), *Lectures and Conferences on Mathematical Statistics and Probability*, 2nd ed. U.S. Department of Agriculture, Washington.
- [95] Neyman, J. (1956), "Note on an Article by Sir Ronald Fisher," *Journal of the Royal Statistical Society*, B, **18**, 288-294.
- [96] Neyman, J. (1977), "Frequentist Probability and Frequentist Statistics," *Synthese*, **36**: 97-131.
- [97] 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.
- [98] Pagan, A.R. (1987), "Three econometric methodologies: a critical appraisal", *Journal of Economic Surveys*, **1**, 3-24. Reprinted in C. W. J. Granger (1990).
- [99] Pagan, A.R. and A. Ullah (1999), *Nonparametric Econometrics*, Cambridge University Press, Cambridge.
- [100] Pearson, K. (1920), "The Fundamental Problem of Practical Statistics," *Biometrika*, **XIII**, 1-16.
- [101] Pearson, E. S. (1955), "Statistical Concepts in the Relation to Reality," *Journal of the Royal Statistical Society, Series B*, **17**, 204-207.
- [102] 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.
- [103] Poole, C. (1987), "Beyond the Confidence Interval," *The American Journal of Public Health*, **77**, 195-199.
- [104] Popper, K. R. (1959), *The Logic of Scientific Discovery*, Hutchinson, London.

- [105] Popper, K. R. (1963), *Conjectures and Refutations*, Routledge and Kegan Paul, London.
- [106] Popper, K. R. (1972), *Objective Knowledge*, Clarendon Press, Oxford.
- [107] Peirce, C. S. (1878), "The Probability of Induction," *Popular Science Monthly*, **12**, 705-718.
- [108] Quine, W. V. (1953), *From the Logical Point of View*, Harvard University Press, Cambridge.
- [109] Quine, W. V. (1960), *World and Object*, The MIT Press, Cambridge.
- [110] Rao, C. R. (2004), "Statistics: Reflections on the Past and Visions for the Future," *Amstat News*, **327**, 2-3.
- [111] Redman, D. A. (1991), *Economics and the Philosophy of Science*, Oxford University Press, Oxford.
- [112] Renyi, A. (1970), *Probability Theory*, North-Holland, Amsterdam.
- [113] Rosenthal, R., R. L. Rosnow, D. B. Rubin (1999), *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach*, Cambridge University Press, Cambridge.
- [114] Salmon, W. (1967), *The Foundations of Scientific Inference*, University of Pittsburgh Press.
- [115] Skyrms, B. (2000), *Choice and Chance: an introduction to inductive logic*, 4th ed., Wadsworth: Thomson Learning.
- [116] Spanos, A., (1986), *Statistical Foundations of Econometric Modelling*, Cambridge University Press, Cambridge.
- [117] Spanos, A. (1988), "Towards a Unifying Methodological Framework for Econometric Modelling", *Economic Notes*, 107-34; reprinted in Granger (1990).
- [118] Spanos, A. (1989), "On re-reading Haavelmo: a retrospective view of econometric modeling", *Econometric Theory*, **5**, 405-429.
- [119] Spanos, A. (1990), "The Simultaneous Equations Model revisited: statistical adequacy and identification", *Journal of Econometrics*, **44**, 87-108.
- [120] Spanos, A. (1995), "On theory testing in Econometrics: modeling with nonexperimental data", *Journal of Econometrics*, 67:189-226.
- [121] Spanos, A. (1999), *Probability Theory and Statistical Inference: econometric modeling with observational data*, Cambridge University Press, Cambridge.
- [122] Spanos, A. (2000), "Revisiting Data Mining: 'hunting' with or without a license," *The Journal of Economic Methodology*, **7**, 231-264.
- [123] Spanos, A. (2001), "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.

- [124] Spanos, A. (2004), "Confidence Curves, Consonance Intervals, P-value Functions and Severity Evaluations," Working Paper, Virginia Tech.
- [125] Spanos, A. (2005), "Misspecification, Robustness and the Reliability of Inference: the simple t-test in the presence of Markov dependence," Working Paper, Virginia Tech.
- [126] Spanos, A. (2006a) "Econometrics in Retrospect and Prospect," in Mills, and Patterson (2006), pp. 3-58.
- [127] Spanos, A. (2006b) "Revisiting the Omitted Variables Argument: Substantive vs. Statistical Adequacy," *Journal of Economic Methodology*, **13**: 179-218.
- [128] 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.
- [129] 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.
- [130] Spanos, A. (2007b), "Curve-Fitting, the Reliability of Inductive Inference and the Error-Statistical Approach," forthcoming *Philosophy of Science*.
- [131] Spanos, A. (2007c), "Statistics and Economics," forthcoming in the *New Palgrave Dictionary of Economics*, 2nd edition, edited by Steven N. Durlauf and Roger E. Backhouse, MacMillan, London.
- [132] Spanos, A. and A. McGuirk (2001), "The Model Specification Problem from a Probabilistic Reduction Perspective," *Journal of the American Agricultural Association*, **83**, 1168-1176.
- [133] Spanos, A. and A. McGuirk (2004), "Revisiting Error Autocorrelation Correction: Common Factor Restrictions and Granger Non-Causality," Virginia Tech working paper.
- [134] Stigler, S. M. (1986), *The History of Statistics: the Measurement of Uncertainty before 1900*, Harvard University Press, Cambridge, Massachusetts.
- [135] Stigum, B. P. (2003), *Econometrics and the Philosophy of Economics*, Princeton University Press, Princeton.
- [136] Suppe, F. (1977), *The Structure of Scientific Theories*, 2nd ed., University of Illinois Press, Urbana.
- [137] Wooldridge, J. M. (2006), *Introductory Econometrics: a modern approach*, Thomson, South-Western.